UDC 519.254 PACS 02.50.Ng 06.20.Dk 07.05.Kf

A Method for Statistical Comparison of Histograms

S. I. Bityukov[*], N. V. Krasnikov[†],
A. N. Nikitenko[‡], V. V. Smirnova[*]

[*] *Institute for High Energy Physics*
1, pl. Nauki, Protvino, Moscow region, Russia, 142281
[†] *INR RAS*
7a, prospekt 60-letiya Oktyabrya, Moscow, Russia, 117312
[‡] *Imperial College Sci., Tech. & Med, London, UK*
[§] *Joint Institute for Nuclear Research*
6, Joliot-Curie str., Dubna, Moscow region, Russia, 141980

The problem of the testing the hypothesis that two histograms are drawn from the same distribution is a very important problem in many scientific researches. There are several approaches to formalize and resolve this problem. Usually, one-dimensional test statistics is used for this purpose. We propose an approach for testing the hypothesis that two realizations of the random variables in the form of histograms are taken from the same statistical population (i.e. two histograms are drawn from the same distribution). The approach is based on the notion "significance of deviation", which has a distribution close to standard normal distribution if both histograms are drawn from the same distribution. This approach allows to estimate the statistical difference between two histograms using multi-dimensional test statistics. The distinguishability of histograms is estimated with the help of the construction a number of clones (rehistograms) of the observed histograms. The approach considered in the paper allows to perform the comparison of histograms with a test more powerful, in the cases considered, than those that use only one test statistic. Also, the probability of correct decision is used as an estimate of the quality of the decision about the distinguishability of histograms.

Key words and phrases: distribution theory and Monte Carlo studies, Measurement and error theory, Data analysis: algorithms and implementation; data management, estimation of parameters, flow of events, hypotheses testing.

1. Introduction

The test of the hypothesis that two histograms are drawn from the same distribution is an important goal in many applications. For example, this task exists for the monitoring of the experimental equipment in particle physics experiments. Let the experimental facility register the flow of events during two independent time intervals $[t_1, t_2]$ and $[t_3, t_4]$. Events from first time interval belong to statistical population of events G_1, events from second time interval belong to statistical population of events G_2. If facility (beam, detectors, data acquisition system, ...) is in norm during both time intervals then the properties of events, registered in the facility during time interval $[t_1, t_2]$, is the same as the properties of events, registered in the facility during time interval $[t_3, t_4]$, i.e. $G_1 = G_2$. If facility is out of norm during one of time intervals then the properties of events from statistical population G_1 differ from the properties of events from statistical population G_2, i.e. $G_1 \neq G_2$. Often the monitoring of the experimental facility is performed with the use of the comparison of histograms, which reflect the properties of events.

Several approaches to formalize and resolve this problem were considered [1]. Recently, the comparison of weighted histograms was developed in paper [2]. Usually, one-dimensional test statistics is used for the comparison of histograms.

Received 30^{th} December, 2013.

The authors are grateful to L. Demortier, T. Dorigo, L.V. Dudko, V.A. Kachanov, L. Lyons, V.A. Matveev, L. Moneta and E. Offermann for the interest and useful comments. The authors would like to thank V. Anikeev, Yu. Gouz, E. Gushchin, A. Karavdina, D. Konstantinov, N. Minaev, A. Popov, V. Romanovskiy, S. Sadovsky and N. Tsirova for fruitful discussions. This work is supported by RFBR grant No.13-02-00363.

In this paper we propose a method which allows to estimate the value of statistical difference between histograms with the use of several test statistics. As example, we consider the case of two test statistics, i.e. bidimensional test statistic.

2. Distribution of Test Statistics

Suppose, there are two histograms $hist_1$ and $hist_2$ (with M bins in each histogram) as a result of the treatment of two independent samples of events. The first histogram is a set of 2M numbers $hist_1: \ \hat{n}_{11} \pm \hat{\sigma}_{11}, \hat{n}_{21} \pm \hat{\sigma}_{21}, \ldots, \hat{n}_{M1} \pm \hat{\sigma}_{M1}$ and the second histogram, correspondingly, is a set of 2M numbers also $hist_2: \ \hat{n}_{12} \pm \hat{\sigma}_{12}, \hat{n}_{22} \pm \hat{\sigma}_{22}, \ldots, \hat{n}_{M2} \pm \hat{\sigma}_{M2}$. The volume of the first sample is N_1, i.e. $N_1 \equiv \sum\limits_{i=1}^{M} \hat{n}_{i1}$ and the volume of the second sample is N_2, i.e. $N_2 \equiv \sum\limits_{i=1}^{M} \hat{n}_{i2}$.

