Histogram comparison used in BESIII software and data validation*

YAO Jian(姚剑)[1;1)] JI Xiao-Bin(季晓斌)[2;2)] LI De-Min(李德民)[1;3)]

[1] Physical Engineering College, Zhengzhou University, Zhengzhou 450001, China
[2] Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China

Abstract The performance of histogram comparison is studied for the various types of hypothesis test methods. The Kolmogorov test is recommended for the software and data validation and the minimum event numbers for different distributions are given in order to get more reliable results. A new bin content comparison method is implemented to deal with the hitmap-like histograms.

Key words histogram comparison, Kolmogorov test, Chi2 test, probability value

PACS 29.85.-c, 02.70.Rr, 02.50.Cw

1 Introduction

Histogram comparison is widely used in the software validation [1] and data quality monitoring at BESIII experiment. There are a number of choices for comparing two histograms, such as the χ^2 test, Geometric test, Kolmogorov-Smirnov test, Anderson-Darling test, Likelihood value test, and so on. There is no single "best" test for all applications [2]. Therefore, it is important to select a reliable and convenient method according to the usage.

In the software validation, the same histograms from two versions of the BOSS [3] release are produced and compared. While in the data quality monitoring, histograms are produced from two different runs during data taking. In most of the cases, the compared histograms are close. We want to know which method holds good for the automatic batch comparison, and how to use it to get a reliable result. So the well-known Kolmogorov-Smirnov (KS) and χ^2 tests are selected as a result of the non-parametric method in the study [4]. Also in the data quality monitoring, many histograms are hitmap-like histograms, a new bin content comparison method is implemented to deal with these histograms. The computations are carried out in the framework of the ROOT [5] package.

2 Study procedure and discriminating power

Two key points in histogram comparison are to find the difference efficiently and to avoid the wrong judgment as much as possible. In the paper, histograms from the same or different parent distributions are studied separately and probability values (p-values) are used in the KS or χ^2 test as a quantitative evaluation.

First, 500 pairs of sample and reference data are created. Each sample and reference data set has N (N = 100, 500, 1000, or 10000) data points taken randomly from the same parent distribution. The parent distribution shapes are chosen to represent some of the distributions that might typically exist in the experiment test data, such as Gaussian, Exponential, and Linear distributions. Compare each of the 500 samples with the corresponding reference by performing the KS or χ^2 test, then, record and plot the calculated p-values.

Figure 1 shows the p-value distribution for Gaussian distribution with the KS test method. As expected, the distribution should be approaching being

Received 3 April 2009, Revised 28 April 2009
* Supported by National Natural Science Foundation of China (10605030)
1) E-mail: yaoj@mail.ihep.ac.cn
2) E-mail: jixb@mail.ihep.ac.cn
3) E-mail: lidm@zzu.edu.cn

flat while the event number is equal to 10000. But actually, it is noticed that the probability distribution is not so flat in the high statistics limit. That is because the KS test is intended for unbinned data, but our research is about binned data. However, there is no real problem with using the KS test for binned data.

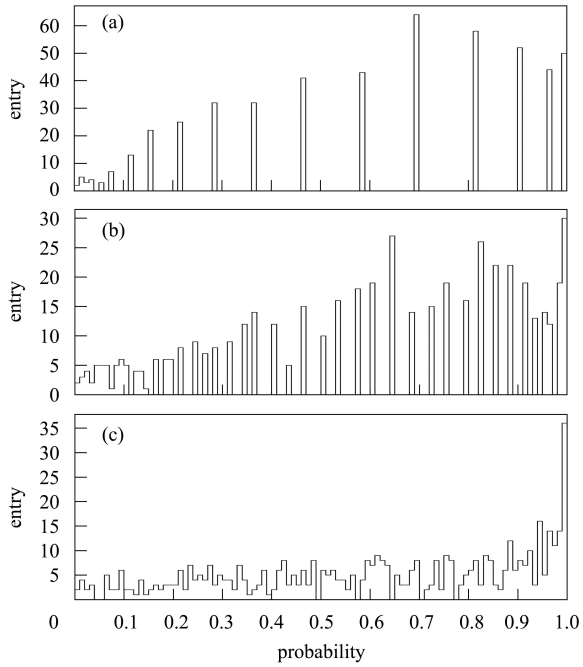

Fig. 1. The p-value distribution of the KS tests for binned data comparing reference data sets to samples where both reference and samples have 100(a), 1000(b), or 10000(c) events.

