

Б. Ю. Лемешко

Использование статистического моделирования для решения проблем применения критериев согласия

УДК: 519.8

DOI: 10.55648/978-5-91434-085-5-2023-117-123

Процесс принятия решения о результатах проверки гипотезы с использованием любого критерия опирается на распределение его статистики. Распределения статистик при проверке простых и сложных гипотез существенно отличаются. При проверке сложных гипотез распределения статистик непараметрических критериев согласия зависят от ряда факторов и часто неизвестны. Более того, распределения статистик существенно зависят от ошибок округления, что исключает возможность применения классических результатов. Единственный выход для решения имеющихся проблем заключается в использовании методов статистического моделирования.

Ключевые слова: проверка гипотез, непараметрические критерии согласия, статистическое моделирование, ошибка 1-го рода, ошибка 2-го рода, ошибки округления.

1. Введение

Курс математической статистики, читаемый в университетах, содержит лишь краткое изложение основ дисциплины и не даёт представления о проблемах применения соответствующих методов на практике. Статистические методы и критерии обладают заданными (асимптотическими) свойствами при выполнении определённых (стандартных) предположений. В приложениях эти предположения могут не выполняться, а в нестандартных условиях свойства применяемых методов и критериев могут существенно изменяться. В литературе учебного характера внимание на нюансах такого характера практически не акцентируется. В учебниках можно встретить неверные рекомендации по использованию и примеры некорректного применения различных критериев проверки гипотез. Не всегда корректно используются критерии проверки гипотез и в популярных программных системах. Всё это в приложениях зачастую приводит к неэффективному применению методов статистического анализа и к неверным статистическим выводам.

В данном случае речь идёт об изменении свойств непараметрических критериев согласия (об изменении распределений статистик) под воздействием различных факторов. В таких ситуациях использование классических результатов, касающихся распределений статистик критериев, как правило, приводит к увеличению вероятностей ошибок или первого, или второго рода. А гарантировать корректность выводов можно только за счёт использования методов статистического моделирования для исследования распределений статистик критериев в нестандартных условиях, что можно делать, опираясь на соответствующее программное обеспечение.

2. Непараметрические критерии согласия при проверке простых гипотез

При проверке гипотез с применением критерия **Колмогорова** [1] рекомендуется использовать статистику с поправкой Большева в форме [2]

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

где $D_n = \max(D_n^+, D_n^-)$, $D_n^+ = \max_{1 \leq i \leq n} \left\{ \frac{i}{n} - F(x_i, \theta) \right\}$, $D_n^- = \max_{1 \leq i \leq n} \left\{ F(x_i, \theta) - \frac{i-1}{n} \right\}$, n – объем выборки; $x_1, x_2, \dots, x_n$ здесь и далее – упорядоченные по возрастанию выборочные значения; $F(x, \theta)$ – функция закона распределения, согласие с которым проверяют. Величина S_K при простой гипотезе в пределе подчиняется закону распределения Колмогорова $K(S)$ [2].

В критерии типа ω^2 **Крамера–Мизеса–Смирнова** статистика имеет вид

$$S_\omega = n\omega_n^2 = \frac{1}{12n} + \sum_{i=1}^n \left\{ F(x_i, \theta) - \frac{2i-1}{2n} \right\}^2. \quad (2)$$

которая при справедливости простой проверяемой гипотезе в пределе подчиняется закону с функцией распределения $a1(s)$, приведенной в [2].

В критерии типа Ω^2 **Андерсона–Дарлингга** [3, 4] используется статистика

$$S_\Omega = -n - 2 \sum_{i=1}^n \left\{ \frac{2i-1}{2n} \ln F(x_i, \theta) + \left(1 - \frac{2i-1}{2n} \right) \ln(1 - F(x_i, \theta)) \right\}. \quad (3)$$

При проверке простой гипотезы эта статистика в пределе подчиняется закону с функцией распределения $a2(s)$ [2].

В критерии **Купера** [5] в качестве меры расхождения используется величина $V_n = D_n^+ + D_n^-$, где D_n^+ , D_n^- находятся так же, как и в критерии Колмогорова, а в качестве статистики используется

$$s = \sqrt{n} \cdot V_n. \quad (4)$$

Для случая проверки простой гипотезы H_0 в [5] приведено её предельное распределение.

Статистика критерия **Ватсона** [6, 7] используется в следующей форме

$$U_n^2 = \sum_{i=1}^n \left(F(x_i, \theta) - \frac{i-1/2}{n} \right)^2 - n \left(\frac{1}{n} \sum_{i=1}^n F(x_i, \theta) - \frac{1}{2} \right)^2 + \frac{1}{12n}. \quad (5)$$

В [6, 7] приведено также асимптотическое распределение статистика U_n^2 , имеющее место при проверке простой гипотезы.

