ТРУДЫ ИНСТИТУТА МАТЕМАТИКИ И МЕХАНИКИ УрО РАН

Том 24 № 2 2018

УДК 519.213

АСИМПТОТИЧЕСКИЙ ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ ДЛЯ ТОЧКИ РАЗРЫВА ПЛОТНОСТИ РАСПРЕДЕЛЕНИЯ

В. Е. Мосягин, Н. А. Швемлер

Рассматривается задача об интервальном оценивании неизвестного параметра $\theta \in \Theta \subset R$ плотности распределения $f(x,\theta)$ (относительно меры Лебега) по выборке $X_1,\ldots,X_n$ большого объема. Предполагается, что в точке $x=\theta$ плотность имеет разрыв первого рода. Доверительный интервал строится по известной оценке максимального правдоподобия θ_n^* и ранее найденной авторами функции распределения $G(x,\theta)$, которая является предельной для последовательности функций распределений нормированных оценок максимального правдоподобия $n(\theta_n^*-\theta)$. Доказывается, что построенный доверительный интервал является асимптотически точным. В статье также описан способ "быстрого" вычисления оценок максимального правдоподобия для точки разрыва плотности.

Ключевые слова: оценивание точки разрыва плотности вероятности, оценки максимального правдоподобия, асимптотический доверительный интервал, предельные распределения статистических оценок.

V. E. Mosyagin, N. A. Shvemler. Asymptotic confidence interval for a discontinuity point of a probability density function.

We consider the problem of interval estimation of an unknown parameter $\theta \in \Theta \subset R$ of a distribution density $f(x,\theta)$ (with respect to the Lebesgue measure) for a sample $X_1,\ldots,X_n$ of large size. It is assumed that the density has a discontinuity of the first kind at the point $x=\theta$. We construct a confidence interval based on a known maximum likelihood estimator θ_n^* and the distribution function $G(x,\theta)$ found by the authors earlier, which is the limit of the sequence of distribution functions of normalized maximum likelihood estimators $n(\theta_n^*-\theta)$. It is proved that the resulting confidence interval is asymptotically exact. We also describe a method for the "fast" calculation of maximum likelihood estimators for a discontinuity point of a density.

Keywords: estimation of a discontinuity point of a probability density, maximum likelihood estimators, asymptotic confidence interval, limiting distributions of statistical estimators.

MSC: 60G51

DOI: 10.21538/0134-4889-2018-24-2-194-199

1. Введение

Пусть $(X_1,\ldots,X_n)$ — вещественная выборка, состоящая из независимых случайных величин, с общей для всех величин плотностью распределения $f(x,\theta)$ (относительно меры Лебега):

$$f(x,\theta)=\begin{cases} f_1(x,\theta), & x<\theta,\\ f_2(x,\theta), & x>\theta. \end{cases} \tag{1.1}$$

Неизвестный параметр $\theta\in\Theta$ с истинным значением θ_0 подлежит оцениванию. Плотность (1.1) предполагается непрерывной по x всюду, за исключением точки θ, в которой она имеет разрыв первого рода:

$$0<q(\theta)=f_1(\theta-0,\theta)<f_2(\theta+0,\theta)=p(\theta),\quad \theta\in\Theta.$$

Условие $p(\theta)>q(\theta)$ принимается для определенности.

Пусть $\{\theta_n^*\}$ обозначают оценки максимального правдоподобия (ОМП), которые определим, следуя [2, с. 360], из равенства:

$$\max\left\{\prod_{i=1}^n f(X_i,\theta_n^*-0),\prod_{i=1}^n f(X_i,\theta_n^*+0)\right\}=\sup_{\theta\in\Theta}\prod_{i=1}^n f(X_i,\theta). \tag{1.2}$$

Если θ_n^* не единственное решение уравнения (1.2), то будем считать ОМП любое решение уравнения (1.2).

Обозначим $t_n^* = n(\theta_n^* - \theta_0)$ — нормированное отклонение ОМП от истинного значения параметра θ_0.

