НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
июль–август 2024 Том 24 № 4 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
July– August 2024 Vol. 24 No 4 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

doi: 10.17586/2226-1494-2024-24-4-654-660
УДК 004.94

Многоуровневое расщепление в методе Монте-Карло для оценки вероятностей редких событий в пермутационных тестах

Владимир Дмитриевич Сухов[1]✉, Геннадий Владимирович Короткевич[2], Алексей Александрович Сергушичев[3]

[1,3] Университет Вашингтона в Сент-Луисе, Сент-Луис, 63110, США
[2,3] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] vdsukhov@ya.ru✉, https://orcid.org/0000-0002-5169-1433
[2] gkorotkevitch@yandex.ru, https://orcid.org/0009-0004-5941-2816
[3] alsergbox@gmail.com, https://orcid.org/0000-0003-1159-7220

Аннотация

Введение. Пермутационные тесты широко применяются при проведении статистического анализа, например, когда нарушаются предположения параметрических тестов или распределение данных неизвестно. Заметим, что в случае применения классических пермутационных тестов могут возникнуть проблемы при попытке оценки вероятностей редких событий с высокой относительной точностью. Это приводит к трудностям при использовании поправки на множественную проверку статистических гипотез. В работе предлагается оригинальный метод оценки произвольно малых *P*-значений в пермутационных тестах, который основан на многоуровневом расщеплении в методе Монте-Карло. **Метод.** Представленный метод включает дробление исходного пространства перестановок на непересекающиеся уровни по значениям статистики. Метод дает возможность свести задачу оценки исходной вероятности редкого события к задаче оценки обычных условных вероятностей для каждого уровня. Использование метода позволяет эффективным образом оценивать искомые *P*-значения, сохраняя баланс между временем работы и уровнем относительной ошибки. **Основные результаты.** Работа метода продемонстрирована в применении к задаче оценки произвольных *P*-значений двухвыборочного теста Колмогорова–Смирнова. Сравнение результатов работы метода с истинными *P*-значениями подтвердило практическую сходимость метода. Показаны примеры превосходства предлагаемого метода над альтернативными асимптотическими подходами. **Обсуждение.** Предлагаемый метод выявил существенный потенциал применения в широком спектре научных областей, таких как системная биология, иммунология и других. Метод может быть адаптирован для использования в различных случаях статистического анализа, который требует работы с вероятностями редких событий в пермутационных тестах.

Ключевые слова

проверка статистических гипотез, *P*-значение, методы Монте-Карло, пермутационные тесты, редкие события

Ссылка для цитирования: Сухов В.Д., Короткевич Г.В., Сергушичев А.А. Многоуровневое расщепление в методе Монте-Карло для оценки вероятностей редких событий в пермутационных тестах // Научно-технический вестник информационных технологий, механики и оптики. 2024. Т. 24, № 4. С. 654–660. doi: 10.17586/2226-1494-2024-24-4-654-660

Multilevel splitting for rare events estimation in permutation tests

Vladimir D. Sukhov[1]✉, Gennady V. Korotkevich[2], Alexey A. Sergushichev[3]

[1,3] University in St. Louis, Saint Louis, 63110, USA
[2,3] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] vdsukhov@ya.ru✉, https://orcid.org/0000-0002-5169-1433
[2] gkorotkevitch@yandex.ru, https://orcid.org/0009-0004-5941-2816
[3] alsergbox@gmail.com, https://orcid.org/0000-0003-1159-7220

Abstract
Permutation tests are widely employed in statistical analysis, especially when the assumptions of parametric tests are violated, or the data distribution is unknown. However, classical permutation tests encounter challenges when attempting to estimate the probabilities of rare events with high relative accuracy, leading to difficulties in applying corrections for multiple hypothesis testing. In this study, we propose an original method for estimating arbitrarily small *P*-values in permutation tests, which is based on multilevel splitting for Monte Carlo method. The proposed method involves splitting the original permutation space into non-overlapping levels based on the statistic values. This approach allows the problem of estimating the original probability of a rare event to be reduced to estimating ordinary conditional probabilities for each level. Utilizing such an approach enables efficient estimation of the desired *P*-values while maintaining a balance between computation time and the level of relative error. The efficacy of the method is demonstrated in its application to the task of estimating arbitrary *P*-values in the two-sample Kolmogorov-Smirnov test. Comparing the method results with true *P*-values has shown practical convergence of the method. Moreover, examples of the superiority of the proposed method over alternative asymptotic approaches have been provided. Thus, the proposed method shows significant potential for application across a broad spectrum of scientific fields, such as systems biology, immunology, and others. Furthermore, the method can be adapted for use in various statistical analysis scenarios that require handling probabilities of rare events in permutation tests.