С ростом объёмов выборок реальные распределения статистик достаточно быстро приближаются к асимптотическим (см. Табл. 1).

Таблица 1. Сходимость реальных распределений статистик к асимптотическим

Критерий	Статистика	Нет отклонения от асимптотического распределения	Можно пренебречь отклонением
Колмогорова	(1)	при $n \geq 20$	при $n \geq 15$
Крамера–Мизеса–Смирнова	(2)	при $n \geq 18$	при $n \geq 12$
Андерсона–Дарлинга	(3)	при $n \geq 10$	при $n \geq 6$
Купера	(4)	при $n \geq 55$	при $n \geq 10$
Ватсона	(5)	при $n \geq 75$	при $n \geq 10$

В таблице указаны объёмы выборок, начиная с которых реальные распределения статистик (1) – (5) при справедливости простой проверяемой гипотезы практически (визуально) не отличаются от асимптотических, а также приведены объёмы выборок, при которых отклонение этих распределений от асимптотических не отражается на статистических выводах. Последнее объясняется тем, что все эти критерии правосторонние, и при имеющемся отличии реального распределения статистики от асимптотического, на правом “хвосте” эти распределения практически совпадают.

3. Непараметрические критерии согласия при сложных гипотезах

При проверке сложных гипотез, когда по этой же выборке оцениваются параметры наблюдаемого закона распределения, все непараметрические критерии согласия теряют свойство “свободы от распределения” [8], и асимптотические распределения статистик $G(S|H_0)$ становятся зависящими от факторов, определяющих “сложность” гипотезы. В частности, на распределение статистики $G(S|H_0)$ критерия влияют следующие факторы:

- (а) – вид закона распределения $F(x, \theta)$, соответствующего проверяемой гипотезе H_0 ;
- (b) – тип оцениваемого параметра;
- (с) – число оцениваемых параметров;
- (d) – используемый метод оценивания параметров [9];
- (е) – в некоторых ситуациях распределения статистик зависят от конкретных значений параметров закона $F(x, \theta)$ (например, от значений параметров формы обобщённого нормального закона [9], гамма-распределения, бета-распределений и т.п.).

Хотя проблемы, связанные с применением непараметрических критериев согласия для проверки сложных гипотез, известны давно [8], с ними крайне редко знакомят в курсах лекций или освещают в учебной литературе. В то же время, отличия в асимптотических распределениях тех же самых статистик при проверке простых и (различных) сложных гипотез настолько существенны, что пренебрегать этим абсолютно недопустимо. Большинство самых грубых ошибок применения непараметрических критериев согласия, приводящих к существенному увеличению вероятностей ошибок 2-го рода, связано именно с незнанием этих обстоятельств.

Разрешение возникших проблем аналитическими методами связано с серьёзными трудностями, а удачные попытки ряда авторов приводили и приводят к положительным результатам лишь частного характера. В то же время успешное использование в этих целях методов статистического моделирования, впервые продемонстрированное в [10, 11], позволяет получать полезные результаты с меньшими затратами. На базе результатов статистического моделирования распределений статистик критериев, имеющих место при проверке различных сложных гипотез, и построенных для них моделей были разработаны рекомендации Госстандарта РФ [12]. Дальнейшее развитие средств вычислительной техники и программного обеспечения исследований позволили уточнить [13, 14, 15] модели и таблицы критических значений распределений статистик непараметрических критериев согласия, представленные ранее в [12], и подготовить руководство [16].

Построение моделей для асимптотических распределений $G(S|H_0)$ статистик критериев ограничено множеством тех параметрических законов $F(x, \theta)$, в случае которых проявляется влияние только различных комбинаций факторов (a) – (d). При наличии влияния фактора (e) возможность исследования неизвестного распределения статистики $G(S|H_0)$ или оценки по этому распределению достигнутого уровня значимости p_v появляется только после нахождения оценки параметра или параметров, от которых зависит $G(S|H_0)$. В такой ситуации применение критерия и реализация процедуры интерактивного исследования возможны только в рамках соответствующего программного обеспечения. Таким примером является программная система [17].

Степень отклонения реальных распределений статистик от асимптотических при проверке сложных гипотез и малых n имеет такой же порядок, как и в случае проверки простых (см. Табл. 1).

4. Влияние ошибок округления на распределения статистик

Все классические результаты, связанные с проверкой простых и сложных гипотез, имеют место в условиях отсутствия ошибок округления. Нет проблем с использованием этих результатов, когда ошибки округления $\Delta \ll \sigma$, где σ – среднеквадратичное отклонение ошибки измерения, и в выборках отсутствуют повторяющиеся значения.

На современном этапе развития приходится сталкиваться с необходимостью анализа выборок очень большого объёма. В таких выборках данные представлены с определёнными ошибками округления Δ и в них присутствует большое количество повторяющихся значений. Вследствие этого с ростом n распределения статистик $G(S_n|H_0, \Delta)$ начинают отклоняться от имеющих место асимптотических $G(S|H_0)$.

