ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2022, том 507, с. 36–39

МАТЕМАТИКА

УДК 519.234.3

О КРИТЕРИЯХ ПРОВЕРКИ ГИПОТЕЗЫ ОБ ЭКВИВАЛЕНТНОСТИ ХВОСТОВ РАСПРЕДЕЛЕНИЙ

© 2022 г. Е. О. Кантонистова[1,*], И. В. Родионов[2,]**

Представлено академиком РАН Д.А. Новиковым
Поступило 15.09.2022 г.
После доработки 23.10.2022 г.
Принято к публикации 30.10.2022 г.

Предложен метод проверки гипотезы об эквивалентности хвоста распределения данных с выбранным хвостом распределения – аналога гипотезы согласия для статистики экстремумов. Метод основан на новом преобразовании данных, переводящем k максимальных порядковых статистик выборки из стандартного равномерного закона $U[0,1]$ в случайные величины, похожие в своем асимптотическом поведении на выборку из $U[0,1]$ размера k. Доказано, что критерии, построенные по предложенному методу, являются состоятельными на максимально широкой альтернативе – отрицании основной гипотезы.

Ключевые слова: хвост распределения, критерий согласия, статистика экстремумов, эквивалентность

DOI: 10.31857/S2686954322600586

1. ВВЕДЕНИЕ

При построении вероятностных моделей данных в различных областях знаний возникают ситуации, когда вероятности редких событий не могут быть качественно описаны в рамках популярных моделей и требуют отдельного анализа. Такие ситуации возникают, в частности, в финансовых и страховых задачах, в задачах надежности и демографических исследованиях, когда "тело" распределения описывается, например, нормальным или логнормальным законом, а хвост – правильно меняющимся распределением. Более того, редкие (экстремальные) события могут быть сами по себе центральным объектом анализа, как, например, при изучении природных катаклизмов и катастроф, в задачах безопасности ядерной энергетики и других. Анализ таких событий является основным предметом изучения стохастической теории экстремумов, см. монографии [1, 2]. В настоящей работе нас будет интересовать статистическая часть этой теории.

По сравнению с оцениванием параметров, методам проверки гипотез в статистике экстремумов посвящено не так много работ, см. обзор [3]. Это во многом связано с тем, что наиболее популярной (и, по сути, безальтернативной для практиков) моделью для хвостов распределений до сих пор является модель обобщенного распределения Парето [1, 2], которая не требует применения аппарата проверки гипотез. Однако в связи с тенденцией к увеличению объемов доступных данных становится возможным рассмотрение более узких (семипараметрических) моделей хвостов распределений, например, таких как модели хвостов распределений вейбулловского и лог-вейбулловского типов. Тем самым, возникает необходимость в разработке статистических критериев для выбора подходящей модели хвоста распределения данных, частным случаем которых являются критерии о принадлежности хвоста распределения какому-то определенному классу.

Для этой цели, по аналогии с классической статистикой, могут быть использованы критерии согласия, в статистики которых вместо неизвестных параметров распределений подставлены их оценки. Тем не менее эта задача в литературе практически не рассматривалась – единственным на текущий момент исследованием, посвященным задаче построения критериев согласия с хвостом распределения, является работа [4]. Однако фактически в этой работе были предложены критерии проверки не гипотезы согласия с хвостом распределения, а гипотезы о том, что хвост распределения пропорционален выбранно-

[1] Национальный исследовательский университет "Высшая школа экономики", Москва, Россия
[2] Институт проблем передачи информации им. А.А. Харкевича Российской академии наук, Москва, Россия
**E-mail: ekantonistova@hse.ru*
***E-mail: vecsell@gmail.com*

му, что может привести к принципиально неверным выводам о хвосте распределения на практике при его использовании. Дополнительной мотивацией к изучению критериев согласия с хвостом распределения служит тот факт, что статистики критериев согласия часто используются для определения оптимального значения высокого уровня при оценивании параметров в рамках статистики экстремумов [5].

