Комбинаторный способ идентификации малой выборки

A combinatorial method of small sample identification

Воловик А.В.
Volovik A.V.

АО «ОДК-Климов», Российская Федерация, Санкт-Петербург
JSC Klimov, Russian Federation, Saint Petersburg

volovik_aleksandr@mail.ru

Воловик А.В.

Резюме. Цель. *Для повышения достоверности принимаемых решений о равномерности распределений по выборкам ограниченного объема разработан комбинаторный метод формирования критерия на основе сочетаний без повторений выборочных значений.* **Методы.** *В статье применяются методы теории вероятностей, математической статистики и комбинаторики.* **Результаты.** *Предложенный критерий обладает высокой эффективностью для различения выборок малого объема при проверке статистически близких гипотез, таких как гипотеза о равномерном законе распределения и гипотеза о бета-распределении 1-го рода.* **Выводы.** *Предлагаемый в статье подход позволяет реализовать процедуру последовательного анализа (обнаружение «разладки» процесса). Такая процедура дает возможность достоверно выявлять «разладку» (отклонение распределения наблюдений от равномерного закона) процесса с достаточной для практики интенсивностью при помощи рекуррентных соотношений.*

Abstract. Aim. *For the purpose of improving the reliability of decisions regarding the uniformity of distributions over samples of limited size, a combinatorial method has been developed for defining a criterion based on simple combinations of sample values.* **Methods.** *The paper uses methods of the probability theory, mathematical statistics, and combinatorics.* **Results.** *The proposed criterion is highly efficient for distinguishing small samples when testing statistically similar hypotheses, such as the hypothesis of a uniform distribution law and the hypothesis of a beta distribution of the first kind.* **Conclusions.** *The approach proposed in the paper enables a sequential analysis procedure (detection of process "imbalance"). This procedure makes it possible to reliably detect the "imbalance" (deviation of the distribution of observations from the uniform law) of a process with a practically sufficient intensity using recurrent relations.*

Ключевые слова: *малая выборка, вариационный ряд, плотность распределения, статистика, гипотеза, достигаемый уровень значимости, последовательный анализ.*
Keywords: *small sample, static series, distribution density, statistics, hypothesis, achieved significance value, sequential analysis.*

Для цитирования: *Воловик А.В. Комбинаторный способ идентификации малой выборки // Надежность. 2024. №2. С. 3-7. https://doi.org/10.21683/1729-2646-2024-24-2-3-7*
For citation: *Volovik A.V. A combinatorial method of small sample identification. Dependability 2024;2:3-7. https://doi.org/10.21683/1729-2646-2024-24-2-3-7*

Поступила: *05.10.2023* / **После доработки:** *05.04.2024* / **К печати:** *10.06.2024*
Received on: *05.10.2023* / **Revised on:** *05.04.2024* / **For printing:** *10.06.2024*

Введение

Роль равномерного распределения вероятностей в задачах разработки стохастических моделей трудно переоценить. Используя вероятностное интегральное преобразование [1], к равномерному закону можно свести любое распределение вероятностей. В этом пространстве некоторые задачи могут быть решены наиболее эффективно, что предопределяет необходимость обоснования исходного допущения о равномерности с требуемой достоверностью.

Сложность решения задач идентификации равномерного распределения в условиях выборок малого объема обусловлена тем обстоятельством, что искомое решение часто в сильной степени зависит от объема выборки и не всегда является достаточно эффективным.

Одним из способов повышения достоверности принимаемых статистических решений по выборкам ограниченного объема может служить комбинаторный метод формирования критерия [2]. Его суть заключается в разработке статистики, которая объединяет комбинации частных решений, получаемых

на основе описанного в [3] вариационного критерия равномерности.

Метод

Для позиционирования предложенного в [3] вариационного критерия равномерности среди существующих и достаточно изученных критериев отклонения распределения от равномерного закона [4] целесообразно исследовать его мощность для случая проверки гипотезы H_0 о равномерном распределении случайной величины X на интервале [0;1] при альтернативной гипотезе H_1 о бета-распределении [4, 5] с плотностью

$$f(x)=\frac{1}{B(\lambda,\mu)}x^{\lambda-1}(1-x)^{\mu-1},\ 0<x<1,\ \lambda>0,\ \mu>0, \quad (1)$$

где $B(\lambda, \mu)$ – бета-функция;

λ, μ – параметры распределения.