В монографии И. А. Ибрагимова, Р. З. Хасьминского [2] при определенных допущениях установлен факт сходимости по распределению $t_n^* \Rightarrow t^*$, где предельная величина t^* является моментом достижения максимума процессом

$$Y(t) = a(\theta_0)t - \nu_+ (p(\theta_0)t) + \nu_- (-q(\theta_0)t) \tag{1.3}$$

с линейным коэффициентом

$$a(\theta) = \frac{p(\theta) - q(\theta)}{\ln(p(\theta)/q(\theta))}, \tag{1.4}$$

$\nu_\pm(t)$ — независимые стандартные пуассоновские процессы при $t \geq 0$, доопределенные нулем на отрицательной оси. Следующие соотношения

$$p(\theta) > a(\theta) > q(\theta) > 0, \quad \theta \in \Theta, \tag{1.5}$$

являются следствием очевидного неравенства $\ln x < x - 1$ при $x > 0$, $x \neq 1$ и обеспечивают процессу (1.3) отрицательный средний снос.

Цель работы – нахождение асимптотического доверительного интервала для точки θ_0 разрыва плотности (1.1). Построение доверительного интервала стало возможным после выхода в свет работы [5] авторов, в которой впервые была найдена функция распределения предельной величины t^*

$$G(x, \theta_0) = \mathbf{P}\left(t^* < x\right) = \lim_{n\to\infty} \mathbf{P}\left(n(\theta_n^* - \theta_0) < x\right).$$

При $x \geqslant 0$

$$\begin{gathered} G(x,\theta) = \frac{(p(\theta) - a(\theta))q(\theta)}{(p(\theta) - q(\theta))a(\theta)} + \int\limits_0^{a(\theta)x} g^+(t,\theta)\,dt, \\ g^+(t,\theta) = \beta(\theta)e^{-p(\theta)t/a(\theta)} \sum_{k=0}^{[t]} \psi(t-k,\theta)\Big(\frac{p(\theta)}{a(\theta)}\Big)^k \Big(\frac{t^k}{k!} - \frac{t^{k-1}}{(k-1)!}\Big). \end{gathered} \tag{1.6}$$

Здесь функция распределения $\psi(z,\theta)$ определена равенством [8, с. 167]

$$\psi(z,\theta) = \Big(1 - \frac{q(\theta)}{a(\theta)}\Big) \sum_{k=0}^{[z]} (-1)^k \Big(\frac{q(\theta)}{a(\theta)}\Big)^k \frac{(z-k)^k}{k!} e^{q(\theta)(z-k)/a(\theta)},$$

а функция $\beta(\theta) > 0$ является при каждом θ единственным положительным корнем уравнения

$$\frac{1 - e^{-\beta(\theta)}}{\beta(\theta)} = \frac{a(\theta)}{p(\theta)}, \tag{1.7}$$

$[\cdot]$ – целая часть числа, заключенного в квадратные скобки.

При $x \leqslant 0$

$$\begin{gathered} G(x,\theta) = \frac{(p(\theta) - a(\theta))q(\theta)}{(p(\theta) - q(\theta))a(\theta)} - \frac{(a(\theta) - q(\theta))q(\theta)}{a^2(\theta)} \int\limits_0^{a(\theta)|x|} g^-(t,\theta)\,dt, \\ g^-(t,\theta) = \sum_{k=[t]+1}^{\infty} \pi_k(\theta) - e^{b(\theta)t} \sum_{k=[t]+1}^{\infty} \pi_k(\theta)e^{-b(\theta)k}. \end{gathered} \tag{1.8}$$

Здесь

$$\pi_k(\theta) = \frac{q(\theta)k}{a(\theta)}^{k-1} \frac{e^{-q(\theta)k/a(\theta)}}{k!}, \quad k = 1, 2, \ldots,$$

$$b(\theta) = \frac{\beta(\theta)(1 - q(\theta)}{p(\theta))}.$$

2. Построение асимптотического доверительного интервала

Построение асимптотического доверительного интервала предполагает вычисление оценки максимального правдоподобия. Пусть $X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)}$ — упорядоченные по возрастанию члены выборки из распределения (1.1). Трудность численного нахождения ОМП заключается в исследовании на максимум функции (см. (1.1), (1.2))

$$\prod_{i=1}^{n} f(X_i, \theta) = \begin{cases} \prod_{i=1}^{n} f_2(X_{(i)}, \theta), \quad \theta \leqslant X_{(1)}, \\ f_1(X_{(1)}, \theta) \prod_{i=2}^{n} f_2(X_{(i)}, \theta), \quad X_{(1)} < \theta \leqslant X_{(2)}, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \prod_{i=1}^{k} f_1(X_{(i)}, \theta) \prod_{i=k+1}^{n} f_2(X_{(i)}, \theta), \quad X_{(k)} < \theta \leqslant X_{(k+1)}, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \prod_{i=1}^{n} f_1(X_{(i)}, \theta), \quad X_{(n)} < \theta, \end{cases}$$