Keywords
statistical hypothesis test, *P*-value, Monte Carlo method, permutation test, rare events

For citation: Sukhov V.D., Korotkevich G.V., Sergushichev A.A. Multilevel splitting for rare events estimation in permutation tests. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2024, vol. 24, no. 4, pp. 654–660 (in Russian). doi: 10.17586/2226-1494-2024-24-4-654-660

Введение

Одной из фундаментальных задач при проведении научных исследований является проверка статистических гипотез, которая позволяет делать выводы о свойствах вероятностных распределений на основе их ограниченных выборок. Важный аспект при проверке статистических гипотез — выбор между параметрическими и непараметрическими тестами. Параметрические тесты предполагают, что данные соответствуют известному параметризованному семейству распределений, например, нормальному распределению, с некоторыми неизвестными параметрами. В свою очередь, непараметрические тесты не требуют каких-либо предположений о распределении для наблюдаемых значений в эксперименте. Основные преимущества непараметрических тестов: отсутствие предположений о вероятностном распределении для данных; применимость для выборок небольших размеров; возможность обработки данных с выбросами.

В настоящей работе рассмотрен специальный класс непараметрических тестов — пермутационные тесты (от англ. permutation — перестановка) [1, 2]. Эти тесты основаны на исследовании случайных перестановок исходных данных для построения распределения статистики критерия, которое используется для последующего принятия или отклонения нулевой гипотезы. В таком случае *P*-значение для теста можно определить как отношение числа перестановок, для которых значение статистики больше или равно значению статистики для исходных наблюдений, к числу всех возможных перестановок. Нулевая гипотеза отклоняется, если *P*-значение меньше некоторого заранее заданного порога значимости, например 0,05. Для проверки этого критерия на практике часто применяются методы Монте-Карло [3, 4]. Эти методы не требуют рассмотрения всех возможных перестановок, а используют их случайное подмножество небольшого фиксированного размера. Методы Монте-Карло зарекомендовали себя на практике за счет простоты их реализации.

Заметим, что в ряде задач требуется оценить экстремально малые *P*-значения с высокой относительной точностью. Примерами областей, где встречаются такие задачи, являются молекулярная динамика, физика элементарных частиц и др. [5, 6]. В данных задачах истинное *P*-значение может быть меньше 10^{-10}, в этом случае применение классических методов Монте-Карло требует огромных вычислительных ресурсов. Возможное решение задачи оценки *P*-значения — использование методов оценки вероятностей редких событий [7]. Одним из таких методов является многоуровневое расщепление в методе Монте-Карло [8, 9]. В отличие от классического метода Монте-Карло, который требует генерирования большого числа перестановок, многоуровневое расщепление позволяет получать необходимую относительную ошибку при использовании значительно меньшего числа перестановок.

Цель работы — решение задачи оценки экстремально малых *P*-значений в пермутационных тестах, когда применение классических методов является непрактичным. Для выполнения поставленной задачи предложен метод, основанный на использовании многоуровневого расщепления в методе Монте-Карло. Осуществлена апробация метода в применении для задачи оценки *P*-значений для двухвыборочного теста Колмогорова–Смирнова.

Постановка задачи вычисления малых *P*-значений в пермутационных тестах

Рассмотрим две выборки $\mathbf{X} = (X_1, X_2, \ldots, X_{N_1})$ и $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_{N_2})$ со значениями из пространства Ω и с функциями распределений F и G соответственно. Пусть задана статистика на множестве вещественных чисел $\mathbb{R}$:

$$D{:}\Omega^{N_1+N_2} \to \mathbb{R},$$

которая не зависит от порядка следования элементов внутри выборок $\mathbf{X}$ и $\mathbf{Y}$.

Тогда сформулируем задачу вычисления P-значения вида $P(D \geq \gamma|H_0)$, где γ — наблюдаемое значение статистики для исходных выборок $\mathbf{X}$ и $\mathbf{Y}$; H_0 — нулевая гипотеза о совпадении распределений $F = G$. На практике часто для этой вероятности неизвестно аналитическое выражение. В связи с этим вместо задачи о вычислении искомого P-значения рассматривают задачу его оценки.

Классический подход решения данной задачи — применение методов Монте-Карло. Суть этих методов заключается в генерировании M случайных перестановок исходных элементов вида:

$$\pi_j: \Omega^{N_1+N_2} \to \Omega^{N_1+N_2}, j \in \{1, 2, \ldots, M\}.$$

Рассмотрим в качестве оценки искомой вероятности величину:

$$\hat{p}_{MC} = \frac{1}{M}\sum_{j=1}^{M} I\{D_j \geq \gamma\},$$

где I — индикаторная функция; D_j — значение статистики, полученное для j-ой перестановки π_j.

Отметим, что недостатком классического подхода является высокая относительная ошибка [10] для P-значений много меньших единицы. Это следует из выражения относительной ошибки для приведенной оценки:

$$\frac{\sqrt{\mathbf{D}(\hat{p}_{MC})}}{\mathbf{E}\hat{p}_{MC}} = \sqrt{\frac{1 - P(D \geq \gamma|H_0)}{MP(D \geq \gamma|H_0)}},$$

где $\mathbf{E}$ и $\mathbf{D}$ — операторы математического ожидания и дисперсии. Следовательно, для получения относительной ошибки, например, в 100 % при оценивании вероятности 10^{-10} необходимо построить выборку размером $M \approx 10^{10}$. Генерирование такого числа случайных перестановок является вычислительно сложной задачей. Для решения задачи оценки малых P-значений в пермутационных тестах необходимы новые методы, которые позволят получать результаты с хорошей относительной точностью даже для таких малых значений.

