

ОБЪЕДИНЕННЫЙ
ИНСТИТУТ
ЯДЕРНЫХ
ИССЛЕДОВАНИЙ
ДУБНА

P10-86-812

П.В.Зрелов, В.В.Иванов

МЕТОД ВЫДЕЛЕНИЯ
МАЛОВЕРОЯТНЫХ СОБЫТИЙ
С ПОМОЩЬЮ КРИТЕРИЯ СОГЛАСИЯ
СМИРНОВА - КРАМЕРА - МИЗЕСА

Направлено на II Международную конференцию
по статистике, Тампере, Финляндия, 1987 г.

1986

ВВЕДЕНИЕ

Критерий Смирнова - Крамера - Мизеса (СКМ) используется при статистической проверке соответствия данных наблюдения простой гипотезе, фиксирующей закон распределения генеральной совокупности, из которой получена выборка. Этот критерий, получивший также название ω^2 (омега-квадрат)-критерия^{/1/}, основывается на непосредственно наблюдаемых, несгруппированных значениях исследуемой случайной величины.

Критерий СКМ основан на статистике

$$\omega_n^2 = \int_{-\infty}^{+\infty} [S_n(x) - F(x)]^2 f(x) dx, \quad (I)$$

где $F(x)$ - функция распределения одномерной случайной величины x с плотностью $f(x)$, а $S_n(x)$ - эмпирическая функция распределения. Соотношение (I) легко преобразуется в равенство (см., например, /2/)

$$\omega_n^2 = \frac{1}{12n^2} + \frac{1}{n} \sum_{k=1}^n \left[F(x_k) - \frac{2k-1}{2n} \right]^2, \quad (2)$$

где $x_1 < x_2 < \dots < x_n$ - упорядоченная выборка из генеральной совокупности случайной величины x .

Обычно критерии согласия используются для проверки гипотез о виде неизвестной функции распределения на основании анализа выборок из генеральной совокупности исследуемой случайной величины^{/1,2,9/}. Особенность критерия СКМ анализировать непосредственно наблюдаемые в каждом испытании значения рассматриваемой случайной величины позволяет исследовать саму выборку на соответствие заранее известному распределению, что представляет наибольший интерес в случае малых объемов выборок. Такой подход дает возможность эффективно применять кри-

терий СКМ в экспериментальной физике при детальном анализе данных наблюдения, когда каждое событие рассматривается в отдельности. Независимость же данного критерия от вида распределения позволяет изучать распределения разнообразных физических величин, используя при этом общие свойства критерия СКМ.

В частности, в работе^{/3/} критерий ω^2 использовался в качестве статистической меры при исследовании эффектов кластеризации в много-частичных конечных состояниях, наблюдаемых в K^-p -соударениях при 13 ГэВ/с. В настоящей работе описан метод, позволивший успешно решить задачу выделения редких событий, связанных с образованием вторичных дейтронов при анализе спектров масс вторичных частиц от фрагментации дейтронов на ядрах мишени^{/4/}. Здесь критерий СКМ был применен для выделения маловероятных событий из анализируемых спектров, полученных в результате одновременного измерения одних и тех же случайных величин различными регистрирующими приборами.

1. ЭКСПЕРИМЕНТАЛЬНЫЙ МАТЕРИАЛ

С помощью спектрометра МАСИХ^{/5/} при бомбардировке дейтронами с импульсом 9 ГэВ/с мишеней из CH_2 , CD_2 и C под углом 140 мрад к оси первичного пучка измерялся импульсный спектр вторичных частиц в интервале от 3,5 до 5,5 ГэВ/с. Основной вклад в спектр в этой области дают протоны, примесь дейтронов составляет примерно 1%.

Идентификация заряженных частиц (протонов или дейтронов) осуществлялась путем измерения их импульсов и времен пролета. Импульсы частиц вычислялись по углам их отклонения в поле анализирующего магнита^{/6/}. Для определения времен пролета использовались сигналы от двух пар сцинтилляционных счетчиков, расстояния между которыми составляли соответственно 16,7 и 21,9 м. В измеренные времена пролета вносились поправки на флуктуацию амплитуд импульсов от фотоумножителей, просматривающих сцинтилляторы счетчиков^{/7/}.