Для построения критериев согласия с хвостом распределения предлагается воспользоваться идеями, которые были применены при построении критериев согласия для цензурированных данных. При разработке последних в литературе преимущественно использовались 2 подхода: усечение статистики стандартного критерия согласия и специальные преобразования данных, см. обзор литературы в работе [6]. Преимущества и недостатки каждого из подходов рассмотрены в работе [4]. Достаточно полное описание критериев, полученных с помощью первого подхода, можно найти в монографии [7]. Однако в случае стандартной постановки статистики экстремумов, а именно, если для статистического анализа используются лишь k максимальных (или минимальных) порядковых статистик выборки размера n, где последовательность $k = k(n)$ удовлетворяет условиям

$$k \to \infty \quad \text{и} \quad k/n \to 0 \quad \text{при} \quad n \to \infty, \tag{1}$$

статистики критериев согласия, построенных в рамках первого подхода, будут стремиться к нулю по вероятности, что делает этот подход бесполезным для нашей задачи. Далее будем полагать, что последовательности $k(n)$ и $k(n)/n$ являются монотонными с некоторого момента.

Ключевой для второго подхода является работа [8], в которой преобразование k максимальных порядковых статистик выборки из стандартного равномерного закона $U[0,1]$, переводящее их в выборку размера k из $U[0,1]$, было впервые использовано для построения критериев согласия по цензурированной выборке. Пусть $U_1,\dots,U_n$ – независимые одинаково распределенные (н.о.р.) случайные величины из распределения $U[0,1]$, а $U_{(1)} \le \dots \le U_{(n)}$ – вариационный ряд этой выборки. Тогда, как показано в [8],

$$Z_{(i)} = U_{(i)}/U_{(k)} \cdot (B_{k,n}(U_{(k)}))^{1/k}, \quad i = 1,\dots,k, \tag{2}$$

являются порядковыми статистиками выборки $\{Z_i\}_{i=1}^k$ из $U[0,1]$, где

$$B_{k,n}(x) = \sum_{j=k}^{n} C_n^j x^j (1-x)^{n-j}. \tag{3}$$

Далее, в рамках второго подхода для проверки гипотезы согласия $H_0 : F = F_0$ по порядковым статистикам $\{X_{(i)}\}_{i=1}^k$ выборки $\{X_i\}_{i=1}^n$ предлагается подставлять в статистики стандартных критериев согласия величины $\{Z_{(i)}\}_{i=1}^k$ вместо $\{U_{(i)}\}_{i=1}^k$, где $U_{(i)} = F_0(X_{(i)})$ и $\{Z_{(i)}\}_{i=1}^k$ получены из $\{U_{(i)}\}_{i=1}^k$ согласно формуле (2). Тем самым, при верности нулевой гипотезы распределения статистик критериев согласия, полученных в рамках второго подхода, будут совпадать с распределениями статистик их стандартных аналогов, что, однако, не будет выполняться в случае верности альтернативной гипотезы. Другие подобные преобразования можно найти в работах [9, 10]. Однако, как выяснилось, критерии согласия с хвостом распределения, которые могут быть получены на основе преобразования (2), фактически проверяют гипотезу об асимптотической пропорциональности хвоста распределения наблюдений с хвостом выбранного распределения и, тем самым, обладают схожими недостатками с критериями, предложенными в работе [4].

В этой работе мы предложим новое преобразование данных, похожее на преобразование (2), результатом которого будут случайные величины, не являющиеся, однако, независимыми и распределенными по закону $U[0,1]$. Тем не менее “эмпирическая функция распределения”, построенная по этим величинам, в случае верности основной гипотезы будет сходиться к функции распределения стандартного равномерного закона. Это свойство данного преобразования позволит нам на его основе предложить метод построения критериев для проверки гипотезы согласия с хвостом распределения, т.е. о том, что хвост распределения асимптотически эквивалентен выбранному. Также мы показываем состоятельность предложенных критериев на максимально широкой альтернативной гипотезе – отрицании основной гипотезы.

