of the theory of classification (review)] / Zavod. Lab. Diagn. Mater. 2009. Vol. 75. N 7. P. 51 – 63 [in Russian].
35. **Shtremel' M. A., Kudrya A. V., Ivashchenko A. V.** Neparametricheskii diskriminantnyi analiz v zadachakh upravleniya kachestvom [Nonparametric discriminant analysis in problems of quality control] / Zavod. Lab. Diagn. Mater. 2006. Vol. 72. N 5. P. 53 – 62 [in Russian].
36. **Koplyarova N. V., Orlov V. I., Sergeeva N. A., Fedosov V. V.** O neparametrichekikh modelyakh v zadachakh diagnostiki élektroradioizdelii [On nonparametric models in the diagnostics electrical radio products] / Zavod. Lab. Diagn. Mater. 2014. Vol. 80. N 7. P. 73 – 77.
37. **Tolcheev V. O.** Modifitsirovannyi i obobshchennyi metod blizhaishego soseda dlya klassifikatsii bibliograficheskikh tekstovykh dokumentov [Modified and generalized method of the nearest neighbor to classify bibliographic text documents] / Zavod. Lab. Diagn. Mater. 2009. Vol. 75. N 7. P. 63 – 70 [in Russian].
38. **Orlov A. I., Tolcheev V. O.** Ob ispol'zovanii neparametricheskikh statisticheskikh kriteriev dlya otsenki tochnosti metodov klassifikatsii (obobshchayushchaya stat'ya) [On the use of nonparametric statistical tests to estimate the accuracy of classification methods (generalizing article)] / Zavod. Lab. Diagn. Mater. 2011. Vol. 77. N 3. P. 58- – 66 [in Russian].
39. **Borodkin A. A., Tolcheev V. O.** Kompleksnaya procedura reduktsii dlya uvelicheniya bystrodeistviya neparametricheskikh metodov klassifikatsii tekstovykh dokumentov [Integrated reduction procedure to increase the speed of nonparametric methods of classification of text documents] / Zavod. Lab. Diagn. Mater. 2011. Vol. 77. N 11. P. 64 – 69 [in Russian].
40. **Borodkin A. A., Tolcheev V. O.** Razrabotka i issledovanie metodov vzveshivaniya blizhaishikh sosedei (na primere klassifikatsii bibliograficheskikh tekstovykh dokumentov) [Development and research of methods of weighing the nearest neighbors (for example, classification of bibliographic text documents)] / Zavod. Lab. Diagn. Mater. 2013. Vol. 79. N 7. P. 70 – 74 [in Russian].

УДК 519.24:543.429.23–42.062

ПРОВЕРКА НОРМАЛЬНОСТИ РАСПРЕДЕЛЕНИЯ И НЕЗАВИСИМОСТИ РЕЗУЛЬТАТОВ ИЗМЕРЕНИЙ ИНТЕГРАЛЬНЫХ ИНТЕНСИВНОСТЕЙ ШИРОКИХ ГРУПП СИГНАЛОВ В СПЕКТРАХ ЯМР ^{1}Н ВЫСОКОГО РАЗРЕШЕНИЯ[1]

© **М. Б. Смирнов**[2]

Статья поступила 21 ноября 2014 г.

Показано, что в ЯМР ^{1}Н высокого разрешения для широких групп сигналов при достаточно высоком уровне шума измерения интегральных интенсивностей в последовательно регистрируемых спектрах независимы. В этом случае распределение ошибок измерения с достаточной точностью соответствует нормальному закону. При низком уровне шума последовательные измерения, выполняемые в течение небольших промежутков времени, не являются независимыми. Для получения пригодного для статистической обработки материала необходимо регистрировать спектры с интервалом более одного часа. Распределение ошибок измерения — бимодальное, вне зависимости от способа коррекции базовой линии. Нормальный закон распределения для них является лишь грубым приближением.

