EPJ Web of Conferences **158**, 06005 (2017) DOI: 10.1051/epjconf/201715806005
QFTHEP 2017

The evaluation of the systematic uncertainties for the finite MC samples in the presence of negative weights

Petr Mandrik[1,⋆]

[1] *NRC «Kurchatov Institute» – IHEP, Protvino*

Abstract. The analysis of results from HEP experiments often involves the estimates of the composition of the binned data samples, based on Monte Carlo simulations of various sources. Due to a finite statistic of MC samples they have statistical fluctuation. This work proposes the method of incorporating the systematic uncertainties due to finite statistics of MC samples with negative weights. The possible approximations are discussed and the comparison of different methods are presented.

1 Introduction

Experimental results in high energy physics are often represented as a binned distribution (histogram) of observed events $\boldsymbol{X} = (X_1, X_2, ...)$, where X_i is a number of events in bin i. The usual method to estimate physical parameters such as particle masses or cross sections from this distribution is to perform some of Bayesian or frequentist analyses based on likelihood function. Likelihood function $\mathcal{L}(\boldsymbol{X}|\boldsymbol{m})$ connect the data with a theoretical model and represente how well the observatios are described by the prediction $\boldsymbol{m} = (m_1, m_2, ...)$. This prediction may depend on several different parameters $\boldsymbol{\pi} = (\pi_1, \pi_2, ...)$: nuisance parameters and parameters of interests. In addition if some signals or background processes are known from Monte-Carlo simulations then the likelihood function depends on template distributions $\boldsymbol{t} = (t_1^a, t_2^a, ...t_1^b, t_2^b, ...t_1^c, t_2^c, ...)$, where t_i^k is a number of events for process k in bin i:

$$\mathcal{L}(\boldsymbol{X}|\boldsymbol{m}) = \mathcal{L}(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{t}) = \prod_i P(X_i|\boldsymbol{\pi}, \boldsymbol{t}_i) = \prod_i P(X_i|\pi_1, \pi_2, ..., t_i^a, t_i^b, ...) \tag{1}$$

This is important task to define an adequate likelihood function and take into account all existed statistical and systematic uncertainties present in the analysis.

Template distributions from Monte-Carlo generators are subject to statistical fluctuations due to finite number of events in samples. The influence of these fluctuations can be expected to be significant in regions of low amounts of Monte-Carlo events. For incorporating such uncertainties into likelihood function Barlow and Beeston proposed a method [1] wherein one for every bin i and every process k introduces a new parameter T_i^k corresponding to unknown expected number of events in infinite statistics limit:

$$\prod_i P(X_i|\boldsymbol{\pi}, \boldsymbol{t}_i) \to \prod_i \left[P(X_i|\boldsymbol{\pi}, \boldsymbol{T}_i) \cdot \prod_k P(t_i^k|T_i^k) \right] \tag{2}$$

⋆e-mail: Petr.Mandrik@ihep.ru

where for constrain $P(t_i^k|T_i^k)$ Barlow and Beeston assumed a Poisson distribution.

On the other hand several of the modern Monte-Carlo generators [2] produce a weighted events with both negative and positive weights. In this case the transformation (2) is not applicable. In this paper we provide a method of incorporating uncertainties due to the finite statistics of Monte-Carlo samples in the presence of negative weights.

2 Likelihood functions for Monte-Carlo samples with negative weights

In simplified form the algorithm of event production in most important example of Monte-Carlo generator with negative weights MadGraph5_aMC@NLO can be described as follow [3]. A cross section of some process σ_{NLO} is calculated by computing the integrals of two functions $F_H(x)$ and $F_S(x)$:

$$\sigma_{NLO} = \int F_H(x)dx + \int F_S(x)dx \tag{3}$$

By definition, the functions $F_H(x)$ and $F_S(x)$ are finite and $F_H(x) + F_S(x) > 0$, but for some values of x the function $F_S(x)$ is negative. Using the absolute values of the integrands $|F_H(x)|$ and $|F_S(x)|$ two set of events are produced - $\{x\}_H$ and $\{x\}_S$ respectly with weights $w_i^{H,S}$ equal to $+1$ for positive values of functions and weight -1 if function is negative, so:

$$\sigma_{NLO} = \frac{\int |F_H(x)|dx}{N_H} \cdot \sum_i^{N_H} w_i^H + \frac{\int |F_S(x)|dx}{N_S} \cdot \sum_i^{N_S} w_i^S \tag{4}$$

where N_H, N_S are thw number of events in corresponding sets.