Многоуровневое расщепление в методе Монте-Карло для оценки малых *P*-значений в пермутационных тестах

Для решения поставленной задачи в настоящей работе исследовано применение многоуровневого расщепления в методе Монте-Карло для оценки малых P-значений в пермутационных тестах. Метод основан на рассмотрении дробления диапазона возможных значений статистики D по некоторым границам l_j (уровням):

$$-\infty = l_0 < l_1 < \cdots < l_t = \gamma.$$

Тогда искомая вероятность может быть переписана в виде:

$$P(D \geq \gamma) = \prod_{j=1}^{t} P(D \geq l_j|D \geq l_{j-1}) = \prod_{j=1}^{t} p_j,$$

где $p_j := P(D \geq l_j|D \geq l_{j-1})$. Таким образом, задачу оценки P-значения можно свести к определению уровней l_j и получению оценок $\hat{p}_j$ для сомножителей p_j.

Предположим, есть возможность получать выборку из условного распределения вида $P(D \in \cdot|D \geq l_{j-1})$. Тогда сформулируем следующий алгоритм для одновременной оценки сомножителей p_j и построения уровней l_j, $j \in \{1, 2, \ldots, t\}$.

Шаг 1. Для j-го уровня генерируется выборка статистик $D_1^j, D_2^j, \ldots D_M^j$ нечетного размера M из условного распределения $P(D \in \cdot|D \geq l_{j-1})$.

Шаг 2. В качестве нового потенциального уровня рассматривается медиана выборки $\tilde{l}_j := \mathrm{med}(D_1^j, D_2^j, \ldots, D_M^j)$.

Шаг 3. Если $\tilde{l}_j > \gamma$, то шаги останавливаются и в качестве l_j принимается значение γ. Иначе, если $\tilde{l}_j = l_{j-1}$, то дальнейшая работа метода завершается с ошибкой. Если оба эти условия не выполняются, то $l_j := \tilde{l}_j$ и происходит переход к шагу 1.

По построению уровней имеем, что оценка $\hat{p}_j \approx 1/2$ для $1 \leq j \leq t - 1$. Получим значение оценки $\hat{p}_t$ в виде:

$$\hat{p}_t = \frac{1}{M}\sum_{i=1}^{M} I\{D_i^t \geq \gamma\}.$$

В результате искомую вероятность оценим следующим выражением:

$$P(D \geq \gamma) \approx \frac{1}{2^{t-1}}\sum_{i=1}^{M}\frac{I(D_i^t \geq \gamma)}{M}.$$

Рассмотрим задачу генерирования выборки из условного распределения $P(D \in \cdot|D \geq l_{j-1})$, где $j \geq 1$. Данная задача соответствует задаче генерирования перестановок π равномерно случайно из множества π: $D(\pi) \geq l_{j-1}$. Отметим, что значение статистики D не зависит от порядка элементов, и, таким образом, все перестановки можно разбить в равномощные классы эквивалентности, соответствующие N_1-сочетаниям из множества $\{X_1, \ldots, X_{N_1}, Y_1, \ldots, Y_{N_2}\}$. Далее, рассматриваемую задачу представим, как генерацию равномерно случайных N_1-сочетаний, для которых значение статистики не меньше l_{j-1}. Для решения этой задачи используем алгоритм Метрополиса–Гастингса [11, 12], который представляет собой один из вариантов методов Монте-Карло по схеме марковской цепи. В рамках данного алгоритма состояниями цепи являются N_1-сочетания из $\{X_1, \ldots, X_{N_1}, Y_1, \ldots, Y_{N_2}\}$, а переход между двумя состояниями возможен, если пересечение пары соответствующих N_1-сочетаний имеет размер $N_1 - 1$.

Заметим, что для получения выборки $D_1^1, D_2^1, \ldots D_M^1$ из распределения $P(D \in \cdot|D \geq l_0)$ на шаге 1 алгоритма достаточно сгенерировать случайные сочетания без накладывания каких-либо условий. Действительно, для любого N_1-сочетания и соответствующему значению статистики заведомо выполнено условие $D_i^1 \geq l_0$, $i \in \{1, 2, \ldots, M\}$ так как $l_0 = -\infty$.

Рассмотрим процесс получения выборки $P(D \in \cdot|D \geq l_{j-1})$ для $j \geq 2$. Пусть имеется выборка $D_1^{j-1}, D_2^{j-1}, \ldots D_M^{j-1} \sim P(D \in \cdot|D \geq l_{j-2})$. В качестве начальных кандидатов для сочетаний на уровне j применим

сочетания, соответствующие следующим значениям статистики:

$$\tilde{D}_i^j = \begin{cases} D_{(M+1-i)}^{j-1}, & i < d, \\ D_{(i)}^{j-1}, & i \geq d, \end{cases}$$

где $d = \lceil M/2 \rceil$ и $D_{(i)}^j$ — i-й элемент из вариационного ряда. По построению имеем, что $\tilde{D}_i^j \geq l_{j-1}$ для всех j. Однако при таком подходе изначально сочетания не являются независимыми. Для решения этой проблемы для каждого N_1-сочетания $\{\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_{N_1}\}$ и его дополнения $\{\tilde{Y}_1, \tilde{Y}_2, \ldots, \tilde{Y}_{N_2}\}$ на шаге $j-1$ выполним некоторое число итераций алгоритма Метрополиса–Гастингса вида:

шаг 1. выбирается случайный индекс $k \in \{1, 2, \ldots, N_1\}$;

шаг 2. выбирается случайный индекс $m \in \{1, 2, \ldots, N_2\}$;

шаг 3. рассматривается сочетание вида: $\{\tilde{X}_1, \ldots, \tilde{X}_{k-1}, \tilde{Y}_m, \tilde{X}_{k+1}, \ldots, \tilde{X}_{N_1}\}$.