Ключевые слова: ЯМР ^{1}Н; распределение ошибок измерения; независимость измерений; нефть; лигнин.

Использование ЯМР высокого разрешения для количественного анализа обусловлено прямой пропорциональностью молярной концентрации вещества μ и интегральной интенсивностью I соответствующих резонансных сигналов или их групп: $\mu = I/k$, где k — так называемый «весовой фактор», в общем случае зависящий от ряда параметров (см., например, [1 – 3]). При анализе сложных смесей, таких как нефть или ее фракции, полимеров типа лигнина и т.п., как правило, индивидуальные компоненты не определяются и задача количественного анализа состоит в измерении доли тех или иных атомов (Н, С и т.д.), входящих в определенные типы структурных единиц молекул, например, водорода ароматических циклов в целом (H_{ap}) или водорода в ароматических циклах моноциклоароматических соединений, от общего числа атомов этого элемента в образце [2 – 4]. Это объясняется тем, что в спектрах таких объектов из-за огромного числа содержащихся в них компонентов сигналы перекрываются, образуя неразрешенные широкие группы (рис. 1). Простейшая ситуация имеет место в рутинном вари-

[1] Работа выполнена при поддержке ООО «Хембридж», Москва, Россия.

[2] ФГБУН Ордена Трудового Красного Знамени Институт нефтехимического синтеза им. А. В. Топчиева РАН, Москва, Россия; e-mail: m1952s@yandex.ru

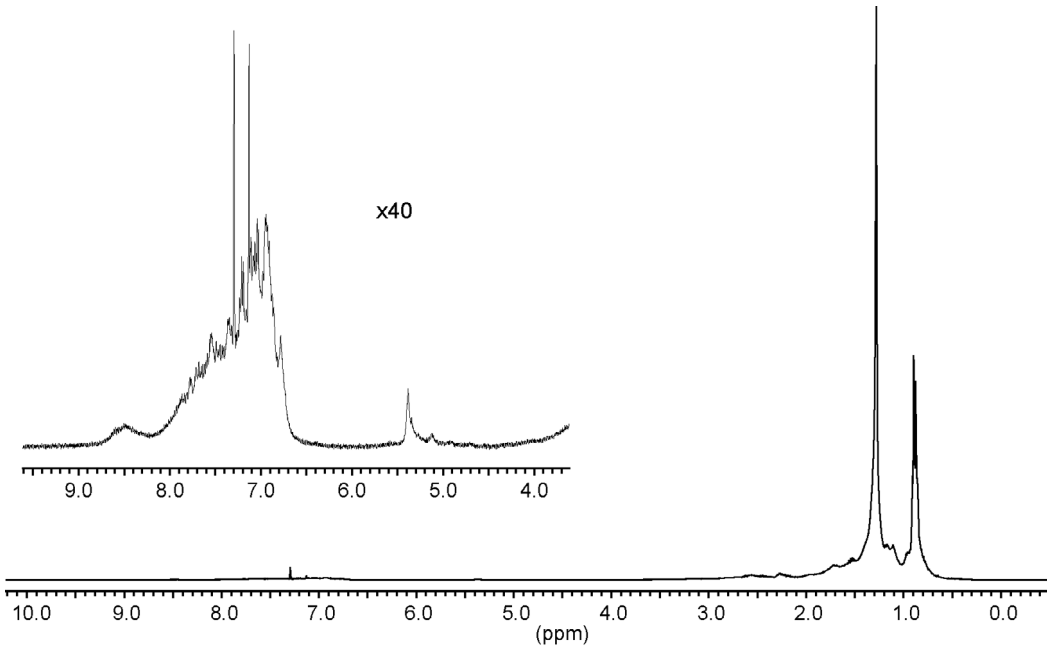

Рис. 1. Спектр ЯМР ^{1}H раствора нефти в $CDCl_3$ после коррекции базовой линии

анте спектроскопии ЯМР ^{1}H, где в принципе реально получать спектры без насыщения сигналов. Здесь для любого типа структур i справедливо: $H_i/H_{об} = I_i/I_{об}$, где $H_i/H_{об}$ — доля водорода в структурах i, которым отвечает группа сигналов с интегральной интенсивностью I_i от общего водорода образца $H_{об}$, $I_{об}$ — общая интегральная интенсивность всех сигналов в спектре.

