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 statistics for both of the hypotheses. Conclusions are made based on the comparison of these distributions and the calculated value of test statistics. Knowledge of distributions of test statistics does not allow in all cases to estimate reliability of the conclusions made.

Let us examine the method allowing quantitative estimation of reliability of the taken decision. As it has been already mentioned, test statistics can be constructed by Monte-Carlo method. Let us examine the simple case of event-by-event histogram accumulation. Number of events in each of the bins of the histogram can be regarded as the realization of random variable with "expected number of events in the bin in question for the sample in question" parameter. In Monte-Carlo simulation it is necessary to either know exactly the parameter of the simulated random variable, or to use experimentally established estimation of this parameter. If the values of "expected number of events in the bin in question for the sample in question" parameters are exactly known for both of the histograms, then they are either completely similar or completely different. The distance between the histograms in this case become senseless. Therefore the uncertainty in the estimation of parameters for each of the bins (at least for one of the histograms) must be extracted from the obtained measured values of random variables. In the general case this is not a straightforward problem but, nevertheless, there exists a class of distributions allowing univalent linking of these uncertainties—the statistically dual distributions. In particular, for self-dual distributions estimation of the "expected number of events in the bin in question for the sample in question" parameter is equal to the observed number of events in the bin and is unbiased. Here the density of the distribution of confidence in the value of the parameter coincides with the distribution of error of measurement of the number of events in the bin of the histogram. Let, for the sake of simplicity, the bins contain enough events in order to approximate the distribution of the errors with normal distribution with average value equal to zero and deviation equal to square root of the number of events in the bin. Then, imitation model of the combination of histograms which could be induced by the flow of events from which the corresponding sample was extracted can be constructed for each of the histograms by Monte-Carlo method. The model in question does not take into account all uncertainties in the evaluated parameters for each of the bins. This procedure, by analogy with generation of replicated sample in the bootstrap methodology, can be called the generation of replicated histogram. Similar technique is used in Refs. [16,17].

Since the number of events in each of the bins is regarded in the method under discussion as the realization of independent random value, then the measured values of the number of events for corresponding bins are compared in the comparison of histograms. The value of "significance of the difference" is the convenient characteristic for such bin-by-bin comparison. Selection of specific representation of this value depends on the solved problem. It is important that if the compared values are the realizations of one and the same random value, then the "significance of the difference" for them is the realization of random value close to the standard normal value. Thus, if the histograms under comparison are compatible, then the distribution of the obtained values of "significances of difference" in each of the bins must also be equal to the standard normal value.

Comparison of methods

Two moments of distributions of significances of differences in the bins of compared histograms, namely, the mean and the square root mean, were used in Ref. [14] as the distance between the histograms, i.e. the test statistics was two-dimensional.

Let two histograms with the number of bins equal to M were obtained as the result of processing of two samples with volumes equal to N_1 and N_2, respectively:

$$hist1 : n_{11} \pm \sigma_{11}, n_{21} \pm \sigma_{21}, \ \ldots, n_{M1} \pm \sigma_{M1} \ \text{ and}$$
$$hist2 : n_{12} \pm \sigma_{12}, n_{22} \pm \sigma_{22}, \ \ldots, n_{M2} \pm \sigma_{M2}.$$

In comparison of these histograms the decision must be made on whether the flows of events $G1$ and $G2$ (from which the processed samples were extracted) belong to one and the same general population, as well as to evaluate the probability of correctness of the decision that they belong to different general populations. Let us introduce the "normalized significance of the difference" in respective bins of the histograms as follows:

$$S_i = (n_{i1} - K \cdot n_{i2})/\left(\sigma_{i1}^2 + K^2 \cdot \sigma_{i2}^2\right)^{1/2}.$$

In this case n_{ik} is the observed number of events in ith bin of histogram k; σ_{ik} is the respective standard deviation; K is the certain normalization factor. Usually, depending on the problem under solution, K is equal to either the ratio of volumes of samples, or to the ratio of durations of time intervals for collection of the samples. Not one-dimensional (as in other methodologies) but, instead, multi-dimensional value is used as the distance between the histograms. In the example examined here the two-dimensional value $SRMS = (S^{\text{cp}}, RMS)$, where S^{cp} is the mean value of the distribution of "normalized significances of difference" and RMS is the standard deviation for this distribution.

