Available online at www.sciencedirect.com

ScienceDirect

Nuclear Energy and Technology 2 (2016) 108–113

www.elsevier.com/locate/nucet

Comparison of histograms in physical research

S.I. Bityukov[a], A.V. Maksimushkina[b,*], V.V. Smirnova[a]

[a] *State Research Center, Institute for High Energy Physics, 1 Ploschad nauki, Protvino, Moscow Reg. 142281, Russia*
[b] *National research nuclear university MEPhI (Moscow Engineering Physics Institute) Kashirskoe sh., 31, Moscow, 115409, Russia*

Available online 24 May 2016

Abstract

Main approaches to the methods of comparison of histograms in physical studies are examined. The term “histogram” was originally introduced by Karl Pierson as the “generalized form of graphic representation” [1]. Histograms are very useful in this canonic application for visual data presentation. However, as of today histograms are often regarded as a purely mathematical object.

Histograms became indispensable tool in different subject fields of science. Besides the scientific data analysis in experimental studies histograms play important role in data base maintenance and in computer “vision” [1]. Accordingly, the goals and methods of histogram processing vary depending on the specific field of application. Histograms are addressed in the resent paper as one of the elements of data processing system used in the analysis of the data collected in the studies conducted on experimental facilities.

Certain methods of histogram comparison are presented and results of comparison are given for three methods (statistical histogram comparison method (SCH), Kolmogorov–Smirnov (KS) method and Anderson–Darling (AD) method) for determination of the possibility to compare histograms during assessment of distinguishability of data samples in the processing of which the histograms were generated.

Keywords: A histogram; The Monte Carlo method; The flow of events; The test statistic.

Introduction

Let a set of disjoint intervals exist. A histogram represents an empirical distribution of the population of the set with realized values of some random variable constructed using the data from the finite sample. Such intervals are usually called the bins. Realization of the random value is called the event.

Analysis of histograms depends on the procedure applied for filling out the histogram. For example, the limiting case is the distribution of brightness in the photo picture. The event in this case is the act of taking the picture. One event—one picture taken and, thus, one histogram is generated. Another limiting case is the construction of histogram if the event is the act of measurement of a random value with entering the obtained result in the histogram. Filling up the histogram is the purpose of independent measurement of the random value with gradual filling up of the histogram, i.e. one sample corresponds to one histogram.

The second approach to plotting histograms is usually applied in physical experiments. Thus, in high energy physics the event is determined by the development of conditions allowing fixing manifestation of interactions caused by impinging particles in the detectors, obtaining respective information by registration electronics in digital form and returning the experimental facility to its initial state ready to react to the emergence of the next event. The flow of registered events is stored in the form of several sets of samples for subsequent processing. Correspondingly, the contents of the histogram bin are called the number of events in the bin. The sum of the numbers of events in all the bins constitutes the histogram volume.

* Corresponding author.
E-mail addresses: Serguei.Bitioukov@cern.ch (S.I. Bityukov), AVMaksimushkina@mephi.ru (A.V. Maksimushkina), Vera.Smirnova@ihep.ru (V.V. Smirnova).

Peer-review under responsibility of National Research Nuclear University MEPhI (Moscow Engineering Physics Institute).

Russian text published: Izvestia Visshikh Uchebnikh Zavedeniy. Yadernaya Energetika (ISSN 0204-3327), 2016, n.1, pp. 81-90.

http://dx.doi.org/10.1016/j.nucet.2016.05.007

There exist a number of problems of general nature in the definition of histograms solution of which also often depends on the specific problem under solution. Such problems are the selection of optimal binning in the histogram and the selection of the model of distribution of errors for the observed value within the histogram bin.

Comparison of histograms

Let us have two histograms given. What is the way to determine whether they are similar or not? And what does it mean—“the similar histograms”? There exist several approaches to solving this problem.