При интегрировании широких групп сигналов возникают специфические проблемы. Во-первых, в нулевой («базовой») линии спектра появляются существенные нелинейные искажения [2, 3]. Программы автоматической коррекции базовой линии оказываются неработоспособными [4]. Во-вторых, коррекция фазы в спектре проводится оператором визуально и из-за отсутствия однозначного критерия (общеизвестного для спектров индивидуальных соединений и простых смесей) возможен определенный произвол, в результате чего появляется субъективная составляющая. Эти два источника и определяют статистическую ошибку измерения (воспроизводимость) площадей не перекрывающихся с другими групп сигналов в спектрах. Для слабых сигналов к ним может добавляться величина отношения сигнал/шум [2, 3]. Прочие ошибки измерения — в основном систематические и для образцов с близким составом практически одинаковы (а именно, при их сравнении ошибки измерения представляют наибольший интерес). Хотя именно на систематических ошибках было сосредоточено основное внимание при анализе ошибок измерений (см. [2, 3] и цитируемую там литературу).

В публикациях, где рассмотрена воспроизводимость измерений, предполагается, что последовательные измерения независимы, статистическая ошибка нормально распределена, относительная погрешность постоянна. Проверка этих положений не проводилось. Существующие оценки среднего квадратического отклонения (СКО) получены в основном более 20 лет назад на оборудовании, которое существенно больше искажало базовую линию, чем спектрометры последних поколений. Объем выборки, по которой определяли СКО, как правило, не указан. С учетом принятых в этой области представлений, скорее всего, он был мал — примерно 10 измерений (см. [2, 3, 5] и цитируемую там литературу). В единственном существующем для таких объектов стандарте ASTMD 5292–99 (измерение $H_{ар}$ в нефтях и их фракциях) соответствующие сведения также отсутствуют. Поэтому целью данной работы явилась проверка с приемлемой значимостью гипотез о независимости последовательных измерений и нормальности распределения статистической ошибки; при выявлении отклонений от этих гипотез — оценка последствий при решении стандартных аналитических задач.

Объектом изучения выбран образец нефти. В спектрах ЯМР ^{1}H нефтей и их фракций имеются три не перекрывающиеся с другими группы сигналов, отвечающие резонансу водорода в ароматических циклах $H_{ар}$ (6,0 – 9,6 м.д.), в изолированных двойных C=C-связях ($H_{дв}$; 4,6 – 5,9 м.д.) и в алифатических фрагментах молекул $H_{ал}$ (>4,5 м.д.) [3 – 5]. Поскольку спектр нормируется к суммарной площади всех сигналов, независимы только две величины. Из содержательных соображений выбирают $H_{ар}$ и $H_{дв}$. Изучаемый образец имел значение $H_{ар}$ несколько меньшее, чем среднее для нефтей (~2,6 % от $H_{об}$) и величину $H_{дв}$, превышающую среднее (~0,12 % от $H_{об}$). Двадцатикратная разница интегральных интенсивностей этих групп сигналов позволяет в одной серии опытов изу-

чить поведение групп сигналов при низком и относительно высоком уровнях шума (рис. 1).

