Сибирский журнал
индустриальной математики
Июль–декабрь, 2001. Том IV, № 2(8)

УДК 519.24

ПОСТРОЕНИЕ ОПТИМАЛЬНЫХ L–ОЦЕНОК ПАРАМЕТРОВ СДВИГА И МАСШТАБА РАСПРЕДЕЛЕНИЙ ПО ВЫБОРОЧНЫМ КВАНТИЛЯМ*)

Б. Ю. Лемешко, Е. В. Чимитова

Исследуются свойства L-оценок параметров сдвига и масштаба в виде линейных комбинаций выборочных квантилей. Предлагается использовать для построения таких оценок выборочные квантили, соответствующие асимптотически оптимальному группированию, при котором минимизируются потери в информации Фишера. Построены таблицы коэффициентов L-оценок параметров для ряда законов распределений, позволяющие просто вычислять оптимальные L-оценки. Методами статистического моделирования показано, что предельными распределениями статистики критерия согласия χ^2 Пирсона в случае использования L-оценок являются χ^2_{k-m-1}-распределения.

Введение. L-оценки параметров распределений, формируемые как линейные комбинации порядковых статистик или выборочных квантилей, обладают двумя важными для широкого практического применения качествами: чрезвычайной простотой вычислений и очень хорошими свойствами робастности.

В данной работе мы остановимся на оценках параметров сдвига и масштаба, вычисление которых базируется на значениях выборочных квантилей [1]. Самой сложной операцией при вычислении таких оценок является сортировка имеющейся выборки по возрастанию с целью определения выборочных квантилей наблюдаемого закона. Интуитивно ясно, что так как L-оценки получаются в виде линейной комбинации выборочных квантилей, то отдельные аномальные наблюдения (очень большие или очень малые), возможно присутствующие в выборке, никоим образом не влияют на значения оценок параметров закона распределения. Такие оценки являются робастными, как и оценки максимального правдоподобия по группированным наблюдениям [2, 3]. Робастность этих оценок подтверждает вид функций влияния Хампеля [4], которые для L-оценок представляют собой ступенчатые, ограниченные по абсолютной величине зависимости [5, 6].

В данном случае основное внимание будет уделено тому, какие квантили закона распределения при заданном их числе следует использовать, чтобы асимптотические свойства рассматриваемых L-оценок были наилучшими. Насколько отличаются свойства оценок при ограниченных объемах выборок от асимптотических? Как отражается использование рассматриваемых оценок на распределении статистики критерия согласия χ^2 Пирсона при проверке сложной гипотезы?

1. Построение L-оценок параметров сдвига и масштаба. Опираясь на асимптотическое распределение выборочных квантилей ($k-1$ квантилей при k интервалах) [7], Дз. Огава в работах [1, 8] получил асимптотическое распределение выборочных квантилей для функции плотности, зависящей только от

*) Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (код проекта 00–01–00913).

параметров расположения μ и рассеяния σ. Им же [1, с. 54–60] методом наименьших квадратов построены «оптимальные линейные несмещенные оценки» параметров сдвига и масштаба, в основе которых лежат значения выборочных квантилей.

Пусть μ и σ — неизвестные параметры сдвига и масштаба закона с функцией распределения $F\left(\frac{x-\mu}{\sigma}\right)$ и функцией плотности $\frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right)$.

При *известном параметре* σ оценка параметра μ имеет вид [1]

$$\tilde{\mu} = \frac{1}{K_1}Z - \sigma\frac{K_3}{K_1}, \tag{1}$$

где

$$Z = \sum_{i=1}^{k} \frac{(f_i - f_{i-1})(f_i\hat{x}_{(i)} - f_{i-1}\hat{x}_{(i-1)})}{P_i}, \tag{2}$$

$$K_1 = \sigma^2 J_\Gamma(\mu) = \sum_{i=1}^{k} \frac{(f_i - f_{i-1})^2}{P_i}, \tag{3}$$

$$K_3 = \sigma^2 J_\Gamma(\mu,\sigma) = \sum_{i=1}^{k} \frac{(f_{i-1} - f_i)(f_{i-1}t_{i-1} - f_i t_i)}{P_i}, \tag{4}$$