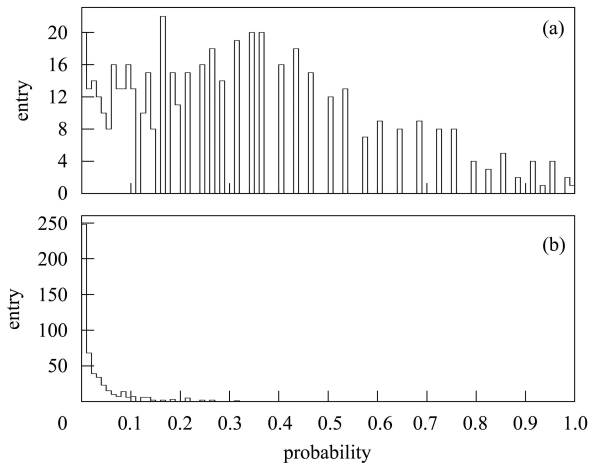

Fig. 2. The p-value distribution of the KS tests for binned data comparing reference data sets (σ=1.0) to samples ((a) σ=0.9; (b) σ=0.8) where both the reference and samples have 1000 events.

Then the similar procedure is repeated for studying histograms from different parent distributions. It is shown that when the Gaussian width of the sample histogram is equal to 0.9, it can not be well discriminated from reference Gaussian distribution with width 1.0 using 1000 events in Fig. 2.

In order to express the discriminating power quantitatively, the following two quantities are used to describe the performance of a given test:

1. False positive fraction: When two distributions are drawn from the same parent function, the fraction of the time that our test (wrongly) determines they are different is called the false-positive fraction.

2. Discrimination efficiency: When two distributions are actually different, the fraction of time that our test finds them different is called the discrimination efficiency.

For a given p-value, both the false positive fraction and the discrimination efficiency can be calculated from the p-value distribution. So the discrimination efficiency versus the false positive fraction curve can be obtained from scanning the p-value from 0 to 1 for a given condition. Several curves for a number of different conditions are shown in Fig. 3. The figure enables us to know how many events are suitable for a kind of distribution during software validation and how much reliability for a given event sample. For example, for detecting about a 10% difference in Gaussian distribution, 7000 event samples give almost 100% discrimination efficiency with a false-positive fraction less than 10%; Linear distributions seem to require large event samples to obtain the same discrimination power, while for Exponential distributions, only 5000 event samples are needed.

The χ^2 test is a commonly used statistic for comparing if two distributions are from the same parent function. Similar figures to Fig. 3 are obtained by replacing the KS test with the χ^2 test. And the results show that the KS test and χ^2 test have a similar discrimination power for high statistics, but when the event number is smaller, the χ^2 test preforms poorly compared with the KS test.

3 Application in software validation

In the software validation, we need to compare histograms from two different releases. Histograms from the previous release are regarded as a reference, and those from the current release are samples. If the calculated p-value from two compared histograms is less than some given value (e.g. 0.1), they are thought to be from different parent distributions, and should be checked by a human being. For example, Fig. 4 is one of the comparing results, and the p-value is