Для того, чтобы использование классических результатов относительно соответствующих критериев (при простых или сложных гипотезах) было корректным, в [18] рекомендуется извлекать из Big Data выборки объёма не превышающего $n_{\max}$, зависящего от Δ , при котором распределение статистики $G(S_{n_{\max}}|H_0, \Delta)$ практически не отличается от асимптотического $G(S|H_0)$. Оценка $n_{\max}$ трудности не представляет, так как очевидно, что она не должна превышать количества уникальных значений случайной величины, которое может присутствовать в выборках из данного закона при данной величине Δ , уменьшенного в 3-4 раза.

В приложениях различного характера часто приходится сталкиваться с выборками, полученными, в том числе, в результате высокоточных измерений, элементы которых отличаются друг от друга буквально единицами последнего десятичного разряда. В таких измерениях “на пределе точности” ошибка округления Δ оказывается достаточно близкой к σ ошибки измерения наблюдаемой величины, а в выборках всегда значительно присутствие повторяющихся значений. Как показывают исследования [19, 20], распределения статистик критериев, применяемых к подобным выборкам, даже не пытаются сходиться к асимптотическим $G(S|H_0)$, а с ростом n только удаляются от них (см. рис. 1 для критерия Ватсона).

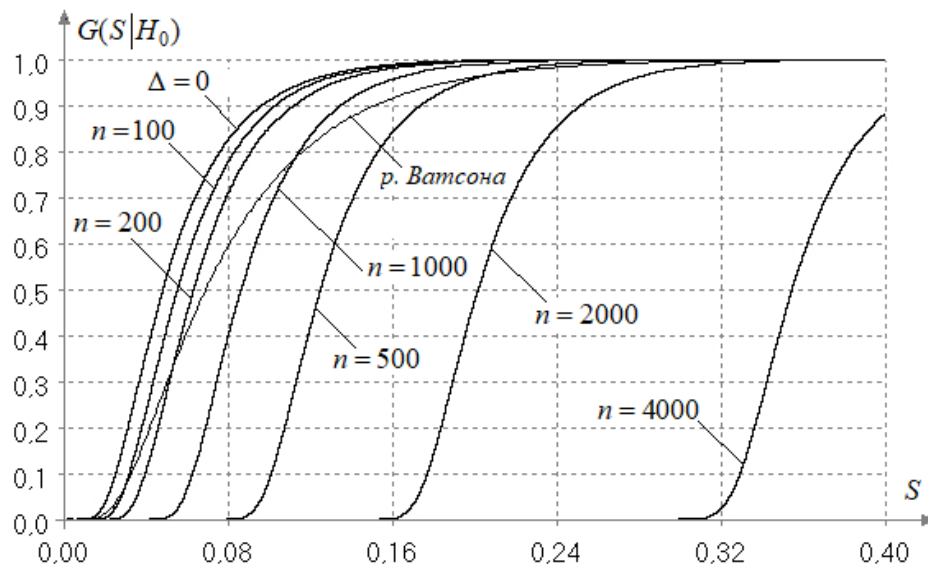


Рис. 1. Зависимость распределения статистики (5) критерия Ватсона от n при справедливости сложной гипотезы H_0 о принадлежности выборки нормальному закону при $\Delta = 0.1\sigma$

Следует констатировать, что наличие влияния ошибок округления для критериев со статистиками (1) – (5) в ситуации проверки простых гипотез приводит к появлению зависимости распределений статистик критериев от n , а при заданной величине Δ – от значения параметра масштаба σ предполагаемого закона [20]. В ситуации проверки сложных гипотез эти зависимости накладываются на зависимости, определяемые факторами (a) – (e).

Подобным же изменениям в связи с влиянием Δ подвергаются распределения статистик других непараметрических критериев согласия, критериев типа χ^2 , специальных критериев нормальности [21].

В заключение подчеркнём, что корректность выводов по всем параметрическим и непараметрическим критериям согласия при Δ близким к σ может быть обеспечена лишь за счёт статистического моделирования распределений $G(S_n|H_0, \Delta)$ или оценивания p_v в этих (нестандартных) условиях. Такая возможность предусмотрена в программной системе [17].