Распределение масс вторичных частиц, полученное с учетом данных от первой пары счетчиков (пролетная база 16,7 м), показано на рис.1. Похожий вид имеет спектр масс, полученный с учетом данных от второй пары счетчиков. В этих распределениях четко выделяется пик при $m \cong 0,94$ ГэВ/с², обусловленный вкладом протонов. Наряду с ним в области $m \cong 1,88$ ГэВ/с² на уровне заметного фона^{*} виден небольшой "прилив"

* Здесь и далее предполагается, что фон образуется частицами, массы которых не отвечают распределениям протонов или дейтронов.

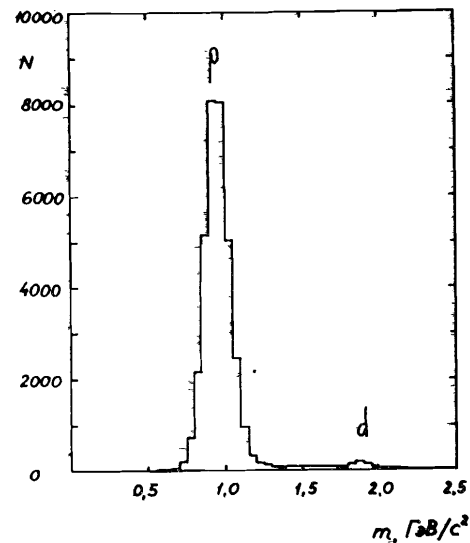


Рис.1. Спектр масс вторичных частиц, полученный с учетом данных от первой пары сцинтилляционных счетчиков.

обусловленный вкладом дейтронов. На рис.3 пунктирной линией приведено распределение масс вторичных частиц, полученное с помощью первой пары счетчиков, для $m > 1,25$ ГэВ/с². Отношение вклада дейтронов в максимуме распределения к уровню фона составляет примерно 3. Выделение дейтронов из рассмотренных спектров существенно затруднено из-за наличия "хвоста" от распределения протонов, а также наличия фона, от которого не удалось полностью избавиться при отборе событий по критериям, указанным в работе^{/8/}.

2. ВЫДЕЛЕНИЕ ДЕЙТРОНОВ С ПОМОЩЬЮ КРИТЕРИЯ СМЕРНОВА - КРАМЕРА - МИЗЕСА

для отбора событий, связанных с дейтронами, можно воспользоваться тем обстоятельством, что для каждой регистрируемой частицы измеряются два времени пролета, t_1 и t_2 , которые при помощи соотношения^{/7/}

$$m_i = \frac{p}{c} \sqrt{\left(\frac{t_i}{t_{0i}}\right)^2 - 1} \quad (i = 1, 2), \quad (3)$$

где p - импульс частицы, t_i - время ее пролета на базе длиной l_i , $t_{0i} = l_i/c$, c - скорость света, m_i - масса частицы,

преобразуются в массы m_1 и m_2 , определяемые соответственно первой и второй парами счетчиков. Ниже будет показано, как на основании совместного анализа измеренных масс m_1 и m_2 можно идентифицировать ре-

гистрируемую частицу значительно более эффективно, чем в случае анализа каждого из спектров масс в отдельности.

Анализ спектров масс (см. рис.1) показывает, что распределения в области протонных пиков содержат около 97% событий, и в этой области их можно аппроксимировать гауссовскими распределениями со средними значениями \bar{m}_i и стандартными отклонениями σ_i ($i = 1, 2$). Гауссовское распределение может быть подвергнуто стандартизации использованием замены переменных:

$$M_i = \frac{m_i - \bar{m}_i}{\sigma_i} \quad (i = 1, 2). \quad (4)$$

Если преобразованию (4) подвергнуть спектры масс, то части этих распределений, обусловленные вкладом протонов, будут описываться одной и той же функцией распределения, которая соответствует плотности $N(0, 1)$. Этот факт, наряду с тем, что переменные m_1 и m_2 в области протонных пиков можно считать независимыми, поскольку их распределения обусловлены физическими процессами, имеющими статистически независимый характер, позволяет исследовать пару значений масс m_1 и m_2 , преобразованных по формуле (4), на принадлежность одному распределению $N(0, 1)$, т.е. можно рассматривать указанную пару величин как выборку из генеральной совокупности объемом $n = 2$ и проверять соответствие этой выборки гипотезе, заданной функцией распределения $F(x)$ плотности $N(0, 1)$ при помощи критерия СКМ. Тогда события (под событием понимается пара преобразованных масс), которые не отвечают гипотезе $F(x)$, т.е. вызванные регистрацией вторичных дейтронов, будут иметь значения ω_2^2 , близкие к предельному для $n = 2$.

На рис.2 приведено полученное в результате обработки эксперимен-

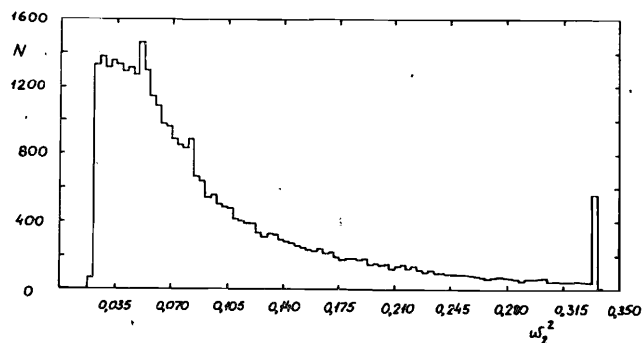


Рис.2. Распределение случайной величины ω_2^2 .

тальных данных распределение величин ω_2^2 *). Гистограмма содержит 35055 событий. В области значений ω_2^2 , больших $\omega_{кр}^2 = 0,333$, на распределении виден четкий пик. Число событий в пике равно 562. Спектр масс вторичных частиц для рассмотренной пары счетчиков, отвечающий событиям, для которых $\omega_2^2 > \omega_{кр}^2$, изображен на рис.3 сплошной линией. Из

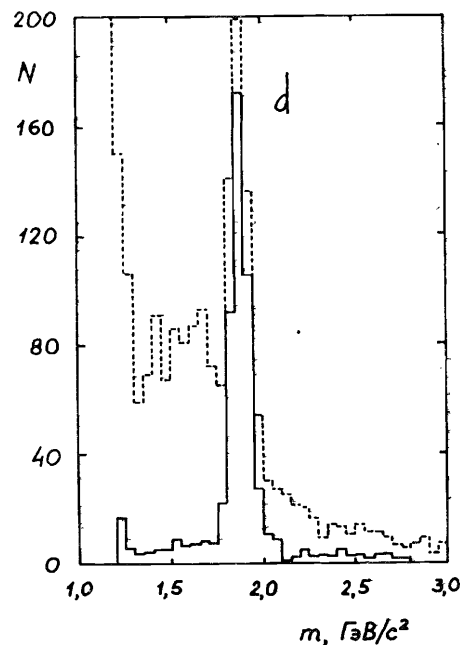


Рис.3. Спектр масс вторичных частиц, полученный с учетом данных от первой пары счетчиков: а) для масс, больших 1,25 ГэВ/c² (пунктирная линия); б) для событий с $\omega_2^2 > 0,333$ (сплошная линия).

этого распределения видно, что над незначительным фоном доминирует пик, вызванный событиями, в которых регистрируемой частицей является дейтрон. Отношение вклада дейтронов в максимуме распределения к уровню фона составляет примерно 17.

3. МОДЕЛИРОВАНИЕ СПЕКТРОВ МАСС ВТОРИЧНЫХ ЧАСТИЦ

Для того, чтобы оценить относительные вклады протонов, дейтронов и фона в критической области распределения ω_2^2 , было проведено моделирование. Далее распределение величин ω_2^2 будет именоваться "распределением ω_2^2 ".

лирование событий, каждое из которых являлось статистически разыгранной парой величин m_1 и m_2 , распределенных в соответствии со спектрами масс вторичных частиц, измеренными на опыте с использованием двух пар счетчиков. При этом спектры представлялись в виде суммы распределений протонов (97% событий), дейтронов (1%) и равномерно распределенного фона (2%). Спектры протонов и дейтронов аппроксимировались гауссовскими распределениями, параметры которых оценивались на основе анализа экспериментальных данных.