Спектры ЯМР ^{1}H раствора нефти в $CDCl_3$ (~1:1 v/v) регистрировали при 313 К на спектрометре DRX-400 (Bruker, ФРГ; 400 МГц) с пятимиллиметровым датчиком без вращения образца в запаянной ампуле. Режим регистрации: время сбора данных — 4 с, релаксационная задержка — 11 с; длительность импульса — 35°; 16 сканирований. Отсчет химических сдвигов проводили от самого интенсивного сигнала, отвечающего резонансу CH_2-групп в середине алкильных цепей, принимая для него δ = 1,280 м.д. Последовательно получено 400 спектров с двумя длительными перерывами во времени. Коррекцию базовой линии выполняли двумя способами. Первый — общепринятый: в областях с нулевой спектральной плотностью расставляется около 10 точек посередине шумовой дорожки с визуальным контролем их положения. В подпрограмме ручной коррекции базовой линии использовали опцию «cubic splines». При втором способе положение точек корректировали так, чтобы величины интегральных интенсивностей на участках с нулевой спектральной плотностью шириной по 0,1 – 0,3 м.д., непосредственно примыкающих к границам интегрируемых областей резонанса $H_{ар}$ (6,00 – 9,60 м.д.) и $H_{дв}$ (5,90 – 4,60 м.д.) [5], не превышали $3 – 5 \cdot 10^{–6}$ от $I_{об}$ (большие значения допускались для участка шириной 0,3 м.д.). В результате из спектра получены по два значения каждой величины: $H_{ар}$, $H_{ар,кор}$ и $H_{дв}$, $H_{дв,кор}$ (индекс «кор» — для второго способа коррекции базовой линии). Анализ первых 200 спектров показал, что по всем рассчитываемым параметрам разница между $H_{дв}$ и $H_{дв,кор}$ намного меньше значимой. Поэтому в дальнейшем измеряли только $H_{дв}$. При интегрировании принимали интегральную интенсивность $H_{ал}$ = = 10 000, так что дискретность оцифровки данных составила менее $10^{–6}$ от $H_{об}$.

Таблица 1. Результаты расчета выборочных параметров распределения $H_{ар}$, $H_{ар,кор}$, $H_{дв}$ и выборочных значений соответствующих статистик критериев согласия

Параметры и критерии	$H_{ар}$	$H_{ар,кор}$	$H_{дв}$	К.З.***
$x_{ср}$*	2,6169	2,6189	0,1178	
СКО*	0,0133	0,0110	0,0023	
g_s	0,26	0,10	–0,05	0,20
e_s	0,44	0,05	0,10	0,41
Размах/СКО	5,94	6,02	5,66	
χ^2 (19)**	12,57	24,15	13,78	26,3
χ^2 (13)**	9,82	17,70	8,56	18,3
χ^2 (9)**	5,95	7,97	2,01	12,6
D_n	0,731	0,470	0,553	0,895
$n\omega^2(1 + 1/2n)$	0,095	0,048	0,034	0,126

* В % от $H_{об}$.

** В скобках — число равновероятных интервалов при группировке данных.

*** Критические значения при p = 0,95.

По опытным данным для $H_{ар}$, $H_{ар,кор}$ и $H_{дв}$ вычисляли выборочные средние арифметические, СКО, коэффициенты эксцентриситета и эксцесса, отношения размаха выборки к СКО. В качестве критериев согласия с нулевой гипотезой о соответствии распределения ошибок измерения нормальному взяты критерии хи-квадрат, типа Колмогорова и типа омега-квадрат. Поскольку во всех случаях проверяется сложная гипотеза, при расчетах значений χ^2 [6] использовали группировку данных по равновероятным интервалам (число интервалов m = 19, 13 и 9) и оценки среднего и СКО, вычисленные по методу минимума хи-квадрат; статистика критерия — хи-квадрат с m – 3 степенями свободы [7]. Критические значения статистики критерия типа Колмогорова заимствованы из работы [8], критерия типа омега-квадрат — из [9]. Расчеты по критерию типа омега-квадрат выполнены согласно [10]. Результаты приведены в табл. 1. Из них следует, что стандартно принимаемые как значимые отклонения от нормальности фиксируются только для $H_{ар}$ по коэффициентам эксцентриситета и эксцесса. Анализ показал, что эти отклонения обусловлены присутствием в выборке одного (для коэффициента эксцесса) и двух (для коэффициента эксцентриситета) наибольших значений. Если отбросить три наибольших значения, получаем g_s = 0,10 и e_s = 0,11. Другими словами, эмпирическая функция распределения по этим показателям не отличима от нормального, за исключением «хвостов». Однако для $H_{ар,кор}$ при разбиении на 19 и 13 интервалов достигаемый уровень значимости критерия хи-квадрат составляет около 0,915 и 0,935 соответственно. Следовательно, с учетом объема выборки вероятность того, что нулевая гипотеза для этой величины неверна, представляется существенной.