$SRMS$ has very straightforward interpretation:

- If $SRMS = (0, 0)$, then the two histograms are identical;
- If $SRMS \approx (0, 1)$, then $G1 = G2$ (if $RMS < 0$, then the samples are partially overlapping, i.e. they are not independent);
- If neither of the above conditions is satisfied, then $G1 \neq G2$.

As it has been already mentioned, distributions of test statistics can be obtained by modeling. Let us examine the results of Monte-Carlo experiment for comparison of the following three methods: statistical comparison of histograms (SCH), Kolmogorov–Smirnov method (KS) and Anderson–Darling method (AD). The purpose is to determine the potential of the three methods for histogram comparison in the

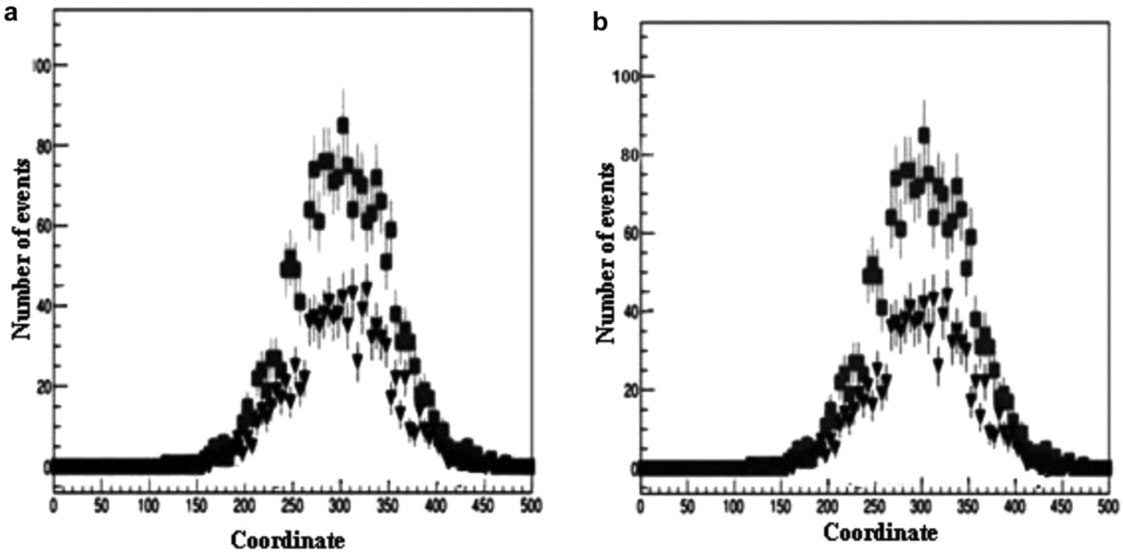

Fig. 1. Examples of distributions of realizations of random variable in the samples from two pairs of flows: a) reference pair of distributions (1000 and 2000 events for *N*(300,50); b) test pair of flows (1000 events for *N*(300,50) and 2000 events for *N*(310,50)).

evaluation of distinguishability of samples in the processing of which the histograms were generated.

Factually, sensitivity of the methods to differences in the flows of events from which the samples were obtained will be compared. Two pairs of independent flows of samples (reference and test flows) consisting of realizations of random variable (each realization is an event) are simulated for this purpose. Volume of each flow is equal to 5000 samples. First flow for each of pairs of flows is the reference flow with volume of the sample equal to 1000 events obtained in the simulation of random variable and satisfying the normal distribution law *N*(300, 50) (Fig. 1a). Selection of the above parameters for the reference distribution was preset by the test program. Second flow from the first pair of flows is also the reference flow with samples with volumes equal to 2000 events (see Fig. 1a). Second flow for the second pair of flows is the test flow with sample with volumes equal to 2000 events with realization of random variable *N*(*X*,*W*), where *X* varies from the value equal to 300 to 310 and *W* varies from 42 to 58.

Distributions of the realized values in the bins of histograms for reference and for test samples from the second pair of flows are shown in Fig. 1b.

