Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Сибирский государственный университет
телекоммуникаций и информатики»
(СибГУТИ)

**МАТЕРИАЛЫ РОССИЙСКОЙ
НАУЧНО-ТЕХНИЧЕСКОЙ
КОНФЕРЕНЦИИ**

«ОБРАБОТКА ИНФОРМАЦИИ И МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ»

22-23 апреля 2021 г.

Новосибирск

ISBN 978-5-91434-061-9

СибГУТИ выражает благодарность всем авторам научных публикаций сборника – сотрудникам, студентам, аспирантам и молодым ученым.

Ответственность за правильность, точность и корректность цитирования, ссылок и перевода, достоверность информации и оригинальность представленных материалов несут их авторы.

СОДЕРЖАНИЕ

ВЫЧИСЛИТЕЛЬНЫЕ СИСТЕМЫ

Баранов А.А., Павский К.В., Новиков П.Л. Ускоренный параллельный алгоритм поиска соседей в атомной кристаллической решетки для систем с распределенной памятью. 5
Берлизов Д. М., Ткачёва Т. А. Применение деревьев принятия решений для выбора оптимального алгоритма трансляционного обмена стандарта MPI. 12
Гайдук П. А., Крюкова Л. П. Метод оценки производительности фазы маркировки в параллельных сборщиках мусора. 19
Гвоздева С.Н. Оценка быстродействия устройства для возведения бинарной матрицы в квадрат. 23
Исанбаев Т. Р., Крюкова Л. П. Тестирование производительности баз данных. 32
Крамаренко К.Е., Курносов М.Г. Анализ потребления памяти при распознавании речи пакетом Kaldi для языковых моделей большой размерности. 37
Курносов М. Г., Токмашева Е. И., Ткачева Т. А. Анализ эффективности алгоритмов барьерной синхронизации на NUMA-системах. 41
Майданов Ю. С., Романюта А. А. Обзор статистических подходов к прогнозированию временных рядов показателей вычислительных систем. 48
Насонова А. О., Курносов М. Г. Оптимизация обработки производных типов данных в операции Scatter библиотеки Open MPI. 53
Ситников. Д.А., Крюкова Л.П., Исследование и сравнение клиент-серверной и P2P архитектуры. 58
Чунихин А. А., Бочкарев Б. В., Крюкова Л.П. Разработка средств анализа эффективности intrinsic-функций в виртуальных Java-машинах. 63

ИНФОРМАТИКА И МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

Алтынбекова Г. Ж., Жукова Е. Д., Неупокоев М. В., Гриф А. М. Решение обратной задачи гравиразведки с использованием аппарата нейронных сетей. 69
Аникеева А. Е. Математическая модель влияния гамма-излучения на параметры оптического волокна. 75
Баргуев С.Г. Приближенное решение начально -краевой задачи о колебаниях системы упругого стержня с твердым телом путем разложения в ряды по системе линейно независимых функций и скаляров. 83
Варданян В.А., Варданян Н.В. Математическое моделирование дисперсионного канала связи на основе частотно-временных преобразований. 88
Васильченко А.А., Керимов И.В., Малахов А.А. Обработка траекторных снимков. 95
Васильченко А.А., Керимов И.В., Второв О.И. Априорная оценка точности траекторных измерений. 103
Воронин Д.В. О детонации в плоской радиальной камере. 114
Гриф А.М. Построение неоднородностей по латерали в виде ячеек Вороного для моделей нефтегазовых месторождений. 121
Данилова Л.Ф., Поддубный Д.И., Андреева Т.И. Технология поддержки образовательного контента в соответствии с профессиональными требованиями на основе классификации данных и построении онтологических моделей. 128
Домников П. А., Иванов М. В., Кошкина Ю. И. Применение векторного метода

конечных элементов для решения двумерных задач индукционного каротажа. 142
Захарова Т.Э. Экспериментальное подтверждение методики определения функциональных зависимостей и параметров определяющих уравнений ползучести и повреждаемости. 147
Казначеева Н.В., Полетайкин А.Н. Оптимизационная модель построения индивидуальной образовательной траектории. 153
Лемешко Б. Ю., Лемешко С. Б. О причинах некорректности выводов при использовании непараметрических критериев согласия. 161
Ляхов О.А. Ошибки агрегирования в сетевых моделях управления проектами. 171
Макаров И. О., Гриф А. М., Патрушев И. И. Сравнение методов интерпретации данных гравиметрии с помощью решения обратной задачи и использования нейронных сетей. 178
Машейченко К. С., Данилова Л. Ф. Информационная технология парсинга учебных заведений. 184
Милешко А.В. Исследование зависимости точности прогноза одномерных временных рядов от длины ряда, при использовании методов прогнозирования, основанных на универсальной мере R. 195
Моргунов А.В., Черемных А.О. Информационные технологии обучения программированию с применением игровых механик геймификации. 204
Моргунов А.В. Оценка эффективности от внедрения информационной системы управления документооборотом научно-образовательного учреждения. 210
Морозова К.И. Использование свёрточных нейронных сетей для восстановления аудиосигналов. 220
Нестеров А.С. Применение ИНС в задаче прогнозирования результатов освоения абитуриентами профессиональной образовательной программы. 226
Павлова У.В. Метод прогнозирования временных рядов на основе конечных автоматов в режиме реального времени. 234
Рубан А.А. Сравнение степенной и показательной функций. 239
Токтошов Г.Ы. Задачи размещения пунктов обслуживания сетей инженерных коммуникаций. 243
Трофимов В. К., Храмова Т. В. Универсальное кодирование источников Мура и Мили при условии неравнозначности длительности кодовых символов. 248
Филимонова Н. А. Применение экспериментально-численной модели потоков данных для расчета пропускной способности локальной сети. 253
Черногорова И.В., Полетайкин А.Н. Реализация цифрового двойника образовательной программы. 259
Четвертакова Ю. С., Черникова О. С. Градиентная процедура параметрической идентификации стохастических нелинейных дискретных систем на основе сигма-точечного фильтра Калмана. 267
Шманина Е. Ю. Использование LSTM-сетей для обнаружения аномалий в поведении пользователя. 274
Шувалова В. И., Полетайкин А. Н. Компьютеризированная подсистема фиксации и мониторинга трудовой деятельности работников завода. 280