Поэтому для всех трех величин построены графики зависимости средней плотности распределения на 19 равновероятных интервалах от $(x_i – x_{ср})/s$ с ОМП оценкой $x_{ср}$ и s. Полученные ломаные приведены на рис. 2 и 3. Распределения $H_{ар}$, $H_{ар,кор}$ — бимодальные. Для первого число значений, попадающих в интервал с $(x_i – x_{ср})/s \sim 0$, вдвое меньше, чем попадающих в интервалы с наибольшей плотностью распределения (соответственно, 14, 30 и 27). Для второго в интервале с минимальной плотностью распределения — 8 значений, с наибольшей — 32 и 28. Распределение же $H_{дв}$ вполне удовлетворительно совпадает с нормальным.

Из сказанного следует, что для $H_{дв}$ использование нормального распределения при статистической обработке опытных данных будет давать адекватные результаты везде, где не используются «хвосты» распределения, т.е. в задачах, в которых разница результатов расчетов, исходящих из нормального распределения и, например, не отличимого от него при n = 400 логистического [8], с содержательной точки зрения не принципиальна. Невозможность использования «хвостов» распределения обусловлена объемом выборки [9]. При измерении $H_{ар}$ в обоих вариантах коррекции

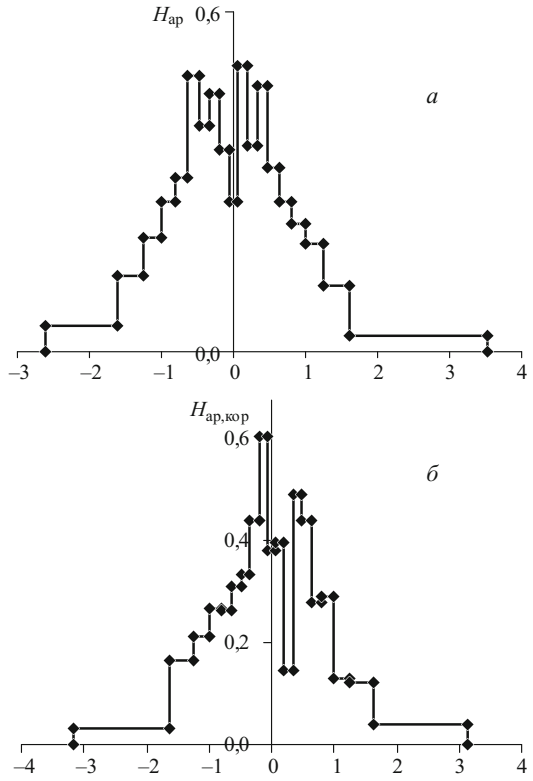

Рис. 2. Зависимости средней плотности распределения на 19 равновероятных интервалах от $(x_i - x_{ср})/s$ с ОМП оценкой $x_{ср}$ и s для $H_{ар}$ и $H_{ар,кор}$