Было выполнено две серии моделирования событий. В первом случае величины m_1 и m_2 разыгрывались независимо друг от друга. Во второй серии случайные величины m_1 и m_2 разыгрывались одновременно в соответствии с одним из распределений: протонов, дейтронов или фона. С целью построения распределения ω^2 значения масс m_1 и m_2 , полученные в результате моделирования, подвергались такой же процедуре обработки, что и данные реального эксперимента.

Распределение случайных величин ω^2 для первого варианта моделирования представлено на рис. 4а, а для второго — на рис. 4б. Из рис. 4а

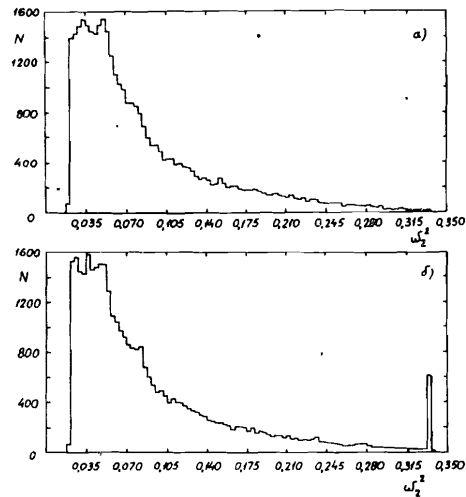


Рис. 4. Распределение значений ω^2 для:
а) независимо разыгранных случайных величин m_1 и m_2 ;
б) "скоррелированных" случайных величин m_1 и m_2 .

видно, что для независимо разыгранных m_1 и m_2 распределение ω^2 непрерывно убывает с ростом ω^2 и при $\omega^2 > \omega_{кр}^2$ становится равным нулю. Во втором случае в распределении ω^2 в области значений $\omega^2 > \omega_{кр}^2$

наблюдается узкий пик, обусловленный в основном разыгранными дейтронами. Такая группировка дейтронных событий вызвана тем, что одновременно для обеих масс, m_1 и m_2 , соответствующие значения функции распределения близки 1, что, как следует из соотношения (2), приводит к значениям ω^2 , близким к предельному $\omega_{кр}^2 = 1/3$. Что же касается разыгранных фоновых событий, то основная их часть распределена в области $\omega^2 \leq \omega_{кр}^2$ и лишь незначительная часть попадает в критическую область $\omega^2 > \omega_{кр}^2$.

Таким образом, выделение маловероятных событий возможно в случае применения критерия СКМ к скоррелированным случайным величинам в смысле их одновременной принадлежности в каждой выборке одному и тому же распределению из числа составляющих анализируемые спектры.

4. ВЫБОР КРИТИЧЕСКОЙ ГРАНИЦЫ КРИТЕРИЯ СМИРНОВА — КРАМЕРА — МИЗЕСА

Особо следует остановиться на вопросе о значении критической границы критерия СКМ. Для выборки размером $n = 2$ критическое значение $Z_{кр} = 2\omega_{кр}^2 = 0,666$ соответствует уровню значимости порядка $0,1\% / 10$. Поэтому, учитывая число выделенных в критической области событий (см. п.2), можно оценить вклад событий, обусловленных протонами, который не превышает 6%. Таким образом, пик в распределении ω^2 (см. рис. 2) при $\omega^2 > \omega_{кр}^2$ вызван в основном регистрацией вторичных дейтронов.

Следует отметить, что в данном случае для определения уровня значимости необходимо пользоваться таблицей функции распределения статистики СКМ для выборки объемом $n = 2$, приведенной в работе /10/, т.к. использование для этой цели асимптотической функции распределения приведет к значительной ошибке в определении уровня значимости, в результате которой классифицировать события с $\omega^2 > \omega_{кр}^2$ будет невозможно. Дело в том, что критическому значению $Z_{кр} = 0,666$ соответствует значение асимптотической функции распределения 0,985 (см., например, /11/), отвечающее уровню значимости 1,5%. Поэтому среди событий, для которых $\omega^2 > \omega_{кр}^2$, может оказаться 1,5% событий от общего числа зарегистрированных в эксперименте, согласующихся с заданной гипотезой (т.е. обусловленных протонами). Число таких событий составляет 526 и примерно совпадает с количеством событий, выделенных в критической области. Таким образом, события с $\omega^2 > \omega_{кр}^2$ не будут статистически обоснованы, и классифицировать их на основании критерия СКМ в случае использования асимптотической функции распределения будет невозможно.

