

# ОБ ОЦЕНИВАНИИ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЙ ПО ИНТЕРВАЛЬНЫМ НАБЛЮДЕНИЯМ

Б. Ю. ЛЕМЕШКО, С. Н. ПОСТОВАЛОВ

*Новосибирский государственный технический университет, Россия*

e-mail: headrd@nstu.nsk.su, post@sun.nstu.nsk.su

The questions of distributions' parameters interval estimation on interval data are considered. The procedures of construction of interval estimates generated by point  $L$ -estimation,  $M$ -estimation and  $MD$ -estimation are proposed. The numerical example is presented.

## 1. Введение

Интервальную выборку [1] случайной величины  $X$

$$\mathbf{X}_n = \{[a_i, b_i] \mid a_i \leq x_i \leq b_i, a_i \in \mathbb{R}, b_i \in \mathbb{R}, i = 1, \dots, n\}, \quad (1)$$

в которой наблюдения  $x_i$  заданы интервалами  $[a_i, b_i]$ , можно рассматривать как  $n$ -мерный параллелепипед в пространстве  $\mathbb{R}^n$ . Тогда выборку, рассматриваемую в классической статистике

$$\mathbf{X}_n = \{x_1, x_2, \dots, x_n\}, \quad (2)$$

можно интерпретировать как точку, принадлежащую этому параллелепипеду. Частным случаем выборки (1), в которой наблюдались значения  $x_i$ , фиксируемые с точностью до интервала  $[a_i, b_i]$ , так, что  $a_i \leq x_i \leq b_i, i = 1, \dots, n$ , будем называть *точечной* и обозначать  $\mathbf{X}_n \in \mathbf{X}_n$ .

Пусть все наблюдения независимы и подчинены одному и тому же распределению  $F(x, \theta)$ . По интервальной выборке (1) требуется найти оценку параметра

$$\hat{\theta} = \hat{\theta}(\mathbf{X}_n). \quad (3)$$

По точечной выборке возможно определение как точечных, так и интервальных оценок параметров. По интервальной выборке также возможно точечное и интервальное оценивание.

## 2. Точечное оценивание

Если нет никакой априорной информации о распределении ошибок измерений, то известны следующие результаты:

— при оценивании по выборке (1) параметра формы гамма-распределения при близких к нулю значениях  $\varepsilon = \max_i (b_i - a_i) > 0$  дисперсия оценки состоит из двух слагаемых, первое из которых зависит от  $\varepsilon$  и не зависит от объема выборки, а второе стремится к нулю при  $n \rightarrow \infty$  [2];

— при проверке согласия фиксированного распределения с интервальной выборкой (1) по критерию Колмогорова, с точки зрения “крайнего пессимиста”, можно отвергнуть любую простую гипотезу при увеличении количества наблюдений [1].

Эти два факта говорят о том, что при точечном оценивании по интервальной выборке могут нарушаться статистические свойства оценок, такие, например, как несмещенность и эффективность, и теряет смысл бесконечное увеличение объема выборки, поскольку статистическая погрешность оценки оказывается существенно меньше погрешности, определяемой погрешностью фиксации наблюдений.

Интервальный характер выборки определяет границы для любой точечной оценки: интервал, которому она принадлежит и который определяется погрешностью наблюдений и методом оценивания. Любая точечная оценка, полученная по интервальной выборке, определяет лишь точку, принадлежащую этому интервалу, но не содержит информации о самом интервале.

## 3. Интервальное оценивание

Введем понятие нестатистической интервальной оценки.

**Определение.** Пусть имеется некоторая точечная оценка параметра по точечной выборке  $\theta^* = \theta^*(X_n)$ . Интервальной оценкой параметра  $\theta$ , порожденной точечной оценкой  $\theta^*(X_n)$ , по интервальной выборке будем называть интервал  $[\underline{\theta}^*, \bar{\theta}^*]$ , границы которого определяются из соотношений

$$\underline{\theta}^* = \min_{X_n \in \mathbf{X}_n} \theta^*(X_n), \quad \bar{\theta}^* = \max_{X_n \in \mathbf{X}_n} \theta^*(X_n). \quad (4)$$

Приведем несколько простых примеров нахождения интервальной оценки параметра по интервальной выборке.