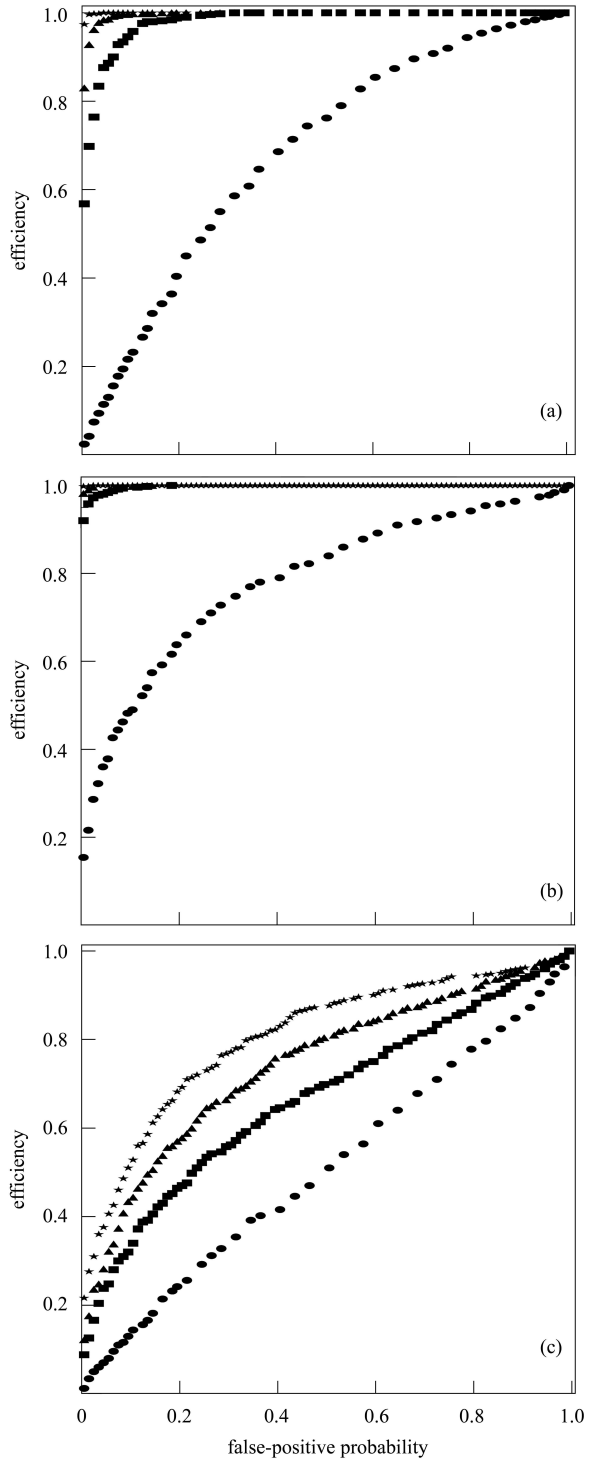

Fig. 3. Efficiency for discriminating Gaussian (a) Exponential (b) and Linear (c) distributions. The results are obtained from KS tests: (stars) $N = 10000$, (triangles) $N = 7000$, (squares) $N = 5000$, (dots) $N = 1000$.

almost equal to 0. The Gaussian fit results show that the difference of width between reference and sample histograms is larger than 10%.

Since the p-value is statistic, a single comparison can not give a definite conclusion. We can just say the two histograms are from one parent distribution or not under some probability. But in the software validation, we can check the p-value distribution from all compared histograms. If most of the histograms are similar, the p-value distribution should be approximately flat. Then we can get the conclusion that this version of software is similar to the previous one.

Figure 5 shows a p-value distribution when we compare BOSS release 6.3.5 and 6.4.0. 77 pairs of histograms are compared with the sample of $J/\psi \to K_S K\pi$. Most of the histograms should be similar according to the figure and the result is confirmed by further checking.

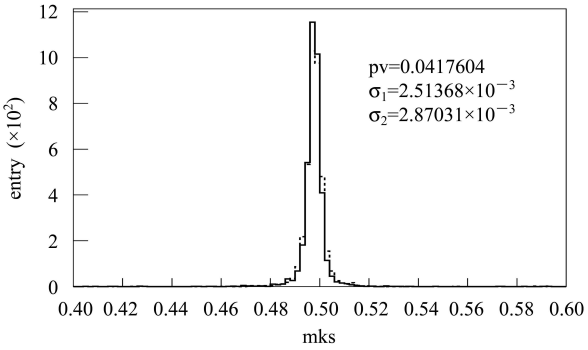

Fig. 4. Histograms comparison and fit results. The histograms represent the mass of K_s in $J/\psi \to K_S K\pi$. The solid line represents the reference histogram (σ_1), and the dot-line represents the sample one (σ_2).