Test statistics is calculated when Anderson–Darling (AD) and Kolmogorov–Smirnov (KS) criteria are used for comparison of histograms. Following this the test statistics is converted into the p-variable or the p-value (*p-value*). P-value is the value used in the testing statistical hypotheses. Usually p-value is equal to the probability that the random value with distribution in question (distribution of test statistics for zero-hypothesis) will have the value which is not less than the factual value of the test statistic. P-value in the case under examination has uniform distribution within the interval [0, 1] if the samples were obtained from one and the same flow of events. If the reference and the test samples were obtained from different flows of events then the distribution

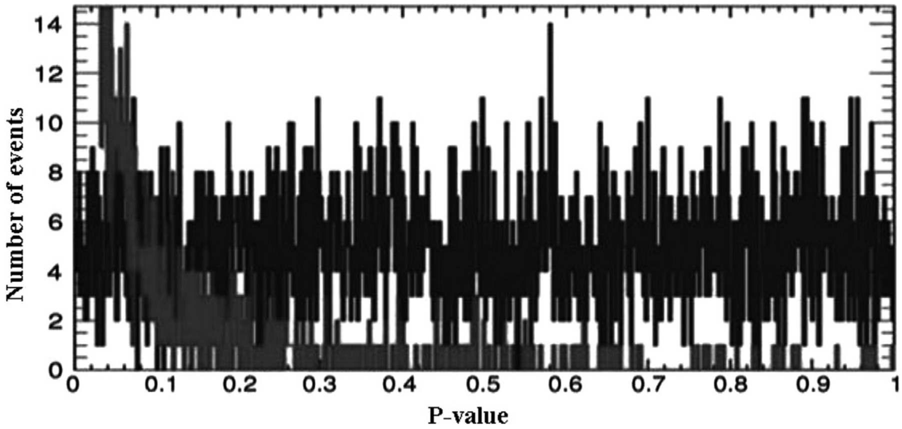

Fig. 2. Distributions of values of p-values (AD-criterion) for two pairs of flows: upper distribution—5000 comparisons for the first pair of flows (reference samples of 1000 and 2000 events for realizations of *N*(300,50)); bottom distribution—5000 comparisons for the second pair of flows (reference sample of 1000 events for *N*(300,50) and test sample of 2000 events for *N*(306,50)).

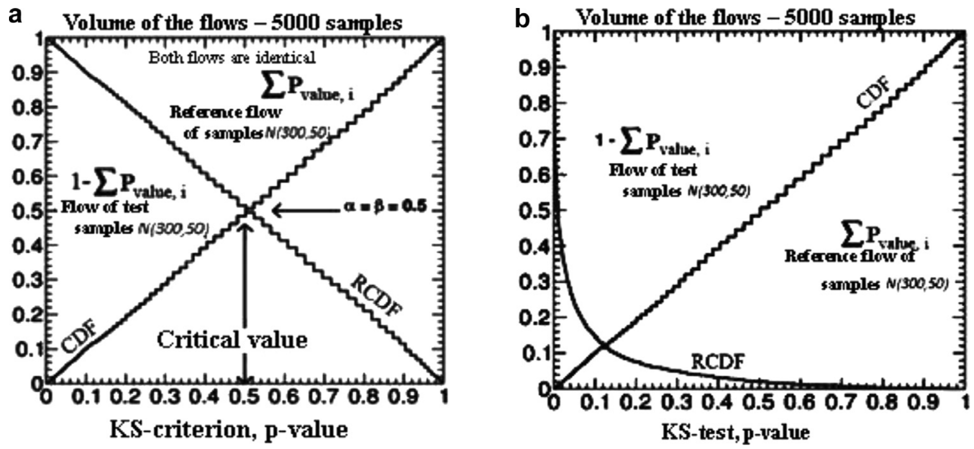

Fig. 3. Kolmogorov–Smirnov criterion. CDF and RCDF for p-values: a) comparison of two reference flows (random variables ~ N(300,50)); b) comparison of samples from the reference flow and samples from the test flow (random variable ~ N(306,50)).

of p-values is concentrated within small values (Fig. 2). 5000 comparisons of pairs of samples taken from different flows were made in each pair of flows of samples.

