

ГАРАНТИРУЮЩЕЕ ДОВЕРИТЕЛЬНОЕ ОЦЕНИВАНИЕ ПАРАМЕТРОВ РЯДА РАСПРЕДЕЛЕНИЙ

Анна Владимировна КАН родилась в городе Москве. Инженер 1-й категории ФГУП ГосНИИАС. Основные научные интересы — в области системного анализа и имитационного моделирования. Автор 12 научных работ.

Anna V. KAN, was born in Moscow. She is a First Class Engineer at the State Research Institute of Aviation Systems (GosNIAS). Her research interests are in systems analysis and simulation. She has published 12 technical papers.

Юрий Сергеевич КАН родился в 1960 г. в городе Истре Московской области. Профессор МАИ. Доктор физ.-мат. наук, доцент. Основные научные интересы — в области стохастического программирования. Автор более 70 научных работ.

Yury S. KAN, D.Sci, was born in 1960, in the Moscow Region. He is a Professor at the MAI. His research interests are in stochastic programming. He has published over 70 technical papers.

Исследуется проблема доверительного оценивания параметров по выборке ограниченного объема в случае, когда для построения асимптотических доверительных интервалов используется центральная предельная теорема. Предлагается с помощью локального неравенства Берри—Эссеена определять поправку к указанным асимптотическим интервалам, превращающую их в доверительные интервалы в обычном смысле для конечной выборки. Такая поправка явно вычисляется в задаче оценивания неизвестной вероятности и в задаче оценивания параметра экспоненциального распределения.

Введение

Задача доверительного оценивания параметров распределений является классической задачей математической статистики [1, 2]. Ее прикладная значимость обусловлена тем, что, во-первых, доверительные интервалы используются при проверке статистических гипотез, например при отладке и тестировании программных комплексов имитационного моделирования, включающих потоки случайных событий [3]. Во-вторых, интервальные оценки решают проблему определения погрешности точечных оценок.

В общем случае задача доверительного оценивания, как правило, не решается точно для выборки конечного объема. Одним из исключений является случай, когда речь идет об оценке параметров нормального распределения. Соответствующие доверительные интервалы для математического ожидания и дисперсии нормальной выборки вошли во многие учебные курсы по математической статистике, например [1, 2]. Поэтому рассматриваемая задача часто решается приближенно. Наиболее известным способом приближенного решения является построение асимптотических доверительных интервалов [1], обеспечивающих выполнение требуемого вероятностного неравенства в асимптотике, когда объем выборки стремится к бесконечности. Обычно такой прием применяется, когда оце-

ниваемый параметр является функцией от математического ожидания и/или дисперсии. В последнем случае для построения асимптотического доверительного интервала обычно используется центральная предельная теорема (ЦПТ) и указанные выше доверительные интервалы для случая нормального распределения. Но на этом пути возникает известная методологическая проблема об использовании асимптотических доверительных интервалов в конкретных задачах, где объем выборки конечен.

В данной статье указанная методологическая проблема решается путем некоторого расширения асимптотических доверительных интервалов для двух частных случаев. В первом случае рассматривается задача оценки неизвестной вероятности (с формальной точки зрения оценивается неизвестный параметр распределения Бернулли), во втором случае — задача оценки параметра экспоненциального распределения, тесно связанная с прикладными задачами моделирования потоков случайных событий с требуемыми свойствами [3]. В обеих задачах строятся доверительные интервалы, называемые ниже гарантирующими. Смысл гарантирующих доверительных интервалов заключается в том, что выполнение требуемого вероятностного неравенства гарантируется для выборки, объем которой превышает некоторую конечную величину.

Необходимо отметить, что в обоих указанных выше частных случаях известны [4] точные доверительные интервалы для выборок конечного объема. Границы точного доверительного интервала для неизвестной вероятности находятся путем решения уравнений Клоппера—Пирсона для определения квантилей биномиального распределения $Bi(n, p)$, где n — объем выборки. Для больших n точное решение этих уравнений затруднительно, поэтому обычно используется их нормальная аппроксимация, основанная на ЦПТ. Границы доверительного интервала для параметра экспоненциального распределения определяются квантилями распределения χ^2 с $2n$ степенями свободы. При больших n точное определение этих квантилей затруднительно, поэтому их определение обычно также осуществляется путем нормальной аппроксимации распределения χ^2 . Предлагаемые ниже гарантирующие доверительные интервалы учитывают погрешность нормальной аппроксимации.