УДК 519.24

О причинах некорректности выводов при использовании непараметрических критериев согласия

Б. Ю. Лемешко, С. Б. Лемешко
Новосибирский государственный технический университет

В основе некорректного применения непараметрических критериев согласия в приложениях в большинстве случаев лежат две причины. Первая причина заключается в том, что при проверке сложных гипотез и оценивании параметров закона по анализируемой выборке используют классические результаты, связанные с проверкой простых гипотез. Вторая причина связана с наличием ошибок округления, которые могут существенно изменять распределения статистик критериев. Асимптотическими результатами при проверке простых и сложных гипотез можно пользоваться при ошибках округления много меньше среднеквадратического отклонения закона распределения ошибок измерения и объёмах выборок n, не превышающих некоторых максимальных значений. При объёмах выборок больших, чем эти максимальные значения, реальные распределения статистик критериев отклоняются от асимптотических в сторону больших значений статистик. Показано, что единственным выходом, обеспечивающим корректность выводов по применяемым критериям, является использование реальных распределений статистик, которые могут находиться в интерактивном режиме.

Ключевые слова: проверка гипотез, непараметрические критерии согласия, простая гипотеза, сложная гипотеза, распределение статистики, ошибки округления, статистическое моделирование, достигнутый уровень значимости, ошибка 1-го рода, ошибка 2-го рода

1. Введение

Непараметрические критерии согласия чаще всего применяются при идентификации модели закона, наилучшим образом соответствующего закону распределения ошибок измерений. В этих целях может использоваться целый ряд существующих критериев, которые могут применяться в условиях проверки простых и сложных гипотез.

Естественно, что в процессе проверки гипотезы H_0 о принадлежности анализируемой выборки закону $F(x,\theta)$ могут совершаться ошибки 1-го рода, когда отклоняется справедливая гипотеза H_0, и ошибки 2-го рода, когда гипотеза H_0 не отклоняется при справедливости некоторой конкурирующей гипотезы H_1. Плохо, когда эти ошибки совершаются вследствие некорректного применения критериев.

Среди множества причин некорректности статистических выводов, получаемых при использовании непараметрических критериев согласия, можно выделить 2 основные.

Первая заключается в использовании при проверке сложных гипотез классических результатов, соответствующих проверке простых гипотез. Этого нельзя делать, так как в случае сложных гипотез распределения статистик критериев существенно отличаются от распределений при простых гипотезах. Как итог, увеличивается вероятность ошибок 2-го рода. Об этой причине известно давно [1], но большинству специалистов, использующих критерии в приложениях, она остаётся неизвестной. Это связано с тем, что университетский курс теории вероятностей и математической статистик, читаемый для нематематических специальностей, абсолютно не ориентирован на применение соответствующих методов в приложениях и о ней, как правило, не упоминает.

Вторая причина связана с ошибками округления, которые сопровождают любой процесс измерений. При ошибках округления соизмеримых с ошибками измерений реальные распределения статистик могут существенно отклоняться от асимптотических (предельных). Если не учитывать этот факт, то резко увеличивается вероятность ошибок 1-го рода.

Цель настоящей работы продемонстрировать действие указанных причин на распределения реальных распределений статистик непараметрических критериев согласия. В частности, показать, как меняются распределения статистик непараметрических критериев согласия в зависимости от вида проверяемой гипотезы (простой или сложной) и величины ошибок округления.

2. О проверке простых гипотез

Все наиболее известные непараметрические критерии согласия предлагались для проверки простых гипотез вида $H_0: F(x) = F(x,\theta)$, где $F(x,\theta)$ – функция распределения вероятностей, с которой проверяют согласие наблюдаемой (упорядоченной) выборки $x_1, x_2, \ldots, x_n$ объёмом n, а θ – известное значение параметра (скалярного или векторного). В процессе проверки вычисляется значение S^* статистики критерия S как некоторой функции от выборки и теоретического закона распределения $F(x,\theta)$. Далее, опираясь на асимптотическое (предельное) распределение $G(S|H_0)$ статистики критерия при справедливости H_0, вычисляют достигнутый уровень значимости $p_{value} = P\{S > S^*\} = 1 - G(S^*|H_0)$. Если p_{value} больше заданного α, гипотеза H_0 не отклоняется.

В критерии **Колмогорова** [2] рекомендуется использовать статистику с поправкой Большева в форме [3]

$$S_K = \sqrt{n}D_n + \frac{1}{6\sqrt{n}} = \frac{6nD_n + 1}{6\sqrt{n}}, \qquad (1)$$

где $D_n = \max\left(D_n^+, D_n^-\right)$, $D_n^+ = \max\limits_{1\le i\le n}\left\{\frac{i}{n} - F(x_i,\theta)\right\}$, $D_n^- = \max\limits_{1\le i\le n}\left\{F(x_i,\theta) - \frac{i-1}{n}\right\}$. При справедливости H_0 статистика (1) подчиняется распределению Колмогорова с функцией распределения $K(s) = \sum\limits_{k=-\infty}^{\infty}(-1)^k e^{-2k^2s^2}$.