Бета-распределение (1) с параметрами $\lambda = 1$ и $\mu = 1$ вырождается в равномерное на интервале [0;1]. Для альтернативной гипотезы H_1 рассмотрим параметры [4] $\lambda = 1{,}5$ и $\mu = 1{,}5$.

Получить аналитическое выражение распределения вариационного критерия [3] для выборки минимального объема $n = 2$

$$\upsilon=\frac{x_{(1)}}{x_{(2)}}, \quad (2)$$

где $x_{(1)} \le x_{(2)}$, представляется возможным только при параметрах $\lambda = 1$ и $\mu = 1$ (равномерный закон [6]). Для альтернативной гипотезы H_1 с параметрами $\lambda = 1{,}5$ и $\mu = 1{,}5$ решение возможно получить численным способом. На рис. 1 приведены графики плотности $g_\upsilon(\upsilon)$ критерия (2) для рассматриваемых гипотез.

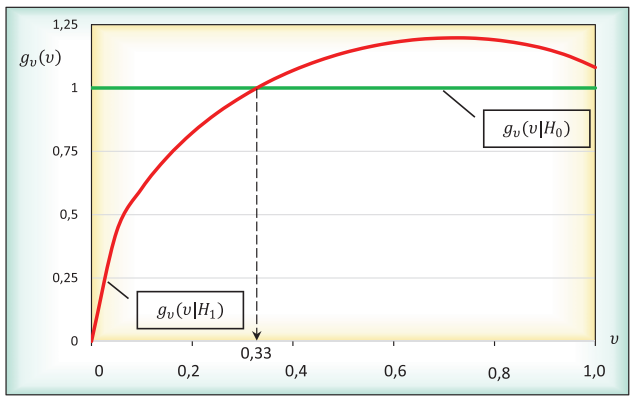

Рис. 1. Плотности распределений $g_\upsilon(\upsilon)$ для гипотез H_0 и H_1

Из рис. 1 видно, что для критерия υ целесообразно выбрать правостороннюю критическую область. Численным решением было установлено, что вероятность отклонения гипотезы H_1 превышает вероятность отклонения гипотезы H_0 при значении $\upsilon \ge 0{,}33$. Это значение можно считать наиболее эффективным ($\max(1 - \beta - \alpha)$, где α и β – ошибки первого и второго рода соответственно) значением критерия $\upsilon_\partial = 0{,}33$. *Это такое теоретически достигаемое значение критерия, для которого* [4, 9]: достигаемый уровень значимости

$$p_{H_0}=\int_{\upsilon_\partial}^{1} g_\upsilon(\upsilon\,|\,H_0)\,d\upsilon; \quad (3)$$

достигаемая мощность

$$p_{H_1}=\int_{\upsilon_\partial}^{1} g_\upsilon(\upsilon\,|\,H_1)\,d\upsilon. \quad (4)$$

Заметим, что υ_∂ совпадает с абсциссой пересечения функций плотностей на рис. 1, что свидетельствует о верности расчетов.

Результаты исследования теоретической достигаемой мощности вариационного критерия для выборок объемом $n = 2\ldots5$ приведены в табл. 1. Там же приведены результаты имитационного моделирования при числе испытаний 10^6.

Анализ табл. 1 показывает, что независимо от объема выборки n результаты различаются не более чем на 0,01–0,03. При этом мощность критерия с увеличением объема выборки практически не растет, что подтверждено имитационными экспериментами. Данный факт можно объяснить уменьшением энтропии двух наименьших порядковых статистик относительно энтропии всей анализируемой выборки при увеличении ее объема. Тем не менее, *теоретически* критерий уверенно ($p_{H_1} > 0{,}7$ *и* $p_{H_0} \gg \alpha = 0{,}1$) различает достаточно близкие гипотезы H_0 и H_1 даже для выборок минимального объема.

На рис. 2 приведены графики функций распределения статистики υ критерия при проверке гипотез H_0 и H_1.

