ГОСУДАРСТВЕННЫЙ НАУЧНЫЙ ЦЕНТР РОССИЙСКОЙ ФЕДЕРАЦИИ

ИНСТИТУТ ФИЗИКИ ВЫСОКИХ ЭНЕРГИЙ

ИФВЭ 2013–21
ОЭФ, ОМВТ

С.И. Битюков, Н.В. Красников[1], А.Н. Никитенко[2], В.В. Смирнова

Метод статистического сравнения гистограмм

Направлено в *ИВУЗ. ЯЭ*

[1] **ФГБУН Институт ядерных исследований РАН, Москва**
[2] Имериал колледж, Лондон

Протвино 2013

УДК 53.088, 519.23 М-24

Аннотация

Битюков С.И. и др. Метод статистического сравнения гистограмм: Препринт ИФВЭ 2013-21. – Протвино, 2013. – 10 с., 5 рис., 2 табл., библиогр.: 6.

В работе дан сравнительный анализ двух методов, позволяющих определить через сравнение двух гистограмм, полученных при обработке двух независимых выборок, являются ли выборки выборками из одного и того же потока событий или выборки взяты из двух разных потоков событий.

Abstract

Bityukov S.I. et al. A method for statistical comparison of histograms: IHEP Preprint 2013-21. – Protvino, 2013. – p. 10, figs. 5, tables 2, refs.: 6.

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. that two histograms are drawn from the same distribution). The approach is based on the notion "significance of deviation". Our approach allows also to estimate the statistical difference between two histograms.

Введение

Во многих энергетических установках поведение характеристик отдельных элементов системы контролируются путем их многократного измерения. Полученные в течение определенного времени распределения измеренных значений часто представляются в виде гистограмм. Важной задачей при контролировании нормальной работы системы является отслеживание возможных изменений в распределениях измеренных характеристик при возникновении неполадок в системе. Одним из методов регистрации изменения в характере работы системы является обнаружение различий в распределениях характеристик элементов системы при сравнении гистограмм, полученных в разные промежутки времени.

Пусть измерения некоторой характеристики в системе производятся в момент времени [*t1,t2*] и поток событий в этот момент времени назовем *G1*. Пусть было зарегистрировано *N1* событий, при обработке которых была получена гистограмма *hist1*. Затем в интервале времени [*t3,t4*] проводились новые измерения объемом *N2* в потоке событий *G2*, результатом которых является гистограмма *hist2*. Если при сравнении гистограмм, выясняется, что различий в потоках событий нет, то есть $G1=G2$, то можно делать вывод, что в системе изменений нет. Если выясняется, что различие в потоках событий есть, то есть $G1\neq G2$, то в системе возможны неполадки.

Существует несколько подходов к задаче сравнения гистограмм [1]. Обычно для этой цели используется одномерная тестовая статистика (например, хи-квадрат [2]). В работе [3] предложен метод статистического сравнения гистограмм, позволяющий использовать многомерную тестовую статистику.

В данной работе проводится сравнение мощности критерия хи-квадрат и мощности критерия, использующего двухмерную тестовую статистику.

Итак, пусть в результате обработки двух выборок объемом *N1* и *N2* получены две гистограммы с числом бинов равным *M*:

hist1: $n_{11}\pm\sigma_{11}$, $n_{21}\pm\sigma_{21}$, ..., $n_{M1}\pm\sigma_{M1}$ и *hist2:* $n_{12}\pm\sigma_{12}$, $n_{22}\pm\sigma_{22}$, ..., $n_{M2}\pm\sigma_{M2}$.

Сравнив эти две гистограммы нужно принять решение о том, равны или нет потоки событий *G1* и *G2*, а также оценить качество решения, то есть вероятность того, что решение правильное.