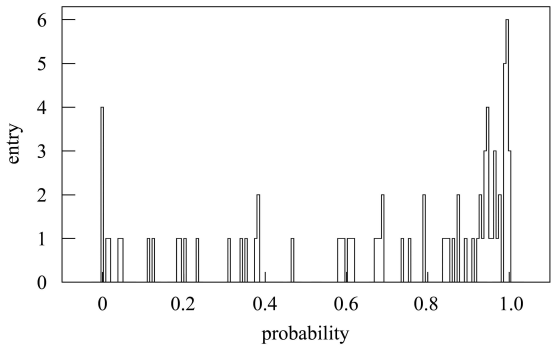

Fig. 5. The p-value distribution between BOSS 6.3.5 and 6.4.0. The p-values are calculated from the histograms filled in the sample $J/\psi \to K_S K\pi$.

4 Comparison method for hitmap-like histograms

In the data quality monitoring, many histograms are hitmap-like histograms. Each bin of this kind of histograms often represents a hardware unit or electronic channel. For most of the cases, there is no obvious interdependence among different bins, and the

number of bins is large. For example, Fig. 6 shows the MDC T hitmap at BESIII during data taking. It has more than 6000 bins, and the distribution is hardly described by an ordinary function. What we are concerned about are the abnormal bins in this kind of histogram, e.g. hot channel or dead channel in the T hitmap. When problems occur, most of the bins are similar to the reference, only a few bins are abnormal. Ordinary histogram comparison methods can not deal with this case. Also since the bin content is continuously increasing during data taking, a simple bin content comparison method can not be used here. So we develop a new bin content comparison method for this kind of histogram. The implementation of the method is described below in detail.

1. The number of events in the ith bin is got, which is named b_{1i} and b_{2i} for the reference and the sample histograms respectively. The errors of b_i are calculated assuming Poisson distribution : $e_i = \sqrt{b_i}$. Then the ratio $b_{3i} = b_{2i}/b_{1i}$ and its error $e_{3i} = b_{3i} \cdot \sqrt{1/b_{1i} + 1/b_{2i}}$ are filled into a new histogram H. If $b_{1i} = 0$, both b_{3i} and e_{3i} are set to 0. And if $b_{2i} \neq 0$, a warning message will be given during comparison. Another case is $b_{2i} = 0$ but $b_{1i} \neq 0$. This may be a normal case when the statistics is low in the bin. We make a conservative assumption and b_{2i} is set to 1 in the calculation.

2. The new histogram H is fitted with a horizontal straight line. The fitted parameter c and its error σ reflect relative scale and fluctuation for the ratio b_{3i} of all bins.

3. Then a new error in the ith bin is defined for the histogram H: $e_{4i} = \sqrt{e_{3i}^2 + \sigma^2}$. It can be used to decide whether the ith is abnormal. For example, we can define the ith bin of the sample histogram is abnormal compared with the reference histogram if $b_{3i} \geqslant c + 5e_{4i}$ or $b_{3i} \leqslant c - 5e_{4i}$, where c is the fitted parameter in Step 2, the ith bin may have some problems, and a warning message will be given.

The advantages of this method are that relative fluctuations among bins are allowed, and almost no wrong warning messages are given when the statistics is low. It is suitable for monitoring hitmap and relative stable histograms during data taking, and finding the possible problems quickly.

The method is tested with real BESIII experiment data. Histograms of T hitmap of all MDC sense wires from different runs are used for testing. In the sample histogram, the 4000th bin is set to be a dead channel (0.5 times normal value) and the 4973th bin is set to be a hot channel (1.5 times normal value) artificially. The warning criterion is $b_{3i} \geqslant c + 5e_{4i}$ or $b_{3i} \leqslant c - 5e_{4i}$, in which c is equal to 0.999294. Fig. 6 shows the result. The 1520th ($b_{1i} = 0, b_{2i} = 0$) and 1535th ($b_{1i} = 0$, $b_{2i} = 0$) bins are dead channel in the reference and sample histograms, so these bins are not abnormal bins, we will not give the warning message about them. The entry of the 6258th bin in both reference and sample histograms is low compared with other bins ($b_{3i} = 0.992675$, $e_{4i} = 0.0227733$). Our method gives the correct warning message, only the 4000th ($b_{3i} = 0.497053$, $e_{4i} = 0.0153814$) and 4973th ($b_{3i} = 1.50074$, $e_{4i} = 0.0246846$) bins are reported as abnormal bins.