**Пример 1.** Оценивание параметра сдвига нормального распределения с плотностью  $f(x, \mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2}$ . Оценка максимального правдоподобия параметра  $\mu$  имеет вид

$$\mu^* = \frac{1}{n} \sum_{i=1}^n x_i.$$

Тогда по (4)

$$\underline{\mu}^* = \min_{x_i \in [a_i, b_i]} \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \sum_{i=1}^n a_i,$$

$$\bar{\mu}^* = \max_{x_i \in [a_i, b_i]} \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \sum_{i=1}^n b_i.$$

**Пример 2.** Оценивание параметра масштаба нормального распределения с плотностью  $f(x, \lambda) = \frac{1}{\lambda\sqrt{2\pi}}e^{-\frac{x^2}{2\lambda^2}}$ . Оценка максимального правдоподобия параметра  $\lambda$  имеет вид

$$\lambda^* = \frac{1}{n} \sum_{i=1}^n x_i^2.$$

Тогда по (4)

$$\begin{aligned} \lambda^* &= \min_{x_i \in [a_i, b_i]} \frac{1}{n} \sum_{i=1}^n x_i^2 = \frac{1}{n} \left\{ \sum_{i=1}^n b_i^2 I_{b_i \leq 0} + \sum_{i=1}^n a_i^2 I_{a_i \geq 0} \right\}, \\ \bar{\lambda}^* &= \max_{x_i \in [a_i, b_i]} \frac{1}{n} \sum_{i=1}^n x_i^2 = \\ &= \frac{1}{n} \left\{ \sum_{i=1}^n a_i^2 I_{b_i \leq 0} + \sum_{i=1}^n b_i^2 I_{a_i \geq 0} + \sum_{i=1}^n \max\{a_i^2, b_i^2\} I_{a_i < 0 < b_i} \right\}, \end{aligned}$$

где  $I_A = 1$ , если условие  $A$  выполнено, и  $I_A = 0$  в противном случае.

В этих двух примерах интервальные оценки имеют аналитически простой вид, но в большинстве случаев определение интервальной оценки в явном виде невозможно. Рассмотрим общие подходы к определению интервальных оценок, порожденных классами точечных  $L$ -, - и  $D$ -оценок.

### 3.1. Интервальные $L$ -оценки

$L$ -оценки формируются как линейные комбинации порядковых статистик:

$$\theta^* = \sum_{i=1}^n c_i x_{(i)}, \quad \sum_{i=1}^n c_i = 1, \quad (5)$$

где  $x_{(i)}$  —  $i$ -я порядковая статистика. Чтобы найти интервальную оценку, порожденную оценкой вида (5), нужно определить интервалы, в границах которых могут находиться порядковые статистики.

Воспользуемся полученными в [1] границами для эмпирической функции распределения, построенной по выборке (1):

$$\underline{F}_n(x) \leq F_n(x) \leq \overline{F}_n(x). \quad (6)$$

Каждый член вариационного ряда

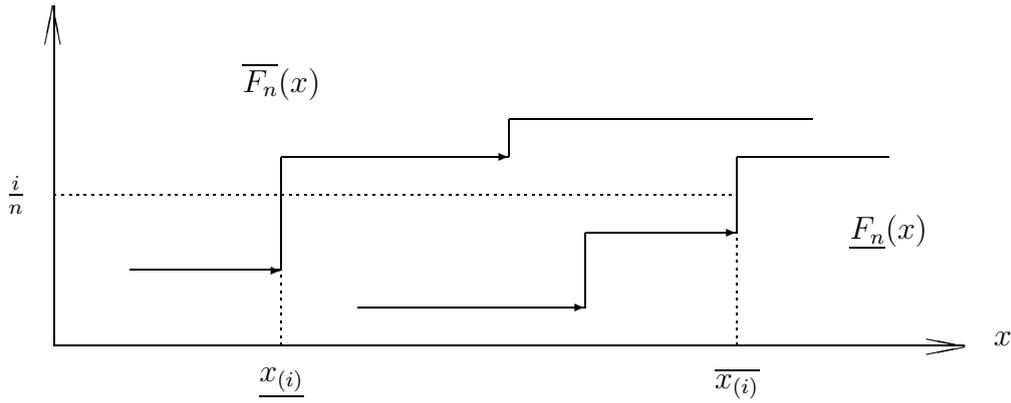
$$x_{(1)} < x_{(2)} < \dots < x_{(n)} \quad (7)$$