которая задается кусочно на $n+1$ интервалах. При большом объеме выборки такой подход приводит к весьма громоздким вычислениям. Объем вычислений можно значительно сократить, если использовать в качестве начальной оценки параметра θ_0 любую состоятельную оценку $\widetilde{\theta_n}$, $\widetilde{\theta_n} \to \theta_0$, методы получения которой можно найти в [1; 7; 10; 11]. По известной состоятельной оценке $\widetilde{\theta_n}$ поиск экстремума функции (1.2) можно вести только на $O(\ln n)$ интервалах $\big(X_{(k)}, X_{(k+1)}\big]$, расположенных слева и справа от точки $\widetilde{\theta_n}$ (см. [3, лемма 1]).

Перейдем к построению асимптотического доверительного интервала для точки θ_0 разрыва плотности (1.1) по найденной ОМП θ_n^*.

Теорема. *Пусть* $\Theta \subset [d_1, d_2] \subset \mathbb{R}$, *а функции* $p(\theta)$ *и* $q(\theta)$ *непрерывны на отрезке* $[d_1, d_2]$ (*см. условие* III *из* [2, с. 324]) .

Для произвольных положительных чисел ε_1, ε_2, *таких что* $\varepsilon_1 + \varepsilon_2 < 1$, *определим квантили* $\lambda_{1-\varepsilon_1}^{(n)}$, $\lambda_{\varepsilon_2}^{(n)} > 0$ *функции распределения* $G(x, \theta_n^*)$ *из уравнений*

$$\begin{aligned} &G(\lambda_{1-\varepsilon_1}^{(n)}, \theta_n^*) = 1 - \varepsilon_1, \\ &G(-\lambda_{\varepsilon_2}^{(n)}, \theta_n^*) = \varepsilon_2. \end{aligned} \tag{2.1}$$

Тогда интервал $[\theta_n^* - \lambda_{1-\varepsilon_1}^{(n)}/n, \theta_n^* + \lambda_{\varepsilon_2}^{(n)}/n]$ *является асимптотическим доверительным интервалом уровня* $1 - \varepsilon_1 - \varepsilon_2$:

$$\lim_{n\to\infty} \mathbf{P}\left(-\lambda_{\varepsilon_2}^{(\mathbf{n})} \leqslant \mathbf{n}(\theta_{\mathbf{n}}^* - \theta_{\mathbf{0}}) \leqslant \lambda_{\mathbf{1}-\varepsilon_1}^{(\mathbf{n})}\right) = \mathbf{1} - \varepsilon_{\mathbf{1}} - \varepsilon_{\mathbf{2}}. \tag{2.2}$$

Д о к а з а т е л ь с т в о. Из равномерной оценки, доказанной в [4, с. 896], вытекает сходимость $(n \to \infty)$

$$\mathbf{P}\left(-\lambda_{\varepsilon_2}^{(\mathbf{n})} \leqslant \mathbf{n}(\theta_{\mathbf{n}}^* - \theta_{\mathbf{0}}) \leqslant \lambda_{\mathbf{1}-\varepsilon_1}^{(\mathbf{n})}\right) - \left(\mathbf{G}(\lambda_{\mathbf{1}-\varepsilon_1}^{(\mathbf{n})}, \theta_{\mathbf{0}}) - \mathbf{G}(-\lambda_{\varepsilon_2}^{(\mathbf{n})}, \theta_{\mathbf{0}})\right) \to \mathbf{0}.$$

Поэтому для доказательства утверждения (2.2) достаточно установить предельные соотношения

$$\begin{aligned} &G(\lambda_{1-\varepsilon_1}^{(n)}, \theta_0) \to 1 - \varepsilon_1, \\ &G(-\lambda_{\varepsilon_2}^{(n)}, \theta_0) \to \varepsilon_2. \end{aligned} \tag{2.3}$$