В настоящей работе рассматриваемые свойства непараметрических критериев согласия мы продемонстрируем на примере критерия Колмогорова, но все выводы можно распространить на критерии Крамера–Мизеса–Смирнова [3], Андерсона–Дарлинга [4, 5], Купера [6], Ватсона [7, 8], а также на критерии Жанга [9, 10], распределения статистик которых зависят от объёмов выборок.

При проверке простых гипотез все перечисленные критерии являются “свободными от распределения”, так как распределения статистик $G(S|H_0)$ не зависят от вида закона $F(x,\theta)$.

При малых объёмах выборок распределения статистик $G(S_n|H_0)$ непараметрических критериев согласия могут несколько отличаться от предельных, что приходится учитывать на практике. Для перечисленных критериев, исключая критерии Жанга, асимптотическими распределениями статистик (вместо реальных распределений $G(S_n|H_0)$), как правило, можно пользоваться при $n \ge 25 \div 30$ [11].

3. О проверке сложных гипотез

При проверке сложных гипотез, когда по той же самой выборке оценивают параметры наблюдаемого закона распределения вероятностей $F(x,\theta)$, все непараметрические критерии согласия теряют свойство “свободы от распределения”. Впервые существование этой проблемы было обозначено в работе [1], которая послужила точкой отсчета для многочисленных попыток её решения. Например, в [12, 13] с использованием статистического моделирования были построены таблицы критических значений для критерия Колмогорова при проверке сложных гипотез относительно нормального и экспоненциального законов. В [14] аналитическими методами аналогичная задача была решена для критерия Крамера–Мизеса–Смирнова. Решению таких же задач были посвящены работы [15, 16]. Жизнь подтвердила, что использование имитационного моделирования является более перспективным направлением при исследованиях распределений статистик критериев, применяемых при проверке различных гипотез.

В наших работах, опирающихся на компьютерные технологии исследований, были построены приближенные модели распределений статистик критериев Колмогорова, Крамера–Мизеса–Смирнова и Андерсона–Дарлинга при проверке различных сложных гипотез относительно ряда законов, часто используемых в приложениях, была показана существенная зависимость распределений статистик от используемого метода оценивания параметров [17]. На базе этих результатов были разработаны рекомендации [18]. Последующие исследования и уточнённые модели распределений статистик для упомянутых выше непараметрических критериев [19-23] послужили основой руководства [11].

При проверке сложной гипотезы на распределение статистики $G(S|H_0)$ критерия влияет совокупность следующих факторов [11]: вид наблюдаемого закона распределения $F(x,\theta)$, соответствующего проверяемой гипотезе H_0; тип оцениваемого параметра и число оцениваемых параметров; используемый метод оценивания параметров [17]; в некоторых ситуациях – конкретное значение параметра (например, значения параметров формы гамма-распределения, бета-распределений, обобщённого нормального закона и других) [11, 22].

Влияние перечисленных факторов на распределения статистик непараметрических критериев согласия (без потери общности) продемонстрируем на статистике (1) критерия Колмогорова. На данном этапе мы не акцентируем внимания на возможном влиянии на распределения статистик $G(S|H_0)$ ошибок округления Δ.

Рис. 1 иллюстрирует изменение распределения статистики (1) в зависимости от типа оцениваемого параметра (сдвига или масштаба) и от числа оцениваемых параметров нормального закона в случае использования оценок максимального правдоподобия (ОМП), а также от метода оценивания. На рисунке показано также распределение статистики для случая оценивания двух параметров нормального закона в результате минимизации статистики (1) (MD-оценки). Как видим, наблюдается существенная зависимость от типа оцениваемого параметра и от используемого метода оценивания. При проверке согласия с другими законами, определяемыми также лишь параметрами сдвига и масштаба, картина будет аналогична. Однако распределения статистик, соответствующие тем же сложным гипотезам, будут отличаться от приведенных на рис. 1, так как распределения $G(S|H_0)$ зависят от вида закона $F(x,\theta)$. Построенные модели распределений статистик критериев при проверке различных сложных гипотез относительно ряда законов $F(x,\theta)$, наиболее часто используемых в различных приложениях и для которых отсутствует зависимость от значения параметра (параметров) формы, приведены в [11, 24].

В [11, раздел 3.5, рис. 3.5] влияние конкретного значения параметра формы на распределение статистики $G(S|H_0)$ при проверке сложной гипотезы демонстрируется на обобщённом нормальном законе с плотностью $f(x)=\dfrac{\theta_2}{2\theta_1\Gamma(1/\theta_2)}\exp\left\{-\left(|x-\theta_0|/\theta_1\right)^{\theta_2}\right\}$, где θ_2 – параметр формы. В частном случае при $\theta_2=2$ обобщённый нормальный закон совпадает с нормальным, а

при $\theta_2 = 1$ – с распределением Лапласа. На упомянутом рисунке 3.5 показаны распределения $G(S|H_0)$ статистики (1) при проверке сложной гипотезы о принадлежности выборки обобщённому нормальному закону в случае оценивания всех 3-х параметров методом максимального правдоподобия в зависимости от значения θ_2. Можно обратить внимание, что с ростом θ_2 распределение $G(S|H_0)$ сначала удаляется от $K(S)$ (до $\theta_2 \approx 1.6$), а затем начинает сближаться с $K(S)$.