2. ОСНОВНЫЕ РЕЗУЛЬТАТЫ

Пусть F – функция распределения. Определим ее правую граничную точку как $x_F^* = \inf\{x : F(x) = 1\}$ и хвостовую функцию распределения как $\overline{F}(x) = 1 - F(x)$. Будем говорить, что хвостовые функции распределения $\overline{F_0}$ и $\overline{F_1}$ эквивалентны (пишем $F_0 \overset{r}{\sim} F_1$), если их правые граничные точки совпадают, т.е. $x^* := x_{F_0}^* = x_{F_1}^*$, и выполнено соотношение

$$\lim_{x \uparrow x^*} \frac{\overline{F_0}(x)}{\overline{F_1}(x)} = 1.$$

Правым хвостом функции распределения F назовем класс эквивалентности $T(\overline{F})$ хвостовых функций распределения по отношению $\overset{r}{\sim}$, т.е. для функции распределения G свойство $G \overset{r}{\sim} F$ эквивалентно выполнению $\overline{G} \in T(\overline{F})$. Далее мы продолжим говорить о правых хвостах распределений, хотя все приведенные ниже рассуждения можно повторить и для левых хвостов. Определим *гипотезу согласия с (правым) хвостом распределения* F_0 как $H_0 : F_1 \overset{r}{\sim} F_0$. Отметим, что в работе [4] гипотеза согласия с хвостом распределения определялась как $H_0^T : F_1(x) = F_0(x)$ *для всех достаточно больших* x, однако стохастическая теория экстремумов является асимптотической наукой, и проверить на практике выполнение соотношения $F_1(x) = F_0(x)$ для всех $x > x_0$, в отличие от более слабого условия $F_1 \overset{r}{\sim} F_0$, может оказаться достаточно проблематичным.

Пусть $(X_1,\ldots,X_n)$ – н.о.р. случайные величины с непрерывной функцией распределения F_1, а функция распределения F_0 тоже непрерывна. Предположим, что мы хотим проверить гипотезу H_0 по этой выборке. Для удобства перейдем к рассмотрению н.о.р. случайных величин $\{U_i\}_{i=1}^n$, где $U_i = F_0(X_i)$, и, тем самым, гипотезе $H_0' : F \overset{r}{\sim} F_{U[0,1]}$, где $F = F_1(F_0^{\leftarrow})$ – функция распределения случайной величины U_1, $F_0^{\leftarrow}(t) = \inf\{x : F_0(x) = t\}$ и $F_{U[0,1]}$ – функция распределения стандартного равномерного закона. Заметим, что выводы о правом хвосте распределения возможно делать только по максимальным членам вариационного ряда выборки, поэтому будем рассматривать только k максимальных членов вариационного ряда выборки $\{U_i\}_{i=1}^n$, где $k < n$.

Введем следующее преобразование k максимальных порядковых статистик выборки $\{U_i\}_{i=1}^n$,

$$V_{(i)} = \frac{1 - U_{(n-i+1)}}{1 - U_{(n-k+1)}} f_{k,n}(1 - U_{(n-k+1)}), \quad i = 1,\ldots,k,$$

где семейство функций $\{f_{k,n}(x), x \in [0,1]\}$ таково, что $f_{k,n}(k/n) \to 1$ и $\limsup_n f_{k,n}(x) \in [0,1)$ для $x = x(n)$ с $\lim_n xn/k \neq 1$, если $k = k(n)$ удовлетворяет (1) при $n \to \infty$. Сохраняя связь с преобразованием (2), положим

$$f_{k,n}(x) = \frac{1}{2}((B_{k,n}(x))^{1/k} + (1 - B_{k+1,n}(x))^{1/k}), \tag{4}$$

где функция $B_{k,n}(x)$ определена в (3). Действительно, функция (4) удовлетворяет условиям, наложенным выше, поскольку, как несложно проверить, $(B_{k,n}(x))^{1/k} \to 1$ для $x = x(n)$ таких, что $\liminf_n xn/k \geq 1$, и $\limsup_n (B_{k,n}(x))^{1/k} < 1$ для $x = x(n)$ таких, что $\limsup_n xn/k < 1$, если $k = k(n)$ удовлетворяет (1) при $n \to \infty$.

