ДОКЛАДЫ АКАДЕМИИ НАУК, 2019, том 488, № 4, с. 358–361

МАТЕМАТИКА

УДК 519.224.24

О ПАРАМЕТРИЧЕСКОМ ОЦЕНИВАНИИ ХВОСТА РАСПРЕДЕЛЕНИЯ

И. В. Родионов

Представлено академиком РАН А.Н. Ширяевым 15.05.2019 г.

Поступило 21.05.2019 г.

В работе предложен общий метод оценивания параметра хвоста распределения, не зависящий от выполнения условий теоремы Гнеденко. Доказана состоятельность введённой оценки и, в более сильных условиях на параметрическое семейство хвостов распределений, её асимптотическая нормальность. Также приведена адаптация метода для оценивания вейбулловского и лог-вейбулловского хвостовых индексов.

Ключевые слова: хвост распределения, параметрическое оценивание, стохастическая теория экстремумов, оценка типа Хилла.

DOI: https://doi.org/10.31857/S0869-56524884358-361

Во многих областях, в частности связанных с финансами, страхованием, телекоммуникациями, возникает задача оценивания хвоста распределения некоторых данных. Если центральные значения функции распределения могут быть промоделированы с помощью стандартных статистических процедур, то для оценивания хвоста распределения применяются специальные методы стохастической теории экстремумов, в рамках которых используются только максимальные наблюдения выборки.

В настоящее время основным методом оценивания хвоста распределения является семипараметрический подход, основанный на теореме Гнеденко–Фишера–Типпета [1]. Пусть $\mathbf{X}_n = (X_1,\ldots,X_n)$ – независимые одинаково распределённые (н.о.р.) случайные величины с функцией распределения (ф.р.) F, обозначим $M_n = \max(X_1,\ldots,X_n)$. Теорема утверждает, что если для некоторых последовательностей констант $\{a_n > 0\}$ и $\{b_n\}$ выполнено

$$\lim_{n\to\infty} P\left(\frac{M_n - b_n}{a_n} \le x\right) = \lim_{n\to\infty} F^n(a_n x + b_n) = G(x) \quad (1)$$

для некоторой невырожденной ф.р. $G(x)$, то найдутся такие константы $a > 0$ и b, что $G(ax+b) = G_\gamma(x)$, где

$$G_\gamma(x) = \begin{cases} \exp(-(1+\gamma x)^{-1/\gamma}),\ 1+\gamma x > 0,\ \gamma \ne 0, \\ \exp(-\exp(-x)),\ x \in \mathbb{R},\ \gamma = 0. \end{cases} \quad (2)$$

Если (1) и (2) выполнены для ф.р. F, то говорят, что F принадлежит области максимального притяжения G_γ, пишем $F \in \mathcal{D}(G_\gamma)$. Параметр γ, называемый индексом экстремального значения, позволяет разделить распределения, удовлетворяющие условиям теоремы Гнеденко, на три класса. Класс $\gamma > 0$, называемый областью максимального притяжения Фреше, содержит распределения со степенными хвостами, тогда как распределения, попавшие в класс $\gamma < 0$ (область максимального притяжения Вейбулла), характеризуются конечной правой точкой распределения, т.е. $x_F^* = \inf\{x:\ F(x) < 1\} < +\infty$. Класс распределений, для которых $\gamma = 0$, называется областью максимального притяжения Гумбеля и содержит распределения как с конечной, так и с бесконечной правой точкой. В основном распределения из этого класса имеют хвосты экспоненциального типа.