$f_i = f(t_i)$, $t = (x-\mu)/\sigma$, $P_i = F(t_i) - F(t_{i-1})$, $f_0 = f_k = 0$. Здесь $\hat{x}_{(i)}$ — оценка по наблюдаемой выборке такой квантили закона, что $F\left(\frac{x_{(i)}-\mu}{\sigma}\right) = F(t_i)$, где t_i — соответствующая квантиль стандартного распределения с нулевым параметром сдвига и единичным масштабным. Через $J_\Gamma(\cdot)$ обозначено количество информации Фишера о соответствующем параметре по группированным данным. В общем случае информационная матрица Фишера о векторе параметров распределения θ по группированным наблюдениям определяется выражением

$$\mathbf{J}_\Gamma(\theta) = \sum_{i=1}^{k} \frac{\nabla P_i(\theta)\nabla^T P_i(\theta)}{P_i(\theta)},$$

где $P_i(\theta) = \int\limits_{x_{(i-1)}}^{x_{(i)}} f(x,\theta)\,dx$ — вероятность попадания наблюдения в интервал. В случае $\theta^T = (\mu,\sigma)$ это будет матрица

$$\mathbf{J}_\Gamma(\mu,\sigma) = \begin{bmatrix} J_\Gamma(\mu) & J_\Gamma(\mu,\sigma) \\ J_\Gamma(\mu,\sigma) & J_\Gamma(\sigma) \end{bmatrix}.$$

При *известном параметре* μ оценка параметра σ определяется выражением [1]

$$\widetilde{\sigma} = \frac{1}{K_2}Y - \mu\frac{K_3}{K_2}, \tag{5}$$

где

$$Y = \sum_{i=1}^{k} \frac{(t_i f_i - t_{i-1}f_{i-1})(f_i\hat{x}_{(i)} - f_{i-1}\hat{x}_{(i-1)})}{P_i}, \tag{6}$$

$$K_2 = \sigma^2 J_\Gamma(\sigma) = \sum_{i=1}^{k} \frac{(t_i f_i - t_{i-1}f_{i-1})^2}{P_i}. \tag{7}$$

При симметричных функциях плотности и симметричных квантилях второе слагаемое в формулах (1) и (5) равно нулю.

Если неизвестны оба параметра, то оценки параметров сдвига и масштаба представлены соотношениями [1]

$$\tilde{\mu} = \frac{1}{\Delta}(K_2 Z - K_3 Y), \tag{8}$$

$$\tilde{\sigma} = \frac{1}{\Delta}(-K_3 Z + K_1 Y), \tag{9}$$

где $\Delta = K_1 K_2 - K_3^2$.

Соотношения (1), (5), (8) и (9) можно преобразовать в совсем простую зависимость [9]. Формулу (1) для оценивания μ при известном σ можно привести к виду

$$\tilde{\mu} = \alpha_0 \sigma + \sum_{i=1}^{k-1} \alpha_i \hat{x}_{(i)}, \tag{10}$$

где

$$\alpha_0 = -\frac{K_3}{K_1}, \quad \alpha_1 = \alpha_1'/K_1 = \left(\frac{f_1^2}{P_1} + \frac{f_1^2 - f_1 f_2}{P_2}\right)/K_1,$$
$$\alpha_i = \alpha_i'/K_1 = \left(\frac{f_i^2 - f_i f_{i-1}}{P_i} + \frac{f_i^2 - f_i f_{i+1}}{P_{i+1}}\right)/K_1, \ i = \overline{2,(k-2)},$$
$$\alpha_{k-1} = \alpha_{k-1}'/K_1 = \left(\frac{f_{k-1}^2 - f_{k-1} f_{k-2}}{P_{k-1}} + \frac{f_{k-1}^2}{P_k}\right)/K_1.$$

А формулу (5) для оценивания σ при известном μ представим в виде

$$\tilde{\sigma} = \beta_0 \mu + \sum_{i=1}^{k-1} \beta_i \hat{x}_{(i)}, \tag{11}$$