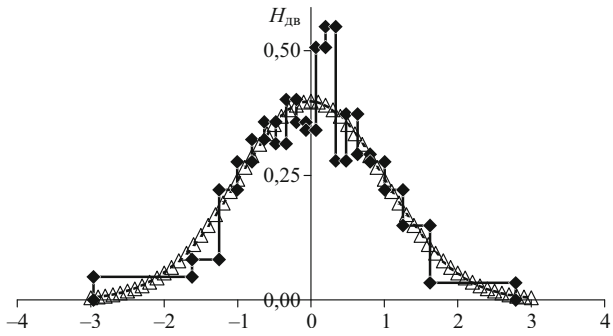

Рис. 3. Зависимости средней плотности распределения при разбиении на 19 равновероятных интервалов от $(x_i - x_{ср})/s$ с ОМП оценкой $x_{ср}$, s для $H_{дв}$ (заштрихованные ромбы и сплошная линия) и нормального распределения (0, 1) (незаштрихованные треугольники)

базовой линии, скорее всего (поскольку моды отстоят примерно на 0,5σ), приближение нормальным распределением является приемлемым при сравнении результатов единичных измерений, когда значима разница ~3σ (см., например, [11]). Хотя и здесь вычисляемый достигаемый уровень значимости, очевидно, можно рассматривать только как грубую оценку. Сравнение по стандартной процедуре данных, полученных усреднением более чем двух измерений, даст явно искаженный результат.

Для выяснения вопроса о независимости последовательных измерений использованы коэффициенты корреляции Спирмена между i-м и $i + 1$-м измерениями $r_s(i, i + 1)$ и критерий серий [8]. Результаты вычислений величин $r_s(i, i + 1)$, числа серий m и достигаемых уровней значимости (p_r, p_m) для нулевой гипотезы (последовательные измерения независимы) приведены в табл. 2. Для $H_{дв}$ гипотеза, очевидно, принимается. При обоих же способах измерения $H_{ар}$ говорить о независимости последовательных измерений не приходится. Особенно четко об этом свидетельствуют величины P_r. Для понимания того, как это может отразиться на практике, получено распределение результатов по числу серий разной длины (табл. 3) и вычислены выборочные средние арифметические $x_{ср,9}(i)$ и СКО $s_9(i)$ для всех последовательных девяток измерений (i — номер первого измерения в девятке; $i = 1, 2, ..., 392$). Результаты для $H_{ар,кор}$ представлены на рис. 4. В табл. 3 приведены также оценки по формуле Муавра – Лапласа для биномиального распределения математического ожидания и СКО числа серий данной длины L.

Из табл. 3 следует, что в последовательных измерениях $H_{ар}$ и $H_{ар,кор}$ намного больше серий с $L > 10$:3 и 4 соответственно. При независимых же измерениях вероятность появления в ряду из 400 измерений трех и более серий с $L > 10$ составляет $\sim 10^{-3}$. На рис. 4 просматриваются квазипериодические колебания получающихся при усреднении величин. Следовательно, поскольку выполненные подряд последовательные измерения не независимы, то при необходимости усреднения по нескольким измерениям их следует проводить с интервалом более одного часа (лучше — несколько часов). Также необходимо проводить измерения для оценки СКО любой широкой группы сигналов, если общее число измерений предполагается небольшим и время накопления каждого спектра мало (минуты и менее). Несоблюдение этого условия приводит к высокой вероятности появления систематической ошибки. Особенно важно, что резко возрастает вероятность кратного занижения СКО. Что касается $H_{дв}$, то никаких отклонений от ожидаемого для независимых последовательных измерений не наблюдается и никаких ограничений на измерение этой величины не требуется.