На рисунке наглядно представлена разность между достигаемыми уровнями значимости и мощности при проверке гипотез H_0 и H_1. Для выборок объемом $n > 2$ это позволяет воспользоваться элементами комбинаторики, т.к. количество k подвыборок объемом $m = 2$ с возвращением из исходной выборки больше, чем n.

Пусть имеется выборка объемом n наблюдений случайной величины X из генеральной совокупности с

Табл. 1. Результаты исследования мощности вариационного критерия

Характеристики		n	2	3	4	5
Наиболее эффективное значение критерия		$\upsilon_{кр}$	0,33	0,36	0,33	0,36
Достигаемый уровень значимости, рассчитанный	аналитически	p_{H_0}	0,67	0,64	0,67	0,64
	имитацией	$\hat{p}_{H_0}$	0,65	0,64	0,64	0,65
Достигаемая мощность критерия, рассчитанная	аналитически	p_{H_1}	0,77	0,75	0,77	0,75
	имитацией	$\hat{p}_{H_1}$	0,74	0,76	0,76	0,77

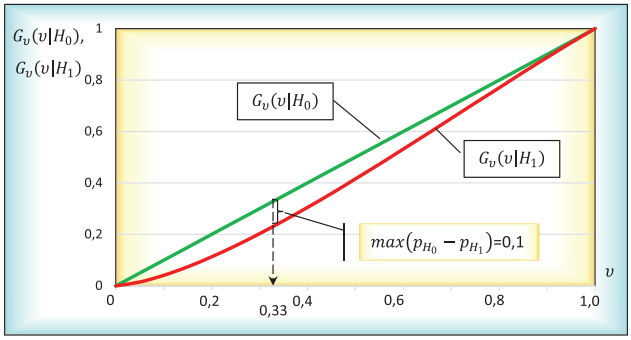

Рисунок 2 – Графики функций распределения статистики υ критерия при проверке гипотез H_0 и H_1

равномерным в интервале [0;1] законом распределения (свойство Ψ). Наибольшее число вариантов подвыборок из нее возможно получить путем сочетаний наблюдений по 2. Тогда из этой выборки последовательно извлекаются с возвращением выборки объемом $m = 2$, из элементов которых формируются порядковые статистики $x_{(1)} \le x_{(2)}$ (вариационный ряд [7]). Из них создаются вариационные критерии аналогично (2)

$$\upsilon_i = \frac{x_{(1)}}{x_{(2)}}, \quad i = 1,2,\ldots k, \tag{5}$$

распределения которых в интервале [0;1] также равномерны [3] (обладают свойством Ψ). Количество k таких критериев равно числу сочетаний без повторений [5]

$$k = C_n^2 = \frac{n!}{(n-2)!2!}. \tag{6}$$

Возможность использования совокупности критериев обусловлена низким уровнем их коррелированности для сочетаний без повторений по два наблюдения, взятых из выборки объемом n [8]. Так, для выборок объемом до $n = 10$ max $corr \approx 0{,}43$. Поэтому предлагается комбинаторный критерий, для которого достигнутый уровень значимости [4] свидетельствует против проверяемой гипотезы. Причем, аргументом у него является количество k подвыборок.

Критерии (5) сравниваются с критическим значением $\upsilon_{кр}$, на основании чего делается вывод о наступлении одного из двух событий:

– i-тая подвыборка также распределена равномерно в интервале [0;1] (обладает свойством Ψ), если $\upsilon_i \le \upsilon_д$;

– оснований считать i-тую подвыборку распределенной равномерно в интервале [0;1] недостаточно (не обладает свойством Ψ), если $\upsilon_i > \upsilon_д$.

Общее число подвыборок, обладающих свойством Ψ в выборке объема n, описывается биномиальным законом распределения [5].

По результатам тестирования подвыборок с помощью критериев υ_i решение принимается в соответствии с правилом: *проверяемая H_0 или альтернативная H_1 гипотезы отклоняются, если эти гипотезы отклоняются при проверке хотя бы одной подвыборки, составленной из исходной выборки путем сочетаний по два наблюдения без повторений.*

Для комбинаторного критерия оценками эффективности будут также достигаемые уровни значимости и мощности. Тогда, в соответствии с приведенным выше правилом, достигаемый уровень значимости при проверке k подвыборок

$$p^*_{H_0} = 1 - \left(1 - p_{H_0}\right)^k. \tag{7}$$

Аналогично достигаемая мощность критерия по k подвыборкам

$$p^*_{H_1} = 1 - \left(1 - p_{H_1}\right)^k. \tag{8}$$

В табл. 2 приведены расчетные значения достигаемых уровней значимости (7) для гипотезы H_0 и мощности (8) для гипотезы H_1 критерия для выборок различного объема n.