известен с точностью до интервала

$$\underline{x}_{(i)} \leq x_{(i)} \leq \overline{x}_{(i)}, \quad (8)$$

где  $\underline{x}_{(i)}$  и  $\overline{x}_{(i)}$  можно определить из (6), так как между вариационным рядом и эмпирической функцией распределения существует взаимно-однозначное соответствие (см. рис. 1) и  $F_n(x_{(i)}) = i/n$ :

$$\underline{x}_{(i)} = \inf\{y \mid y = \overline{F}_n^{-1}(i/n)\}, \quad \overline{x}_{(i)} = \sup\{y \mid y = \underline{F}_n^{-1}(i/n)\}. \quad (9)$$

Рис. 1. Определение границ  $i$ -го члена вариационного ряда.

Подставим (5) в (4) и, учитывая (7)–(9), получим две задачи линейного программирования размерности  $n$ , решение которых даст нам верхнюю и нижнюю границы  $L$ -оценки.

**Пример 3.** К классу  $L$ -оценок относится часто используемая для оценивания параметра положения выборочная медиана:

$$m = \begin{cases} x_{((n+1)/2)}, & n \text{ — нечетное,} \\ \frac{1}{2} (x_{(n/2)} + x_{(n/2+1)}), & n \text{ — четное.} \end{cases}$$

Соответствующая интервальная оценка будет иметь вид

$$\underline{m} = \begin{cases} \underline{x}_{((n+1)/2)}, & n \text{ — нечетное,} \\ \frac{1}{2} (\underline{x}_{(n/2)} + \underline{x}_{(n/2+1)}), & n \text{ — четное.} \end{cases}$$

$$\overline{m} = \begin{cases} \overline{x}_{((n+1)/2)}, & n \text{ — нечетное,} \\ \frac{1}{2} (\overline{x}_{(n/2)} + \overline{x}_{(n/2+1)}), & n \text{ — четное.} \end{cases}$$

### 3.2. Интервальные $M$ -оценки

$M$ -оценки получаются в результате минимизации функционала [3]

$$\theta^*(X_n) = \arg \min_{\theta} M(X_n, \theta) = \arg \min_{\theta} \sum_{i=1}^n \rho(x_i, \theta), \quad (10)$$

где  $\rho(x_i, \theta)$  — функция потерь. Пусть

$$\underline{M}(\theta) = \min_{X_n \in \mathbf{X}_n} M(X_n, \theta) = \sum_{i=1}^n \min_{x_i \in [a_i, b_i]} \rho(x_i, \theta), \quad (11)$$

$$\overline{M}(\theta) = \max_{X_n \in \mathbf{X}_n} M(X_n, \theta) = \sum_{i=1}^n \max_{x_i \in [a_i, b_i]} \rho(x_i, \theta) - \quad (12)$$

соответственно верхняя и нижняя границы  $M(X_n, \theta)$  по всем возможным точечным выборкам  $X_n \in \mathbf{X}_n$ . Тогда

$$\underline{M}(\theta) \leq M(X_n, \theta) \leq \overline{M}(\theta), \quad \forall X_n \in \mathbf{X}_n, \quad (13)$$

и, следовательно,

$$\min_{\theta} \underline{M}(\theta) \leq \min_{\theta} M(X_n, \theta) \leq \min_{\theta} \overline{M}(\theta), \quad \forall X_n \in \mathbf{X}_n. \quad (14)$$

Обозначим

$$\overline{M} = \min_{\theta} \overline{M}(\theta) \quad (15)$$

и

$$\underline{M} = \min_{\theta} \underline{M}(\theta) \quad (16)$$

максимальное и минимальное значение  $M(\theta)$ .