Таким образом, в ЯМР ^{1}H высокого разрешения для широких групп сигналов при достаточно высоком уровне шума измерения интегральных интенсивностей в последовательно регистрируемых спектрах независимы. Распределение ошибок измерения с дос-

Таблица 2. Значения коэффициентов корреляции Спирмена между i-м и $i + 1$-м измерениями $r_s(i, i + 1)$, число серий m и достигаемые уровни значимости (p_r, p_m) для проверки нулевой гипотезы о независимости последовательных измерений

Параметры	$H_{ар}$	$H_{ар,кор}$	$H_{дв}$
$r_s(i, i + 1)$	0,144	0,159	–0,015
p_r	$4{,}4 \cdot 10^{-3}$	$1{,}7 \cdot 10^{-3}$	0,76
m	184	177	207
p_m	0,099	0,019	0,58

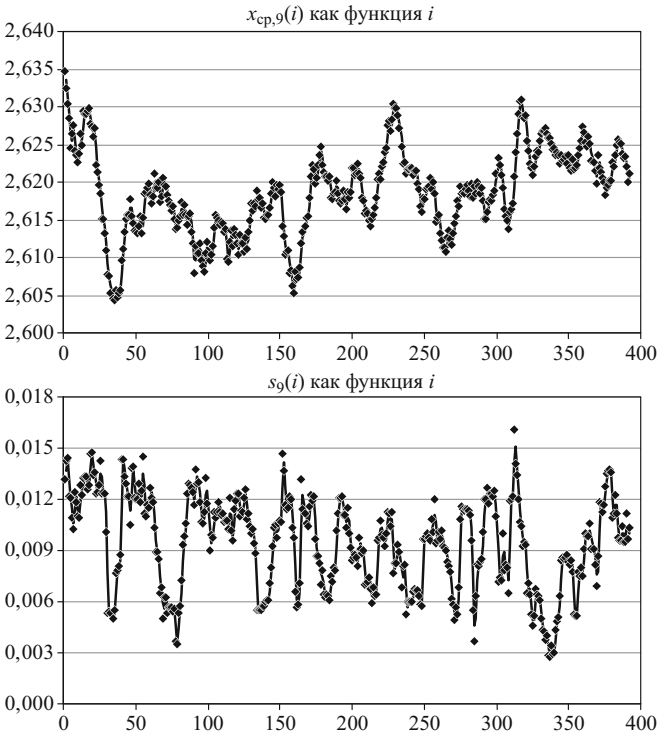

Рис. 4. Зависимости выборочных средних арифметических $x_{ср,9}(i)$ и СКО $s_9(i)$ для всех последовательных девяток измерений $H_{ар,кор}$ (i — номер первого измерения в девятке; i = = 1, 2, ..., 392)

таточной для аналитических целей точностью соответствует нормальному. При низком уровне шума последовательные измерения, выполняемые в течение небольших промежутков времени, не являются независимыми. Для получения пригодного для статистической обработки материала необходимо регистрировать спектры с интервалом более одного часа (желательно — несколько часов). Распределение ошибок измерения — бимодальное, вне зависимости от способа коррекции базовой линии. Нормальный закон распределения для них является лишь грубым приближением. Его можно использовать, анализируя только значения, для которых значима разница ~3σ и более, причем вычисленный уровень значимости следует рассматривать как грубую оценку.

ЛИТЕРАТУРА

1. **Сергеев Н. М.** Спектроскопия ЯМР. — М.: Изд. МГУ, 1981. — 280 с.
2. **Смирнов М. Б., Крапивин А. М.** Методика анализа углеводородных фрагментов высших фракций нефти с помощью спектроскопии ЯМР. — В кн.: Методы исследования состава органических соединений нефти и битумоидов. — М.: Наука, 1985. С. 138 – 181.
3. **Калабин Г. А., Каницкая Л. В., Кушнарев Д. Ф.** Количественная спектроскопия ЯМР природного органического сырья и продуктов его переработки. — М.: Химия, 2000. — 408 с.
4. **Смирнов М. Б., Ванюкова Н. А.** / Нефтехимия. 2014. Т. 54. № 1. С. 17 – 28.
5. **Калабин Г. А., Полонов В. М., Смирнов М. Б., и др.** / Нефтехимия. 1986. Т. 26. № 4. С. 435 – 463.