The most of methods for the histograms comparison use single test statistic as a "distance measure" for the consistency of two samples of events (see, for example [1]).

We propose[1] to use test statistics $\hat{S}_i, \quad i = 1, \ldots, M$ (significances of deviation[2]) for each bin for the histograms comparison. In the case of two observed histograms we consider the significance of deviation of the following type:

$$\hat{S}_i = \frac{\hat{n}_{i1} - K\hat{n}_{i2}}{\sqrt{\hat{\sigma}_{i1}^2 + K^2 \hat{\sigma}_{i2}^2}}\,. \qquad (1)$$

Here $K = \dfrac{N_1}{N_2}$ is a coefficient of the normalization. We use two first statistical moments $\bar{S} = \left(\sum\limits_{i=1}^{M} \hat{S}_i\right)/M$, and $RMS = \sqrt{\left(\sum\limits_{i=1}^{M} (\hat{S}_i - \bar{S})^2\right)/M}$. If condition $G_1 = G_2$ (G_1 and G_2 are taken from the same flow of events) takes place then test statistics ($\hat{S}_i, \quad i = 1, \ldots, M$) obey the distribution which close to the standard normal distribution $\mathcal{N}(0,1)$. Correspondingly, the distribution of these test statistics is close to standard normal distribution too. In this case our bidimensional test statistic ("distance measure between two observed histograms") $SRMS = (\bar{S}, RMS)$ has a clear interpretation:

– if $SRMS = (0,0)$ then histograms are identical;
– if $SRMS \approx (0,1)$ then $G_1 = G_2$ (if $\bar{S} \approx 0$ and $RMS < 1$ then the overlapping exists between samples);
– if previous relations are not valid then $G_1 \neq G_2$.

Note that the relation

$$RMS^2 = \frac{\hat{\chi}^2}{M} - \bar{S}^2\,, \quad \hat{\chi}^2 = \sum_{i=1}^{M} \hat{S}_i^2 \qquad (2)$$

[1] Some details are in ref. [3].

[2] In paper [4] several types of significances of deviation (or significance of an enhancement [5]) between two values were considered:

A. expected significance of deviation between two expected realizations of random variables (for example, S_{c12} [4]);

B. significance of deviation between the observed value and expected realization of random variable (for example, S_{cP} [4]);

C. significance of deviation between two observed values.

As shown (in particular, in paper [4]), many of these significances obey the distribution close to the standard normal distribution if both values are taken from the same statistical population. In this paper the significance of type C is considered.

shows that test statistic $\hat{\chi}^2$ is a combination of two test statistics RMS and $\bar{S}$.

3. Rehistogramming

An accuracy of the estimation of statistical moments depends on the number of bins M in histograms and observed values in bins. The accuracy can be estimated via Monte Carlo experiments. Two models of the statistical populations (pseudo populations) can be produced. Each of models represents one of the histograms.

In considered below example for each of histograms we produced 4999 clones by the Monte Carlo simulation for each bin i of histogram k using the normal distribution $\mathcal{N}(\hat{n}_{ik}, \hat{\sigma}_{ik})$, $i = 1, \dots, M$, $k = 1, 2$. As a result there are 5000 pairs of histograms for comparisons. The comparison is performed for each pair of histograms (5000 comparisons in our example). The distribution of the significances $\hat{S}_i$ is obtained as a result of each comparison. The moments of this distribution are calculated (in our case $\bar{S}$ and RMS). It allows to estimate the errors in determination of statistical moments.

This procedure can be named as “rehistogramming” in analogy with ”resampling” in the bootstrap method [6].