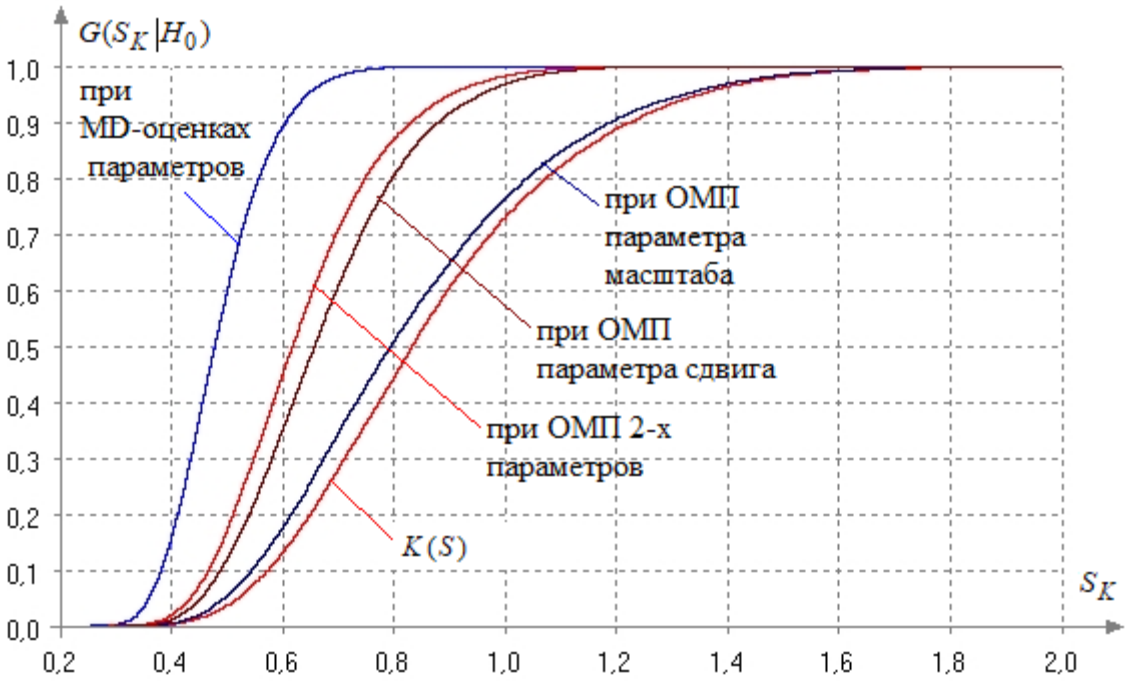

Рис. 1. Предельные распределения статистики (1) при справедливости простой и сложных гипотез о принадлежности выборки нормальному закону при отсутствии ошибок округления

В подобных ситуациях (наличия зависимости $G(S|H_0)$ от значений параметра или параметров формы) исключается возможность заранее построить модель (асимптотического или предельного) закона, так как значение параметра мы узнаём только в процессе проверки гипотезы (при оценивании). Отсюда следует, что распределения статистик применяемых критериев должны находиться (моделироваться) в интерактивном режиме в ходе проводимого статистического анализа [25], а затем использоваться при формировании вывода по итогам проверки сложной гипотезы.

Картина, представленная на рис. 1, показывает, что если в ситуации проверки сложной гипотезы использовать классические результаты, касающиеся проверки простых гипотез, существенно увеличивается вероятность ошибок 2-го рода.

4. Влияние ошибок округления на распределения статистик

Любые измерения фиксируются с некоторой погрешностью округления. Очевидность возможного влияния ошибок округления на статистические выводы признавалась многими авторами. Вопрос оставался лишь в том, как ошибки округления сказываются на распределениях статистик критериев. В наших работах [26-28] показано, как в условиях соизмеримости ошибок округления Δ и среднеквадратичного отклонения ошибок измерения σ изменяются распределения статистик различных статистических критериев.

В данном случае, без потери общности, в условиях соизмеримости Δ и σ проиллюстрируем влияние Δ на распределения $G(S|H_0)$ статистики (1) критерия Колмогорова на примере проверки гипотезы о принадлежности выборок нормальному закону распределения.

На рис. 2 в условиях справедливости простой гипотезы H_0 показана зависимость распределения статистики (1) от величины $\Delta = w\sigma$ при $n = 50$. Очевидно, что отклонением $G(S_{50}|H_0)$ от $K(S)$ можно пренебречь лишь при $\Delta < 0.1\sigma$.

На рис. 3, также при простой гипотезе H_0 и фиксированной ошибке округления $\Delta = 0.1\sigma$ показана зависимость $G(S_n|H_0)$ от объёмов выборок n.

Как можно видеть, при проверке простых гипотез и относительно небольших объёмах выборок наличие ошибок округления приводит не только к отклонению $G(S_n|H_0)$ от распределения $K(S)$, но и делает $G(S_n|H_0)$ достаточно дискретным. Заметим, что этого не происходит в таких ситуациях, например, с распределениями статистик критериев Крамера–Мизеса–Смирнова и Андерсона–Дарлинга.