Empirical cumulative distribution function (CDF) was constructed following this for the first (reference) of the obtained distributions (рис. 3). Empirical inverse cumulative distribution function (RCDF) was constructed for the second (test) distribution (see Fig. 3). Critical value for the equal tailed test is determined by the point of intersection of lines CDF and RCDF. Consequently, errors of the first type (α) and the second type (β) in testing the hypothesis of compatibility of histograms versus the alternative hypothesis that the histograms were constructed in the processing of samples taken from different flows of events are approximately equal ($\alpha \approx \beta$). This allows characterizing the difference between the reference and the test flows by the value $(\alpha + \beta)/2$.

In the case of statistical comparison of histograms (SCH criterion) histogram is plotted for each sample (see Fig. 1). Following this contents of the histograms for respective pairs of samples are used for bin-by-bin calculation of the “normalized significance of the differences” S_i, ($i = 1, \ldots, M$), determination of mean value S^{mean} and the root mean square RMS values of significances of the differences. (S^{mean}, RMS) two-dimensional distribution of the obtained values is constructed for each of the pairs of flows of samples (reference and test samples) (Fig. 4). Using the equal tailed test the critical line is found and errors of the first and second types characterizing the probability of making erroneous decisions in the selection of the main and the alternative hypotheses are calculated. Mean error of decision making $(\alpha + \beta)/2$ is also used for comparison of the criterion under discussion with AD and KS criteria.

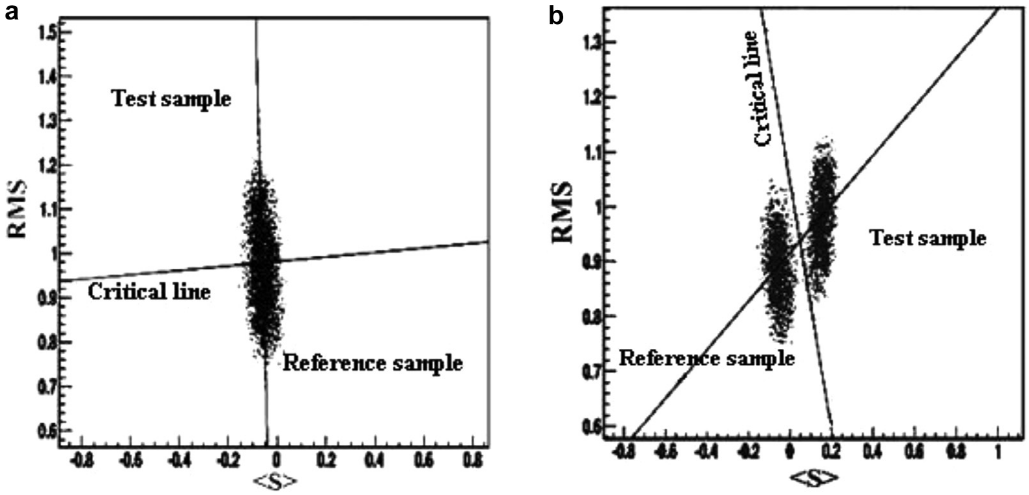

Fig. 4. Distributions (S^{mean}, RMS) for 5000 comparisons of pairs of histograms: a) spot of results for reference histograms below the critical line, spot of results of comparison of test samples (center of the distribution is changed, N(308, 50)) with reference samples above the critical line; b) spot of results for reference histograms to the left from the critical line, spot of results of comparison of test samples (width of the distribution is changed, N(300, 44)) with reference samples to the right from the critical line.