Расстояние между гистограммами

Большинство методов сравнения гистограмм используют в качестве меры различимости гистограмм некоторое «расстояние между гистограммами». Так, например, в методе χ^2 расстояние между двумя гистограммами

$$\chi^2 = \sum_{i=1}^{M} \frac{\left(\frac{n_{i1}}{N_1} - \frac{n_{i2}}{N_2}\right)^2}{\frac{n_{i1}}{N_1^2} + \frac{n_{i2}}{N_2^2}} = \sum_{i=1}^{M} S_i^2,$$

где S_i в случае пуассоновских потоков событий (*G1* и *G2*) можно назвать «нормализованной значимостью различия» значения в бине *i* в первой гистограмме и значения в бине *i* во второй гистограмме.

Отметим, что в предлагаемом в работе [3] методе также используется S_i, но несколько иначе, чем в методе χ^2. Существуют и другие «расстояния» (см. [1]).

Распределение тестовых статистик

Предлагается использовать статистические моменты распределения $S_i, i = 1, M$. Это распределение, состоящее из *M* значений, при условии, что *G1=G2*, близко к стандартному нормальному распределению, поскольку каждая реализация S_i случайной величины «нормализованная значимость различия» значений в бине *i* является реализацией стандартной нормальной величины.

Таким образом, в качестве расстояния между гистограммами предлагается не одномерная величина, как в других методах, а многомерная. Конкретно, в рассмотренном

примере, двухмерная *SRMS* = ($\bar{S}$, *RMS*), где $\bar{S} = \frac{\sum_{i=1}^{M} S_i}{M}$ есть среднее значение распределения «нормализованных значимостей различия», а *RMS*=$\sqrt{\frac{\sum_{i=1}^{M}(S_i-\bar{S})^2}{M}}$ – среднее квадратическое отклонение этого распределения.

SRMS имеет ясную интерпретацию:

– если *SRMS*=(0,0), то две гистограммы идентичны;

– если *SRMS*≈(0,1), то *G1=G2* (если *RMS* < 1, то выборки частично перекрываются, то есть они не независимы);

– если вышеупомянутые условия не выполняются, то *G1≠G2*.

Отметим, что существует взаимосвязь между средним, среднеквадратичным и значением хи-квадрат:

$$RMS^2 = \frac{\chi^2}{M} - \bar{S}^2,$$

где $\chi^2 = \sum_{i=1}^{M} S_i^2$. Данная взаимосвязь указывает на то, что тест-статистика χ^2 является комбинацией двух тест-статистик RMS и $\bar{S}$.

Нормализованная значимость различия

Рассмотрим модель, в которой гистограммы *hist1* и *hist2* определены следующим образом: случайная переменная «содержимое бина *i*» подчиняется нормальному распределению

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

Здесь ожидаемое число событий в *i*-ом бине *k*-ой гистограммы есть n_{ik} и дисперсия σ_{ik}^2 также равна n_{ik}. Данную модель можно рассматривать как некоторое приближение распределения Пуассона нормальным распределением.

Для случая сравнения двух наблюденных гистограмм введем нормализованную значимость различия в двух соответствующих бинах гистограмм

$$\hat{S}_i = \frac{n_{i1} - K n_{i2}}{\sqrt{\sigma_{i1}^2 + K^2 \sigma_{i2}^2}}.$$

В данном случае n_{ik} это наблюденное значение в бине *i* гистограммы *k*, σ_{ik} соответствующее стандартное отклонение и K некоторый коэффициент нормализации, отражающий отношение интегральных светимостей эксперимента при наборе сравниваемых выборок. Обычно, в зависимости от задачи, K равно либо отношению объемов выборок, либо отношению длительностей временных интервалов набора выборок.

Пример

Рассмотрим пример, в котором ожидаемое число событий n_{i1} в бине *i* и ожидаемое значение дисперсии σ_{i1}^2 первой гистограммы равно номеру бина *i*, а во второй гистограмме $n_{i2}= \sigma_{i2}^2=\frac{i}{2}$ (верхние гистограммы на Рис. 1). На нижних гистограммах показано, соответственно, слева нормализованные значимости различия для каждого бина гистограмм, справа распределение нормализованных значимостей различия.