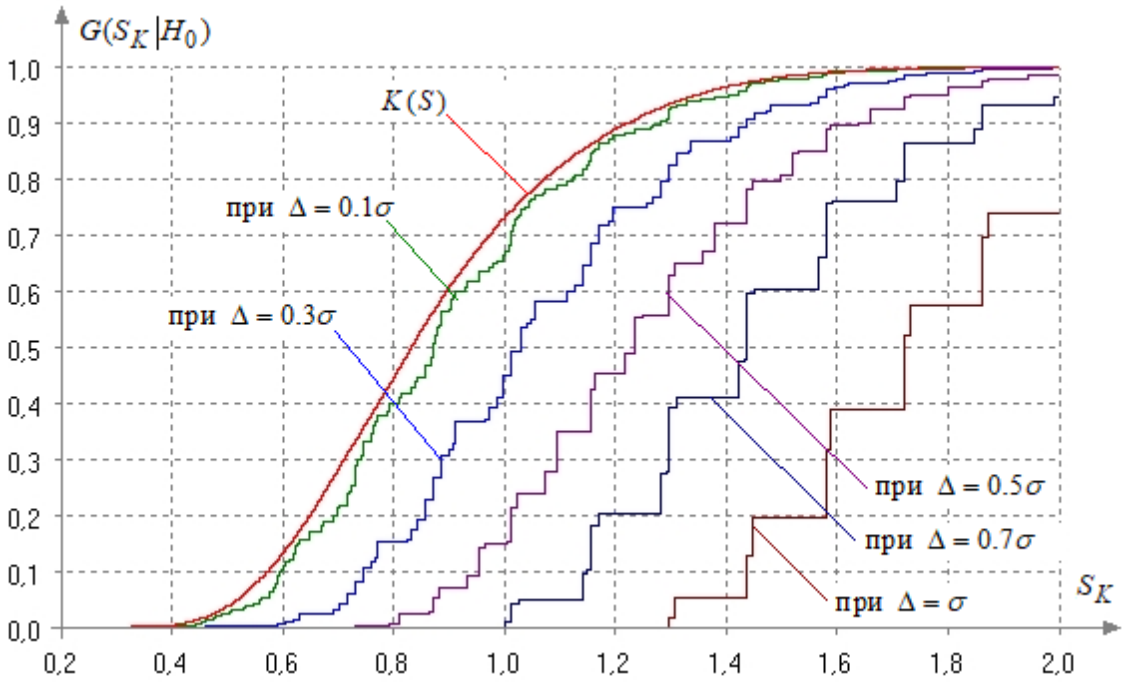

Рис. 2. Зависимость распределения статистики (1) от ошибки округления Δ при справедливости простой гипотезы H_0 о принадлежности выборки нормальному закону при $n = 50$

Какие выводы следуют из представленных результатов исследований?

Во-первых, можно видеть, что наличие ошибок округления приводит к появлению зависимости $G(S|H_0)$ от n.

Во-вторых, признание самого факта наличия округлений в данных исключает возможность использования предельного распределения $K(S)$ в качестве распределения статистики (1) в условиях больших выборок.

В-третьих, в условиях соизмеримости Δ и σ и при относительно небольших объёмах выборок распределения $G(S_n|H_0)$ статистики (1) могут значительно отличаться от $K(S)$, что полностью исключает возможность при проверке гипотезы использовать классические результаты.

В-четвёртых, проведенные исследования показали, что вследствие округлений потеря свойства “свободы от распределения” происходит и в условиях проверки простых гипотез. В частности, при одном и том же соотношении $\Delta = w\sigma$ между Δ и σ наблюдаемого симметричного закона степень отклонения $G(S_n|H_0)$ от $K(S)$ увеличивается в случае законов с более тяжёлыми “хвостами” (по сравнению с нормальным законом).

В ситуации проверки сложной гипотезы при использовании ОМП для оценки 2-х параметров нормального закона мы имеем аналогичную картину влияния Δ на распределения статистики (1) (см. рис. 4 и 5).

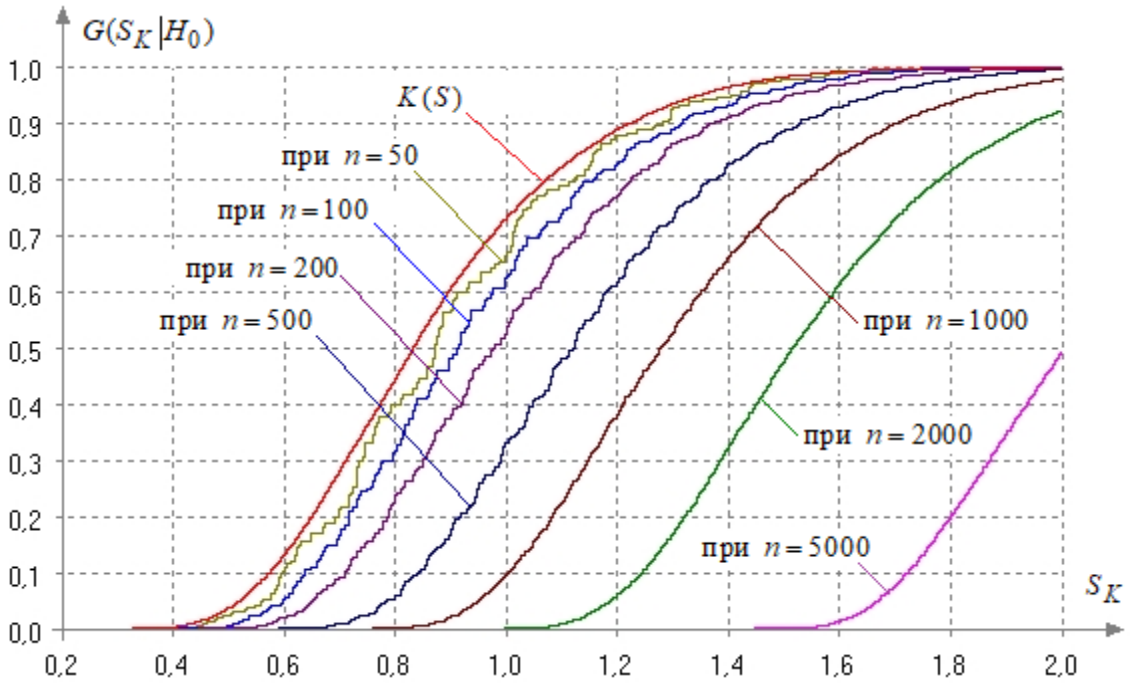

Рис. 3. Зависимость распределения статистики (1) от объёма выборки n при справедливости простой гипотезы H_0 о принадлежности выборки нормальному закону и ошибке округления $\Delta = 0.1\sigma$