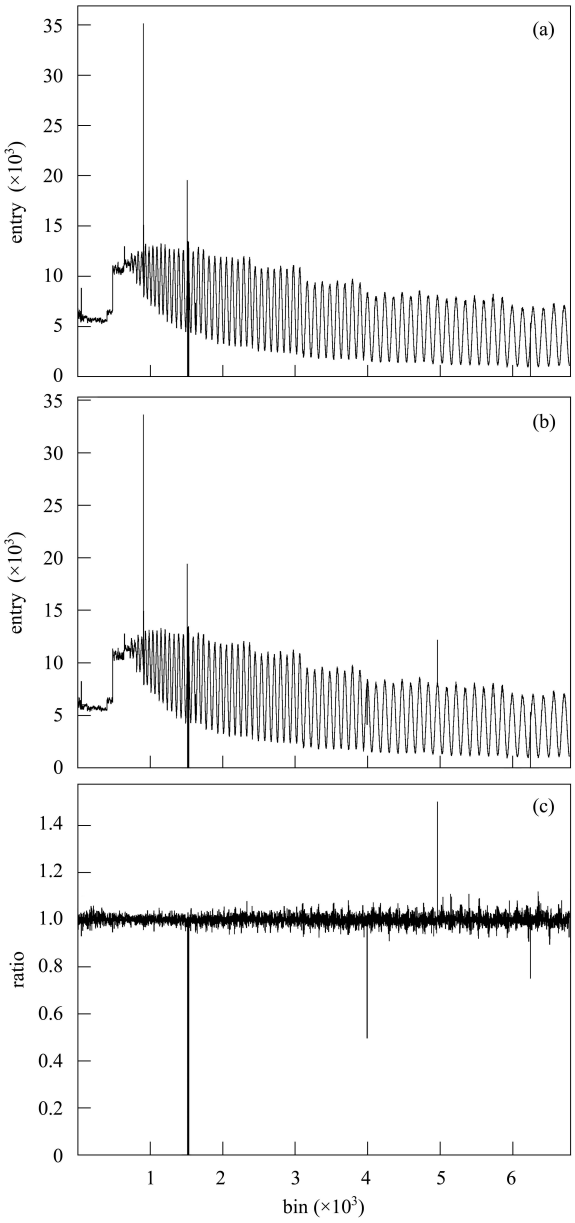

Fig. 6. MDC T hitmaps obtained in the BESIII data taking. (a) reference; (b) sample (the 4000th and 4973th bin are modified.); (c) the ratio of (b) to (a).

5 Summary and conclusions

Two widely-used histogram comparison methods, KS test and χ^2 test, are studied. We studied their discrimination power quantitatively and found that the KS test is preferred over the χ^2 test in most cases. To get a suitable discriminating power, the critical event number for different distributions is different. 7000 events can be used to distinguish the overall shape changes for the distributions of Gaussian or Exponential, however more than 10000 events are needed for the distribution which changes slowly (like Linear distribution). This is the recommended setting in the software validation.

In order to find the problem in hitmap-like histograms, a new method is developed. It can find the abnormal bins in the sample histogram compared with the reference. According to our test, the performance of the method is satisfactory, the bins manually set abnormal are picked out, and there is no wrong warning message even under the low statistics. It is useful to find the dead or hot channels during data taking.

We are grateful to Prof. Zhu Yongsheng from IHEP and Associate Prof. Huang Xingtao from Shandong University for their helpful discussions and suggestions.

References

1 QIN Ya-Hong, LI De-Min, JI Xiao-Bin. Nucl. Elec. & Dete. Tech., 2008, **28**: 1163 (in Chinese)
2 Porter F C. arXiv:0804.0380. Testing Consistency of Two Histograms
3 http://boss.ihep.ac.cn/
4 http://root.cern.ch/
5 CMS NOTE 2005/000, March 6, 2006