В случае выполнения условий теоремы Гнеденко хвост функции распределения оценивается с использованием представления, возникающим в теореме Пикандса–Балкема–де Хаана [2, 3]: при $u \to x_F^*$ условное распределение случайной величины $X_1 - u$ при условии $X_1 > u$ приближается обобщённым законом Парето $GPD(\sigma,\gamma)$,

$$P(X_1 > u + y \mid X_1 > u) - (1 + \gamma y/\sigma)^{-1/\gamma} \to 0,\ \ y > 0,$$

где $\sigma = a + \gamma(u - b)$, а выражение $(1+\gamma x)^{-1/\gamma}$ при $\gamma = 0$ стоит понимать как e^{-x}. Тогда в качестве состоятельной оценки ф.р. F случайной величины X_1 можно выбрать

$$F^*(x) = \begin{cases} F_n(x),\ \ x \le u, \\ 1 - \zeta_u^* \left(1 + \dfrac{\gamma_n^*(x-u)}{\sigma_n^*}\right)^{-1/\gamma_n^*},\ \ x > u, \end{cases}$$

где $F_n(x)$ – эмпирическая функция распределения выборки $\mathbf{X}_n$ или любая другая состоятельная оценка

Математический институт им. В.А. Стеклова Российской Академии наук, Москва

Институт проблем управления им. В.А. Трапезникова Российской Академии наук, Москва

E-mail: vecsell@gmail.com

F, ζ_u^* – оценка вероятности $P(X_1 > u)$ и γ_n^*, σ_n^* – оценки параметров γ и σ соответственно. В качестве порога u, как правило, выбирается $X_{(n-k)}$, $(n-k)$-я порядковая статистика выборки $\mathbf{X}_n$, $k < n$, в этом случае $\zeta_u^* = k/n$. Отсюда возникает задача оптимального выбора параметра k [4]. Оценки параметров γ и σ можно получить с помощью аналога метода максимального правдоподобия [5]. Другие оценки индекса экстремального значения γ предложены, в частности, в работах [2, 6]. Подробнее о статистике экстремумов можно узнать в монографиях [7, 8].

Однако указанный подход имеет свои недостатки. Так, существует большой класс распределений, например распределения с хвостом логарифмического типа или такие распределения, как пуассоновское и геометрическое, для которых не выполнены условия теоремы Гнеденко, поэтому применение рассмотренного подхода не представляется возможным. Данный метод хорошо работает для распределений, принадлежащих областям максимального притяжения Фреше и Вейбулла, так как они хорошо описываются посредством индекса экстремального значения γ, однако с помощью него невозможно различить распределения из области максимального притяжения Гумбеля, поскольку $\gamma = 0$ для всех распределений из этой области. Тем самым, возникает необходимость в общем методе оценивания хвоста распределения, не зависящем от выполнения условий теоремы Гнеденко. Настоящая работа является первым шагом в решении данной задачи.

Пусть ф.р. F является непрерывной, а также $x_F^* = +\infty$. Заметим, что для предлагаемого в работе метода достаточно, чтобы F была непрерывна, начиная с некоторого $x_0 \in \mathbb{R}$. Будем говорить, что хвост непрерывной ф.р. G легче хвоста непрерывной ф.р. H с условием $x_G^* = x_H^* = +\infty$, если

$$\lim_{x\to\infty} \frac{1 - G(x)}{1 - H(x)} = 0.$$

Если указанный предел равняется 1, то скажем, что G и H имеют одинаковый хвост. Далее будем полагать, что все рассматриваемые функции распределения непрерывны и обладают бесконечной правой точкой.

О п р е д е л е н и е . Будем говорить, что ф.р. G и H удовлетворяют условию $B(G, H)$, если для некоторых $\varepsilon > 0$ и x_0 функция

$$\frac{(1 - G(x))^{1-\varepsilon}}{1 - H(x)} \text{ не возрастает при } x > x_0.$$

Легко видеть, что при выполнении данного условия хвост ф.р. G легче хвоста ф.р. H. Далее, назовём параметрическое семейство хвостов функций распределения $\mathcal{F} = \{F_\theta,\ \theta \in \Theta\}$, $\Theta \subset \mathbb{R}$, упорядоченным по параметру θ, если для всех $\theta_1, \theta_2 \in \Theta, \theta_1 < \theta_2$, выполнено условие $B(F_{\theta_1}, F_{\theta_2})$ или для всех $\theta_1,\ \theta_2 \in \Theta,\ \theta_1 < \theta_2$, выполнено $B(F_{\theta_2}, F_{\theta_1})$. Предположим, что хвост ф.р. F выборки $\mathbf{X}_n$ принадлежит параметрическому семейству $\mathcal{F}$, упорядоченному по параметру θ. Предложим общий метод оценивания параметра θ хвоста ф.р. F. Для этой цели рассмотрим статистику

$$R_{k,n}(\theta) = \ln(1 - F_\theta(X_{(n-k)})) - \frac{1}{k}\sum_{i=n-k+1}^{n} \ln(1 - F_\theta(X_{(i)})), \tag{3}$$