Табл. 2. Расчетные достигаемые уровни значимости и мощности

n	2	3	4	5
k	1	3	6	10
$p^*_{H_0}$	0,670	0,964	0,999	1,000
$p^*_{H_1}$	0,770	0,988	1,000	1,000

В результате имитационного моделирования процедуры проверки гипотез H_0 и H_1 с помощью предлагаемого критерия получены оценки достигнутых уровней значимости $\hat{p}^*_{H_0}$ и мощности $\hat{p}^*_{H_1}$, приведенные в табл. 3.

Табл. 3. Оценки достигнутых уровней значимости и мощности

n	2	3	4	5
k	1	3	6	10
$\hat{p}^*_{H_0}$	0,669	0,953	0,999	1,000
$\hat{p}^*_{H_1}$	0,767	0,981	1,000	1,000

Анализ табл. 2 и 3 свидетельствует о практическом соответствии оценок, полученных в результате имитации, и рассчитанных теоретически достигнутых уровней значимости и мощности критерия.

На рис. 3 приведены диаграммы достигаемых уровней значимости и мощности.

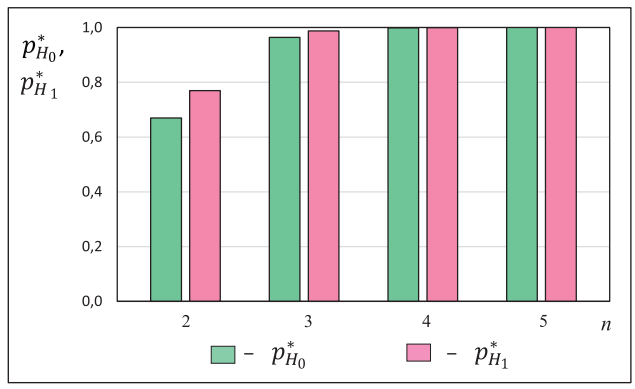

Рис. 3. Диаграммы достигаемых уровней значимости и мощности

Из рис. 3 наглядно видна их быстрая сходимость к 1 при увеличении объема исследуемой выборки до $n = 4$. Причем, при $n > 3$ достигаемые уровни значимости и мощности практически равны 1, что свидетельствует о высокой эффективности критерия для идентификации особо малых выборок.

Результат

Для демонстрации методики применения предлагаемого критерия рассмотрим задачу проверки выборки объемом $n = 3$ из генеральной совокупности с равномерным распределением в интервале [0;1] на равномерность (гипотеза H_0).

Пусть в результате наблюдений получена выборка

$$x = \{0,02; 0,079; 0,106\}. \quad (9)$$

Количество подвыборок $k = 3$ объемом $m = 2$. Матрица подвыборок примет вид

$$\begin{bmatrix} 0,02 & 0,079 \\ 0,02 & 0,106 \\ 0,079 & 0,106 \end{bmatrix}. \quad (10)$$

Вектор значений вариационных критериев (5)

$$\upsilon = [0,255 \quad 0,191 \quad 0,749]. \quad (11)$$

Для равномерного закона распределения $F(\upsilon_i) = \upsilon_i$, поэтому оценка достигнутого уровня значимости (7)

$$p^*_{H_0} = 1 - \prod_{i=1}^{3} (1 - \upsilon_i) = 0,85. \quad (12)$$

Достигнутый уровень значимости (12) намного больше используемого обычно уровня $\alpha = 0,1$. Поэтому отклонить гипотезу H_0 оснований недостаточно.