В случае, если значение статистики сочетания на шаге 3 алгоритма Метрополиса–Гастингса является большим либо равным значения l_{j-1}, то данная замена элементов сохраняется. В противном случае замена отклоняется.

Основное свойство алгоритма Метрополиса–Гастингса — сходимость распределения цепи к стационарному при росте числа итераций к бесконечности. Однако для применения алгоритма на практике требуется правило останова, позволяющее достигнуть достаточной сходимости при ограниченном времени работы.

Предложим правило останова, зависящее от параметра α. Итерации алгоритма, выполняются параллельно для каждого сочетания, и повторяются до тех пор, пока число успешных замен T не станет большим или равным $T \geq \alpha \cdot N_1 \cdot M$ (т. е. доля успешных замен для каждого сочетания не меньше, чем $\alpha \cdot N_1$). Параметр α позволяет контролировать баланс между временем работы предлагаемого метода и степенью независимости и равномерности полученной выборки (большее значение соответствует лучшему качеству выборки).

Многоуровневое расщепление в методе Монте-Карло для двухвыборочного теста Колмогорова–Смирнова

Приведем пример использования предложенного метода и рассмотрим его применение для распределения статистики непараметрического двухвыборочного теста Колмогорова–Смирнова [13, 14]. Данный тест определяется в следующем виде.

Пусть имеются две выборки $X_1, X_2, \ldots, X_{N_1}$ и $Y_1, Y_2, \ldots, Y_{N_2}$, для которых вычислена статистика Колмогорова–Смирнова:

$$\gamma = \sup_x |F_{1,N_1}(x) - F_{2,N_2}(x)|,$$

где F_{1,N_1} и F_{2,N_2} — соответствующие эмпирические функции распределения для исследуемых выборок. Требуется вычислить вероятность следующего вида:

$$P(D \geq \gamma),$$

где D — случайная величина значений статистик Колмогорова–Смирнова, построенных на случайных перестановках исходных наблюдений.

Для оценки искомой вероятности, согласно предложенному методу, рассмотрим дробление исходного пространства значений статистики D и перепишем P-значение в виде:

$$P(D \geq \gamma) = \prod_{j=1}^{t} P(D \geq l_j | D \geq l_{j-1}) = \prod_{j=1}^{t} p_j.$$

На практике удобно перейти от оценки исходной вероятности к оценке логарифма вероятности. Тогда эту оценку можно представить в виде суммы логарифмов оценок для каждого уровня:

$$\log P(D \geq \gamma) \approx \sum_{j=1}^{t} \log \hat{p}_j.$$

В этом случае общая оценка является случайной величиной, которую — как сумму t независимых случайных величин $\log \hat{p}_j$ — можно приблизить нормальным распределением. Параметры данного распределения — математическое ожидание и дисперсия — могут быть оценены в виде суммы соответствующих параметров распределений на каждом уровне.

Для оценки логарифма сомножителя p_j воспользуемся свойствами непрерывных распределений. Во-первых, для непрерывной случайной величины η, с заданной функцией распределения $F_\eta(x) = P(\eta < x)$, случайная величина $F_\eta(x)$ имеет стандартное равномерное распределение. Во-вторых, m-я порядковая статистика выборки размера M из стандартного равномерного распределения является случайной величиной из бета-распределения $B(m, M + 1 - m)$. Наконец, математическое ожидание логарифма случайной величины $\xi \in B(m, M + 1 - m)$ выразим следующим образом:

$$\mathbf{E}(\log \xi) = \psi(m) - \psi(M + 1),$$

где ψ — дигамма-функция.

Рассмотрим выборку $D_1^j, D_2^j, \ldots D_M^j$ из распределения с функцией распределения $F^j(D') := P(D < D' | D \geq l_{j-1})$, полученную на уровне j.

Тогда

$$p_j = P(D \geq D_{M_j} | D \geq l_{j-1}) = 1 - P(D < D_{M_j} | D \geq l_{j-1}) = 1 - F^j(D_{M_j}),$$

где $M_j = \sum_{i=1}^{M} I(D_i^j < l_j)$.

Предположим, что расхождение между функцией распределения $F_{U_i^j}$ случайной величины $U_i^j = F^j(D_i^j)$ и функцией распределения G стандартного равномерного распределения ограничено некоторым малым значением ε:

$$\sup_x |F_{U_i^j}(x) - G(x)| \leq \varepsilon << 1. \quad (1)$$

Тогда запишем выражение для оценки p_j в виде:

$$p_j = 1 - F^j(D_{M_j}) \approx 1 - U_{(M_j)}^j = U_{(M - M_j)}^j,$$

где $U_{(M_j)}^j$ — M_j-я порядковая статистика из выборки U_i^j, $1 \leq i \leq M$.