Данный пример демонстрирует, что распределение значений S_i близко к стандартному нормальному распределению. Нужно отметить, что любая произвольная перестановка бинов одновременно в обеих гистограммах не меняет распределение в правой нижней гистограмме.

Отметим, что все Монте Карло вычисления и представление гистограмм выполнены в рамках системы *ROOT* [4].

Генерация повторной гистограммы

Следующий шаг является важным в данном методе сравнения гистограмм. По аналогии с генерацией повторной выборки в методе бутстреп [5] он может быть назван генерацией повторной гистограммы. Для каждой из сравниваемых гистограмм создается определенное количество подобных гистограмм (клонов) в соответствии с рассматриваемой здесь моделью, а именно, значение в каждом бине клонируемых гистограмм разыгрывалось в соответствии с законом $\mathbf{N}(n_{ik}, \sigma_{ik})$. Это позволяет создать две имитационные модели генеральных совокупностей гистограмм для сравниваемых гистограмм. Так, в рассмотренном ниже примере было смоделировано 49999 клонов для каждой из гистограмм и, затем, было проведено 50000 сравнений пар полученных гистограмм. В ходе каждого сравнения строилось распределение нормализованных

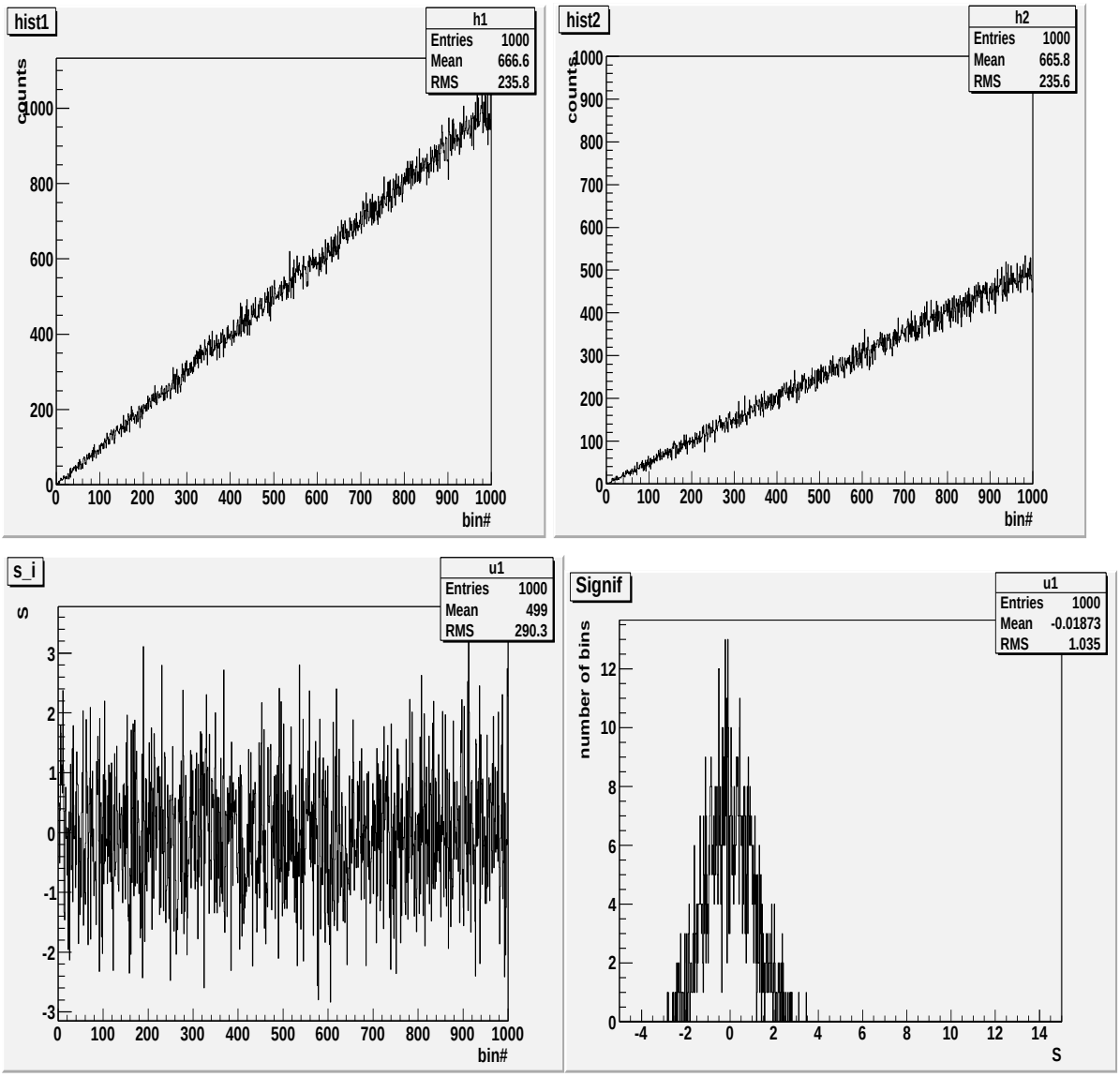

Рис. 1. Треугольные распределения (K=2, M=1000): вверху две сравниваемые гистограммы, внизу слева нормализованные значимости различия для каждого бина гистограмм, внизу справа распределение нормализованных значимостей различия.