Запишем очевидные представления, вытекающие из (2.1):

$$G(\lambda^{(n)}_{1-\varepsilon_1},\theta_0)=G(\lambda^{(n)}_{1-\varepsilon_1},\theta^*_n)+\Big(G(\lambda^{(n)}_{1-\varepsilon_1},\theta_0)-G(\lambda^{(n)}_{1-\varepsilon_1},\theta^*_n)\Big)$$
$$=1-\varepsilon_1+\Big(G(\lambda^{(n)}_{1-\varepsilon_1},\theta_0)-G(\lambda^{(n)}_{1-\varepsilon_1},\theta^*_n)\Big), \quad (2.4)$$

$$G(-\lambda^{(n)}_{\varepsilon_2},\theta_0)=\varepsilon_2+\Big(G(-\lambda^{(n)}_{\varepsilon_2},\theta_0)-G(-\lambda^{(n)}_{\varepsilon_2},\theta^*_n)\Big). \quad (2.5)$$

Если предположить равномерную непрерывность семейства функций распределения $G(x,\theta)$ на множестве $(x,\theta)\in(-\infty,\infty)\times[d_1,d_2]$, то в силу состоятельности ОМП ($\theta^*_n\to\theta_0$) выражения в скобках в (2.4), (2.5) будут стремиться к нулю, что приведет нас к соотношениям (2.3) и доказательству теоремы.

Итак, сосредоточимся на доказательстве указанной выше равномерной непрерывности. Из непрерывности функций $p(\theta)$ и $q(\theta)$ на компакте $[d_1,d_2]$, условий (1.5) и определении функций (1.4) и (1.7) следует существование не зависящих от θ констант $\delta_1,\dots,\delta_5>0$ таких, что для всех $\theta\in\Theta\subset[d_1,d_2]$ будут справедливы ограничения

$$\begin{aligned}&\frac{p(\theta)}{a(\theta)}\geqslant\delta_1>1>\delta_2\geqslant\frac{q(\theta)}{a(\theta)},\\&0<\delta_3\leqslant a(\theta)\leqslant\delta_4,\quad \beta(\theta)\leqslant\delta_5.\end{aligned} \quad (2.6)$$

В [6, теорема 2] для функции $g^+(t,\theta)$ из (1.6) получена оценка

$$g^+(t,\theta)\leqslant\beta(\theta)e^{-\Lambda(p(\theta)/a(\theta))a(\theta)t},\quad t\geqslant0,\quad \theta\in[d_1,d_2],$$

где функция $\Lambda(z)=z-1-\ln z$ положительна и возрастает при $z>1$. Следовательно, $\Lambda(p(\theta)/a(\theta))\geqslant\Lambda(\delta_1)>0$, $\theta\in[d_1,d_2]$, что вместе с другими ограничениями (2.6) приводит к равномерной по θ оценке

$$g^+(t,\theta)\leqslant C_1e^{-c_1t},\quad t\geqslant0,\quad \theta\in[d_1,d_2], \quad (2.7)$$

с константами $C_1=\delta_5$ и $c_1=\delta_3\Lambda(\delta_1)$.

Для функции $g^-(t,\theta)$ из (1.8) также воспользуемся соответствующей оценкой из [6, теорема 2]

$$g^-(t,\theta)\leqslant\beta(\theta)a(\theta)e^{-\Lambda(q(\theta)/a(\theta))a(\theta)|t|},\quad t\geqslant0,\quad \theta\in[d_1,d_2].$$

Здесь $q(\theta)/a(\theta)<1$, а так как функция $\Lambda(z)$ убывает при $0<z<1$, то из (2.6) выводим неравенство $\Lambda(q(\theta)/a(\theta))\geqslant\Lambda(\delta_2)>0$, $\theta\in[d_1,d_2]$ и, следовательно,

$$g^-(t,\theta)\leqslant C_2e^{-c_2t},\quad t\geqslant0,\quad \theta\in[d_1,d_2] \quad (2.8)$$

с константами $C_2=\delta_4\delta_5$, $c_2=\delta_3\Lambda(\delta_2)$.