4. Distinguishability of Histograms

The estimation of the distinguishability of histograms is performed with the use of hypotheses testing. “A probability of correct decision” $(1-\tilde{\kappa})$ about distinguishability of hypotheses [7] is used as a measure of the potential in distinguishing of two flows of events (G_1 and G_2) via comparison of histograms ($hist_1$ and $hist_2$).

It is a probability of the correct choice between two hypotheses “the histograms are produced by the treatment of events from the same event flow (the same statistical population)” or “the histograms are produced by the treatment of events from different event flows”. The value $1 - \tilde{\kappa}$ characterizes the distinguishability of two histograms.

For $1 - \tilde{\kappa} = 1$ the distinguishability of histograms is 100%, i.e. histograms are produced by the treatment of events from different event flows.

For $1 - \tilde{\kappa} = 0$ we can’t distinguish the histograms, i.e. histograms are produced from the same event flow.

The probability of correct decision $1 - \tilde{\kappa}$ is a function of type I error (α) and the type II error (β) testing, namely[1]

$$1 - \tilde{\kappa} = 1 - \frac{\alpha + \beta}{2 - (\alpha + \beta)}\,. \tag{3}$$

5. Example

Let us consider a simple model with two histograms in which the random variable in each bin obeys the normal distribution

$$\varphi(x_{ik}|n_{ik}) = \frac{1}{\sqrt{2\pi}\sigma_{ik}}\, \mathrm{e}^{-\frac{(x_{ik}-n_{ik})^2}{2\sigma_{ik}^2}}\,.$$

Here the expected value in the bin i is equal to n_{ik} (in this example $n_{i1} = i$) and the variance σ_{ik}^2 is also equal to n_{ik}. k is the histogram number ($k = 1, 2$). This model can be considered as the approximation of Poisson distribution by normal distribution.

[1]The type I error α is the probability to accept the alternative hypothesis if the main hypothesis is correct. The type II error β is the probability to accept the main hypothesis if the alternative hypothesis is correct.

All calculations, Monte Carlo experiments and histograms presentation in this paper are performed using ROOT code [8]. Histograms are obtained from independent samples.

The example with histograms produced from the same events flow during unequal independent time ranges (Fig. 1) shows that the standard deviation of the distribution in the picture (right, down) can be used as an estimator of the statistical difference between histograms (this distribution is close to $\mathcal{N}(0,1)$).