значимостей различия в бинах и определялось среднее и среднеквадратическое полученного распределения, а также тест-статистика χ^2. Полученные величины используются для проверки гипотезы о равенстве потоков *G1* и *G2*, определения ошибок I-го (α) и II-го (β) рода и оценки вероятности правильного решения.

Различимость гистограмм

Различимость гистограмм можно оценить с помощью некоторой функции ошибок I-го (α) и II-го (β) рода, которая фактически является вероятностью правильного решения. Если эта величина равна 1, то гистограммы 100% различимы. Если же эта величина равна 0, то гистограммы неразличимы и можно утверждать, что потоки событий равны, то есть *G1=G2*. Если критическая область (критическая величина, критическая линия, критическая поверхность, …) выбрана корректно, то есть выполнено условие $\alpha + \beta \leq 1$, то вероятность правильного решения определяется как [6]

$$1 - \frac{\alpha+\beta}{2-(\alpha+\beta)}.$$

Рассмотрим Случай А: гистограммы получены из независимых выборок одной генеральной совокупности (Рис. 2). Проведем процесс генерации повторной гистограммы, а также сравнения полученных пар гистограмм для определения величин среднего и среднеквадратического в распределении нормализованных значимостей различия в каждой паре сравниваемых гистограмм. Полученные распределения тестовых статистик $T_{\chi^2} = \sqrt{\frac{\chi^2}{M}}$ и $RMS\&S$ показаны на Рис. 3 (соответственно слева и справа).