1. Основные определения и вспомогательные результаты

Пусть $Z_n = (X_1, \dots, X_n)$ — случайная выборка объема n из распределения $F(x; \theta)$, т.е. случайный вектор с независимыми компонентами, распределение каждой из которых определяется функцией

$$F(x; \theta) = P_\theta(X_1 \leq x),$$

где P_θ — вероятность, зависящая от неизвестного скалярного параметра θ .

Определение 1 [2]. *Доверительным интервалом уровня β для неизвестного параметра θ называется интервал $[a(Z_n), b(Z_n)]$ со случайными концами $a(Z_n)$ и $b(Z_n)$, обеспечивающий выполнение вероятностного неравенства*

$$P_\theta(a(Z_n) \leq \theta \leq b(Z_n)) \geq \beta. \quad (1)$$

Величина β называется *доверительной вероятностью*.

Определение 2. *Асимптотическим доверительным интервалом уровня β для неизвестного параметра θ называется интервал $[a(Z_n), b(Z_n)]$, если*

$$\inf \lim_{n \rightarrow \infty} P_\theta(a(Z_n) \leq \theta \leq b(Z_n)) \geq \beta,$$

где $\inf \lim$ — нижний предел.

Определение 3. *Гарантирующим доверительным интервалом уровня β для неизвестного параметра θ называется интервал $[a(Z_n), b(Z_n)]$, если найдется натуральное число n_0 , не зависящее от θ и такое, что неравенство (1) выполнено для всех $n > n_0$.*

Пусть X_1 имеет конечные математическое ожидание $M[X_1] = m$ и дисперсию σ^2 . Рассмотрим нормированную сумму

$$Y_n = \frac{X_1 + \dots + X_n - n \cdot m}{\sigma \sqrt{n}}. \quad (2)$$

В соответствии с ЦПТ

$$F_n(y) = P(Y_n \leq y) \rightarrow \Phi(y) \quad (3)$$

при $n \rightarrow \infty$, где

$$\Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^y \exp\{-x^2/2\} dx$$

— функция распределения стандартного нормального распределения $N(0,1)$. Для дальнейшего принципиальное значение будет иметь следующий результат [5, с. 82], на который ниже будем ссылаться как на локальную теорему Берри—Эссеена.

Теорема 1. *Если X_1, \dots, X_n, \dots — последовательность независимых, одинаково распределенных случайных величин с математическим ожиданием m и дисперсией σ^2 , то*

$$|F_n(y) - \Phi(y)| \leq A \frac{M[|X_1 - m|^3]}{\sigma^3 \sqrt{n} (1 + |y|^3)},$$

где $A = 0,7655$.

Пусть X — случайная величина с функцией распределения $F(x) = P(X \leq x)$. Обозначим посредством $[X]_\alpha$ α -квантиль распределения этой случайной величины, т.е.

$$[X]_\alpha = \min\{x: F(x) \geq \alpha\}.$$

Теорема 2. *Если $X \sim N(0,1)$, то для любого $\lambda \in (0,1)$ $[X]_{\alpha+\lambda(1-\alpha)} - [X]_\alpha \rightarrow 0$ при $\alpha \rightarrow 1$.*

Доказательство. Так как $X \sim N(0,1)$, то функция распределения $F(x)$ непрерывна и стро-

го возрастает. Поэтому для любого $\alpha \in (0, 1)$ α -квантиль $[X]_\alpha$ является корнем уравнения $F(x) = \alpha$. Далее, функция $F(x)$ вогнута в области $x > 0$, в которой лежат все α -квантили для $\alpha > 1/2$. Поэтому для всех $x > 0$ график функции $y = F(x)$ находится ниже касательной $y = \hat{F}(x)$ к этому графику, построенной в точке $[X]_{\alpha+\lambda(1-\alpha)}$.