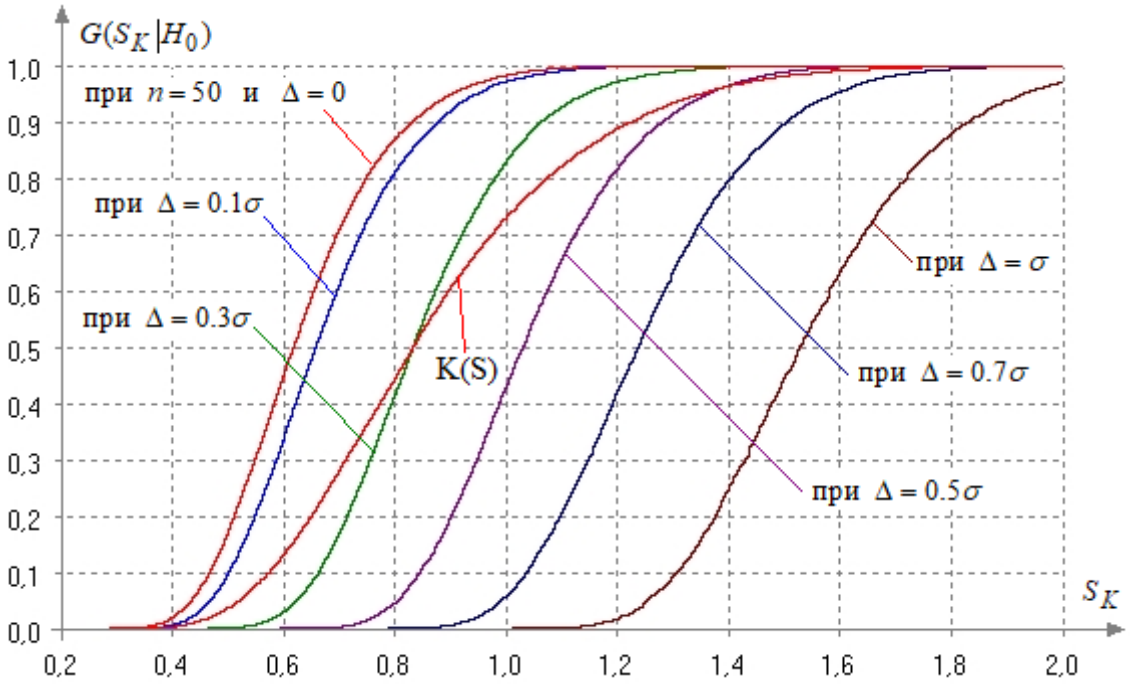

Рис. 4. Зависимость распределения статистики (1) от ошибки округления Δ при справедливости сложной гипотезы H_0 о принадлежности выборки нормальному закону (в случае ОМП) при $n = 50$

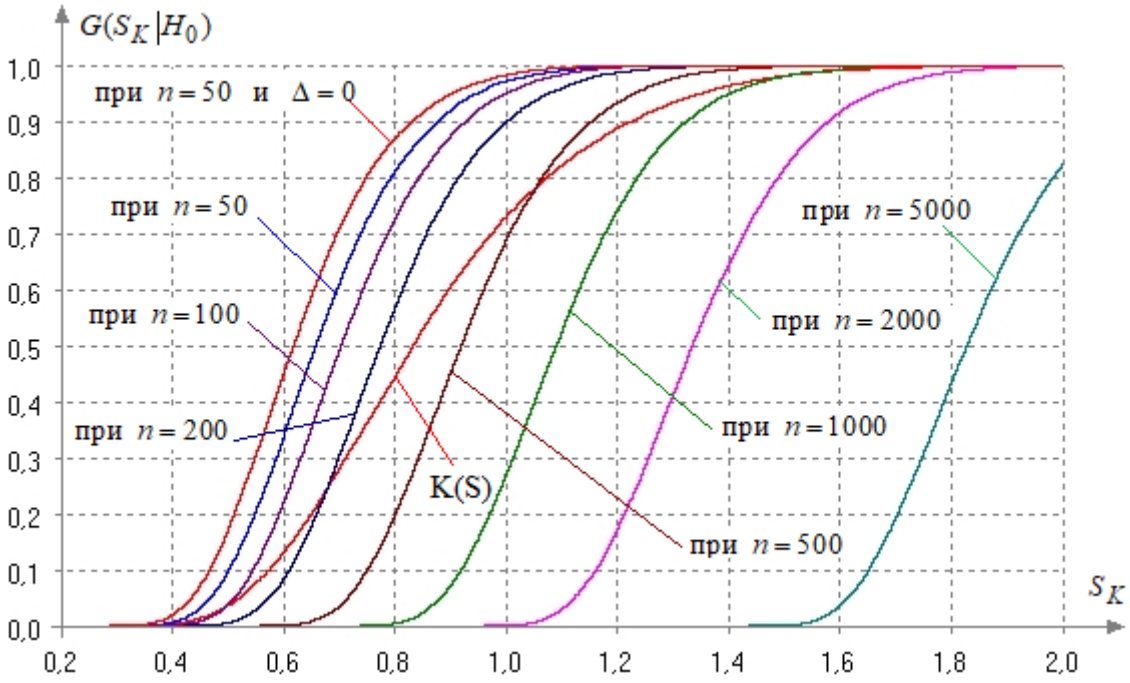

Рис. 5. Зависимость распределения статистики (1) от объёма выборки n при справедливости сложной гипотезы H_0 о принадлежности выборки нормальному закону (в случае ОМП) при $\Delta = 0.1\sigma$

На примере выборки измерений ширины лепестка ириса разноцветного объёмом $n = 50$, представленной ниже и заимствованной в работе [29], проверим сложную гипотезу о её принадлежности нормальному закону.