где $X_{(1)} \le \ldots \le X_{(n)}$ – вариационный ряд выборки $\mathbf{X}_n$. Данная статистика является обобщением статистики $R_{k,n}$, предложенной автором в работе [9] для решения задачи различения двух разделимых классов хвостов распределений. Этот результат позволяет выбрать подходящее параметрическое семейство $\mathcal{F}$ в рамках нашей задачи, см. также [10, 11]. Следующая теорема утверждает, что оценка

$$\theta_{k,n}^* = \arg\{\theta\colon R_{k,n}(\theta) = 1\} \tag{4}$$

является состоятельной для параметра θ.

Т е о р е м а 1 . *Пусть $X_1, \ldots, X_n$ – н.о.р. случайные величины с ф.р. $F = F_{\theta_0}$, семейство хвостов распределений $\mathcal{F} = \{F_\theta,\ \theta \in \Theta\}$, $\Theta \subset \mathbb{R}$ является упорядоченным по параметру θ и $F_\theta(x)$ непрерывна по x и θ. Предположим, что последовательность $k = k(n)$, $k \in \mathbb{N}$, удовлетворяет условиям*

$$k \to \infty,\ \frac{k}{n} \to 0 \ \textit{при}\ n \to \infty.$$

Тогда

$$\theta_{k,n}^* \xrightarrow{P_{\theta_0}} \theta_0.$$

З а м е ч а н и е . Условие упорядоченности параметрического семейства $\mathcal{F}$ по параметру θ гарантирует единственность почти наверное решения уравнения $R_{k,n}(\theta) = 1$ относительно θ. Тем самым, оценка $\theta_{k,n}^*$ определена корректно.

Для того чтобы оценка $\theta_{k,n}^*$ была асимптотически нормальной, на параметрическое семейство $\mathcal{F}$ необходимо наложить более сильные условия. Обозначим $S(x, \theta) = \ln(1 - F_\theta(x))$. Легко видеть, что функция $S(x, \theta)$ дифференцируема по θ тогда и только тогда, когда $F_\theta(x)$ дифференцируема по θ. Обозначим также