Литература

1. *Kolmogoroff A. N.* Sulla determinazione empirica di una legge di distribuzione // *G. Ist. Ital. attuar.*, 1933, vol. 4, no. 1, pp. 83–91.
2. *Большев Л. Н., Смирнов Н. В.* Таблицы математической статистики. М. : Наука, 1983. 416 с.
3. *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, no. 2, pp. 193–212. <https://www.jstor.org/stable/2236446>
4. *Anderson T. W., Darling D. A.* A test of goodness of fit // *Journal of the American Statistical Association*, 1954, vol. 29, pp. 765–769.
5. *Kuiper N. H.* Tests concerning random points on a circle // *Proc. Koninkl. Nederl. Akad. Van Wetenschappen*, 1960, Series A, vol. 63, pp. 38–47.
6. *Watson G. S.* Goodness-of-fit tests on a circle. I. // *Biometrika*, 1961, vol. 48, no. 1/2. pp. 109–114. <https://doi.org/10.2307/2333135>
7. *Watson G. S.* Goodness-of-fit tests on a circle. II. // *Biometrika*, 1962, vol. 49, no. 1/2. pp. 57–63. <https://doi.org/10.2307/2333467>
8. *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, pp. 189–211. <https://www.jstor.org/stable/2236876>
9. *Лемешко Б. Ю., Маклаков А. А.* Непараметрические критерии при проверке сложных гипотез о согласии с распределениями экспоненциального семейства // *Автометрия*. 2004. № 3. С. 3–20.
10. *Lilliefors H. W.* On the Kolmogorov-Smirnov test for normality with mean and variance unknown // *Journal of the American Statistical Association*, 1967, vol. 62, pp. 399–402. <https://doi.org/10.1080/01621459.1967.10482916>
11. *Lilliefors H. W.* On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown // *Journal of the American Statistical Association*, 1969, vol. 64, pp. 387–389. <https://doi.org/10.1080/01621459.1969.10500983>
12. *P 50.1.037–2002.* Рекомендации по стандартизации. Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Часть II. Непараметрические критерии. М.: Изд-во стандартов, 2002. 64 с.
13. *Лемешко Б. Ю., Лемешко С. Б.* Модели распределений статистик непараметрических критериев согласия при проверке сложных гипотез с использованием оценок максимального правдоподобия. Ч. I // *Измерительная техника*. 2009. № 6. С. 3–11.
14. *Лемешко Б. Ю., Лемешко С. Б.* Модели распределений статистик непараметрических критериев согласия при проверке сложных гипотез с использованием оценок максимального правдоподобия. Ч. II // *Измерительная техника*. 2009. № 8. С. 17–26.
15. *Lemeshko B. Yu., Lemeshko S. B. and 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, no. 3, pp. 460–471. <https://doi.org/10.1080/03610920903140148>
16. *Лемешко Б. Ю.* Непараметрические критерии согласия: Руководство по применению: монография. М.: ИНФРА–М, 2014. 163 с. <https://doi.org/10.12737/11873>
17. Статистический анализ интервальных наблюдений одномерных непрерывных случайных величин «Интервальная статистика 5.4»: свидетельство о гос. регистрации программы для ЭВМ № 2018666213 / Лемешко Б.Ю., Лемешко С.Б., Блинов П.Ю., Веретельникова И.В., Новикова А.Ю. – Заявка № 2018663206; заявл. 22.11.2018; зарег. 13.12.2018. <https://ami.nstu.ru/~headrd/ISW.htm> (дата обращения: 01.03.2023)

18. Лемешко Б. Ю., Лемешко С. Б., Семёнова М. А. К вопросу статистического анализа больших данных // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2018. № 44. С. 40–49. <https://doi.org/10.17223/19988605/44/5>
19. 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. <https://doi.org/10.17223/19988605/53/5>
20. Лемешко Б. Ю., Лемешко С. Б. Непараметрические критерии согласия при проверке нормальности в условиях округления измерений // Системы анализа и обработки данных. 2022. № 2 (86). С. 21–38. <https://doi.org/10.17212/2782-2001-2022-2-21-38>.
21. Лемешко Б. Ю. Критерии проверки отклонения распределения от нормального закона. Руководство по применению : монография. 2-е изд., перераб. и доп. Москва : ИНФРА-М, 2023. 353 с. <https://doi.org/10.12737/1896110>

Лемешко Борис Юрьевич

д.т.н., профессор, профессор кафедры теоретической и прикладной информатики НГТУ (630073, Новосибирск, пр. К. Маркса, 20), тел. (383) 346-06-00, e-mail: lemeshko@ami.nstu.ru.

Using Statistical Simulating to Solve the Problems of Applying Goodness of fit Tests

B. Yu. Lemeshko

The process of deciding on the results of testing a hypothesis using any test relies on the distribution of its statistics. The distributions of statistics for testing simple and composite hypotheses are significantly different. When testing composite hypotheses, the distributions of the statistics of nonparametric goodness-of-fit tests depend on a number of factors and are often unknown. Moreover, the distributions of statistics are highly dependent on rounding errors, which excludes the possibility of applying the classical results. The only way to solve the existing problems is to use statistical simulating methods.

Keywords: hypothesis testing, nonparametric goodness of fit tests, statistical simulating, type 1 error, type 2 error, rounding errors.