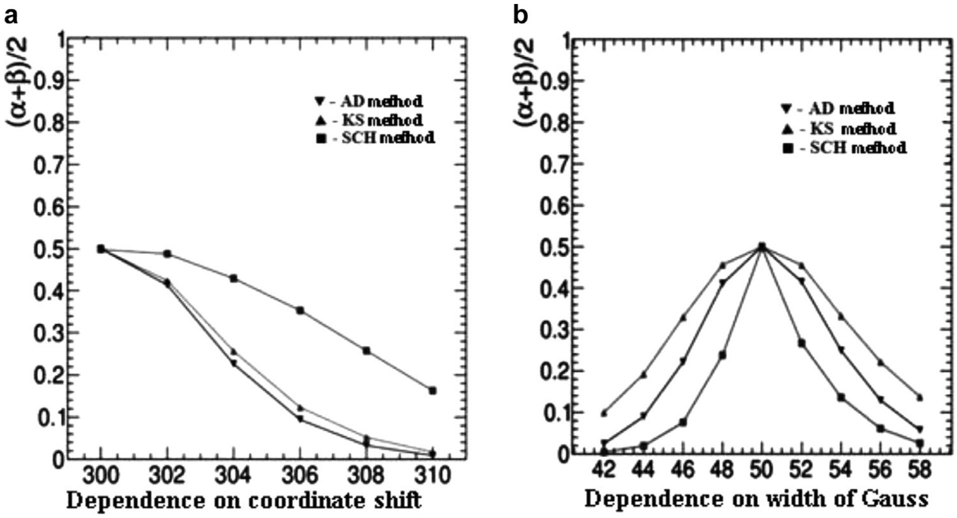

Fig. 5. a) Variation of mean error of decision making in the equal tailed test with varying mathematical expectation for the random variable in the test flow of events. AD- and KS-criteria allow distinguishing the flows of events with higher probability. b) Variation of mean error of decision making with varying width of the distribution of the random variable. Here the SCH criterion works better.

Results of comparison of three methods for histogram comparison are presented in Fig. 5. Results of the present study demonstrated that Anderson–Darling and Kolmogorov–Smirnov criteria better distinguish the flows in which random variables have differences in mathematical expectation. At the same time the method of statistical comparison of histograms better distinguish the flows in which random variables have differences in the widths of distributions. However, the method of statistical comparison of histograms [14] is multi-dimensional and gives the possibility to include any of the one-dimensional test statistics commonly used in the comparisons of histograms as additional dimensions. For instance, inclusion of Anderson–Darling test statistics as the third component of already three-dimensional test statistics in the method of statistical comparison of histograms resolves the problems with distinguishability of flows of events having differences in the values of mathematical expectation. This is a serious advantage of the method in question with respect to other methods discussed in the present review.

Acknowledgments

Authors are grateful to V.A. Kachanov, Yu.A. Korovin and N.V. Krasnikov for their support to the study and to N.A. Korneeva for the fruitful discussions. Results of the study were obtained under financial support of the Ministry of Education and Science of the Russian Federation within the framework of the project implemented under agreement no. 14.610.21.0004 dated October 17, 2014, identifier PNIER RFMEFI61014×0004.

References

[1] Y. Ioannidis, in: Proceedings 2003 VLDB Conference, 2003, pp. 19–30.
[2] S.-H. Cha, S.N. Srihari, Pattern Recognit. 35 (6) (2002) 1355–1370.
[3] A.N. Kolmogorov, Ann. Math. Stat. 12 (4) (1941) 461–463.
[4] S. Kullback, Information Theory and Statistics, Wiley, New York, 1959.
[5] J. Rosenthal, SIAM Rev. 37 (1995) 387–485.
[6] W. Cochran, Ann. Math. Stat. 23 (3) (1952) 315–342.
[7] T. Kailath, IEEE Trans. Commun. Technol. 15 (1) (1967) 52–60.
[8] E. Hellinger, J. die reine angew. Math. 1909 (136) (1909) 210 (in German).
[9] T.W. Anderson, D.A. Darling, Ann. Math. Stat. 23 (2) (1952) 193.
[10] Gretton A., Borgwardt K., Rasch M.J., Scholkopf B., Smola A.J., A Kernel Method for Two-sample Problem, arXiv:0805.2368, 2008.
[11] H.B. Mann, D.R. Whitney, Ann. Math. Stat 18 (1) (1947) 50.
[12] H. Bandemer, W. Nather, Fuzzy Data Analysis, Kluwer Academic Publishers, Dordrecht, 1992.
[13] P. Luuka, M. Collan, in: J.M. Alonso, H. Bustince, M. Reformat (Eds.), Proceedings of IFSA-EUSFLAT2015, Atlantis Press, 2015.
[14] Bityukov S., Krasnikov N., Nikitenko A., Smirnova V., A Method for Statistical Comparison of Histograms, arXiv:1302.2651, 2013.
[15] B. Krupanek, R. Bogacz, Przeglad Elektrotechniczny (11) (2014) 32.
[16] Y. Cao, L. Petzold, J. Comput. Phys. 212 (1) (2006) 6–24.
[17] Xu K.-M., Using the Bootstrap Method for a Statistical Significance Test of Differences between Summary Histograms, NASA Technical Reports Server, ID: 20080015431, 2006.