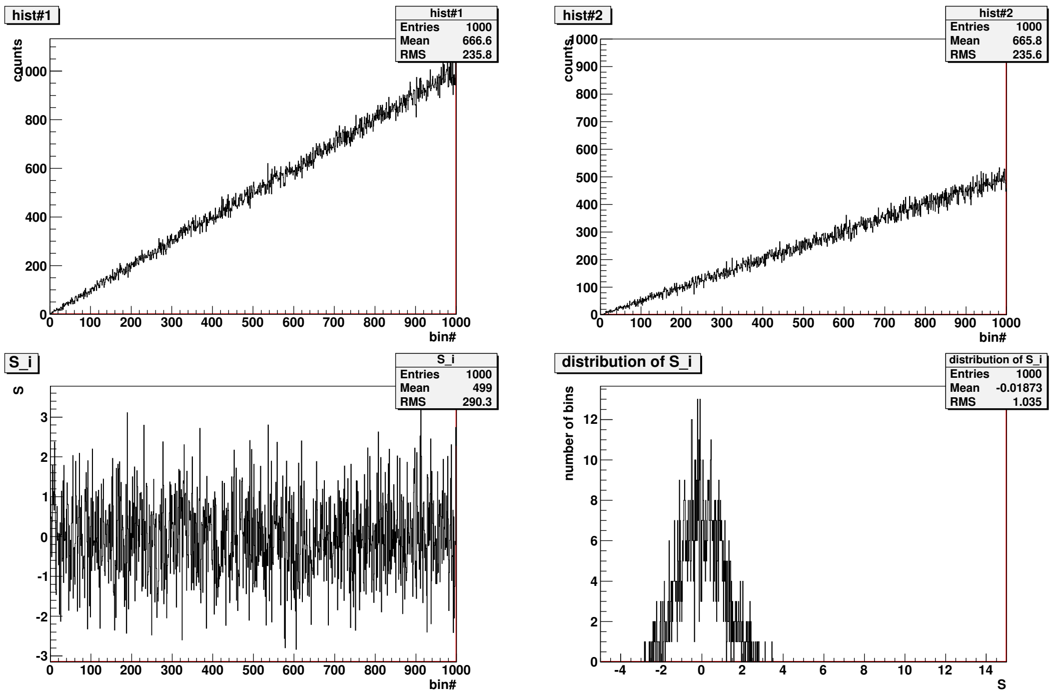

Figure 1. Triangle distributions in histograms ($M = 1000$, $K = 2$): the observed values $\hat{x}_{i1}$ in the first histogram (left, up), the observed values $\hat{x}_{i2}$ in the second histogram (right, up), observed normalized significances $\hat{S}_i$ bin-by-bin (left, down) and the distribution of observed normalized significances (right, down)

At first we consider the Case A (Fig. 2) when both histograms ($hist1$ and $hist2$) are obtained from the same statistical population. The distributions of test statistic $T_{\chi^2} = \sqrt{\hat{\chi}^2/M}$ and test statistic RMS versus $\bar{S}$ are produced during 5000 comparisons of histograms (by the use of rehistogramming).

After that, the content of second histogram (hist2) was changed (Case B), namely, the expected content of left bin of histogram was increased from $n_{12} = 1.0$ up to $n_{12} = 8.5$, the expected content of right bin of histogram was decreased from $n_{M2} = 300.0$ up to $n_{M2} = 292.5$, the expected content of other bins was changed to conserve linear dependence between contents in bins. The result of the rehistogramming for the Case B is shown in Fig. 3. One can see that distributions of test statistic $T_{\chi^2} = \sqrt{\hat{\chi}^2/M}$ and test statistic RMS versus $\bar{S}$ are shifted.

The probability of correct decision as a measure for distinguishability of two histograms is determined by the comparison of distributions for the Case A and corresponding distributions for the Case B. The critical value $T_{\text{critical}} = 1.06$ is used for comparison of one-dimensional T_{χ^2} distributions. The critical line ($S_{\text{critical}} = 1.2 \cdot RMS_{\text{critical}} - 1.36$) is used for comparison of two-dimensional $RMS\&\bar{S}$ distributions. The results are presented in Tab. 1.

For χ^2 method the probability of the correct decision $(1 - \kappa)$ about the Case realization (A or B) is equal to 87.26%. For the other method the probability of the