Соответственно, в качестве оценки логарифма p_j используем выражение:

$$\log p_j \approx \psi(M - M_j) - \psi(M + 1).$$

Аналогично определим оценку для последнего уровня $l_j = t$. В результате получим полное выражение для оценки:

$$\log P(D \geq \gamma) \approx \sum_{j=1}^{t} (\psi(M - M_j) - \psi(M + 1)).$$

Соответственно получим оценку дисперсии случайной величины, используя формулу для дисперсии логарифма случайной величины бета-распределения: $\mathbf{D}(\log\xi) = \psi_1(m) - \psi_1(M + 1)$, где ψ_1 — тригамма-функция.

Заметим, что по построению для каждого уровня $m \geq (M + 1)/2$, возможно оценить дисперсию на каждом уровне сверху выражением $\psi_1\left(\frac{M+1}{2}\right) - \psi_1(M + 1)$ в силу монотонности тригамма-функции на положительной полуоси вещественных чисел $\mathbb{R}_+$. Выполним суммирование дисперсии для каждого уровня, получим общую оценку дисперсии и, следовательно, оценку для стандартного отклонения:

$$sd = \sqrt{t \cdot \left(\psi_1\left(\frac{M+1}{2}\right) - \psi_1(M+1)\right)}.$$

Таким образом, предложенный метод позволяет оценить логарифм произвольно малого P-значения для случая двухвыборочного теста Колмогорова–Смирнова. Также введенное выражение для стандартного отклонения может быть использовано для приближения 95 % доверительного интервала в предположении нормальности оценки логарифма P-значения:

$$(\log\hat{p} - 2\text{sd}, \log\hat{p} + 2\text{sd}),$$

где $\log\hat{p}$ — оценка искомого логарифма.

Анализ практической сходимости оценок P-значений для двухвыборочного теста Колмогорова–Смирнова

Рассмотрим вопрос практической сходимости результатов работы предлагаемого метода при применении его для двухвыборочного теста Колмогорова–Смирнова. Для этого при заданных значениях N_1, N_2 и γ исследуем 100 независимых запусков метода для оценки логарифма искомой вероятности. В качестве критерия оценки работы метода зафиксируем число раз, когда истинное значение $\log p$ принадлежит используемому приближению 95 % доверительного интервала $(\log\hat{p} - 2\text{sd}, \log\hat{p} + 2\text{sd})$.

Рассмотрим случай $N_1 + N_2$, который равен 1001, а значения N_1 выберем из множества {50, 100, 250, 500}. Для каждого значения N_1 изучим значения статистики γ, которым соответствуют P-значения следующих порядков: 10^{-10}, 10^{-20}, 10^{-30}, 10^{-40}, 10^{-50}. Результаты работы метода для данных параметров приведены на рис. 1, *a*. По оси абсцисс отложены значения параметра α, а по оси ординат — доля доверительных интервалов, содержащих истинное значение. Для определения истинных P-значений используем библиотеку SciPy [15] языка программирования Python. В результате выполненного исследования можно утверждать о практической сходимости полученных оценок при значениях параметра $\alpha \geq 1$. При этом в более чем 95 % случаев оказалось, что истинное значение логарифма вероятности принадлежит приближению 95 % доверительного интервала для полученных оценок.

В отличие от случая $N_1 + N_2 = 1001$, при $N_1 + N_2 = 1000$ нарушается предположение о малости расхождения между функциями распределений в выражении (1) для двухвыборочного теста Колмогорова–Смирнова. На практике это выражается в том, что на шаге 3 предложенного алгоритма для одновременной оценки сомножителей p_j и построения уровней l_j возникает ситуация, когда $\tilde{l}_j = l_{j-1}$, что приводит к завершению метода с ошибкой. Такие случаи выходят за пределы применимости метода. На рис. 1, *b* представлены результаты работы