1.4	1.5	1.5	1.3	1.5	1.3	1.6	1.0	1.3	1.4
1.0	1.5	1.0	1.4	1.3	1.4	1.5	1.0	1.5	1.1
1.8	1.3	1.5	1.2	1.3	1.4	1.4	1.7	1.5	1.0
1.1	1.0	1.2	1.6	1.5	1.6	1.5	1.3	1.3	1.3
1.2	1.4	1.2	1.0	1.3	1.2	1.3	1.3	1.1	1.3

Измерения в см зафиксированы с ошибкой округления $\Delta = 0.1$. С подобными результатами измерений часто сталкиваются в различных приложениях. ОМП параметров сдвига и масштаба нормального закона, вычисленные по выборке, $\hat{\mu} = 1.3260$, $\hat{\sigma} = 0.1958$.

В таблице 1 приведены значения статистик используемых критериев и оценки p_{value}, вычисленные по распределениям статистик критериев, имеющих место в отсутствие округлений (при $\Delta = 0$) и при ошибке округления $\Delta = 0.1$.

Таблица 1. Результаты проверки гипотезы о нормальности

№ п/п	Критерий	Статистика	p_{value}	
			$\Delta = 0$	$\Delta = 0.1$
1	Колмогорова	1.065	0.010	0.453
2	Крамера–Мизеса–Смирнова	0.154	0.023	0.467
3	Андерсона–Дарлинга	0.975	0.015	0.319

4	Купера	1.886	0.001	0.432
5	Ватсона	0.153	0.014	0.451
6	Жанга Z_A	3.387	0.050	0.214
7	Жанга Z_C	12.31	0.055	0.226
8	Жанга Z_K	2.563	0.008	0.283

По разности значений p_{value} можно судить о различии соответствующих распределений статистик. И при $\Delta = 0.1$ гипотеза о нормальности не отклоняется.

5. Заключение

Таким образом, при идентификации закона распределения ошибок измерений с использованием непараметрических критериев согласия наряду со сложностью проверяемой гипотезы следует учитывать возможное влияние ошибок округления.

Ошибки округления всегда сопровождают процесс измерений. В ситуации, когда $\Delta << \sigma$ и n меньше некоторого $n_{\max}$, зависящего от n и σ, влиянием Δ на $G(S_n|H_0)$ можно пренебречь, Но при $n > n_{\max}$ реальное распределение статистики отклоняется от асимптотического. В таком случае использование асимптотического распределения приводит к увеличению вероятностей ошибок 1-го рода, т.е. к отклонению справедливой гипотезы H_0. При соизмеримости Δ и σ такая ситуация может иметь место и при малых объёмах выборок, а с ростом n она будет только усугубляться.

Единственным выходом, способным обеспечивать корректность выводов по применяемым критериям в нестандартных условиях (при проверке сложных гипотез и влиянии Δ на $G(S_n|H_0)$), является использование реальных распределений статистик этих критериев (имеющих место в этих нестандартных условиях). Эта задача должна решаться в интерактивном режиме (в процессе проверки) и опираться на компьютерные технологии исследования и аппарат математической статистики.

Литература

1. *Kac M., Kiefer J., Wolfowitz J.* On tests of normality and other tests of goodness of fit based on distance methods // The Annals of Mathematical Statistics. – 1955. – Vol. 26. – P. 189–211.
2. *Kolmogoroff A.N.* Sulla determinazione empirica di una legge di distribuzione // Giornale del Istituto Italiano degli Attuari. – 1933. – Vol. 4. – No. 1. – P. 83–91.
3. *Большев Л.Н., Смирнов Н.В.* Таблицы математической статистики. – М. : Наука, 1983. – 416 с.
4. *Anderson T.W., Darling D.A.* Asymptotic theory of certain “Goodness of fit” criteria based on stochastic processes // The Annals of Mathematical Statistics. – 1952. – Vol. 23. – P. 193–212.
5. *Anderson T.W., Darling D.A.* A test of goodness of fit // Journal of the American Statistical Association. – 1954. – Vol. 29. – P. 765–769.
6. *Kuiper N.H.* Tests concerning random points on a circle // Proc. Konikl. Nederl. Akad. Van Wettenschappen. Series A. – 1960. – Vol. 63. – P.38-47.
7. *Watson G.S.* Goodness-of-fit tests on a circle. I. // Biometrika. – 1961. – Vol. 48. – No. 1-2. – P. 109-114.
8. *Watson G.S.* Goodness-of-fit tests on a circle. II. // Biometrika. – 1962. – Vol. 49. – No. 1-2. – P.57-63.
9. *Zhang J.* Powerful goodness-of-fit and multi-sample tests // PhD Thesis. York University, Toronto. 2001. – 113 p. URL: http://www.collectionscanada.gc.ca/obj/s4/f2/dsk3/ftp05/NQ66371 .pdf (дата обращения 28.02.2021).
10. *Zhang J.* Powerful goodness-of-fit tests based on the likelihood ratio // Journal of the Royal Statistical Society: Series B. – 2002. – Vol. 64. – № 2. – P.281-294.