где

$$\beta_0 = -\frac{K_3}{K_2}, \quad \beta_1 = \beta_1'/K_2 = \left(\frac{t_1 f_1^2}{P_1} + \frac{t_1 f_1^2 - t_2 f_1 f_2}{P_2}\right)/K_2,$$
$$\beta_i = \beta_i'/K_2 = \left(\frac{t_i f_i^2 - t_{i-1} f_i f_{i-1}}{P_i} + \frac{t_i f_i^2 - t_{i+1} f_i f_{i+1}}{P_{i+1}}\right)/K_2, \ i = \overline{2,(k-2)},$$
$$\beta_{k-1} = \beta_{k-1}'/K_2 = \left(\frac{t_{k-1} f_{k-1}^2 - t_{k-2} f_{k-1} f_{k-2}}{P_{k-1}} + \frac{t_{k-1} f_{k-1}^2}{P_k}\right)/K_2.$$

Аналогично формулы (8) и (9) можно преобразовать к виду

$$\tilde{\mu} = \sum_{i=1}^{k-1} \gamma_i \hat{x}_{(i)}, \tag{12}$$

$$\tilde{\sigma} = \sum_{i=1}^{k-1} \nu_i \hat{x}_{(i)}, \tag{13}$$

где $\gamma_i = \big(\alpha_i' K_2 - \beta_i' K_3\big)/\Delta$, $\nu_i = \big(-\alpha_i' K_3 + \beta_i' K_1\big)/\Delta$.

2. Выбор квантилей t_i стандартного распределения и вычисление α_i, β_i, γ_i, ν_i. Все рассмотренные выше оценки параметров асимптотически эффективны [1], и их асимптотические дисперсии определяются количеством информации Фишера по группированным данным, а в случае векторного параметра — соответствующей информационной матрицей

$$D(\theta) = n^{-1}\mathbf{J}_{\Gamma}^{-1}(\theta). \tag{14}$$

Коэффициенты α_i, β_i, γ_i, ν_i зависят от граничных точек t_i (квантилей стандартного распределения). Очевидно, что так как рассматриваемые оценки асимптотически эффективны, то использование квантилей (граничных точек интервалов), соответствующих асимптотически оптимальному группированию, при котором минимизируются потери в информации Фишера, связанные с группированием [10, 11], обеспечит оптимальные свойства этих оценок [9] — минимум асимптотической дисперсии, а в случае оценивания сразу двух параметров — минимум обобщенной асимптотической дисперсии. Несложно вычислить значения α_i, β_i, γ_i, ν_i при асимптотически оптимальном группировании и сформировать таблицы соответствующих коэффициентов. И если в случае больших выборок мы будем выбирать $\hat{x}_{(i)}$ таким образом, чтобы $n \approx nP_i$, где P_i соответствует вероятности попадания в интервал при асимптотически оптимальном группировании, используя соответственно формулы (10)–(13) с полученными коэффициентами, то получим оптимальные оценки.

Отметим, что в частных случаях решение такой задачи рассматривалось в ряде работ. В [1, 12] рассматривались оценки параметров для нормального распределения, в [1] — для однопараметрического экспоненциального распределения, в [13] — для двухпараметрического экспоненциального распределения, в [14] — для параметров логистического распределения, в [15] — для параметров распределения Коши, в [16] — для параметров распределения экстремальных значений. Приближенный подход к решению такой задачи рассматривался в [17]. Причем в случае одновременного оценивания параметров μ и σ оптимальные наборы граничных точек определялись исходя из минимума величины $D[\tilde{\mu}] + cD[\tilde{\sigma}]$, $c = 1, 2, 3\ldots$, а не минимума $\det \mathbf{J}_{\Gamma}^{-1}(\tilde{\mu}, \tilde{\sigma})$.

Опираясь на построенную нами совокупность таблиц асимптотически оптимального группирования [11, 18], значения коэффициентов α_i, β_i, γ_i, ν_i для параметров законов распределений, упоминаемых в данной статье, получены в [9] (64 таблицы) и вместе с таблицами асимптотически оптимального группирования (58 таблиц) доступны читателям журнала на WEB-сайте http://www.ami.nstu.ru/~headrd/.