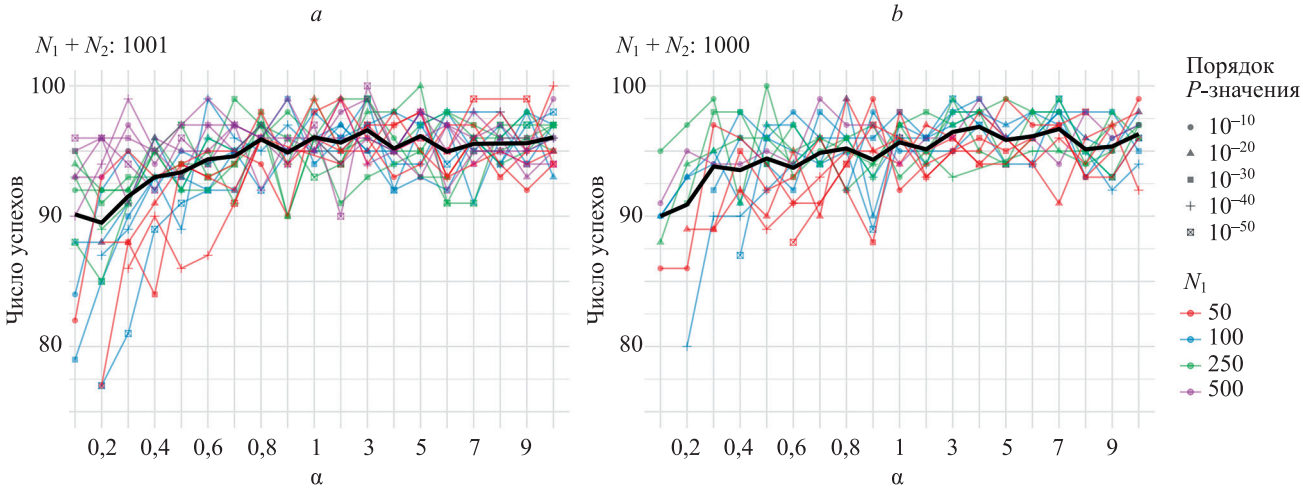

Рис. 1. Результаты работы метода в зависимости от параметра α при $N_1 + N_2 = 1001$ (*a*) и $N_1 + N_2 = 1000$ (*b*). Усреднение результатов для различных комбинаций параметров N_1, N_2 и γ (черная кривая)

Fig. 1. The results of the method performance depending on the parameter α for $N_1 + N_2 = 1001$ (*a*) and $N_1 + N_2 = 1000$ (*b*)

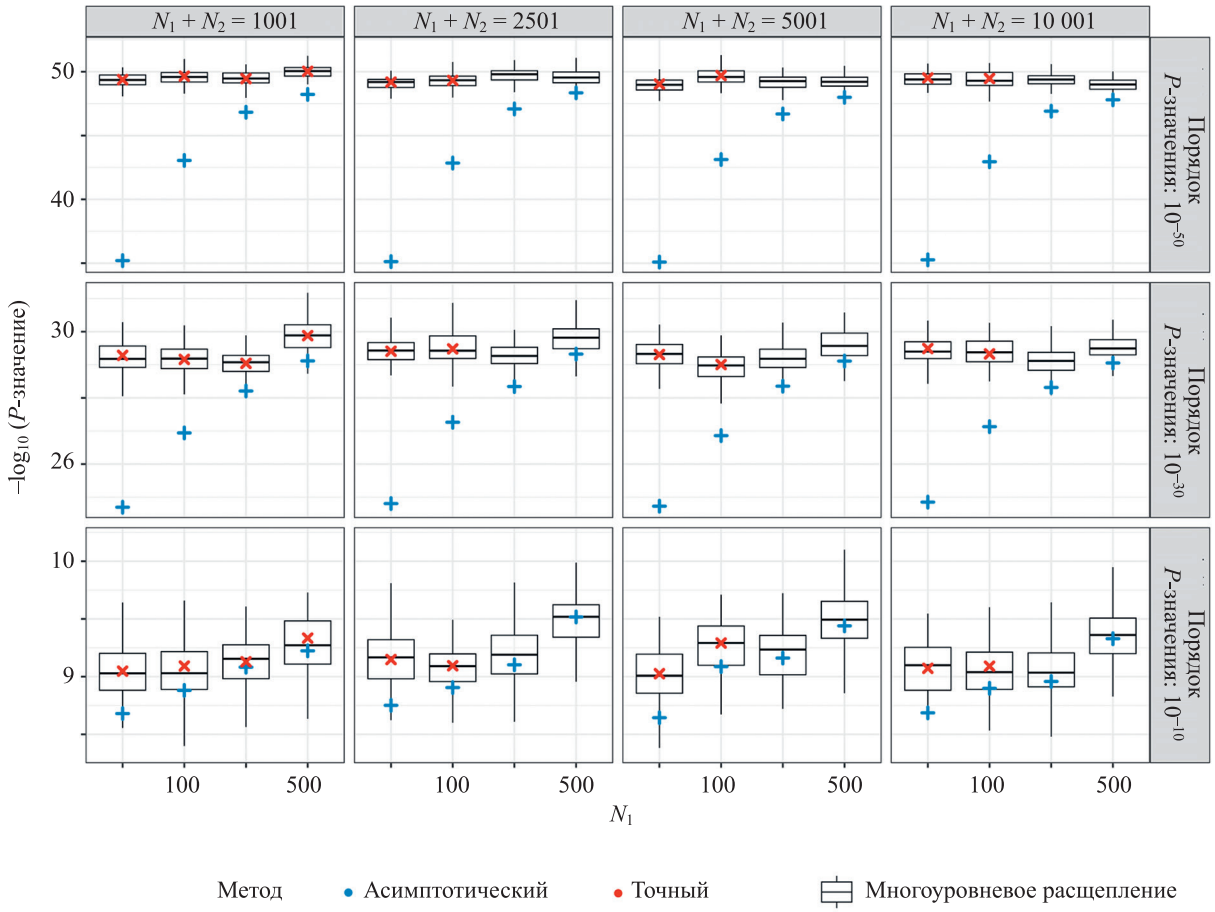

Рис. 2. Сравнение результатов работы предложенного метода с асимптотическим и точным методами из библиотеки SciPy

Fig. 2. Comparison of the method results with asymptotic and exact methods from the SciPy package

метода в случаях, когда метод завершился без ошибок. Заметим, что, как и на рис. 1, *a*, истинное значение логарифма вероятности принадлежит приближению 95 % доверительного интервала для полученных оценок.