Зафиксируем какую-нибудь точечную выборку  $X_n \in \mathbf{X}_n$ . Пусть минимум функционала  $M(X_n, \theta)$  достигается при каком-то значении  $\theta^*$ :

$$M(X_n, \theta^*) = \min_{\theta} M(X_n, \theta). \quad (17)$$

Тогда из (13) и (14) следует, что

$$\underline{M}(\theta^*) \leq M(X_n, \theta^*) \leq \overline{M}. \quad (18)$$

Отсюда

$$\underline{M}(\theta^*) \leq \overline{M}. \quad (19)$$

Таким образом мы получили следующую теорему.

**Теорема.** *Множество*

$$T = \{\theta \mid \underline{M}(\theta) \leq \overline{M}\} \quad (20)$$

*содержит все возможные точечные  $M$ -оценки (10) при  $X_n \in \mathbf{X}_n$ .*

Точечную оценку, при которой достигается минимум (15)

$$\min_{\theta} \overline{M}(\theta) = \overline{M}(\theta_{\text{песс}}), \quad (21)$$

можно назвать оценкой “крайнего пессимиста”, а оценку, при которой достигается минимум (16)

$$\min_{\theta} \underline{M}(\theta) = \underline{M}(\theta_{\text{опт}}), \quad (22)$$

можно назвать оценкой “крайнего оптимиста” (рис. 2).

Теорема описывает допустимое множество оценок параметров, но не дает метода его построения. Рассмотрим случай, когда  $\theta$  — скалярный параметр и функционал  $M(X_n, \theta)$  является выпуклым. Тогда множество  $T$  представляет собой замкнутый интервал, границы которого определяются из уравнения  $\underline{M}(\theta) = \overline{M}$  (см. рис. 2). Это уравнение можно свести к двум задачам нелинейного программирования:

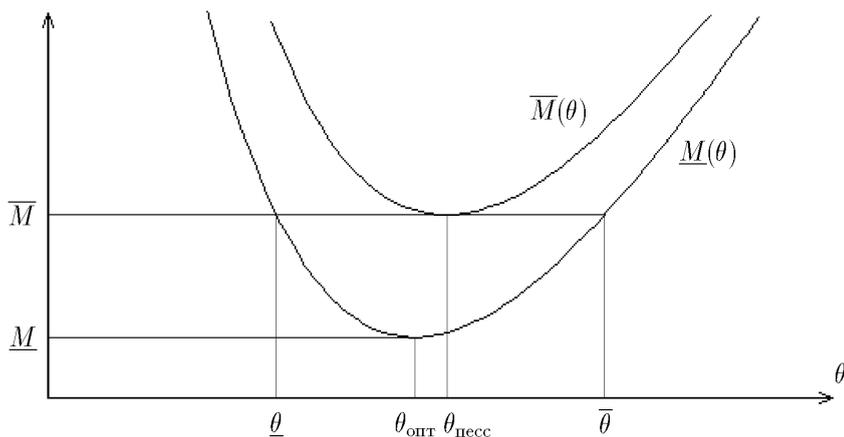
$$\underline{\theta}^* = \arg \min_{\theta < \theta_{\text{опт}}} (\underline{M}(\theta) - \overline{M})^2, \quad (23)$$

$$\overline{\theta}^* = \arg \min_{\theta > \theta_{\text{опт}}} (\underline{M}(\theta) - \overline{M})^2. \quad (24)$$

Частным случаем  $M$ -оценок являются оценки максимального правдоподобия, когда функция потерь имеет вид

$$\rho(x, \theta) = -\ln f(x, \theta),$$

где  $f(x, \theta)$  — функция плотности распределения.

Рис. 2. Нахождение интервальной  $M$ -оценки.