Назовем $F_{k,n}^*(x) = k^{-1}\sum_{i=1}^k I(V_{(i)} \leq x)$ эмпирической функцией распределения набора случайных величин $\{V_{(i)}\}_{i=1}^k$. Далее будем предполагать, что F дважды дифференцируема. Для формулировки результатов этой работы введем следующее условие, являющееся классическим для стохастической теории экстремумов: скажем, что функция распределения F удовлетворяет условию фон Мизеса [2], если

$$\lim_{x \uparrow x^*} \frac{(1 - F(x))F''(x)}{(F'(x))^2} = -\gamma - 1, \tag{5}$$

где γ – индекс экстремального значения. В частности, в случае верности гипотезы H_0' $\gamma = -1$ и $x^* = 1$.

Далее, обозначим через D пространство Скорохода, т.е. пространство непрерывных справа функций, имеющих предел слева, на $[0,1]$. Следующая теорема является развитием классического результата Донскера, Колмогорова и Скорохода о сходимости нормированной разности эмпирической и теоретической функций распределения к броуновскому мосту (гауссовскому процессу $B(t)$ на отрезке $[0,1]$ с нулевой функцией среднего и ковариационной функцией $r(s,t) = \min(s,t) - st$).

Теорема 1. *Пусть $\{U_i\}_{i=1}^n$ – н.о.р. случайные величины с функцией распределения F, удовлетворяющей условию* (5). *Предположим, что выполнена гипотеза H_0'. Пусть последовательность $k = k(n)$ удовлетворяет условиям* (1) *и*

$$\sqrt{k}\left|\frac{1 - F(k/n)}{k/n} - 1\right| \to 0 \quad \text{при} \quad n \to \infty. \tag{6}$$

Тогда процесс $y(t) = \sqrt{k}|F_{k,n}^*(t) - t|$ слабо сходится в D к броуновскому мосту $B(t)$ при $n \to \infty$.

Данная теорема позволяет строить критерии проверки гипотезы согласия H_0' на основе статистик классических критериев согласия. В качестве примера рассмотрим критерии Колмогорова и Андерсона-Дарлинга; разумеется, приведенные ниже рассуждения могут быть адаптированы и для других критериев согласия. Обозначим через

$$D_{k,n}^* = \sup_{t\in[0,1]}|F_{k,n}^*(t) - t|, \quad W_{k,n}^* = \int_0^1 \frac{(F_{k,n}^*(t) - t)^2}{t(1-t)}dt$$

модификации статистик критериев согласия Колмогорова и Андерсона-Дарлинга проверки гипотезы $H_0^* : F = F_{U[0,1]}$ соответственно. Пусть $\alpha \in [0,1]$. Тогда, как следует из Теоремы 1, в случае выполнения условий (1) и (6) правила

$$\text{если}\quad \sqrt{k}D^*_{k,n} > K_{1-\alpha},\quad \text{то отвергнуть } H'_0, \quad (7)$$

$$\text{если}\quad kW^*_{k,n} > W_{1-\alpha},\quad \text{то отвергнуть } H'_0 \quad (8)$$

являются критериями проверки гипотезы H'_0, асимптотически имеющими уровень значимости α, где $K_{1-\alpha}$ и $W_{1-\alpha}$ – квантили уровня $1-\alpha$ распределений Колмогорова и Андерсона-Дарлинга соответственно. Следующая теорема позволяет утверждать, что критерии (7) и (8) являются состоятельными на альтернативе $H'_1 : H'_0$ неверна.

Теорема 2. *Пусть* $\{U_i\}_{i=1}^n$ *– н.о.р. случайные величины с функцией распределения* F, *удовлетворяющей условию* (5). *Предположим, что верна гипотеза* H'_1, *а последовательность* $k = k(n)$ *удовлетворяет* (1). *Тогда для всех* $t \in (0,1)$, *кроме, может быть, одной точки*

$$\sqrt{k}|F^*_{k,n}(t) - t| \to \infty, \quad n \to \infty.$$

В данной работе предложен метод построения критериев проверки гипотезы согласия с хвостом распределения H'_0. Метод основан на новом преобразовании данных, который переводит k максимальных членов вариационного ряда выборки из стандартного равномерного закона в случайные величины, близкие по своему поведению к вариационному ряду выборки из стандартного равномерного закона размера k. В отличие от первой работы [4], посвященной данной задаче, в статье предложены критерии в точности для проверки гипотезы согласия с хвостом распределения и доказывается их состоятельность на максимально широкой альтернативе – отрицании основной гипотезы.