Отметим, что, когда рассматриваемый метод успешно завершил свое выполнение для обоих случаев $N_1 + N_2 = 1001$ и $N_1 + N_2 = 1000$, получена практическая сходимость метода при значениях параметра $\alpha \geq 1$.

Проведенный анализ подтвердил, что установка значения параметра $\alpha = 1$ является необходимым условием для достижения практической сходимости. Рассмотрим результаты работы метода при постоянном значении $\alpha = 1$. В качестве суммы $N_1 + N_2$ используем значения из набора {1001, 2501, 5001, 10 001}, оставляя множество значений для N_1 неизменным. Для каждого значения N_1 возьмем значения статистики γ, соответствующие P-значениям следующих порядков 10^{-10}, 10^{-30}, 10^{-50}. Результаты работы метода при 100 независимых запусках для всех возможных комбинаций входных данных показаны на рис. 2. «Ящики с усами» отображают распределение полученных оценок для логарифма искомой вероятности. Также на графике представлены результаты работы точного и асимптотического методов из библиотеки SciPy.

Заметим, что использование точного метода ограничено и не применимо для всех возможных входных данных. В то время как результаты работы асимптотического метода демонстрируют систематическую ошибку, проявляющуюся в случаях, когда удается получить истинные значения для логарифмов P-значений. При этом в отличие от асимптотического приближения результаты работы предложенного метода характеризуются хорошим соответствием между средней оценкой и истинным значением.

Результаты выполненных исследований подтвердили высокую степень соответствия между полученными в работе оценками и истинными значениями. Также обнаружено отсутствие избыточной систематической ошибки, которая характерна для результатов при применении асимптотического метода. Эти наблюдения говорят о надежности и эффективности представленного метода в сравнении с альтернативными асимптотическими подходами.

Заключение

Многоуровневое расщепление в методе Монте-Карло показало свою эффективность в задаче оценки вероятностей редких событий в пермутационных тестах. Применение данного метода решает проблемы, связанные с оценкой экстремальных вероятностей хвостов распределений, которые имеют важное значение при проведении проверки гипотез во множестве научных областей.

Применив метод к пермутационным тестам, получены оценки искомых P-значений с высокой относительной точностью даже для небольших выборок и сложной структуры данных. Полученные результаты показали превосходство многоуровневого расщепления в методе Монте-Карло над традиционными методами Монте-Карло, а также асимптотическими подходами при работе с редкими событиями. Предложенный метод открывает новые возможности для получения статистических выводов в различных областях, таких как системная биология, иммунология и др.

Литература

1. Good P. Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses. Springer Science & Business Media, 2013.
2. Pesarin F., Salmaso L. Permutation Tests for Complex Data: Theory, Applications and Software. John Wiley & Sons, 2010. 448 p.
3. Hammersley J. Monte Carlo Methods. Springer Science & Business Media, 2013. 178 p.
4. Kalos M.H., Whitlock P.A. Monte Carlo Methods. John Wiley & Sons, 2009. 215 p.
5. Trendelkamp-Schroer B., Noé F. Efficient estimation of rare-event kinetics // Physical Review X. 2016. V. 6. N 1. P. 011009. https://doi.org/10.1103/physrevx.6.011009
6. Lestang T., Ragone F., Bréhier C.-E., Herbert C., Bouchet F. Computing return times or return periods with rare event algorithms // Journal of Statistical Mechanics: Theory and Experiment. 2018. V. 2018. N 4. P. 043213. https://doi.org/10.1088/1742-5468/aab856
7. Caron V., Guyader A., Zuniga M.M., Tuffin B. Some recent results in rare event estimation // ESAIM: Proceedings. 2014. V. 44. P. 239–259. https://doi.org/10.1051/proc/201444015
8. L'Ecuyer P., Demers V., Tuffin B. Splitting for rare-event simulation // Proc. of the 2006 Winter Simulation Conference. 2006. P. 137–148. https://doi.org/10.1109/wsc.2006.323046
9. Glasserman P., Heidelberger P., Shahabuddin P., Zajic T. Multilevel splitting for estimating rare event probabilities // Operations Research. 1999. V. 47. N 4. P. 585–600. https://doi.org/10.1287/opre.47.4.585
10. Botev Z.I., Kroese D.P. An efficient algorithm for rare-event probability estimation, combinatorial optimization, and counting // Methodology and Computing in Applied Probability. 2008. V. 10. N 4. P. 471–505. https://doi.org/10.1007/s11009-008-9073-7
11. Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller A.H., Teller E. Equation of state calculations by fast computing machines // The Journal of Chemical Physics. 1953. V. 21. N 6. P. 1087–1092. https://doi.org/10.1063/1.1699114
12. Hastings W.K. Monte Carlo sampling methods using Markov chains and their applications // Biometrika. 1970. V. 57. N 1. P. 97–109. https://doi.org/10.2307/2334940
13. Kolmogorov A. Sulla determinazione empirica di una legge di distribuzione // Giornale dell'Istituto Italiano degli Attuari. 1933. V. 4. P. 83–91.
14. Smirnoff N. Sur les Écarts de la courbe de distribution empirique // Matematicheskii Sbornik. 1939. V. 48. N 1. P. 3–26.
15. Virtanen P., Gommers R., Oliphant T.E., Haberland M., Reddy T., Cournapeau D., Burovski E., Peterson P., Weckesser W., Bright J. et. al. SciPy 1.0: fundamental algorithms for scientific computing in Python // Nature Methods. 2020. V. 17. N 3. P. 261–272. https://doi.org/10.1038/s41592-019-0686-2