Пусть теперь дана выборка такого же объема $n = 3$

$$y = \{0,584; 0,039; 0,766\} \quad (13)$$

из генеральной совокупности с бета-распределением (гипотеза H_1) с параметрами $\lambda = 1,5$ и $\mu = 1,5$. Матрица подвыборок примет вид

$$\begin{bmatrix} 0,584 & 0,039 \\ 0,584 & 0,766 \\ 0,039 & 0,766 \end{bmatrix}. \quad (14)$$

Вектор значений вариационных критериев (5)

$$\upsilon = [0,067 \quad 0,762 \quad 0,051]. \quad (15)$$

Оценка достигнутого уровня значимости (7)

$$p^*_{H_1} = 1 - \prod_{i=1}^{3} (1 - \upsilon_i) = 0,79. \quad (16)$$

Заметим, что оценка достигнутого уровня значимости (16) при проверке гипотезы H_1 ниже аналогичной оценки (12) при проверке гипотезы H_0. Это свидетельствует о том, что критерий позволяет достаточно уверенно $p^*_{H_0} - p^*_{H_1} = 0,85 - 0,79 = 0,06$ различить статистически близкие гипотезы H_0 и H_1.

Комбинаторный критерий целесообразно использовать и при реализации последовательного анализа. В этом случае исходная выборка в общем виде запишется следующим образом

$$x = \{x_1; x_2; \ldots; x_n\}. \quad (17)$$

Матрица подвыборок будет формироваться по следующей рекуррентной схеме

$$[x_1 \ x_2] + x_3 \to \begin{bmatrix} x_1 & x_2 \\ x_1 & x_3 \\ x_2 & x_3 \end{bmatrix} + x_4 \to \begin{bmatrix} x_1 & x_2 \\ x_1 & x_3 \\ x_1 & x_4 \\ x_2 & x_3 \\ x_2 & x_4 \\ x_3 & x_4 \end{bmatrix} + \ldots + \to \begin{bmatrix} x_1 & x_2 \\ \ldots & \ldots \\ x_1 & x_n \\ x_2 & x_3 \\ \ldots & \ldots \\ x_2 & x_n \\ x_3 & x_4 \\ \ldots & \ldots \\ x_{n-1} & x_n \end{bmatrix}. \quad (18)$$

То есть, при каждом новом появлении случайной величины x_n в матрице подвыборок добавляется ее $n - 1$ сочетание с предыдущими наблюдениями.

Наиболее быстродействующую и эффективную процедуру последовательного анализа можно реализовать путем обнаружения «разладки процесса» при снижении достоверности анализа ниже заданного уровня. Суть ее заключается в том, что полученное на каждом шаге процедуры значение статистики (2) является случайной величиной, равномерно распределенной в интервале [0;1], и содержит информацию обо всех предыдущих наблюдениях. Поэтому дальнейший анализ с ней и поступившей новой случайной величиной аналогичен предыдущему шагу.

Таким образом, комбинаторный критерий позволяет реализовать преимущества вариационного критерия равномерности при анализе выборок объемом $n > 2$. При этом мощность критерия значительно возрастает, а простота использования не выдвигает каких-либо дополнительных требований. Кроме того, в силу наивысшей энтропии равномерного закона распределения [10] представляется целесообразным формирование и других положений и критериев математической статистики на его основе. Это позволит получать некоторые решения в аналитическом виде с более высокой эффективностью по сравнению с традиционными.

Заключение

Использование подвыборок объемом $m = 2$ из исходной выборки малого объема путем сочетания наблюдений без повторений при $n > 3$ увеличивает количество k частных критериев. Построенный на основе их совокупности комбинаторный критерий обладает более высокими достигнутыми уровнями значимости и мощности при различении даже весьма близких гипотез. Такой критерий может служить основой эффективной процедуры последовательного анализа малой выборки.

Библиографический список

1. Джонсон Н., Лион Ф. Статистика и планирование эксперимента в технике и науке. Методы обработки данных / Под ред. Э.К. Лецкого. М.: Мир, 1980. 610 с.

2. Виноградова М.С., Кандаурова И.Е., Ткачева О.С. Комбинаторный метод вычисления вероятностей // Modern European Researches. 2021. No 3 (T.1). Pp. 67-79.

3. Воловик А.В. Вариационный критерий равномерности // Надежность. 2023. № 1. С. 52-55. DOI: 10.21683/1729-2646-2023-23-1-52-55.

4. Лемешко Б.Ю., Блинов П.Ю. Критерии проверки отклонения распределения от равномерного закона. Руководство по применению. Новосибирск: НГТУ, 2015. 182 с.