In this way in the infinite statistics limit the prediction of any observable in any intervals $[x_i, x_i + \Delta x]$ of histograms from MadGraph5_aMC@NLO can be only positive, but in the case of finite statistics the prediction could get negative values in some bins. On the other hand the events with negative and positive weights should be treated in the same way during analyses and pass the same cuts to keep the correct cross section value (4). Further we assume that the last condition is satisfied, so for the finite number of generated Monte-Carlo events the probability of obtaining a given one in the case of only positive or negative weights is described by multinomial distribution, which is usually approximated by multiplication of independent Poisson distributions (see for example, [4]).

Let us consider a simple case of single bin and only one generated process with total number of event t. If t^+ is a sum of all positively weighted events from Monte-Carlo samples and t^- is a sum of all negative, then t is a difference of two Poissonian quantities t^+ and t^- and described by Skelleman distribution:

$$P(t) = \mathcal{S}(t|T^+, T^-) = \sum_{s=max(0,t)}^{\infty} \mathcal{P}(s|T^+) \cdot \mathcal{P}(s - t|T^-) = e^{-(T^+ + T^-)} \Big(\frac{T^+}{T^-}\Big)^{t/2} \mathcal{I}_t(2\sqrt{T^+T^-}) \tag{5}$$

where $\mathcal{I}_t$ is a modified Bessel functions of the first kind, $\mathcal{P}$ is Poisson distribution, T^+ and T^- are the parameters of Skellenam distribution, corresponding to unknown "true" prediction of negative and positive events from MC generator.

Using the equations (5) in (2) we obtain the following transformation rule for taking into account uncertainties due to the finite statistics of Monte-Carlo samples in the presence of negative weights in likelihood function:

$$\prod_i P(X_i|\boldsymbol{\pi}, \boldsymbol{t}_i) \to \prod_i \Big[P(X_i|\boldsymbol{\pi}, \boldsymbol{T}_i) \cdot \prod_k \mathcal{S}(t_i^k|T_i^{k+}, T_i^{k-})\Big] \tag{6}$$

where the new parameters are related by the equation $T_i^k = T_i^{k+} - T_i^{k-}$.

The constrain on parameters T_i^k, T_i^{k+}, T_i^{k-} in formula (6) may be improved by an independent treatment of values t^+ and t^- in analyses. In this case we get:

$$P(t) = \mathcal{P}(t^+|T^+) \cdot \mathcal{P}(t^-|T^-) \tag{7}$$

and from (2) with (7):

$$\prod_i P(X_i|\boldsymbol{\pi}, \boldsymbol{t}_i) \to \prod_i \Big[P(X_i|\boldsymbol{\pi}, \boldsymbol{T}_i) \cdot \prod_k \mathcal{P}(t_i^{k+}|T_i^{k+}) \cdot \mathcal{P}(t_i^{k-}|T_i^{k-}) \Big] \tag{8}$$

The number of extra parameters in the transformation rule (8) is equal to 2× *number of processes* × *number of bins*. We can decrease the number of parameters by using the method of maximum likelihood function Indeed, if $\mathcal{L}$ is a likelihood function with transformation (7) then for bin i one gets:

$$\begin{aligned} -\ln \mathcal{L}_i = -\ln P(X_i|\boldsymbol{\pi}, \boldsymbol{T}_i) - \sum_k \Big(- T_i^{k+} + t_i^{k+} \cdot \ln T_i^{k+} - \ln (t_i^{k+}!) \Big) - \\ - \sum_k \Big(- T_i^{k-} + t_i^{k-} \cdot \ln T_i^{k-} - \ln (t_i^{k-}!) \Big) \end{aligned} \tag{9}$$

The requirement of an extremum gives the following system of equations:

$$\begin{cases} -\dfrac{\partial \ln \mathcal{L}}{\partial T_i^{k+}} = 1 - \dfrac{\partial \ln P(X_i|\boldsymbol{\pi}, \boldsymbol{T}_i)}{\partial T_i^{k+}} - \dfrac{t_i^{k+}}{T_i^{k+}} = 0 \\ -\dfrac{\partial \ln \mathcal{L}}{\partial T_i^{k-}} = 1 - \dfrac{\partial \ln P(X_i|\boldsymbol{\pi}, \boldsymbol{T}_i)}{\partial T_i^{k-}} - \dfrac{t_i^{k-}}{T_i^{k-}} = 0 \end{cases} \tag{10}$$

This system (10) in some cases may be solved analytically for the parameters T_i^{k-}, T_i^{k+}, or numerically with some fixed values of the remaining parameters.