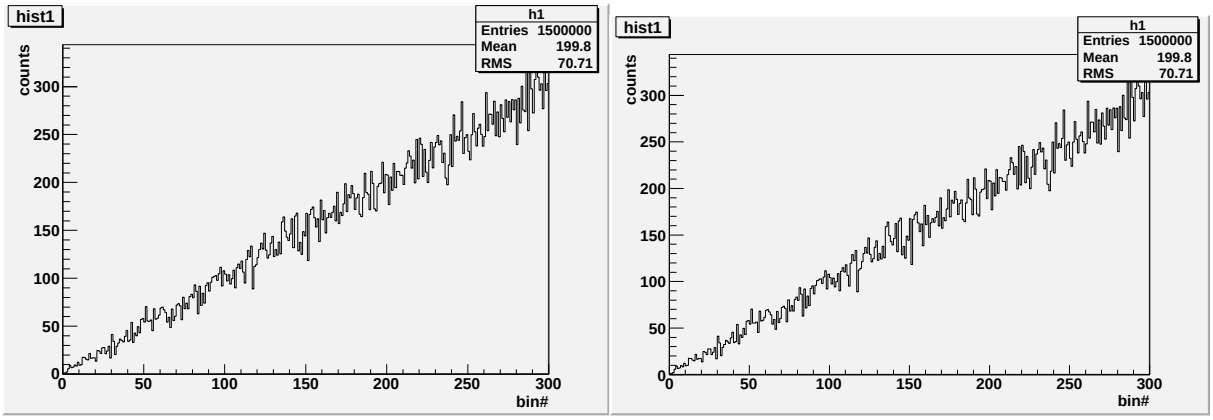

Рис. 2 Случай А (*G1=G2*): исходные сравниваемые гистограммы получены при обработке независимых выборок равного объема из одного и того же потока событий (K=1,M=300).

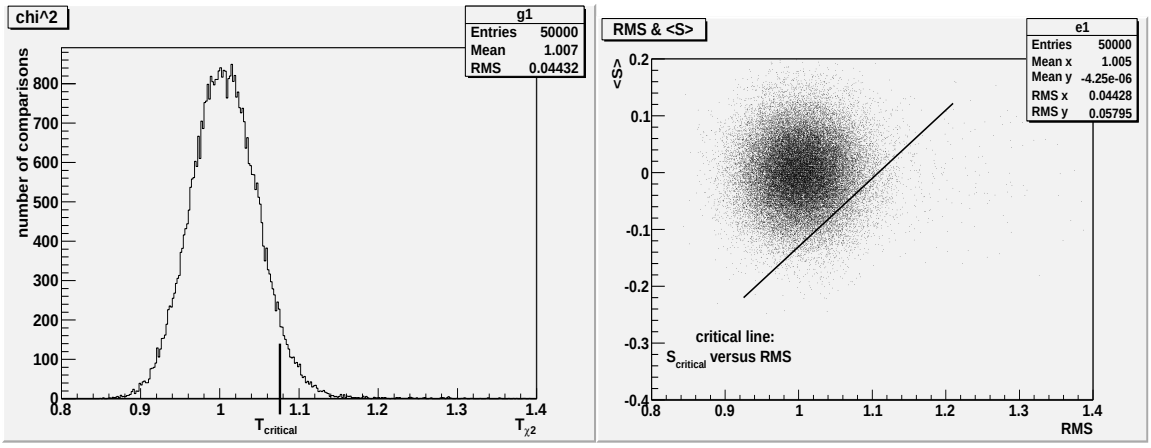

Рис. 3 Случай A: 50000 сравнений. Слева распределение тест-статистики $\sqrt{\frac{\chi^2}{N}}$. Tcritical – критическая точка определяющая 5% уровень значимости теста при основной гипотезе *G1=G2* против альтернативы *G1≠G2*. Справа распределение тест-статистики *SRMS*. Сплошная линия это критическая линия также определяющая 5% уровень значимости двумерного теста при основной гипотезе *G1=G2* против альтернативы *G1≠G2*.

Нужно отметить, что генерацию повторной гистограммы для Случая A необходимо проводить перед сравнением гистограмм, полученных при обработке выборок из различных потоков событий при заданном уровне значимости критерия. Она фактически является самокалибровкой метода, а именно, она нужна для нахождения критической области при проверке гипотез (критическая точка *T*critical на левом распределении Рис. 3 и критическая линия на правом распределении Рис. 3 выбраны из условия 5% уровня значимостей критерия). Критическую область в данном методе также возможно построить исходя из требования максимума вероятности правильного решения.

Рассмотрим Случай B: гистограммы получены из выборок разных генеральных совокупностей (Рис. 4).