5. ИСКЛЮЧЕНИЕ ФОНА ИЗ ВЫДЕЛЕННЫХ СОБЫТИЙ

Для исключения фона из выделенных событий можно эти события повторно проанализировать с помощью критерия СКМ. При этом распределения от обеих пар счетчиков в области дейтронных пиков можно аппроксимировать гауссовскими распределениями, которые преобразуются к $N(0,1)$ с помощью соотношения (4). Функция распределения плотности $N(0,1)$ используется в качестве гипотезы при анализе указанных событий с помощью критерия СКМ, при этом отбираются события, согласующиеся с этой гипотезой, а попадающие в критическую область отбрасываются. В качестве критического значения было выбрано $Z_{кр} = 2\omega_{кр}^2 = 0,548$, отвечающее уровню значимости 1%. На рис.5 представлен измеренный первой парой счетчиков спектр масс вторичных частиц, для которых $\omega_2^2 \leq \omega_{кр}^2$. Из

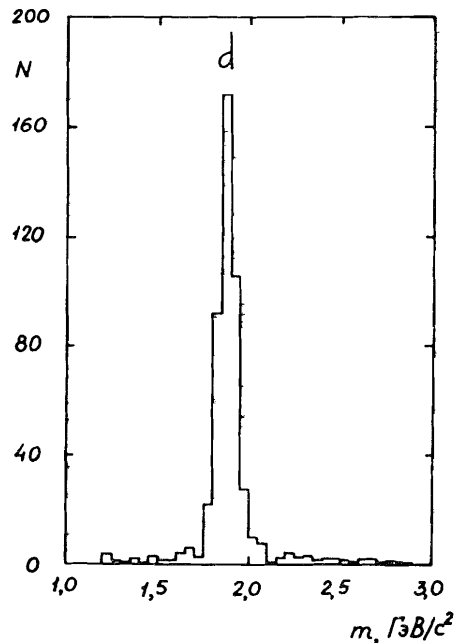


Рис.5. Спектр масс вторичных частиц, полученный с учетом данных от первой пары счетчиков, после повторного применения критерия СКМ.

рисунка видно, что в результате повторного применения критерия СКМ отношение вклада дейтронов в максимуме распределения к уровню фона составило примерно 80.

ЗАКЛЮЧЕНИЕ

В настоящей работе предложен метод выделения маловероятных событий, суть которого заключается в следующем:

а) анализируемые спектры преобразуются таким образом, чтобы вклады доминирующих распределений можно было аппроксимировать одной и той же функцией распределения $F(x)$;

б) каждая выборка, составленная из величин, одновременно принадлежащих преобразованным спектрам, исследуется на соответствие гипотезе, заданной функцией $F(x)$ при помощи критерия согласия Смирнова - Крамера - Мизеса; при этом маловероятным событиям (см. п.3), не согласующимся с этой гипотезой, отвечают большие значения ω^2 , что позволяет выделить их в критической области соответствующего распределения статистики СКМ;

в) выделенные в критической области события подвергаются повторному анализу согласно п.п. а) и б) с той лишь разницей, что отбираются события, согласующиеся с заданной гипотезой, а попавшие в критическую область - отбрасываются; тем самым производится дополнительное исключение фона из исследуемых спектров.

На примере выборки $n = 2$ показано, что при работе с малыми выборками уровень значимости, отвечающий заданной критической границе, необходимо определять, используя таблицы функции распределения для конкретной выборки, т.к. использование для этих целей асимптотической функции может приводить к трудностям в классификации событий из-за ошибки в определении уровня значимости.

Применение описанного метода при анализе экспериментальных данных от фрагментации дейтронов высокой энергии на ядрах мишени позволило эффективно выделить редкие события, связанные с образованием вторичных дейтронов.