11. *Лемешко Б.Ю.* Непараметрические критерии согласия: Руководство по применению : монография. – М.: ИНФРА-М, 2014. – 163 с. DOI: 10.12737/11873
12. *Lilliefors H.W.* On the Kolmogorov-Smirnov test for normality with mean and variance unknown // Journal of the American Statistical Association. – 1967. – Vol. 62. – P. 399–402.
13. *Lilliefors H.W.* On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown // Journal of the American Statistical Association. – 1969. – Vol. 64. – P. 387–389.
14. *Мартынов Г.В.* Критерии омега-квадрат. – М. : Наука, 1978. – 80 с.
15. *Тюрин Ю.Н.* О предельном распределении статистик Колмогорова–Смирнова для сложной гипотезы // Изв. АН СССР. Сер. мат. – 1984. – Т. 48, № 6. – С. 1314–1343.
16. *Тюрин Ю.Н., Саввушкина Н.Е.* Критерии согласия для распределения Вейбулла–Гнеденко // Изв. АН СССР. Сер. Техн. кибернетика. – 1984. – № 3. – С. 109–112.
17. *Лемешко Б.Ю., Постовалов С.Н.* О зависимости распределений статистик непараметрических критериев и их мощности от метода оценивания параметров // Завод. лаб. Диагностика материалов. – 2001. – Т. 67, № 7. – С. 62–71.
18. Р 50.1.037–2002. Рекомендации по стандартизации. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Ч. II. Непараметрические критерии. – М. : Изд-во стандартов, 2002. – 64 с.
19. *Lemeshko B.Yu., Lemeshko S.B.* Distribution models for nonparametric tests for fit in verifying complicated hypotheses and maximum-likelihood estimators. P. 1 // Measurement Techniques. – 2009. – Vol. 52. № 6. – P. 555–565.
20. *Lemeshko B.Yu., Lemeshko S.B.* Models for statistical distributions in nonparametric fitting tests on composite hypotheses based on maximum-likelihood estimators. P. II // Measurement Techniques. – 2009. – Vol. 52. № 8. – P. 799–812.
21. *Lemeshko B.Yu., Lemeshko S.B., Postovalov S.N.* Statistic Distribution Models for Some Nonparametric Goodness-of-Fit Tests in Testing Composite Hypotheses // Communications in Statistics – Theory and Methods. – 2010. – Vol. 39. № 3. – P. 460–471.
22. *Lemeshko B.Yu., Lemeshko S.B.* Models of Statistic Distributions of Nonparametric Goodness-of-Fit Tests in Composite Hypotheses Testing for Double Exponential Law Cases // Communications in Statistics - Theory and Methods – 2011. – Vol. 40. – No. 16. – P. 2879-2892.
23. *Lemeshko B.Yu., Gorbunova A.A.* Application of nonparametric Kuiper and Watson tests of goodness-of-fit for composite hypotheses // Measurement Techniques. – 2013. – Vol. 56. – № 9. – P.965-973.
24. *Лемешко Б.Ю.* Статистический анализ данных, моделирование и исследование вероятностных закономерностей. Компьютерный подход : монография / Б.Ю. Лемешко, С.Б. Лемешко, С.Н. Постовалов, Е.В. Чимитова. – Новосибирск : Изд-во НГТУ, 2011. – 888 с.
25. *Lemeshko B.Yu.*, *Lemeshko S.B., Rogozhnikov A.P.* Interactive investigation of statistical regularities in testing composite hypotheses of goodness of fit // Statistical Models and Methods for Reliability and Survival Analysis : monograph. – Wiley-ISTE, 2013. – Chap. 5. – P. 61–76.
26. *Лемешко Б.Ю., Лемешко С.Б.* Влияние округления на свойства критериев проверки статистических гипотез // Автометрия. 2020. – Т. 56, № 3. – С. 35-45. DOI: 10.15372/AUT20200305
27. *Лемешко Б.Ю., Лемешко С.Б.* О влиянии ошибок округления на распределения статистик критериев согласия // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2020. – № 53. – С. 47-60. DOI: 10.17223/19988605/53/5
28. *Lemeshko B.Y., Lemeshko S.B.* About the effect of rounding on the properties of tests for testing statistical hypotheses // Journal of Physics: Conference Series. – 2021. – Vol. 1715. 012063. DOI: 10.1088/1742-6596/1715/1/012063
29. *Fisher R.A.* The Use of Multiple Measurements in Taxonomic Problems // Annals of Eugenics. 1936. – Vol. 7. – P. 179-188

Лемешко Борис Юрьевич

Профессор кафедры прикладной и теоретической информатики НГТУ, д.т.н., профессор (630073, Новосибирск, пр. Карла Маркса, 20), тел. (383) 346-06-00, e-mail: Lemeshko@ami.nstu.ru, http://www.ami.nstu.ru/~headrd/

Лемешко Станислав Борисович

С.н.с. Центра статистических технологий НГТУ, к.т.н., (630073, Новосибирск, пр. Карла Маркса, 20), тел. (383) 346-06-00, e-mail: skyer@mail.ru

On the reasons for the incorrectness of conclusions when using nonparametric goodness-of-fit tests

B. Yu. Lemeshko, S. B. Lemeshko

Novosibirsk State Technical University

In most cases, the incorrect application of nonparametric goodness-of-fit criteria in applications is based on two reasons. The first reason is that when testing composite hypotheses and evaluating the parameters of the law for the analyzed sample, the classical results associated with testing simple hypotheses are used. The second reason is associated with the presence of round-off errors, which can significantly change the distributions of test statistics. Asymptotic results when testing simple and composite hypotheses can be used when rounding errors are much less than the standard deviation of the distribution law of measurement errors and sample sizes n not exceeding some maximum values. For sample sizes larger than these maximum values, the real distributions of test statistics deviate from asymptotic ones towards larger statistics values. It is shown that the only way out that ensures the correctness of conclusions according to the applied criteria is the use of real distributions of statistics, which can be in an interactive mode.

Keywords: hypothesis testing, nonparametric goodness-of-fit tests, simple hypothesis, composite hypothesis, distribution of statistics, round-off errors, statistical modeling, achieved level of significance, error of the 1st kind, error of the 2nd kind