The another way to decrease the number of parameters related to finite statistics of Monte-Carlo is known as Barlow-Beeston "light" transformation [5]. As the statistical uncertainties for each source in each bin are independent they may be combined and be represented approximately by single effective parameter per bin M_i:

$$\prod_i P(X_i|m_i) \to \prod_i P(X_i|M_i) \cdot P(m_i|M_i) \tag{11}$$

Usually, in this approximation for $P(m_i|M_i)$ usually a Gaussian constrain $\mathcal{G}(m_i|M_i, \sigma_i)$ is used, where the value of σ_i are calculated by propagation of the Monte-Carlo statistical uncertainties in bin i with fixed values of the remaining parameters.

For histograms with negative weights the transformation (11) has the form:

$$\prod_i P(X_i|m_i^+ - m_i^-) \to \prod_i P(X_i|M_i^+ - M_i^-) \cdot P(m_i^+|M_i^+) \cdot P(m_i^-|M_i^-) \tag{12}$$

The number of extra parameters is equal to 2× *number of bins in histogram*.

For the likelihood function with transformation (12) a system of equations similar to (10) can be obtained. Using the Gaussian constrain one gets:

$$\mathcal{L}_i = \mathcal{P}(X_i|M_i^+ - M_i^-) \cdot \mathcal{G}(m_i^+|M_i^+, \sigma_i^+) \cdot \mathcal{G}(m_i^-|M_i^-, \sigma_i^-) \tag{13}$$

$$-\ln \mathcal{L}_i = -\Big[-(M_i^+ - M_i^-) + X_i \cdot \ln(M_i^+ - M_i^-) - \ln X_i!\Big] - \\ -\Big[\frac{(M_i^+ - m_i^+)^2}{2(\sigma_i^+)^2} - \ln \sigma_i^+ \sqrt{2\pi}\Big] - \Big[\frac{(M_i^- - m_i^-)^2}{2(\sigma_i^-)^2} - \ln \sigma_i^- \sqrt{2\pi}\Big] \quad (14)$$

$$\begin{cases} -\dfrac{\partial \ln \mathcal{L}}{\partial M_i^+} = 1 - \dfrac{X_i}{M_i^+ - M_i^-} - \dfrac{M_i^+ - m_i^+}{(\sigma_i^+)^2} = 0 \\ -\dfrac{\partial \ln \mathcal{L}}{\partial M_i^-} = 1 + \dfrac{X_i}{M_i^+ - M_i^-} - \dfrac{M_i^- - m_i^-}{(\sigma_i^-)^2} = 0 \end{cases} \quad (15)$$

3 The performance of methods

In this section few results of study the proposed transformations for taking into account uncertainties due to the finite statistics of Monte-Carlo samples are given. The source code was implemented with statistical package SHTA [6].

From the simple single-bin single-channel model:

$$\mathcal{L}_0 = \mathcal{P}(X|\pi \cdot (C + T^+ - T^-)) \cdot \mathcal{P}(t^+|T^+) \cdot \mathcal{P}(t^-|T^+) \quad (16)$$

a set of events (X, t^+, t^-) may be generated for the fixed values of parameters π, T^+, T^-. Here the constant C is introduced in order to avoid a long tail in posterior distribution of π from $T^+ - T^- \sim 0$.

To estimate the parameter of interest π we use three different likelihood functions. First of all a naive approach without incorporating uncertainties due to the finite statistics of Monte-Carlo samples:

$$\mathcal{L}_n = \mathcal{P}(X|\pi \cdot (C + t^+ - t^-)) \quad (17)$$

a likelihood function with transformation (8):

$$\mathcal{L}_p = \mathcal{P}(X|\pi \cdot (C + T^+ - T^-)) \cdot \mathcal{P}(t^+|T^+) \cdot \mathcal{P}(t^-|T^-) \cdot \mathcal{H}(T^+ - T^-) \quad (18)$$

and similar one but with Gaussian approximation for multiplication of two Poissons:

$$\mathcal{L}_g = \mathcal{P}(X|\pi \cdot (C + T)) \cdot \mathcal{G}(t^+ - t^-, T, \sqrt{t^+ + t^-}) \cdot \mathcal{H}(T) \quad (19)$$

where $\mathcal{H}$ is a Heaviside function.

