Multidimensional Comparison of Histograms

S.I. BITYUKOV
Department of experimental physics
Institute for high energy physics National Research Centre "Kurchatov institute"
Protvino, Moscow region, 142281
RUSSIA
serguei.bitioukov@cern.ch   http://cern.ch/bityukov

N.V. KRASNIKOV
Theory physics department
Institute for nuclear research RAS
prospekt 60-letiya Oktyabrya 7a, Moscow 117312
RUSSIA
Nikolai.Krasnikov@cern.ch

A.V. MAKSIMUSHKINA
Faculty of Natural Science, Department of general and special physics
National research nuclear university MEPhI (Moscow engineering physics institute)
Kashirskoe sh., 31, Moscow, 115409
RUSSIA
AVMaksimushkina@mephi.ru

V.V. SMIRNOVA
Department of mathematics and computing
Institute for high energy physics National Research Centre "Kurchatov institute"
Protvino, Moscow region, 142281
RUSSIA
Vera.Smirnova@ihep.ru

Abstract: - Methods of histograms comparison in physical researches are considered. A new approach for the multidimensional comparative analysis of histograms is proposed.

Key-Words: - Distribution theory, Monte Carlo method, measurements and theory of errors, data analysis

1 Introduction

The term ``histogram" was coined by the famous statistician Karl Pearson to refer to a ``common form of graphical representation" [1]. Histograms are very useful in their canonical visual representation, but today histograms are considered as purely mathematical objects.

Histograms are used in different scientific fields. Besides physics data analyses, histograms play a very important role in databases, image processing, computer vision [1]. Correspondingly, goals and methods of the treatment of histograms are varied in dependence to the area of application. In this paper histograms are considered in frame of tasks related to physical experiments.

2 Histogram

Let us call the appearance of the realization of the random variable (or random variables) as the event. Suppose, there is given a set of non-overlapping intervals. A histogram represents the frequency distribution of data which populates those intervals. This distribution is obtained during data processing of the sample taken from the flow of events. These intervals usually are called as bins.

The filling procedure of a histogram influences the analysis of histogram. There are two extreme cases. The first case: one event produces one histogram. For example, the distribution of brightness in a photo is a result of data processing of one event. Here one sample consists from one event and one event is one photo.

ISBN: 978-1-61804-354-2

The second case: one event is one measurement of random variable and resulting value is put to histogram. The filling of the histogram is a chain of independent measurements with gradual filling of the histogram. The second case is used, usually, in physical experimental researches for data processing. Correspondingly, the content of the bin is called the number of events in the bin, the sum of contents of bins in histogram is a volume of the histogram.

Common issues of are in construction of histograms, for example, the choice of optimal binning and the choice of the model for distribution of errors for observed values in the bins.

2 Comparison of Histograms

Given two histograms, how do we assess whether they are similar or not? What does it means "similar"? Several standard procedures exist for this task.

Suppose, a reference histogram is known. Usually, the proximity of test histogram and reference histogram is measured via a test statistic, that provides the quantitative expression of the ``distance'' between histograms [2]. The smaller the distance the more similar they are.

2.1 ``Distance'' between histograms

There are several definitions of distance in the literature, for example, the Kolmogorov distance [3], the Kullback-Leibner [4] distance, the total variation distance [5], the chi-square distance [6] and so on. Usually, it is the some test statistics, distribution of which can be calculated via formulae or constructed by Monte Carlo method. Other approach is based on the fact that a histogram of a measurement provides the basis for an empirical estimate of the probability density function (pdf) [7]. Computing the distance between two pdfs can be regarded as the same as computing the Bayes (or minimum misclassification) probability. This is equivalent to measuring the overlap between two pdfs as the distance. Sometimes, the Bhattacharyya distance [8] (or Hellinger distance [9]) is used as the distance between two pdfs. Note, that the Kolmogorov distance [3], the Anderson-Darling distance [10], the Kullback-Leibner distance [4] also allow to compare samples of events without their presentation in form of histograms. Recently, the test based on the maximum mean discrepancy (MMD) [11] was appeared. The important approach for comparison of histograms is tests based on ranks and/or permutations (Mann-Whitney [12], …). In the vector approach, a histogram is treated as a fixed-dimensional vector. Hence standard vector norms such as city block, Euclidean or intersection can be used as distance measures [13]. Similarity measures can be used in the comparing histograms. For example, the method of modulo similarity [14] is based on Lukasiewicz logic [15].

2.2 Testing of consistency of histograms or distinguishability of histograms

Also, a goal of histogram comparison is a testing of their consistency [16] or vice versa of their distinguishability [17]. Consistency here is the statement that both histograms are produced during data processing of independent samples which are taken from the same flow of events (or from the same population of events). In paper [18] is proposed approach which allows to estimate the distinguishability of histograms and, correspondingly, the distinguishability of parent events flows. The method is based on the statistical comparison of histograms. The multidimensional test statistic is used as a distance between histograms. The modification [19] of this method is used for the detection of the changing of parameters in the context of wireless transmission.