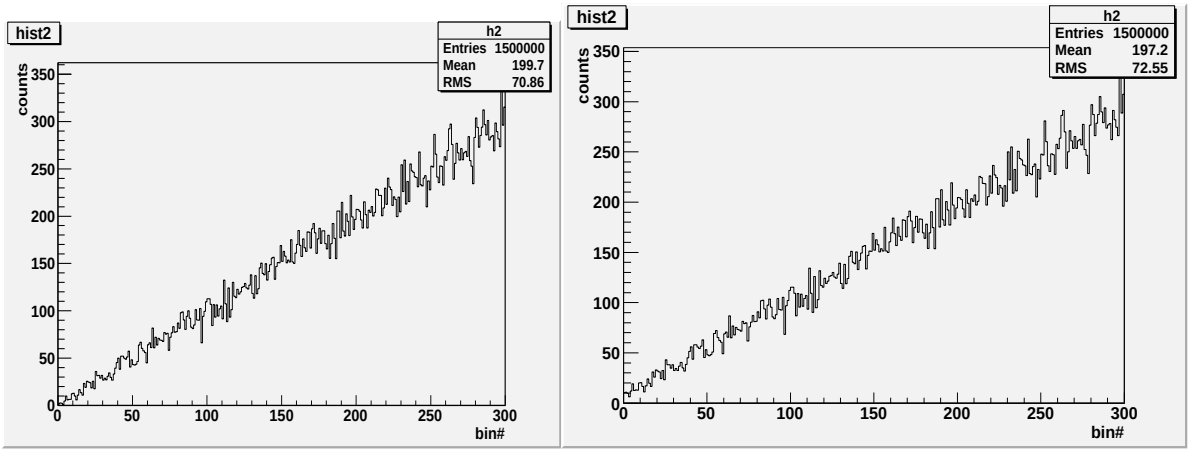

Рис. 4. Случай B (*G1≠G2*): исходные сравниваемые гистограммы получены при обработке независимых выборок равного объема из разных потоков событий (K=1,M=300). Угол наклона во второй гистограмме уменьшен на 5% (левый край приподнят, правый – опущен).

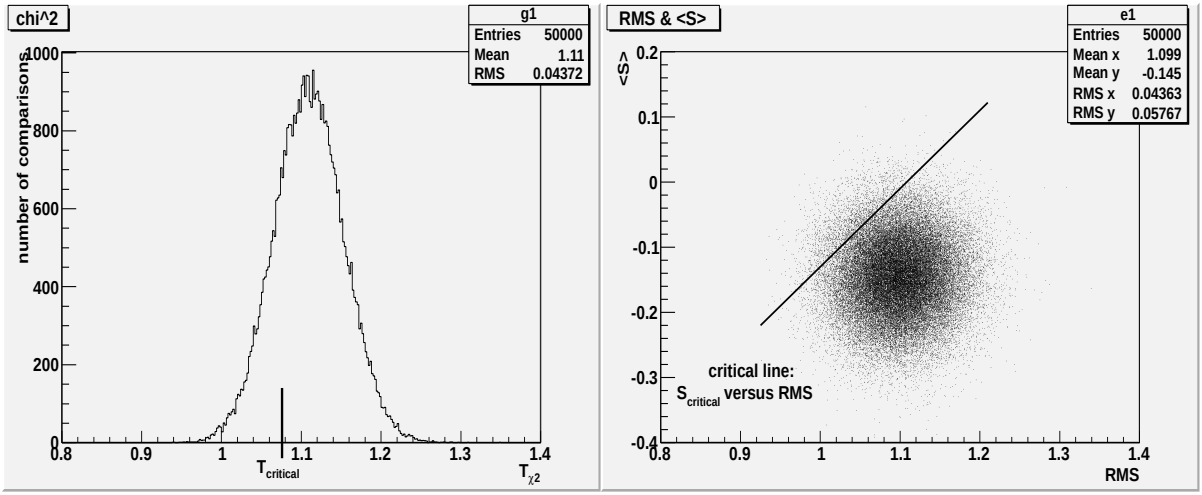

Рис. 5 Случай B: 50000 сравнений. Слева распределение тест-статистики $\sqrt{\frac{\chi^2}{M}}$. *T*critical – критическая точка определяющая 5% уровень значимости теста при основной гипотезе *G1=G2* против альтернативы *G1≠G2*. Справа распределение тест-статистики *SRMS*. Сплошная линия это критическая линия также определяющая 5% уровень значимости теста при основной гипотезе *G1=G2* против альтернативы *G1≠G2*.