The generated set of toy data from (16) is used to perform a Bayesian inference (see for example [7]) with non-informative flat priors for all parameters. The posterior probability density functions for parameters were obtained from likelihood functions (17), (19), (18) and the confidence intervals were found. The results for two different set of initial values of π, T^+, T^- are presented in the table 1.

From the table 1 we can see that the difference between solution with multiplication of two Poisson and its Gaussian approximation do not exceed one percent. On the other hand without taking into account the uncertainties due to the finite statistics of Monte-Carlo samples in likelihood function (17) the accuracy of measurements falls significantly. For example, only in two out of three experiments the correct interval will be obtained for $CL = 2\sigma$ and first set of parameters.

Larger number of bins in such test models provide additional constrain on the parameter of interest and without considering the increase of the computational complexity may only improve results.

Table 1. Percent of toy data for which the confidence interval with confidence level CL from posterior distribution includes the true value of parameter π. The uncertainties are evaluated by averaging the results from different sets of toy data.

$\mathcal{L}_0$ parameters	CL	$\mathcal{L}_n$	$\mathcal{L}_g$	$\mathcal{L}_p$
$\pi = 3, T^+ = 12, T^- = 4$	1σ	34.71 ± 0.79	67.79 ± 0.75	67.98 ± 0.72
	2σ	62.82 ± 0.70	95.25 ± 0.27	96.14 ± 0.22
$\pi = 3, T^+ = 9, T^- = 7$	1σ	26.41 ± 0.57	77.52 ± 0.63	78.51 ± 0.63
	2σ	49.73 ± 0.71	98.18 ± 0.15	98.21 ± 0.15

4 Conclusion

In this work a method of incorporating the systematic uncertainties due to finite statistics of MC samples with negative weights is presented. The influence of this statistical uncertainty can be expected to be high in regions of low amounts of Monte Carlo events and they must be included into the fit. The proposed transformation (8) and its simplified version (12) can be used to construct the correct likelihood function. While using the Gaussian approximation of multiplication of two Poisson distribution in (8) or (12) leads to the known expressions used in different statistical packages [8][9] in different forms, the choice of specific form of likelihood function depends on the analysis and in some cases the more accurate proposed methods can improve the results.

Acknowledgments

I wish to thank L. Dudko, S. Slabospitskii and G. Vorotnikov for useful discussions.

References

[1] R.J. Barlow, C. Beeston, Comput. Phys. Commun. **77**, 219 (1993)
[2] A. Buckley, J. Butterworth, S. Gieseke, D. Grellscheid, S. Höche, H. Hoeth, F. Krauss, L. Lönnblad, E. Nurse, P. Richardson et al., Physics Reports **504**, 145 (2011)
[3] S. Frixione, B.R. Webber, Journal of High Energy Physics **2002**, 029 (2002)
[4] C. Walck, *Hand-book on statistical distributions for experimentalists* (1996), `http://www.fysik.su.se/~walck/suf9601.pdf`
[5] J.S. Conway, *Incorporating Nuisance Parameters in Likelihoods for Multisource Spectra*, in *Proceedings, PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding, CERN,Geneva, Switzerland 17-20 January 2011* (2011), pp. 115–120, `1103.0354`, `http://inspirehep.net/record/891252/files/arXiv:1103.0354.pdf`
[6] P. Mandrik, *SHTA - package for experimental statistical data analyses in high energy physics*, `https://github.com/pmandrik/shta/tree/negative_weights` (2017)
[7] G. D'Agostini, CERN-99-03, CERN-YELLOW-99-03 (1999)
[8] T. Müller, J. Ott, J. Wagner-Kuhr, CMS Internal Note CMS-IN 017 (2010)
[9] K. Cranmer, G. Lewis, L. Moneta, A. Shibata, W. Verkerke (ROOT Collaboration), Tech. Rep. CERN-OPEN-2012-016, New York U., New York (2012), `http://cds.cern.ch/record/1456844`