Поэтому $[X]_\alpha > \hat{x}_\alpha$, где \hat{x}_α — корень уравнения $\hat{F}(x) = \alpha$. Уравнение указанной касательной имеет вид

$$\hat{F}(x) = \alpha + \lambda(1 - \alpha) + \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{[X]_{\alpha+\lambda(1-\alpha)}^2}{2}\right\} (x - [X]_{\alpha+\lambda(1-\alpha)}).$$

Приравняв это выражение к α и решая полученное уравнение относительно x , находим

$$\hat{x}_\alpha = [X]_{\alpha+\lambda(1-\alpha)} - \lambda(1 - \alpha)\sqrt{2\pi} \exp\left\{-\frac{[X]_{\alpha+\lambda(1-\alpha)}^2}{2}\right\}.$$

Из $[X]_\alpha > \hat{x}_\alpha$ следует, что

$$[X]_{\alpha+\lambda(1-\alpha)} - [X]_\alpha \leq [X]_{\alpha+\lambda(1-\alpha)} - \hat{x}_\alpha = \lambda\sqrt{2\pi}f(\alpha),$$

где $f(\alpha) = \frac{1 - \alpha}{\exp\left\{-\frac{[X]_{\alpha+\lambda(1-\alpha)}^2}{2}\right\}}$; α -квантиль $[X]_\alpha$,

как функция от α , монотонно не убывает. Следовательно,

$$0 \leq [X]_{\alpha+\lambda(1-\alpha)} - [X]_\alpha.$$

Поэтому, если показать, что

$$L = \lim_{\alpha \rightarrow 1} f(\alpha) = 0,$$

то теорема будет доказана. Функция $f(\alpha)$ имеет структуру дроби, в которой числитель и знаменатель стремятся к 0 при $\alpha \rightarrow 1$. Применяя правило Лопиталю, получаем

$$L = \lim_{\alpha \rightarrow 1} \frac{1}{[X]_{\alpha+\lambda(1-\alpha)} \exp\left\{-\frac{[X]_{\alpha+\lambda(1-\alpha)}^2}{2}\right\} \frac{\partial}{\partial \alpha} [X]_{\alpha+\lambda(1-\alpha)}}.$$

Так как $[X]_{\alpha+\lambda(1-\alpha)}$ является корнем уравнения $F(x) = \alpha + \lambda(1 - \alpha)$, то, применяя теорему о производной неявной функции, находим

$$\frac{\partial}{\partial \alpha} [X]_{\alpha+\lambda(1-\alpha)} = (1 - \lambda)\sqrt{2\pi} \exp\left\{-\frac{[X]_{\alpha+\lambda(1-\alpha)}^2}{2}\right\}.$$

Поэтому

$$L = \frac{1}{(1 - \lambda)\sqrt{2\pi}} \lim_{\alpha \rightarrow 1} \frac{1}{[X]_{\alpha+\lambda(1-\alpha)}} = 0$$

в силу того, что $[X]_\alpha \rightarrow \infty$ при $\alpha \rightarrow 1$.

Теорема доказана.

2. Гарантирующий доверительный интервал для параметра экспоненциального распределения

В данном разделе решается задача доверительного оценивания параметра θ экспоненциального распределения $E(\theta)$ по выборке X_1, \dots, X_n . Рассмотрим точечную оценку искомого параметра в виде

$$\hat{\theta}_n = \frac{n}{X_1 + \dots + X_n}. \quad (4)$$

Для построения гарантирующего доверительного интервала для θ достаточно при заданной доверительной вероятности β определить величины ε и n_0 такие, чтобы для всех $n > n_0$ было выполнено вероятностное неравенство

$$P_\theta \left(\left| \frac{\theta}{\hat{\theta}_n} - 1 \right| \leq \varepsilon \right) \geq \beta. \quad (5)$$

Тогда искомым интервал можно определить как

$$\left[\hat{\theta}_n(1 - \varepsilon), \hat{\theta}_n(1 + \varepsilon) \right], \text{ т.е. } a(Z_n) = \hat{\theta}_n(1 - \varepsilon),$$

$$b(Z_n) = \hat{\theta}_n(1 + \varepsilon).$$

Так как для экспоненциального распределения справедливо $m = \sigma = \theta^{-1}$, то с учетом (2) и (4) неравенство (5) равносильно следующему:

$$P_{\theta}(|Y_n| \leq \varepsilon\sqrt{n}) \geq \beta. \quad (6)$$

Это неравенство выполнено, если справедливо более сильное неравенство

$$P_{\theta}(-\varepsilon\sqrt{n} < Y_n \leq \varepsilon\sqrt{n}) \geq \beta. \quad (7)$$