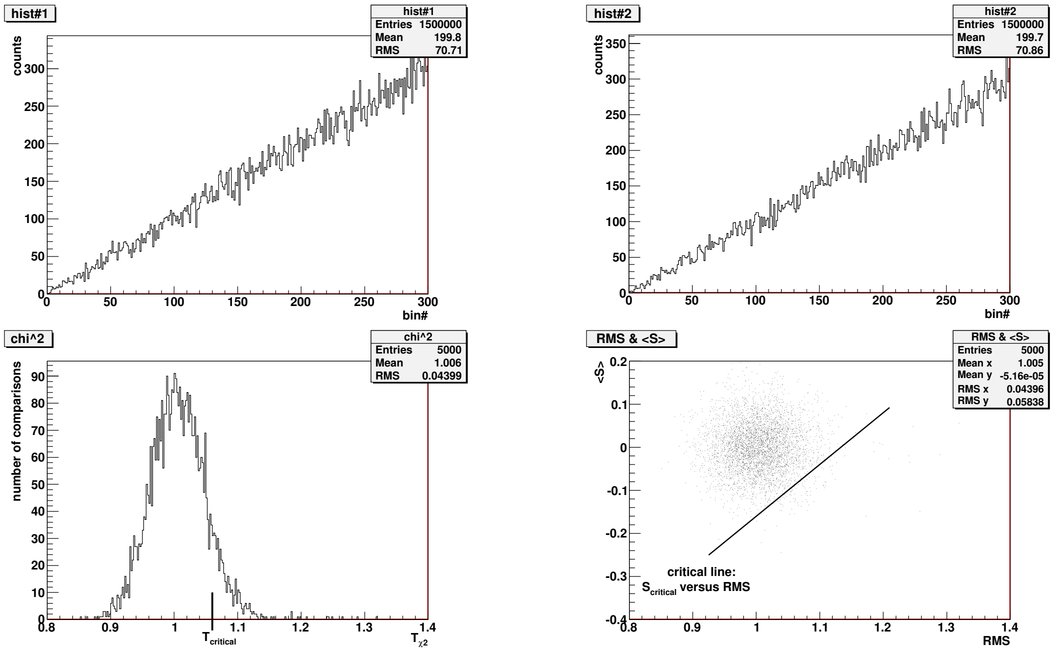

Figure 2. Case A: input histograms (triangle distributions, $M = 300$, $K = 1$) $hist1$ (left, up), $hist2$ (right, up) and T_{χ^2} (left, down), RMS & $\bar{S}$ (right, down) of the distribution of significances for 5000 comparisons for input histograms and their clones

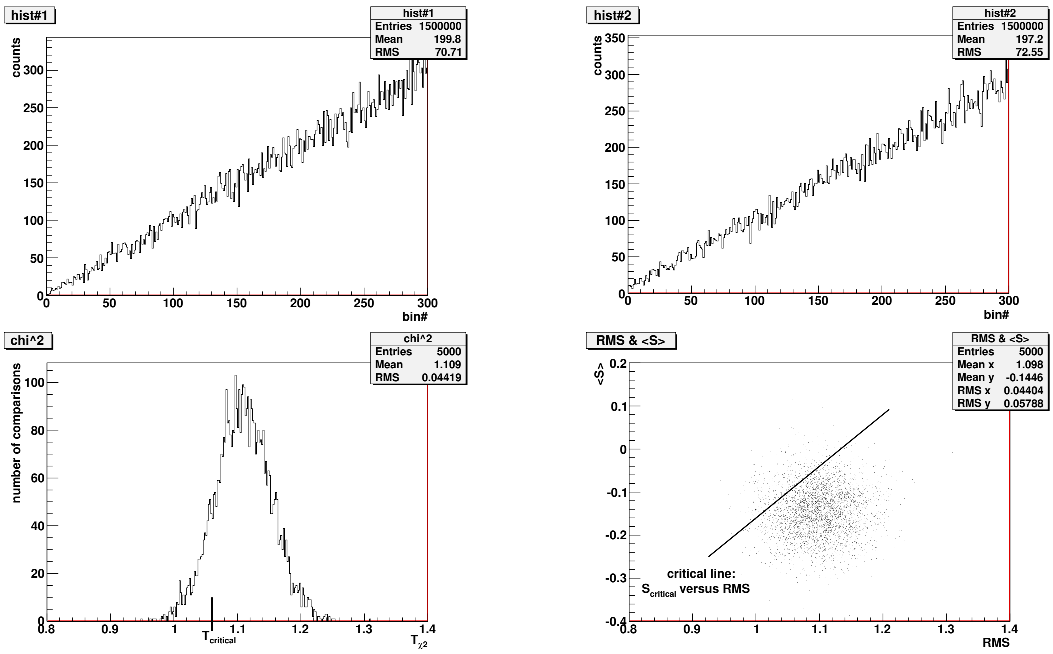

Figure 3. Case B: input histograms (the triangle distribution and the trapezoidal distribution, M=300, K=1) hist1 (left, up), hist2 (right, up) and T_{χ^2} (left, down), RMS & $\bar{S}$ (right, down) of the distribution of significances for 5000 comparisons for input histograms and their clones

Table 1

The quality of the hypothesis testing about distinguishability of two different histograms for two methods of comparison histograms

Distribution of T_{χ^2}			Distribution of $RMS\&S$		
	In reality			In reality	
Accepted	Case A	Case B	Accepted	Case A	Case B
Case A	4543	673	Case A	4843	456
Case B	457	4327	Case B	121	4544
$1-\kappa$	α	β	$1-\kappa$	α	β
0.8726	0.0914	0.1346	0.9388	0.0242	0.0912

correct decision $(1-\kappa)$ about the Case realization (A or B) is equal to 93.88%. One can see that the method, which uses RMS and $\bar{S}$, gives better distinguishability of histograms than the χ^2 method. Note that we use only two moments of the significance distributions (the first initial moment $(\bar{S})$ and the square root from the second central moment (RMS)) for the estimation of distinguishability of histograms.