5. Корн Г., Корн Т. Справочник по математике. Для научных работников и инженеров. М.: Наука, 1974. 832 с.

6. Способ сигнализации наличия стружки в масле и устройство для его реализации: пат. 2791174 Рос. Федерация. № 2022116675 / Воловик А.В.; заявл. 20.06.2022; опубл. 03.03.2023, Бюл. № 7. 9 с.

7. Джонсон Н. Одномерные непрерывные распределения. Ч. 2. М.: БИНОМ. Лаборатория знаний, 2010–2012. 600 с.

8. Нейросетевой анализ нормальности малых выборок биометрических данных с использованием хи-квадрат критерия и критериев Андерсона – Дарлинга / В.И. Волчихин [и др.] // Инженерные технологии и системы. 2019. Т. 29. № 2. С. 205-217. DOI: 10.15507/2658-4123.029.201902.205-217

9. ГОСТ Р 50779.10-2000 Статистические методы. Вероятность и основы статистики. Термины и определения. М.: Стандартинформ, 2005. IV, 41 с.

10. Ивченко Б.П., Мартыщенко Л.А., Табухов М.Е. Управление в экономических и социальных системах. Системный анализ. Принятие решений в условиях неопределенности. СПб.: «Нордмед-Издат», 2001. 248 с.

References

1. Johnson N., Leone F. Statistics and Experimental Designs and Engineering and the Physical Sciences. Mathods of Data Processing. Moscow: Mir; 1980.

2. Vinogradova M.S., Kandaurova I.E., Tkachiova O.S. [A combinatorial method of calculating probabilities]. *Modern European Researches* 2021;3(1):67-79. (in Russ.)

3. Volovik A.V. Variational criterion of evenness. *Dependability* 2023;23(1):52-55. DOI: 10.21683/1729-2646-2023-23-1-52-55. (in Russ.)

4. Lemeshko B.Yu., Blinov P.Yu [Criteria for testing a distribution for deviation from a uniform law. An application guide]. Novosibirsk: NSTU; 2015. (in Russ.)

5. Korn G., Korn T. Mathematical handbook. Moscow: Nauka; 1974.

6. Volovik A.V. A method of signalling the presence of shavings in oil and a device to implement it: patent 2791174 Russian Federation. No. 2022116675; submitted 20.06.2022; published 03.03.2023, Bul. no. 7.

7. Johnson N. Continuous univariate distributions. Vol. 2. Moscow: BINOM. Laboratoria znaniy; 2010-2012.

8. Volchikhin V.I. The neural network analysis of normality of small samples of biometric data through using the Chi-square test and Anderson–Darling criteria. *Inzhenernyye tekhnologii i sistemy* 2019;29(2):205-217. DOI: https://doi.org/10.15507/2658-4123.029.201902.205-217. (in Russ.)

9. GOST R 50779.10-2000 Statistical methods. Probability and general statistical terms. Terms and definitions. Moscow: Standartinform; 2005. (in Russ.)

10. Ivchenko B.P., Martyshchenko L.A., Tabukhov M.E. [Control in economic and social systems. Systems analysis-making under uncertainty]. Saint Petersburg: Nordmed-Izdat; 2001. (in Russ.)

Сведения об авторе

Воловик Александр Васильевич – кандидат технических наук, ведущий инженер-конструктор АО «ОДК-Климов», Санкт-Петербург, Российская Федерация, e-mail: volovik_aleksandr@mail.ru

About the author

Alexander V. Volovik, Candidate of Engineering, Lead Design Engineer, JSC Klimov, Saint Petersburg, Russian Federation, e-mail: volovik_aleksandr@mail.ru.

Вклад автора в статью

Воловик А.В. Предложен критерий согласия на основе совокупности комбинаций вариационного ряда, составленного из наблюдений в выборке малого объема. Проведены эксперименты по статистической оценке эффективности критерия по отношению к близкой альтернативе при разных объемах выборки, начиная с минимальной. Сделан вывод о целесообразности использования критерия для последовательного анализа и проверки гипотез о других законах распределения с помощью вероятностного интегрального преобразования.

Конфликт интересов

Автор заявляет об отсутствии конфликта интересов.