Таблицы коэффициентов для формул вида (10)–(13) сформированы для нормального распределения, логистического распределения с функцией плотности

$$f(x) = \frac{\pi}{\sigma\sqrt{3}} \exp\left\{-\frac{\pi(x-\mu)}{\sigma\sqrt{3}}\right\} \Bigg/ \left[1 + \exp\left\{-\frac{\pi(x-\mu)}{\sigma\sqrt{3}}\right\}\right]^2,$$

распределения Коши с плотностью

$$f(x) = \frac{\sigma}{\pi[\sigma^2 + (x-\mu)^2]},$$

распределения наименьшего экстремального значения с плотностью

$$f(x) = \frac{1}{\sigma} \exp\left\{\frac{x-\mu}{\sigma} - \exp\left(\frac{x-\mu}{\sigma}\right)\right\},$$

распределения наибольшего экстремального значения с плотностью

$$f(x) = \frac{1}{\sigma} \exp\left\{-\frac{x-\mu}{\sigma} - \exp\left(\frac{x-\mu}{\sigma}\right)\right\}.$$

При этом в зависимости от того, известен ли один из параметров или неизвестны оба, наборам коэффициентов α_i, β_i и паре γ_i, ν_i соответствуют свои таблицы асимптотически оптимального группирования. В частности, для нормального распределения асимптотически оптимальные граничные точки t_i для случая одновременного оценивания двух параметров представлены в табл. 1, а соответствующие вероятности — в табл. 2. Полученные значения коэффициентов γ_i, ν_i приведены в табл. 3, 4.

Для распределений экспоненциального с плотностью

$$f(x) = \frac{1}{\sigma} \exp\left\{-\frac{x-\mu}{\sigma}\right\},$$

модуля нормального вектора ($m = 1 \div 9$) с плотностью

$$f(x) = \frac{2(x-\mu)^{m-1}}{(2\sigma^2)^{m/2}\Gamma(m/2)} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\},$$

частными случаями которого являются распределения полунормальное $m = 1$, Рэлея $m = 2$ и Максвелла $m = 3$, таблицы коэффициентов α_i, β_i, γ_i, ν_i опираются на таблицы асимптотически оптимального группирования *только относительно масштабного параметра* σ. Это связано с тем, что область определения этих случайных величин зависит от параметра сдвига μ и, следовательно, в этом случае теряет смысл максимизация соотношения (3) для построения асимптотически оптимальных граничных точек относительно этого параметра.

Симметричность коэффициентов в формулах (10)–(13) для симметричных распределений определяется симметричностью оптимальных граничных точек интервалов. Для параметров масштаба при известном параметре сдвига и четном k задача асимптотически оптимального группирования обычно имеет два решения с несимметричными значениями квантилей. В таких случаях пара этих решений зеркальна относительно центра симметрии распределения. Поэтому не единственным будет оптимальный набор коэффициентов в формулах (11). Таким образом, не оправдывается предположение о симметричности оптимальных порядковых статистик для параметра σ нормального распределения, высказанное в [19].

Значения $\hat{x}_{(i)}$, фигурирующие в формулах (10)–(13), следует выбирать из условия

$$X_{([nP^i])} \le \hat{x}_{(i)} \le X_{([nP^i]+1)},$$

где $X_{(l)}$ — члены вариационного ряда $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$, построенного по исходной выборке, $P^i = \sum\limits_{j=1}^{i} P_j$, $[\cdot]$ означает целую часть числа, а P_j выбираются из соответствующей строки таблицы оптимальных вероятностей. Например, в качестве $\hat{x}_{(i)}$ могут быть взяты средние значения между соответствующими соседними членами вариационного ряда.

Пример. Для нормального распределения при $k = 9$ соотношение (10) принимает вид (см. на упомянутом выше WEB-сайте)

$$\tilde{\mu} = 0,056339(\hat{x}_{(1)} + \hat{x}_{(8)}) + 0,111523(\hat{x}_{(2)} + \hat{x}_{(7)}) + 0,154649(\hat{x}_{(3)} + \hat{x}_{(6)}) + 0,177489(\hat{x}_{(4)} + \hat{x}_{(5)}),$$