ИСТОЧНИК ФИНАНСИРОВАНИЯ

Работа над разделами 2 и 3 выполнена И.В. Родионовым за счет гранта Российского научного фонда (проект № 21-71-00035) в Институте проблем передачи информации им. А.А. Харкевича Российской академии наук.

СПИСОК ЛИТЕРАТУРЫ

1. *Beirlant J., Goegebeur Y., Teugels J., Segers J.* Statistics of Extremes: Theory and Applications. N.Y.: Wiley. 2004. 498 p. https://doi.org/10.1002/0470012382
2. *de Haan L., Ferreira A.* Extreme Value Theory: An Introduction. N.Y.: Springer Verlag. 2006. 417 p. https://doi.org/10.1007/0-387-34471-3—8-3-642-33483-2
3. *Hüsler J., Peng L.* Review of testing issues in extremes: in honor of Professor Laurens de Haan // Extremes. 2008. V. 11. № 1. P. 99–111. https://doi.org/10.1007/s10687-007-0052-0
4. *Kantonistova E.O., Rodionov I.V.* Analogues of classical goodness-of-fit tests for distribution tails // Doklady Mathematics. 2021. V. 103. I. 1. P. 35–38. https://doi.org/10.1134/S1064562421010063
5. *Danielsson J., Ergun L.M., de Haan L., De Vries C.* Tail Index Estimation: Quantile Driven Threshold Selection. Available at SSRN: https://ssrn.com/abstract'17478. 2016. https://doi.org/10.2139/ssrn.2717478
6. *Goldmann C., Klar B., Meintanis S. G.* Data transformations and goodness-of-fit tests for type-II right censored samples // Metrika. 2015. V. 78. P. 59–83. https://doi.org/10.1007/s00184-014-0490-z
7. *D'Agostino R.B., Stephens M.A.* Goodness-of-Fit Techniques. New York: Marcel Dekker. 1986. 576 p. https://doi.org/10.1201/9780203753064
8. *Michael J.R., Schucany W.R.* A new approach to testing goodness of fit for censored samples // Technometrics. 1979. V. 21. P. 435–441. https://doi.org/10.1080/00401706.1979.10489813
9. *Lin C.-T., Huang Y.-L., Balakrishnan N.* A New Method for Goodness-of-Fit Testing Based on Type-II Right Censored Samples // IEEE Transactions in Reliability. 2008. V. 57. № 4. P. 633–642. https://doi.org/10.1109/TR.2008.2005860
10. *Fischer T., Kamps U.* On the existence of transformations preserving the structure of order statistics in lower dimensions // Journal of Statistical Planning and Inference. 2011. V. 141. P. 536–548. https://doi.org/10.1016/j.jspi.2010.06.028

ON PROCEDURES FOR TESTING EQUIVALENCE OF DISTRIBUTION TAILS

E. O. Kantonistova[a] and I. V. Rodionov[b]

[a] *National Research University Higher School of Economics, Moscow, Russia*

[b] *Institute for Information Transmission Problems (Kharkevich Institute) of RAS, Moscow, Russia*

Presented by Academician of the RAS D.A. Novikov

We propose the method for testing the hypothesis about the equivalence of distribution tail of the observed data and certain distribution tail, it is the analogue of goodness-of-fit hypothesis for statistics of extremes. The method is based on the new data transformation, moving up k largest order statistics of a sample from standard uniform distribution $U[0,1]$ to random variables asymptotically similar to a sample from $U[0,1]$ of size k. We prove that tests built by the proposed method are consistent on the widest alternative – specifically, on the negation of the main hypothesis.

Keywords: distribution tail, goodness-of-fit test, statistics of extremes, equivalence