Левая часть неравенства (7) равна $F_n(\varepsilon\sqrt{n}) - F_n(-\varepsilon\sqrt{n})$, где функция $F_n(y)$ распределения нормированной суммы определена формулой (3). Поэтому на основании локальной теоремы Берри—Эссеена можно утверждать, что неравенство (7) справедливо, если справедливо более сильное неравенство

$$\Phi(\varepsilon\sqrt{n}) - \Phi(-\varepsilon\sqrt{n}) - 2\delta_n \geq \beta, \quad (8)$$

где

$$\delta_n = A \frac{M_{\theta}[|X_1 - m|^3]}{\theta^{-3}\sqrt{n}(1 + |\varepsilon\sqrt{n}|^3)}.$$

Элементарные (хотя и довольно громоздкие) выкладки приводят к следующему выражению:

$$M_{\theta}[|X_1 - m|^3] = a\theta^{-3},$$

где

$$a = \frac{12}{e} - 2 \approx 2,415.$$

Поэтому величина

$$\delta_n = \frac{Aa}{\sqrt{n}(1 + |\varepsilon\sqrt{n}|^3)}$$

не зависит от θ . Эта величина может быть оценена сверху следующим образом:

$$\delta_n \leq \frac{Aa}{\varepsilon^3 n^2} = \delta_n^*.$$

Отметим, что $\Phi(x) - \Phi(-x) = 2\Phi_0(x)$, где функция

Лапласа $\Phi_0(x)$ определена выражением

$$\Phi_0(x) = \frac{1}{\sqrt{2\pi}} \int_0^x \exp\{-t^2/2\} dt.$$

Поэтому неравенство (8) выполнено, если выполнено неравенство

$$2\Phi_0(\varepsilon\sqrt{n}) - 2\delta_n^* \geq \beta. \quad (9)$$

Следует отметить, что если бы погрешность δ_n изначально не учитывалась, то вместо неравенства (9) было бы получено неравенство

$$2\Phi_0(\varepsilon\sqrt{n}) \geq \beta,$$

приводящее к построению асимптотического доверительного интервала

$$\left[\hat{\theta}_n \left(1 - \frac{x_{\beta/2}}{\sqrt{n}} \right), \hat{\theta}_n \left(1 + \frac{x_{\beta/2}}{\sqrt{n}} \right) \right], \quad (10)$$

где $x_{\beta/2}$ — корень уравнения $\Phi_0(x) = \beta/2$.

Обратимся к исследованию неравенства (9). Будем рассматривать только такие ε и n , для которых справедлива оценка

$$\delta_n^* \leq \gamma, \quad (11)$$

где γ — константа из интервала $(0, (1-\beta)/2)$. Тогда неравенство (9) выполнено, если

$$2\Phi_0(\varepsilon\sqrt{n}) \geq \beta + 2\gamma.$$

Последнее справедливо, если

$$\varepsilon = \frac{x_{(\beta/2)+\gamma}}{\sqrt{n}}. \quad (12)$$

Это выражение решает поставленную задачу. Величина n_0 , указанная в определении 3, находится из условия (11), если в выражение для δ_n^* подставить ε в соответствии с формулой (12). Имеем

$$\frac{Aa}{x_{(\beta/2)+\gamma}^3 \sqrt{n}} \leq \gamma,$$

откуда

$$n_0 = \left[\frac{A^2 a^2}{\gamma^2 x_{(\beta/2)+\gamma}^6} \right] + 1. \quad (13)$$

Очевидно, что знаменатель дроби в (13) монотонно возрастает по γ . Это приводит к увеличению n_0 при уменьшении γ . С другой стороны, уменьшение γ приводит согласно (12) к уменьшению длины гарантирующего доверительного интервала. При этом оказывается, что если γ устремить к нулю, то гарантирующий доверительный интервал превращается в асимптотический, но при этом величина n_0 уходит в бесконечность.