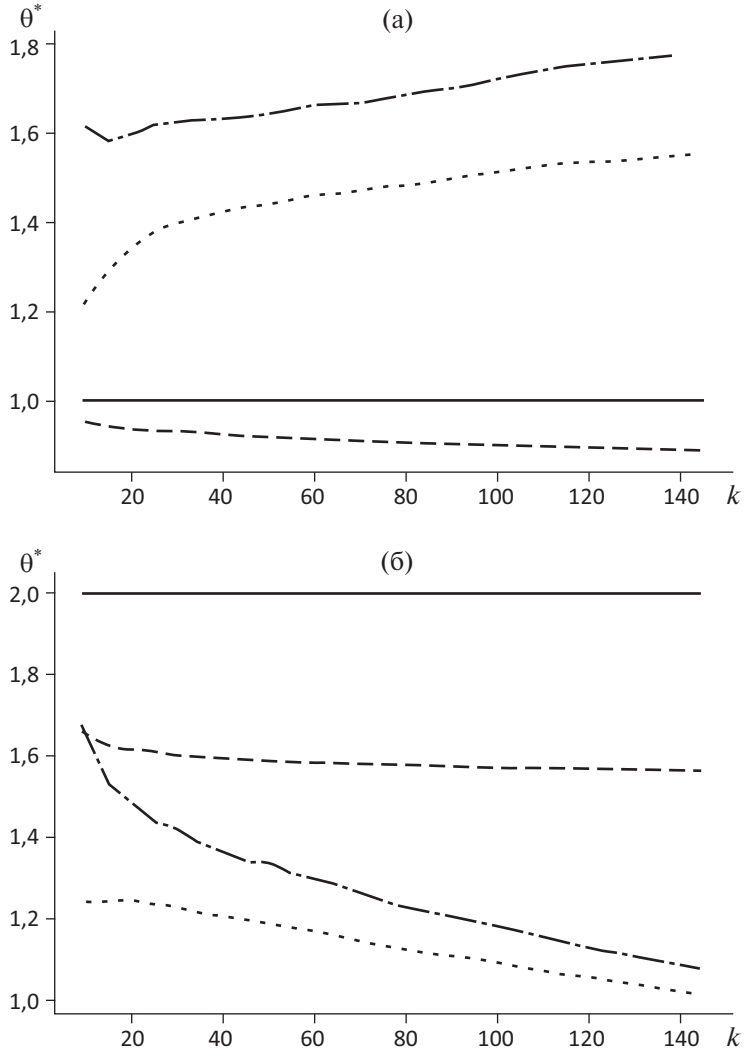

Рис. 1. Сравнение оценок θ_j^*, $j = 0, 1, 2$, на выборках из распределений Г(3, 1) (а) и $N(0, 1)$ (б).

$$I(q,\theta) = \frac{1}{1 - F_\theta(q)} \int_q^\infty \frac{\partial S(x,\theta)}{\partial \theta} dF_\theta(x) - \frac{\partial S(q,\theta)}{\partial \theta}.$$

Рассмотрим следующие условия регулярности:

А1. Функция $\dfrac{\partial S(x,\theta)}{\partial \theta}$ непрерывна по паре переменных (x, θ) при $x > x_0$;

А2. Для всех $\theta_0 \in \Theta$ существует $x_1(\theta_0)$ такое, что функция $\dfrac{\partial S(x,\theta)}{\partial \theta}$ монотонна по θ в некоторой окрестности θ_0 при $x > x_1(\theta_0)$.

А3. Для всех $\theta_0 \in \Theta$ существует $x_2(\theta_0)$ такое, что выполнено $| I(x,\theta_0) | < \infty$ при $x > x_2(\theta_0)$.

В следующей теореме установлены условия асимптотической нормальности оценки (4).

Т е о р е м а 2. *Пусть выполнены условия* А1–А3. *Тогда в предположениях теоремы* 1,

$$\sqrt{k} I(X_{(n-k)}, \theta_{k,n}^*)(\theta_{k,n}^* - \theta_0) \xrightarrow{d_{\theta_0}} \xi \sim N(0, 1)$$

для всех $\theta_0 \in \Theta$.

Предложенный нами метод может быть использован не только в рассмотренной постановке, но и в задаче оценивания вейбулловского и лог-вейбулловского индексов. Семейства распределений вейбулловского и лог-вейбулловского типа образуют важные классы в области максимального притяжения Гумбеля. Будем говорить, что ф.р. F имеет хвост типа Вейбулла, если существует такое $\theta > 0$, что для всех $\lambda > 0$

$$\lim_{x \to \infty} \frac{\ln(1 - F(\lambda x))}{\ln(1 - F(x))} = \lambda^\theta.$$

Параметр θ в этом случае называется вейбулловским хвостовым индексом. Легко видеть, что данный индекс задаёт поведение хвоста ф.р. F, и чем индекс больше, тем легче хвост. Тем самым задача оценивания вейбулловского хвостового индекса является первостепенной в случае, если требуется оценить хвост вейбулловского типа. Далее, если для некоторой ф.р. F ф.р. $F(e^x)$ имеет хвост типа Вейбулла с индексом θ, то в таком случае скажем, что F имеет лог-вейбулловский хвост, а θ будем называть лог-вейбулловским хвостовым индексом. Заметим, что распределение лог-вейбулловского типа принадлежит области максимального притяжения Гумбеля лишь в случае $\theta > 1$. Оценки вейбулловского хвостового индекса, в частности, были предложены в работах [12–14], тогда как задача оценивания лог-вейбулловского хвостового индекса, насколько нам известно, в литературе до сих пор не рассматривалась.