6. Conclusions

The considered approach allows to perform the comparison of histograms in more details than methods which use only one test statistics. Our method can be used in tasks of monitoring of the equipment during experiments.

The main items of the consideration are

- the normalized significance of deviation provides us the distribution which is close to $\mathcal{N}(0,1)$ if $G_1 = G_2$;
- the rehistogramming provides us the tool for an estimation of the accuracy in the determination of statistical moments and, correspondingly, for testing the hypothesis about distinguishability of histograms;
- the probability of correct decision gives us the estimator of the decision quality.

References

1. *Porter F.* Testing Consistency of Two Histograms. — 2008.
2. *Gagunashvili N. D.* Chi-Square Tests for Comparing Weighted Histograms // Nucl. Instrum. Meth. — 2010. — Vol. 614, No 2. — Pp. 287–296.
3. *Bityukov S. I., Krasnikov N. V., Nikitenko A. N., Smirnova V. V.* A Method for Statistical Comparison of Histograms. — 2013.
4. Two Approaches to Combining Significances / S. I. Bityukov, N. V. Krasnikov, A. N. Nikitenko, V. V. Smirnova // PoS (ACAT08). — 2008. — P. 118.
5. *Frodesen A. G., Skjeggestad O., Toft H.* Probability and Statistics in Particle Physics. — Bergen-Oslo-Tromsø: Universitetsforlaget, 1979.
6. *Efron B.* Bootstrap Methods: Another Look at the Jackknife // Annals of Statistics. — 1979. — Vol. 7, No 1. — Pp. 1–26.
7. *Bityukov S. I., Krasnikov N. V.* Distinguishability of Hypotheses // Nucl. Instrum. Meth. — 2004. — Vol. 534, No 1–2. — Pp. 152–155.
8. *Brun R., Rademaker F.* ROOT — an Object Oriented Data Analysis Framework // Nucl.Instrum.Meth. — 1997. — Vol. 389, No 1–2. — Pp. 81–86.

УДК 519.254 PACS 02.50.Ng 06.20.Dk 07.05.Kf

Об одном методе статистического сравнения гистограмм

С. И. Битюков*, Н. В. Красников†, А. Н. Никитенко‡, В. В. Смирнова*

* *ФГБУ «ГНЦ РФ ИФВЭ» НИЦ «Курчатовский Институт»*
пл. Науки, д. 1, Протвино, Московская область, Россия, 142281
† *ФГБУН ИЯИ РАН*
пр-кт 60-летия Октября, д.7а, Москва, В-312, Россия, 117312
‡ *Империал колледж, Лондон, Великобритания*
§ *Объединённый институт ядерных исследований*
ул. Жолио-Кюри, д. 6, Дубна, Московская область, Россия, 141980

Задача проверки статистической гипотезы о том, что две гистограммы получены при обработке событий, взятых из одной и той же генеральной совокупности событий, ключевая во многих научных исследованиях. Существует несколько подходов к решению данной задачи. Обычно используется одномерная тестовая статистика. Мы предлагаем новый подход к проверке гипотезы о том, что две реализации случайной величины, представленные в виде гистограмм, получены при обработке событий, берущихся из одной и той же генеральной совокупности. Подход основан на понятии «значимость различия». Данная величина вычисляется для каждого бина гистограмм и подчиняется распределению, близкому к стандартному нормальному распределению, если обе гистограммы получены при обработке событий, взятых из одной генеральной совокупности. Предлагаемый метод позволяет определить статистическую разницу между гистограммами при помощи многомерной тест статистики. Различимость гистограмм оценивается через генерацию повторных (подобных) гистограмм для каждой из исходных гистограмм. Данный метод позволяет использовать более мощные критерии различимости гистограмм, чем методы использующие одномерную тест-статистику. Предлагается использовать понятие «вероятность правильного решения» в утверждении о различимости гистограмм как оценку качества принимаемого решения.

Ключевые слова: теория распределений, метод Монте-Карло, теория ошибок, анализ данных, обработка событий, оценивание параметров распределений, поток событий, проверка гипотез.