Таблица 3. Число серий разной длины в последовательных результатах измерений $H_{ар}$, $H_{ар,кор}$ и $H_{дв}$

Размер серий, шт.	$H_{ар}$	$H_{ар,кор}$	$H_{дв}$	M*	σ*
1	99	92	106	99,5	8,66
2	45	45	49	49,6	6,61
3	16	17	32	24,8	4,84
4	7	5	9	12,3	3,48
5	5	6	5	6,2	2,48
6	2	2	4	3,1	1,76
7	2	0	1	1,5	1,25
8	2	5	1	0,76	0,88
9	2	0	0	0,38	0,62
10	1	1	0	0,19	0,44
11	0	1	0	0,09	0,31
12	0	0	0	0,05	0,22
13	2	2	0	0,02	0,16
14	1	0	0	0,01	0,11
18	0	1	0	$9{,}0 \cdot 10^{-5}$	$9{,}8 \cdot 10^{-3}$

* Оценки математического ожидания (M) и СКО (σ) по формуле Муавра – Лапласа для биномиального распределения.

6. **Мирвалиев М., Никулин М. С.** / Заводская лаборатория. 1992. Т. 58. № 3. С. 52 – 58.
7. **Большев Л. Н., Смирнов Н. В.** Таблицы математической статистики. — М.: Наука, 1983. — 416 с.
8. **Орлов А. И.** Прикладная статистика. — М.: Экзамен, 2006. — 672 с.
9. **Мартынов Г. В.** Критерии омега-квадрат. — М.: Наука, 1978. — 78 с.
10. **Орлов А. И.** / Заводская лаборатория. 1985. Т. 51. № 1. С. 60 – 62.
11. **Смирнов М. Б., Ванюкова Н. А.** / Нефтехимия. 2014. Т. 54. № 5. С. 360 – 370.

REFERENCES

1. **Sergeev N. M.** Spektroskopiya YaMR [NMR Spectroscopy]. — Moscow: Izd. MGU, 1981. — 280 p. [in Russian].
2. **Smirnov M. B., Krapivin A. M.** Metodika analiza uglevodorodnykh fragmentov vysshikh fraktsii nefti s pomoshch'yu spektroskopii YaMR / Metody issledovaniya sostava organicheskikh soedinenii nefti i bitumoidov. — Moscow: Nauka, 1985. P. 138 – 181 [in Russian].
3. **Kalabin G. A., Kanitskaya L. V., Kushnarev D. F.** Kolichestvennaya spektroskopiya YaMR prirodnogo organicheskogo syr'ya i produktov ego pererabotki. — Moscow: Khimiya, 2000. — 408 p. [in Russian].
4. **Smirnov M. B., Vanyukova N. A.** / Neftekhimiya. 2014. Vol. 54. N 1. P. 17 – 28 [in Russian].
5. **Kalabin G. A., Polonov V. M., Smirnov M. B., et al.** / Neftekhimiya. 1986. Vol. 26. N 4. P. 435 – 463 [in Russian].
6. **Mirvaliev M., Nikulin M. S.** / Zavod. Lab. 1992. Vol. 58. N 3. P. 52 – 58 [in Russian].
7. **Bol'shev L. N., Smirnov N. V.** Tablitsy matematicheskoi statistiki. — Moscow: Nauka, 1983. — 416 p. [in Russian].
8. **Orlov A. I.** Prikladnaya statistika. — Moscow: Ékzamen, 2006. — 672 p. [in Russian].
9. **Martynov G. V.** Kriterii omega-kvadrat. — Moscow: Nauka, 1978. — 78 p. [in Russian].
10. **Orlov A. I.** / Zavod. Lab. 1985. Vol. 51. N 1. P. 60 – 62 [in Russian].
11. **Smirnov M. B., Vanyukova N. A.** / Neftekhimiya. 2014. Vol. 54. N 5. P. 360 – 370 [in Russian].