### 3.3. Интервальные $MD$ -оценки

$D$ -оценки получаются, если в выражении (10) в качестве минимизируемого функционала взять расстояние между эмпирической функцией распределения  $F_n(x)$  и теоретической  $F(x, \theta)$ . Интервальные  $MD$ -оценки получаются аналогично интервальным  $M$ -оценкам, если учесть выражение (6), задающее границы возможного расположения эмпирической функции распределения. В качестве меры близости эмпирической и теоретической функций распределения можно использовать статистики непараметрических критериев согласия. Оценки границ  $\underline{M}(\theta)$  и  $\overline{M}(\theta)$  для статистик критериев Колмогорова, Смирнова,  $\omega^2$  и  $\Omega^2$  Мизеса получены в [4].

## 4. Свойства интервальных оценок

Различные точечные оценки параметров порождают в общем случае различные интервальные оценки. Естественным критерием для сравнения различных интервальных оценок является длина интервала  $\Delta\theta = \bar{\theta} - \underline{\theta}$ .

**Пример 4.** Выборка, состоящая из 200 интервальных наблюдений, была смоделирована следующим образом: середины интервалов определялись по реализациям трех случайных величин  $\zeta_i$ ,  $\xi_i$  и  $y_i$

$$x_i = (1 + \zeta_i)y_i + \xi_i,$$

где  $\zeta_i$  распределена равномерно на интервале  $[-\delta, \delta]$ ,  $\xi_i$  распределена равномерно на интервале  $[-\varepsilon, \varepsilon]$ ,  $y_i$  распределена по нормальному закону с параметром сдвига  $\mu = 0$  и параметром масштаба  $\sigma = 1$ . Границы интервалов определялись по формулам

$$a_i = x_i - \delta|x_i| - \varepsilon, \quad b_i = x_i + \delta|x_i| + \varepsilon, \quad (25)$$

где  $\varepsilon = 0.01$  — абсолютная, а  $\delta = 0.01$  — относительная погрешности наблюдения.

Интервальная гистограмма для полученной выборки приведена на рис. 3. Высота каждого столбца гистограммы является интервалом, так как при группировании выборки число наблюдений, попадающих в интервал группирования, не однозначно. В то же время ни

нижняя, ни верхняя граница интервальной гистограммы не удовлетворяют условию нормировки, а лишь задают миноранту и мажоранту для гистограммы с теми же граничными точками, построенную по любой точечной выборке  $X_n \in \mathbf{X}_n$ .

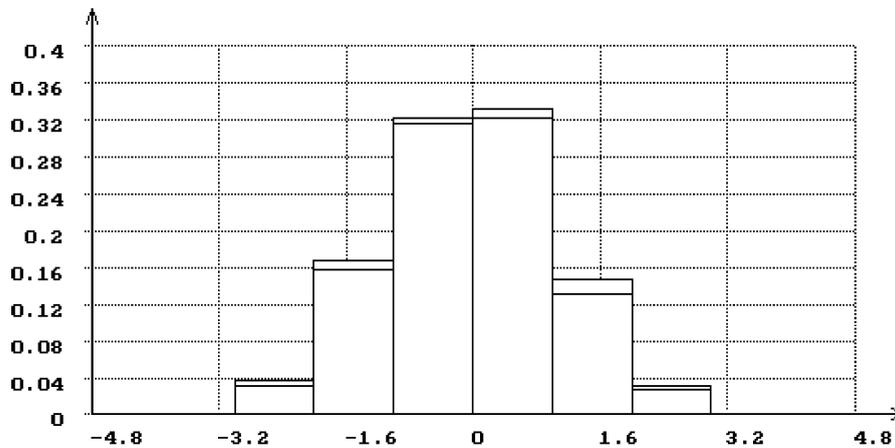


Рис. 3. Интервальная гистограмма для интервальной выборки объемом 200 наблюдений.