References

1. Good P. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer Science & Business Media, 2013.
2. Pesarin F., Salmaso L. *Permutation Tests for Complex Data: Theory, Applications and Software*. John Wiley & Sons, 2010, 448 p.
3. Hammersley J. *Monte Carlo Methods*. Springer Science & Business Media, 2013, 178 p.
4. Kalos M.H., Whitlock P.A. *Monte Carlo Methods*. John Wiley & Sons, 2009, 215 p.
5. Trendelkamp-Schroer B., Noé F. Efficient estimation of rare-event kinetics. *Physical Review X*, 2016, vol. 6, no. 1, pp. 011009. https://doi.org/10.1103/physrevx.6.011009
6. Lestang T., Ragone F., Bréhier C.-E., Herbert C., Bouchet F. Computing return times or return periods with rare event algorithms. *Journal of Statistical Mechanics: Theory and Experiment*, 2018, vol. 2018, no. 4, pp. 043213. https://doi.org/10.1088/1742-5468/aab856
7. Caron V., Guyader A., Zuniga M.M., Tuffin B. Some recent results in rare event estimation. *ESAIM: Proceedings*, 2014, vol. 44, pp. 239–259. https://doi.org/10.1051/proc/201444015
8. L'Ecuyer P., Demers V., Tuffin B. Splitting for rare-event simulation. *Proc. of the 2006 Winter Simulation Conference*, 2006, pp. 137–148. https://doi.org/10.1109/wsc.2006.323046
9. Glasserman P., Heidelberger P., Shahabuddin P., Zajic T. Multilevel splitting for estimating rare event probabilities. *Operations Research*, 1999, vol. 47, no. 4, pp. 585–600. https://doi.org/10.1287/opre.47.4.585
10. Botev Z.I., Kroese D.P. An efficient algorithm for rare-event probability estimation, combinatorial optimization, and counting. *Methodology and Computing in Applied Probability*, 2008, vol. 10, no. 4, pp. 471–505. https://doi.org/10.1007/s11009-008-9073-7
11. Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller A.H., Teller E. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 1953, vol. 21, no. 6, pp. 1087–1092. https://doi.org/10.1063/1.1699114
12. Hastings W.K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 1970, vol. 57, no. 1, pp. 97–109. https://doi.org/10.2307/2334940
13. Kolmogorov A. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 1933, vol. 4, pp. 83–91.
14. Smirnoff N. Sur les Écarts de la courbe de distribution empirique. *Matematicheskii Sbornik*, 1939, vol. 48, no. 1, pp. 3–26.
15. Virtanen P., Gommers R., Oliphant T.E., Haberland M., Reddy T., Cournapeau D., Burovski E., Peterson P., Weckesser W., Bright J. et. al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 2020, vol. 17, no. 3, pp. 261–272. https://doi.org/10.1038/s41592-019-0686-2

Авторы

Сухов Владимир Дмитриевич — исследователь, Университет Вашингтона в Сент-Луисе, Сент-Луис, 63110, США, sc 57219925767, https://orcid.org/0000-0002-5169-1433, vdsukhov@ya.ru

Короткевич Геннадий Владимирович — ассистент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0009-0004-5941-2816, gkorotkevitch@yandex.ru

Сергушичев Алексей Александрович — кандидат технических наук, доцент, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация; профессор, Университет Вашингтона в Сент-Луисе, Сент-Луис, 63110, США, sc 55772694000, https://orcid.org/0000-0003-1159-7220, alsergbox@gmail.com

Authors

Vladimir D. Sukhov — Researcher, Washington University in St. Louis, Saint Louis, 63110, USA, sc 57219925767, https://orcid.org/0000-0002-5169-1433, vdsukhov@ya.ru

Gennady V. Korotkevich — Assistant, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0009-0004-5941-2816, gkorotkevitch@yandex.ru

Alexey A. Sergushichev — PhD, Assistant Professor, ITMO University, Saint Petersburg, 197101, Russian Federation; Professor, Washington University in St. Louis, Saint Louis, 63110, USA, sc 55772694000, https://orcid.org/0000-0003-1159-7220, alsergbox@gmail.com

Статья поступила в редакцию 12.03.2024
Одобрена после рецензирования 26.06.2024
Принята к печати 23.07.2024

Received 12.03.2024
Approved after reviewing 26.06.2024
Accepted 23.07.2024