Если в качестве γ выбрать середину интервала $(0, (1-\beta)/2)$, т.е. положить $\gamma = (1-\beta)/4$, то выражения (12) и (13) приобретут вид:

$$\varepsilon = \frac{x_{(1+\beta)/4}}{\sqrt{n}};$$

$$n_0 = \left[\frac{16A^2 a^2}{(1-\beta)^2 x_{(1+\beta)/4}^6} \right] + 1. \quad (14)$$

Некоторые значения n_0 , вычисленные по формуле (14) для различных значений β , представлены в табл. 1. В табл. 2 для различных значений β представлена величина

$$\Delta_\beta = \frac{x_{(1+\beta)/4} - x_{\beta/2}}{x_{\beta/2}} \cdot 100\%,$$

характеризующая относительную разность между длинами гарантирующего и асимптотического доверительного интервалов. Видно, что эта величина уменьшается с ростом β . Более того, она стремится к 0, хотя и медленно, в силу доказанной выше теоремы 2.

Таблица 1

β	0,9	0,95	0,99
n_0	97	174	1111

Таблица 2

β	0,9	0,95	0,99
Δ_β	19	14	9

3. Гарантирующий доверительный интервал для вероятности

Рассмотрим задачу построения гарантирующего доверительного интервала для вероятности $\theta = P(A)$ случайного события A по наблюдениям этого события в схеме Бернулли из n экспериментов G_1, \dots, G_n . Таким образом, формально речь идет о доверительном оценивании параметра θ распределения Бернулли $Bi(1, \theta)$ по выборке Z_n , элементами которой являются бернуллиевы случайные величины X_i , равные 1, если A происходит в i -м эксперименте G_i , и равные 0 в противном случае.

В качестве точечной оценки вероятности рассмотрим частоту успехов

$$\hat{\theta}_n = \frac{X_1 + \dots + X_n}{n}.$$

Известно [2], что частота имеет математическое ожидание θ и дисперсию

$$\sigma_n^2 = \frac{\theta(1-\theta)}{n}.$$

Поэтому

$$Y_n = \frac{\hat{\theta}_n - \theta}{\sigma_n} = \frac{X_1 + \dots + X_n - n\theta}{\sqrt{n\theta(1-\theta)}} \quad (15)$$

является нормированной суммой вида (2) с $m = \theta$ и $\sigma = \sqrt{\theta(1-\theta)}$.

Искомый гарантирующий доверительный интервал будем искать в виде $[\hat{\theta}_n - \varepsilon, \hat{\theta}_n + \varepsilon]$, т.е. определим из условия

$$P_\theta \left(\left| \hat{\theta}_n - \theta \right| \leq \varepsilon \right) \geq \beta,$$

которое с учетом (15) равносильно вероятностному неравенству

$$P_\theta \left(\left| Y_n \right| \leq \frac{\varepsilon\sqrt{n}}{\sigma} \right) \geq \beta. \quad (16)$$

Это неравенство верно, если справедливо более сильное неравенство

$$P_\theta \left(-\frac{\varepsilon\sqrt{n}}{\sigma} < Y_n \leq \frac{\varepsilon\sqrt{n}}{\sigma} \right) \geq \beta. \quad (17)$$

Используя ЦПТ и локальную теорему Берри—Эссеена, заключаем, что (17) справедливо при выполнении более сильного неравенства

$$2\Phi_0 \left(\frac{\varepsilon\sqrt{n}}{\sigma} \right) - 2\delta_n \geq \beta, \quad (18)$$

где

$$\delta_n = \frac{Ac(\theta)}{\sigma^3\sqrt{n} \left(1 + \left| \frac{\varepsilon\sqrt{n}}{\sigma} \right|^3 \right)}, \quad c(\theta) = M_\theta \left[\left| X_1 - \theta \right|^3 \right].$$

Несложные вычисления приводят к выражению $c(\theta) = \sigma^2 (\theta^2 + (1-\theta)^2)$. Так как θ — вероятность, то $\theta \in [0, 1]$. Поэтому $\sigma^2 \leq 1/4$, $\theta^2 + (1-\theta)^2 \leq 1$. Отсюда вытекает, что $c(\theta) \leq 1/4$ и

$$\delta_n \leq \delta_n^* = \frac{A}{4\varepsilon^3 n^2}. \quad (19)$$

Таким образом, как и в предыдущем разделе, погрешность Берри—Эссеена удалось оценить сверху величиной, не зависящей от θ .