Let us assume that the reference histogram is known. Often closeness between the reference and the tested histograms is measured using certain test statistics ensuring quantitative expression of the "distance” between the histograms [2]. The smaller is this distance, the more similar are the histograms. There exist several definitions of such distances in reference literature, for example, distance according to Kolmogorov [3], Kullback–Leibler distance [4], total variation of a function [5], chi-square-distance [6]. Usually these are the test statistics distribution of which can be set by formulas or using Monte-Carlo method. Another way is to convert the histograms into probability density functions and to perform comparison of these densities. The latter approach is based on the assumption that the histograms are obtained in the measurements of random variables providing the basis for the assessment of empirical probability density distribution. Calculation of the distance between the two densities can be regarded similarly to the calculation of Bayesian probability. For example, *Bhattacharyya distance* [7] or Hellinger distance [8] are used as the distance between two statistical ensembles. It has to be mentioned that distances according to Kolmogorov [2], according to Anderson–Darling [9], according to Kolmogorov [3] Kullback–Leibler [4] also allow comparing the initial samples without their representation in histogram form. However, this is a somewhat different problem.

There exists as well a new maximum average distance method [10]. The methodology based on the ranking or permutations (Mann–Whitney method [11]) and, in some cases, vector approach as well, are also used in histogram comparison. Histograms are regarded as vectors with the preset bin number size while the distance between them is estimated in Euclidean or Minkowskian metrics [12]. Sometimes similarity measure (*similarity*) is introduced in some logical scheme, for instance, such approach is addressed in Ref. [13] based on the Lukasiewicz logic.

Important task in histogram comparison is testing their compatibility or, vice versa, testing their distinguishability. Statement that both histograms are the result of processing of independent samples obtained from the same flow of events (or, which is the same, are taken from the same total population of events) is understood as the compatibility. Method is suggested in Ref. [14] allowing estimating distinguishability of histograms and, correspondingly, the distinguishability of initial flows of events according to the samples collected within them. The method is based on the statistical comparison of histograms; multidimensional test statistics is suggested to be used as the distance between histograms. Modification of the method under discussion for registering changes in parameters of information flows in problems of wireless data transmission is presented in Ref. [15].

If the purpose of comparison of histograms is to test their compatibility, then the problem is reduced to testing hypotheses where the main hypothesis $H0$ will be the statement that histograms were obtained in the processing of independent samples taken from one and the same flow of events, and the alternative hypothesis $H1$ will be the statement that histograms were obtained in the processing of samples taken from different flows of events. Selection of the main hypothesis and the alternative hypothesis depends on the specific problem addressed. Having determined the critical area for decision making and having made the choice between $H0$ and $H1$ probabilities can be estimated of errors of the first type (α) and the second type (β). First type error is the probability to make choice in favor of hypothesis $H1$ while hypothesis $H0$ is correct. Second type error is the probability to make choice in favor of hypothesis $H0$ while hypothesis $H1$ is correct. Selection of significance level α allows estimating the strength of the test $1-\beta$. Usually the significance level is established at the level of 1, 5 or 10%. If the hypotheses are equisignificant, then other combinations of α and β can be used. For example, the value of relative uncertainty in decision making $(\alpha+\beta)/(2-(\alpha+\beta))$ can be applied in the problem of distinguishability of flows of events. Average error of decision making $(\alpha+\beta)/2$ works when the “equal tailed” test is used. This is associated with the fact that in working with discrete distributions it is usually difficult to obtain absolute equality between α and β.

Other purposes for comparison of histograms also exist. Thus, search for abnormal structures in the histogram under testing which are not present in the reference histogram is a very important problem in particle physics. Bib-by-bin comparison of histograms is the possible solution of such problem. In this case probability is calculated that average values in the bins are similar and presence or absence of abnormal structures in the histogram is determined based on that.

Comparison of histograms is usually subdivided in the comparison of normalization of histograms and the comparison of shapes of histograms. Comparison of shape of histograms often depends on the normalization and, therefore, a combination of two tests is applied. In the simplest case normalization is estimated from the general considerations. These can be the ratio of volumes of the compared samples corrected by the additional knowledge (for instance, efficiency of registration of events during collection of data samples), or the ratios of time spent on the collection of the compared samples with constant flow of events. Methods of comparison of distributions are usually applied in the comparisons of shapes of histograms.

Testing hypotheses of compatibility or distinguishability of histograms requires knowledge of distributions of test statis-