Экспоненциальные оценки (2.7), (2.8) обеспечивают равномерную по θ сходимость интегралов $\displaystyle\int_0^\infty g^\pm(t,\theta)\,dt$, что влечет непрерывность семейства функций распределений $G(x,\theta)$, определенных формулами (1.6), (1.8), на любом компактном множестве $(x,\theta)\in[-N,N]\times[d_1,d_2]$. Следовательно, $G(x,\theta)$ будет здесь и равномерно непрерывна. Ввиду того что

$$\int\limits_N^\infty g^\pm(t,\theta)\,dt\leqslant Ce^{-cN},$$

равномерная непрерывность $G(x,\theta)$ распространяется и на множество $(x,\theta)\in(-\infty,\infty)\times[d_1,d_2]$. □

ОМП и доверительный интервал (ДИ) для параметра с истинным значением $\theta_0 = 0.45$

Обьем выборки	ОМП	Уровень ДИ	
		0.95	0.80
50	0.4587	[0.4262; 0.4698]	[0.4314; 0.4643]
100	0.4568	[0.4336; 0.4652]	[0.4425; 0.4611]
300	0.4532	[0.4428; 0.4622]	[0.4465; 0.4593]
500	0.4491	[0.4476; 0.4585]	[0.4481; 0.4573]

В заключение приведем пример вычисления построения асимптотического доверительного интервала для неизвестного параметра θ треугольной плотности вида

$$f(x,\theta) = \begin{cases} \dfrac{x}{2\theta^2}, & 0 \leqslant x \leqslant \theta, \\ \\ \dfrac{3(1-x)}{2(1-\theta)^2}, & \theta < x \leqslant 1, \end{cases}$$

с истинным значением $\theta_0 = 0.45$.

Для нахождения оценки максимального правдоподобия и построения асимптотического доверительного интервала разработан комплекс программ на языке MATLAB [9]. Особенностью разработанного комплекса является эффективный алгоритм нахождения ОМП, которая вычисляется по состоятельной оценке параметра θ_0, что делает алгоритм ее нахождение менее трудозатратным.

Результаты вычислений примера сведены в таблицу выше.

СПИСОК ЛИТЕРАТУРЫ

1. **Боровков А.А.** Оценки момента разладки по большим выборкам при неизвестных распределениях // Теория вероятностей и ее применения. 2008. Т. 53. С. 437–457.
2. **Ибрагимов И.А., Хасьминский Р.З.** Ассимптотическая теория оценивания. М.: Наука, 1979. 528 с.
3. **Мосягин В.Е.** Асимптотическое представление для процесса отношения правдоподобия в случае разрывной плотности // Сиб. мат. журн. 1994. Т. 35, № 2. С. 416–423.
4. **Мосягин В.Е.** Оценка скорости сходимости распределений нормированных оценок максимального правдоподобия в случае разрывной плотности // Сиб. мат. журн. 1996. Т. 37, №4, С. 895–903.
5. **Мосягин В.Е., Швемлер Н.А.** Распределение момента максимума разности двух пуассоновских процессов с отрицательным линейным сносом // Сиб. электрон. мат. изв. 2016. Т. 13. С. 1229–1248. doi: 10.17377/semi.2016.13.096 .
6. **Мосягин В.Е., Швемлер Н.А.** Локальные свойства предельного распределения статистической оценки точки разрыва плотности // Сиб. электрон. мат. изв. 2017. Т. 14. С. 1307–1316. doi: 10.17377/semi.2017.14.111 .
7. **Мосягин В.Е., Швемлер Н.А.** Метод усредненного минимума для нахождения состоятельной оценки точки разрыва плотности распределения // Ломоносовские чтения на Алтае: фундаментальные проблемы науки и образования: тез. докл. Междунар. конф. / АлтГУ. Барнаул, 2017. С. 453–455.
8. **Скороход А.В.** Случайные процессы с независимыми приращениями. М.: Наука, 1964. 280 с.
9. **Швемлер Н.А., Мосягин В.Е.** Программа для оценки неизвестного параметра точки разрыва плотности распределения // Свид. № 2017660490 от 22.09.2017 Федеральной службы по интеллектуальной собственности РОСПАТЕНТ.
10. **Borovkov A.A., Linke Yu. Yu.** Change–point problem for large samples and incomplete information on distributions // Math. Methods Statist. 2005. Vol. 14, no. 4. P. 404–430.