Чтобы оценить вейбулловский хвостовой индекс с помощью предложенного метода, выберем в (3) параметрическое семейство $\mathcal{F}^W = \{F_\theta^W(x),\ \theta > 0\}$, где $F_\theta^W(x) = 1 - \exp(-x^\theta),\ \ x > 0$. Для оценивания лог-вейбулловского хвостового индекса стоит выбрать параметрическое семейство $\mathcal{F}^{LW} = \{F_\theta^{LW}(x),\ \theta > 0\}$, где $F_\theta^{LW}(x) = 1 - \exp(-(\ln x)^\theta),\ x > 1$. На рисунках 1а и 1б приведены средние значения оценки $\theta_0^* = \theta_{k,n}^*$ (штриховая линия), а также оценок θ_1^* [12] (пунктирная линия) и θ_2^* [14] (штрих-пунктирная линия) вейбулловского хвостового индекса, смоделированные для выборок размера $n = 1000$ из распределений $\Gamma(3,1)$ и $N(0,1)$, в зависимости от k, количество выборок равно $m = 300$. Сплошной линией обозначено истинное значение вейбулловского хвостового индекса, оно равно 1 для гамма-распределения и 2 для нормального распределения. Заметим, что гамма-распределение и нормальное распределение не принадлежат классу $\mathcal{F}^W$, что свидетельствует об устойчивости предложенного метода к нарушению его предположений.

Источник финансирования. Работа выполнена при поддержке гранта РНФ, проект 19–11–00290.

СПИСОК ЛИТЕРАТУРЫ

1. *Gnedenko B.V.* // Annals of Mathematics. 1943. V. 44. P. 423–453.
2. *Pickands III J.* // Annals of Statistics. 1975. V. 3. P. 119–131.
3. *Balkema A. A., de Haan L.* // Annals of Probability. 1974. V. 2. P. 792–804.
4. *de Haan L., Peng L.* // Statistica Neerlandica. 1998. V. 52. P. 60–70.
5. *Smith R.L.* // Annals of Statistics. 1987. V. 15. P. 1174–1207.
6. *Hill B.* // Annals of Statistics. 1975. V. 3. P. 1163–1174.
7. *de Haan L., Ferreira A.* Extreme Value Theory: An Introduction. New York: Springer Verlag. 2006. 417 p.
8. *Beirlant J., Goegebeur Y., Teugels J., Segers J.* Statistics of Extremes: Theory and Applications. N.Y.: Wiley, 2004. 498 p.
9. *Родионов И.В.* // Проблемы передачи информации. 2018. В. 54. № 2. С. 29–44.
10. *Родионов И.В.* // Теория вероятностей и ее применения. 2018. В. 63. № 2. С. 402–413.
11. *Родионов И.В.* // Теория вероятностей и ее применения. 2018. В. 63. № 3. С. 447–467.
12. *Beirlant J., Broniatowski M., Teugels J.L., Vynckier P.* // J. of Statistical Planning and Inference. 1995. V. 45. P. 21–48.
13. *Balakrishnan N., Kateri M.* // Statistical and Probability Letters. 2008. V. 78. P. 2971–2975.
14. *Gardes L., Girard S., Guillou A.* // J. of Statistical Planning and Inference. 2009. V. 141. № 4. P. 429–444.

INFERENCES ON PARAMETRIC ESTIMATION OF DISTRIBUTION TAILS

I. V. Rodionov

Steklov Mathematical Institute of Russian Academy of Sciences, Moscow, Russian Federation
Trapeznikov Institute of Control Sciences of the Russian Academy of Sciences, Moscow, Russian Federation

Presented by Academician of the RAS A.N. Shiryaev May 15, 2019

Received May 21, 2019

We propose a general method of parameter estimation of a distribution tail, that does not depend on the fulfillment of the conditions of Gnedenko theorem. We prove the consistency of the proposed estimator and its asymptotic normality under the stronger conditions imposed on the parametric family of distribution tails. Additionally, the adaptation of the proposed method to Weibull and log-Weibull tail indices estimation is provided.

Keywords: distribution tail, parametric estimation, extreme value theory, Hill-type estimator.