Ввиду того, что задачи такого рода часто встречаются в экспериментальной физике (см., например, /3,9/), критерий СКМ может играть важную роль в вопросах статистического анализа данных наблюдения наряду с наиболее часто используемыми F -, t - и χ^2 -критериями. Приведенный в работе пример можно считать типичным для указанного класса задач.

Авторы благодарят Л.С.Ажгиря, Г.А.Ососкова и Н.И.Чернова, обсуждения с которыми стимулировали появление настоящей работы.

ЛИТЕРАТУРА

1. Мартынов Г.В. Критерии омега-квадрат. "Наука", М., 1978.
2. Соболев И.М. Численные методы Монте-Карло. "Наука", М., 1973.
3. Ludlam T., Slansky R. A Statistical Measure of Clustering in Multiparticle Final States. Physics Department, Yale University New Haven, Connecticut, 06520, 1973.
4. Ажгирей Л.С., Взоров И.К., Жмыров В.Н. и др. ОИЯИ, PI-86-729, Дубна, 1986.
5. Ажгирей Л.С., Взоров И.К., Жмыров В.Н. и др. В кн.: Труды совещания по исследованиям в области релятивистской ядерной физики. ОИЯИ, Д2-82-568, Дубна, 1982, с.83.
6. Ажгирей Л.С., Будкин Л.В., Взоров И.К. и др. ОИЯИ, I3-86-52, Дубна, 1986.
7. Ажгирей Л.С., Зарубин П.И., Иванов В.В. и др. ОИЯИ, PI-85-749, Дубна, 1985.
8. Ажгирей Л.С., Взоров И.К., Иванов В.В. ОИЯИ, PI0-12655, Дубна, 1979.
9. Статистические методы в экспериментальной физике. Пер. с англ. Под ред. А.А.Тяпкина. Атомиздат, М., 1976.
10. Зрелов П.В., Иванов В.В. ОИЯИ, PI0-86-547, Дубна, 1986.
- II. Anderson T.W., Darling D.A. Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. Ann. Math. Statist., 1952, v.23, p.193.

Рукопись поступила в издательский отдел
16 декабря 1986 года.

Зрелов В.П., Иванов В.В. Метод выделения маловероятных событий с помощью критерия согласия Смирнова - Крамера - Мизеса	P10-86-812
Предложен метод выделения маловероятных событий при помощи критерия согласия Смирнова - Крамера - Мизеса для случая, когда анализируемые распределения представляют собой одновременные измерения одних и тех же случайных величин различными приборами. Применение описанного метода при анализе спектров масс вторичных частиц от фрагментации высокоэнергетических дейтронов на ядрах мишени позволило эффективно выделить редкие события, связанные с образованием вторичных дейтронов. Показано, что при работе с малыми выборками уровень значимости, отвечающий заданной критической границе, необходимо определять, используя таблицы функции распределения для конкретной выборки, т.к. использование для этих целей асимптотической функции может приводить к трудностям в классификации выделенных событий из-за ошибки в определении уровня значимости.	
Работа выполнена в Лаборатории вычислительной техники и автоматизации ОИЯИ.	
Препринт Объединенного института ядерных исследований. Дубна 1986	

Перевод О.С.Виноградовой

Zrelov P.V., Ivanov V.V. Small Probability Event Separation Method Based on Smirnov - Cramer - Von Mises Test	P10-86-812
The method for small probability event separation based on the Smirnov - Cramer - Von Mises test is suggested for the case when distributions analysed are simultaneous measurements of the same random variables by different instruments. Application of this method for analysing mass spectra of secondary particles produced by high energy deuteron fragmentation on target nuclei allowed one to separate efficiently rare events in the processes with secondary deuteron production. It is shown that in the case of small size samples the significance level corresponding to selected critical level is to be determined using cumulative function tables for specific sample since the asymptotic function used for these purposes can lead to the difficulties in separated event classification due to an error in significance level determination.	
The investigation has been performed at the Laboratory of Computing Techniques and Automation, JINR.	
Preprint of the Joint Institute for Nuclear Research. Dubna 1986	