Неравенство (18) выполнено, если выполнено неравенство

$$2\Phi_0\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) - 2\delta_n^* \geq \beta. \quad (20)$$

Потребуем, чтобы

$$\delta_n^* \leq \gamma, \quad (21)$$

где $\gamma \in (0, (1-\beta)/2)$, начиная с некоторого номера n_0 , который будет определен ниже. Усилим с учетом (21) неравенство (20):

$$2\Phi_0\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right) - 2\gamma \geq \beta. \quad (22)$$

Учтем теперь, что оцениваемый параметр θ — вероятность. По этой причине $\theta \in [0, 1]$, откуда $\sigma \leq 1/2$. С учетом этого усилим (22):

$$2\Phi_0(2\varepsilon\sqrt{n}) - 2\gamma \geq \beta,$$

откуда находим, что

$$\varepsilon = \frac{x_{(\beta/2)+\gamma}}{2\sqrt{n}} \quad (23)$$

решает поставленную задачу, где величина $x_{(\beta/2)+\gamma}$ та же, что и в предыдущем разделе. Если в (23) положить формально $\gamma = 0$, то получим асимптотический доверительный интервал.

Для нахождения величины n_0 подставим (23) в (19) и рассмотрим условие (21). В результате получаем, что

$$n \geq \frac{4A^2}{x_{(\beta/2)+\gamma}^6 \gamma^2},$$

откуда

$$n_0 = \left\lceil \frac{4A^2}{x_{(\beta/2)+\gamma}^6 \gamma^2} \right\rceil + 1. \quad (24)$$

Для $\gamma = (1-\beta)/4$ выражения (23) и (24) приобретают вид

$$\varepsilon = \frac{x_{(1+\beta)/4}}{2\sqrt{n}}, \quad (25)$$

$$n_0 = \left\lceil \frac{64A^2}{x_{(1+\beta)/4}^6 (1-\beta)^2} \right\rceil + 1. \quad (26)$$

Табл. 3 аналогична приведенной в предыдущем разделе табл. 1.

Таблица 3

β	0,9	0,95	0,99
n_0	67	119	761

Выводы

Построены гарантирующие доверительные интервалы для неизвестного параметра экспоненциального распределения и для неизвестной вероятности. Предложенная методика может быть легко применена для решения задач построения односторонних (право- и левосторонних) гарантирующих доверительных интервалов для указанных параметров.

Summary

A problem under consideration is parametric estimation based on some sample of finite size in the case where the Central Limit Theorem (CLT) is applied to construct asymptotic confidence intervals. It is suggested to construct guaranteeing confidence intervals taking into account an inaccuracy of the CLT normal approximation. Such intervals are constructed explicitly for some unknown parameter of the exponential distribution and also for some unknown probability.

Библиографический список

1. *Ивченко Г.И., Медведев Ю.И.* Математическая статистика. — М.: Высшая школа, 1992.
2. *Кибзун А.И., Горяинова Е.Р., Наумов А.В.* Теория вероятностей и математическая статистика: Базовый курс с примерами и задачами: Учебное пособие. — 2-е изд., испр. и доп. — М.: Физматлит, 2005.
3. *Егорова В.П., Зубкова И.Ф., Кан А.В., Кухтенко В.И.* Синтез детерминированных и случайных потоков воздушного движения в составе комплекса имитационного моделирования системы ОрВД // Третья Всероссийская научно-практическая конференция ИММОД-2007. Сб. докладов. СПб., 2007. С. 111-116.
4. *Кобзарь А.И.* Прикладная математическая статистика. Для инженеров и научных работников. — М.: Физматлит, 2006.
5. *Королюк В.С., Портенко Н.И., Скороход А.В., Турбин А.Ф.* Справочник по теории вероятностей и математической статистике. — 2-е изд. — М.: Наука, 1985.

Работа выполнена при финансовой поддержке РФФИ, грант № 05-08-17963.

Московский авиационный институт
Статья поступила в редакцию 15.12.2007