По интервальной выборке оценивался параметр масштаба нормального распределения по трем методам: максимального правдоподобия, минимума статистики Колмогорова и минимума статистики  $\omega^2$  Мизеса. Полученные точечные и интервальные оценки приведены в таблице. Наихудшим из трех методов оказался метод максимального правдоподобия, так как длина интервала оценки получилась наибольшей, а наилучшим оказался метод минимума статистики Колмогорова. Отметим, что все три интервальные оценки накрыли “истинное” значение параметра, то есть то значение параметра, при котором моделировалась выборка.

Свойства интервальных оценок во многом зависят от свойств точечных оценок, по которым они строятся. Если точечная оценка робастна, т. е. нечувствительна к наличию аномальных наблюдений в выборке [5], то следует ожидать, что соответствующая ей ин-

Оценивание параметра масштаба нормального распределения по интервальной выборке объемом 200 наблюдений

Метод оценивания	Класс оценки	Точечные оценки			Интервальная оценка	Длина интервала
		по срединам интервалов	“крайнего оптимиста”	“крайнего пессимиста”		
Метод максимального правдоподобия	<i>M</i> -оценка	1.0680	1.0493	1.0867	[0.8798, 1.2811]	0.4013
Метод минимума статистики Колмогорова	<i>MD</i> -оценка	1.0276	1.0272	1.0278	[0.9774, 1.0946]	0.1172
Метод минимума статистики $\omega^2$ Мизеса	<i>MD</i> -оценка	1.0897	1.0809	1.0965	[0.9544, 1.2323]	0.2779

тервальная оценка будет иметь меньшую длину, чем интервальная оценка, порожденная точечной оценкой, не обладающей свойством робастности. Исследование функций влияния Хэмпела показало, что оценки максимального правдоподобия по негруппированным данным для большинства используемых на практике распределений не являются робастными [6]. Этим можно объяснить тот факт, что в рассмотренном примере метод максимального правдоподобия оказался наихудшим.

Длина интервала  $\Delta\theta$  слабо зависит от объема выборки и не стремится к нулю при увеличении числа наблюдений. Нетрудно показать, что если интервальные наблюдения получены по (25), а интервальная оценка максимального правдоподобия параметра масштаба нормального распределения вычисляется так же, как в примере 2, то  $\Delta\theta \geq \frac{8\varepsilon^2}{1-\delta} > 0$ .

## 5. Заключение

По интервальной выборке можно находить как точечные, так и интервальные оценки. Интервальные оценки параметров отражают интервальную неопределенность в задании исходных данных. В случае одного параметра интервальное оценивание сводится к решению задач линейного (при поиске  $L$ -оценок) и нелинейного (при поиске  $M$ -оценок и  $MD$ -оценок) программирования. Длина интервала  $(\bar{\theta} - \underline{\theta})$  существенно зависит от метода оценивания и слабо зависит от объема выборки.

## Список литературы

- [1] ЛЕМЕШКО Б. Ю., ПОСТОВАЛОВ С. Н. О решении задач статистического анализа по интервальным наблюдениям. *Вычисл. технологии*, **2**, №1, 1997, 28–36.
- [2] ОРЛОВ А. И. О влиянии погрешностей наблюдений на свойства статистических процедур (на примере гамма-распределения) В *“Стат. методы оценивания и проверки гипотез: Межвуз. сб. науч. трудов”*. Пермский ун-т, Пермь, 1988, 45–55.
- [3] ЛЕМАН Э. *Теория точечного оценивания*. Наука, М., 1991.
- [4] ЛЕМЕШКО Б. Ю., ПОСТОВАЛОВ С. Н. Статистический анализ наблюдений, имеющих интервальное представление. В *“Сб. науч. трудов НГТУ”*, Новосибирск, №1, 1996, 3–12.
- [5] ХЬЮБЕР П. *Робастность в статистике*. Мир, М., 1984.
- [6] ЛЕМЕШКО Б. Ю., ПОСТОВАЛОВ С. Н. К вопросу о робастности оценок по группированным данным. В *“Сб. науч. трудов НГТУ”*, Новосибирск, №2, 1996, 9–18.

Поступила в редакцию 31 октября 1997 г.