11. **Hawkins D.L.** A simple least squares method for estimating a change in mean // Comm. Statist. Simulation. 1986. Vol. 15. P. 655–679.

Поступила 31.03.2018

Мосягин Вячеслав Евгеньевич
канд. физ.-мат. наук, доцент
Тюменский государственный университет,
г. Тюмень
e-mail: vmosyagin@mail.ru

Швемлер Наталья Александровна
ассистент
Тюменский государственный университет,
г. Тюмень
e-mail: shvemler.natalya@mail.ru

REFERENCES

1. Borovkov A.A. Large sample change-point estimation when distributions are unknown. *Theory Probability Appl.*, 2009, vol. 53, no. 3, pp. 402–418. doi: 10.1137/S0040585X97983729 .
2. Ibragimov I.A., Khas'minski R.Z. *Statistical estimation. Asymptotic theory.* N Y: Springer-Verlag, 1981, 403 p. doi: 10.1007/978-1-4899-0027-2 . Original Russian text published in Ibragimov I.A., Khas'minskii R.Z. *Asimptoticheskaya teoriya otsenivaniya.* Moscow, Nauka Publ., 1979, 528 p.
3. Mosyagin V.E. Asymptotic representation of the process of the likelihood ratio in the case of a discontinuous density. *Sib. Math. J.*, 1994, vol. 35, no. 2, pp. 375–382. doi: 10.1007/BF02104785 .
4. Mosyagin V.E. Estimation of the convergence rate for the distributions of normalized maximum likelihood estimators in the case of a discontinuous density. *Sib. Math. J.*, 1996, vol. 37, no. 4, pp. 788–796. doi: 10.1007/BF02104670 .
5. Mosyagin V.E., Shvemler N.A. Distribution of the time of attaining the maximum for the difference of the two Poisson's processes with negative linear drift. *Sib. Electron. Math. Reports*, 2016, vol. 13, pp. 1229–1248 (in Russian). doi: 10.17377/semi.2016.13.096 .
6. Mosyagin V.E., Shvemler N.A. Local properties of the limiting distribution of the statistical estimator for jump point of a density. *Sib. Electron. Math. Reports*, 2017, vol. 14, pp. 1307–1316 (in Russian). doi: 10.17377/semi.2017.14.111 .
7. Mosyagin V.E., Shvemler N.A. The average minimum method for construction of consistent estimates for jump point of a probability density. *Abstr. Internet. Conf. "Lomonosovskie chteniya na Altae: fundamental'nye problemy nauki i obrazovaniya".* [*Lomonosov's Readings in the Altai: Fundamental Problems of Science and Education*], Barnaul, 2017, pp. 453–455 (in Russian).
8. Skorokhod A.V. *Stochastic processes with independent increments.* Dordrecht: Kluwer Acad. Publ. Group, 1991, 279 p. ISBN: 0-7923-0340-7 . Original Russian text (1st ed.) published in Skorokhod A.V., *Sluchainye protsessy s nezavisimymi prirashcheniyami*, Moscow, Nauka Publ., 1964, 280 p.
9. Shvemler N.A., Mosyagin V.E. Program for estimating the unknown parameter of the point of discontinuity in the probability density. *Registered Rospatent Certificate* No. 2017660490 dated 22.09.2017 (in Russian).
10. Borovkov A.A., Linke Yu. Yu. Change–point problem for large samples and incomplete information on distributions. *Math. Methods Statist.*, 2005, vol. 14, no. 4, pp. 404–430.
11. Hawkins D.L. A simple least squares method for estimating a change in mean *Comm. Statist. Simulation*, 1986, vol. 15, pp. 655–679.

The paper was received by the Editorial Office on March 31, 2018.

Vyacheslav Evgen'evich Mosyagin, Cand. Sci. (Phys.-Math.), Tyumen State University, Tyumen, 625003 Russia, e-mail: vmosyagin@mail.ru .

Natal'ya Aleksandrovna Shvemler, Tyumen State University, Tyumen, 625003 Russia,
e-mail: shvemler.natalya@mail.ru .