Итак, после генерации повторной гистограммы при 5% уровне значимости теста проверка гипотезы о том, что *G1=G2*, показала: мощность одномерного теста (Рис. 5,

слева) равна 0.7797, мощность двумерного теста (Рис. 5, справа) равна 0.9574. Детали приведены в Таблице 1 и Таблице 2.

Таблица 1. Одномерный тест

Тест-статистика $\sqrt{\frac{\chi^2}{n}}$	В действительности		Мощность теста	Вероятность правильного решения
Отобрано	Случай A	Случай B		
Случай A	47499	11014		
Случай B	2501	38986		
	α	β	1 - β	1 - $\hat{\beta}$
	0.05	0.2203	0.7797	0.8437

Результаты проверки гипотезы в одномерном случае показали, что при 5% уровне значимости теста на основную гипотезу *G1=G2* вероятность правильно выбрать альтернативную гипотезу равна 77.97%. При этом вероятность сделать правильный выбор вне зависимости от выбора гипотезы равна 84.37%.

Таблица 2. Двумерный тест

Тест-статистика *SRMS*	В действительности		Мощность теста	Вероятность правильного решения
Отобрано	Случай A	Случай B		
Случай A	47502	2132		
Случай B	2498	47868		
	α	β	1 - β	1 - $\hat{\beta}$
	0.05	0.0426	0.9574	0.9515

Результаты проверки гипотезы в двумерном случае показали, что при 5% уровне значимости теста на основную гипотезу *G1=G2* вероятность правильно выбрать альтернативную гипотезу равна 95.74%. При этом вероятность сделать правильный выбор вне зависимости от выбора гипотезы равна 95.15%.

Заключение

Предложенный подход с использованием двумерной тест-статистики позволяет значительно усилить мощность критерия при проверке гипотез по сравнению с методами, использующими одномерные тест-статистики.

Предложенный метод может быть использован для контроля работоспособности оборудования во время функционирования установки.

Основные этапы предложенной процедуры:

– введение в рассмотрение нормализованной значимостей различия в бинах гистограмм позволяет построить распределение значимостей различия, которое близко к стандартному нормальному распределению в случае выполнения условия *G1=G2*;

– генерация повторной гистограммы позволяет провести оценку точностных характеристик метода при сравнении конкретных гистограмм;

– вероятность правильного решения является удобной оценкой качества принятого решения о различимости двух гистограмм.

Список литературы

[1] *Porter F.* Testing consistency of two histograms. arXiv:0804.0380 – 2008.

[2] *Gagunashvili N.D.* Chi-square tests for comparing weighted histograms// Nucl.Instr.&Meth. – 2010. – A614. P. 287-296.

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

[4] *Brun R., Rademaker F.* ROOT – An object oriented data analysis framework// Nucl.Instr.&Meth. – 1997 – A389. P. 81-86.

[5] *Efron B.* Bootstrap methods: another look at the jackknife// Annals of Statistics - 1979 – 7. P. 1-26.

[6] Bityukov S.I., Krasnikov N.V., Distinguishability of Hypotheses// Nucl.Inst.&Meth. – 2004 – A534. P. 152-155.

Рукопись поступила 15 ноября 2013 г.

С.И. Битюков и др.

Метод статистического сравнения гистограмм.

Препринт отпечатан с оригинала-макета, подготовленного авторами.

Подписано к печати 04.12.2013. Формат 60 × 84/16. Цифровая печать.
Печ.л. 0, 87. Уч.– изд.л. 1,15. Тираж 80. Заказ 48. Индекс 3649.

ФГБУ ГНЦ ИФВЭ
142281, Протвино Московской обл., пл. Науки, 1.

Индекс 3649

П Р Е П Р И Н Т 2013-21, ИФВЭ, 2013