If the goal of the comparison of histograms is the check of their consistency, then task is reduced to hypotheses testing: main hypothesis H0 (histograms are produced during data processing of samples taken from the same flow of events) against alternative hypothesis H1 (histograms are produced during data processing of samples taken from different flows of events). In principle, the choice between main and alternative hypothesis depends on the task. The determination of critical area allows to estimate Type I error (α) and Type II error (β) in decision about choice between H0 and H1. The Type I error is a probability of mistake if done choice is H1, but H0 is true. The Type II error is a probability of mistake if done choice is H0, but H1 is true. The selection of a significance level (α) allows to estimate the power of the test ($1-\beta$). Usually, values of significance level are 10%, 5%, 1%. If both hypotheses are equivalent, then other combinations of the α and β are used. For example, in task about distinguishability of the flows of events works a relative uncertainty $(\alpha+\beta)/(2-(\alpha+\beta))$ [20]. Under the test of equal tails [21] the mean error $(\alpha+\beta)/2$ can be used.

2.3 Other goals of comparison of histograms

Many other goals of comparison of histograms exist.

For example, the search for anomalous structures in test histogram in comparison with reference histogram is a very important task in particle physics. Possible solution is the comparison of the contents of two histograms, bin by bin. In this case the probability that both bins were produced from a distribution with the same mean is calculated.

Also, the method for sorting events of multiparticle production according to the anisotropy of their momentum distribution by the use of histograms is presented in paper [22].

2.4 Comparison of normalization and comparison of shape

The histograms comparison can usually be decomposed into comparison of normalization and comparison of shape. Sometimes the normalization and the shape are not independent, so the decomposition till works but it becomes more difficult to come up with a meaningful combination of the two tests. In the simplest case, normalization can be estimated by common suppositions. It may be the ratio of the volumes of samples corrected due to any additional knowledge (for example, efficiencies of registration of events). It may be the ratio of times for gathering samples and so on. A vast amount of statistical literature is devoted to the theme of shape comparisons (see, for example, [23]).

2.5 “Rehistogramming”

The hypotheses testing require the knowledge of the distribution of test statistics. As mentioned above the distribution of test statistics can be constructed by Monte Carlo. Let us consider the simple case of the filling of histograms - event-by-event in frame of the method of statistical comparison of histograms [18,17]. The number of events in each bin of histogram can be considered as a realization (observed value) of the random variable with parameter “the expected number of events in given bin of histogram for given sample”. The knowledge of uncertainty of the observed value in the case of statistically dual distributions [24] allows to describe the uncertainty of the corresponding value of parameter. If we work with Poisson flows of events, then uncertainty of the parameter obeys the gamma distribution. If we work with Gaussian approximation, then the distribution of uncertainty obeys the normal distribution. As a result we can use the Monte Carlo method for construction of two imitation models of the possible histograms sets. These two sets of histograms imitate the two general populations (two models) which provided us two histograms for comparison. This procedure can be named as "rehistogramming" , similar to "resampling" in bootstrap technique [25]. The first imitation population (the first set of histograms) is used for construction of the distribution of test statistics for the case of H0 hypothesis. The second imitation population (the second set of histograms) is used for construction of the distribution of test statistics for the case of H1 hypothesis. The comparison of these distributions allows to estimate the uncertainty in the hypotheses testing [18, 17]. The similar approaches for histograms comparison is described in papers [26, 27] too.

2.6 “Significance of the difference”

The convenient characteristic for comparison of histograms is a distribution of the “significances of the difference”. The “significance of the difference” is calculated for corresponding pairs of bins of the comparing histograms. The choice of type of “significance of the difference” depends on the task [28]. If the comparing histograms are taken from the same population of histograms (or the corresponding samples are taken from the same flow of events), the distribution of “significances of the difference” is close to standard normal distribution.

2.7 Multidimensional comparison

As mentioned above, the method of statistical comparison of histograms [18, 17] is a multidimensional method. It allows to include any one-dimensional test statistic as an additional component of multidimensional test statistic in the frame of the method. For example, the including of the Anderson-Darling test statistic into this method as additional component of the multidimensional test statistic allows to improve the distinguishability of histograms.

3 Conclusion

Possible approaches for the comparative analysis of histogram are considered. As shown, there is no single best test for all applications. It means that before application any test must be checked with care.

As a good solution of the problem with the comparison of histograms for the distinguishing of flows of events under studing we propose a combined use of several tests in frame of multidimensional test statistic.

The work was supported by the Ministry of Education and Science RF (Agreement on October 17, 2014 N 14.610.21.0004, id. PNIER RFMEFI61014X0004). The authors would like to thank Prof. V. Kachanov, Prof. Yu. Korovin, Prof. N.

Krasnikov, Dr. S. Gleizer and Dr. N. Korneeva for helpful discussions.

References:

[1] Y. Ioannidis, The history of histograms (abridged), *Proceedings 2003 VLDB Conference*, pp.19-30, 2003.

[2] S.-H. Cha, S.N. Srihari, On measuring the distance between histograms, *Pattern Recognition*, Vol.35, 2002, pp. 1355-1370.

[3] A.N. Kolmogorov, Confidence limits for an unknown distribution function, *Ann.Math.Stat.* Vol.12, 1941, pp.461-463.

[4] S. Kullback, *Information Theory and Statistics*, Wiley, New York, 1959.

[5] J. Rosenthal, Convergence rates for Markov chains, *SIAM Rev.* Vol.37, 1995, pp. 387-485.

[6] W. Cochran, The chi-square test of goodness of fit, *Ann.Math.Stat.*, Vol.23, 1952, pp. 315-342.

[7] R.O. Duda, P.E. Hart, *Pattern Classification and Scene Analysis*, 1st Edition, Wiley, New York, 1973.

[8] T. Kailath, The divergence and Bhattacharyya distance measures in signal selection, *IEEE Trans.Commun.Technol. COM-15*, Vol.1, 1967, pp. 52-60.

[9] E. Hellinger, Neue Begrundung der Theorie quadratischer Formen von unendlichvielen Veranderlichhen, *Journal fur die reine und angewandte Mathematik, (Crelle's Journal)* Vol. 1909, Issue 136, pp. 210-271.

[10] T.W. Anderson, D.A. Darling, Asymptotic theory of certain «goodness of fit» criteria based on stochastic processes, *Ann. Math. Statist.*, Vol.23, 1952, pp. 193-212.

[11] A. Gretton, K. Borgwardt, M.J. Rasch, B. Scholkopf, A.J. Smola, A Kernel method for two-sample problem, *arXiv:0805.2368*, 2008.

[12] H.B. Mann, D.R. Whitney, On a test of whether one of two random variables is stochastically larger than the other, *Annals of Mathematical Statistics*, Vol.18, 1947, pp. 50-60.

[13] H. Bandemer, W. Nather, *Fuzzy data analysis*, Kluwer academic publishers, Dordrecht, 1992.

[14] P. Luuka, M. Collan, Modulo similarity in comparing histograms, Proc. of IFSA-EUSFLAT2015, Eds. J.M. Alonso, H. Bustince, M. Reformat, Atlantis Press, 2015, pp. 393-397.

[15] J. Lukasiewiicz, *Selected Works*, North-Holland Publishing co., Amsterdam, 1970.

[16] F. Porter, Testing Consistency of Two Histograms, *arXiv:0804.0380*, 2008.

[17] S. Bityukov, N. Krasnikov, A. Nikitenko, V. Smirnova, On the distinguishability of histograms, Eur.Phys.J.Plus, Vol.128, No.143, 2013, pp.1-6.

[18] S. Bityukov, N. Krasnikov, A. Nikitenko, V. Smirnova, A method for statistical comparison of histograms, *arXiv:1302.2651*, 2013.

[19] B. Krupanek, R. Bogacz, Comparison algorithm of multimodal histograms from wireless transmission, *Przeglad Electrotechniczny*, Vol. 11, 2014, pp. 32-34.

[20] S.I. Bityukov, N.V. Krasnikov, Distinguishability of hypotheses, *Nucl.Instr.& Meth. A*, Vol.534, 2004, pp. 152-155.

[21] S.I. Bityukov, N.V. Krasnikov, *The use of statistical methods for the search for new physics at the LHC* (in Russian), Second edtion, Krasand, Moscow, 2014.

[22] R. Kopecna, B. Tomasik, Event shape sorting, *arXiv:1506.06776*, 2015.

[23] O. Thas, *Comparing Distributions*, Springer Series in Statistics, 2010.

[24] S. Bityukov, N. Krasnikov, S. Nadarajah, V. Smirnova, Statistically Dual Distributions and Estimations, *Applied Mathematics*, Vol.5, 2014, pp. 963-968.

[25] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall, New York, 1993.

[26] Y. Cao, L. Petzold, Accuracy limitations and the measurement of errors in the stochastic simulation of chemically reacting systems, *J. of Computational Physics*, Vol.212, 2006, pp. 6-24.

[27] K.-M. Xu, Using the bootstrap method for a statistical significance test of differences between summary histograms, *NASA Technical Reports Server*, ID: 20080015431, 2006.

[28] S. Bityukov, N. Krasnikov, A. Nikitenko, V. Smirnova, Two approaches to Combining Significances, *Proceedings of Science (ACAT08)* Vol.118, 2008, pp. 1-12.