соотношение (11) (см. WEB-сайт)

$$\tilde{\sigma} = 0,031157(\hat{x}_{(8)} - \hat{x}_{(1)}) + 0,072629(\hat{x}_{(7)} - \hat{x}_{(2)}) + 0,116643(\hat{x}_{(6)} - \hat{x}_{(3)}) + 0,147029(\hat{x}_{(5)} - \hat{x}_{(4)}),$$

Таблица 1

Оптимальные граничные точки интервалов группирования в виде $t_i = (x_{(i)} - \mu)/\sigma$ для одновременного оценивания двух параметров нормального распределения и проверки согласия по критерию χ^2 Пирсона и соответствующие значения относительной асимптотической информации A

k	t_1	t_2	t_3	t_4	t_5	t_6	t_7
3	-1,1106	1,1106	-	-	-	-	-
4	-1,3834	0,0	1,3834	-	-	-	-
5	-1,6961	-0,6894	0,6894	1,6961	-	-	-
6	-1,8817	-0,9970	0,0	0,9970	1,8817	-	-
7	-2,0600	-1,2647	-0,4918	0,4918	1,2647	2,0600	-
8	-2,1954	-1,4552	-0,7863	0,0	0,7863	1,4552	2,1954
9	-2,3188	-1,6218	-1,0223	-0,3828	0,3828	1,0223	1,6218
10	-2,4225	-1,7578	-1,2046	-0,6497	0,0	0,6497	1,2046
11	-2,5167	-1,8784	-1,3602	-0,8621	-0,3143	0,3143	0,8621
12	-2,5993	-1,9028	-1,4914	-1,0331	-0,5334	0,0	0,5334
13	-2,6746	-2,0762	-1,6068	-1,1784	-0,7465	-0,2669	0,2669
14	-2,7436	-2,1609	-1,7092	-1,3042	-0,9065	-0,4818	0,0
15	-2,8069	-2,2378	-1,8011	-1,4150	-1,0435	-0,6590	-0,2325

t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}	A
-	-	-	-	-	-	-	0,4065
-	-	-	-	-	-	-	0,5527
-	-	-	-	-	-	-	0,6826
-	-	-	-	-	-	-	0,7557
-	-	-	-	-	-	-	0,8103
-	-	-	-	-	-	-	0,8474
2,3188	-	-	-	-	-	-	0,8753
1,7578	2,4225	-	-	-	-	-	0,8960
1,3602	1,8784	2,5167	-	-	-	-	0,9121
1,0331	1,4914	1,9028	2,5993	-	-	-	0,9247
0,7465	1,1784	1,6068	2,0762	2,6746	-	-	0,9348
0,4818	0,9065	1,3042	1,7092	2,1609	2,7436	-	0,9430
0,2325	0,6590	1,0435	1,4150	1,8011	2,2378	2,8069	0,9498

Таблица 2

Вероятности попадания наблюдений в интервалы при асимптотически оптимальном группировании в случае одновременного оценивания двух параметров нормального распределения или проверки согласия по критерию χ^2 Пирсона и значения относительной асимптотической информации A

k	P_1	P_2	P_3	P_4	P_5	P_6	P_7
3	0,1334	0,7332	0,1334	-	-	-	-
4	0,0833	0,4167	0,4167	0,0833	-	-	-
5	0,0449	0,2004	0,5094	0,2004	0,0449	-	-
6	0,0299	0,1295	0,3406	0,3406	0,1295	0,0299	-
7	0,0197	0,0833	0,2084	0,3772	0,2084	0,0833	0,0197
8	0,0141	0,0587	0,1431	0,2841	0,2841	0,1431	0,0587
9	0,0102	0,0422	0,1009	0,1976	0,2982	0,1976	0,1009
10	0,0077	0,0317	0,0748	0,1438	0,2420	0,2420	0,1438
11	0,0059	0,0243	0,0567	0,1074	0,1823	0,2468	0,1823
12	0,0047	0,0190	0,0442	0,0829	0,1392	0,2100	0,2100
13	0,0037	0,0152	0,0352	0,0652	0,1085	0,1670	0,2104
14	0,0030	0,0124	0,0283	0,0524	0,0862	0,1327	0,1850
15	0,0025	0,0101	0,0232	0,0427	0,0698	0,1066	0,1532

P_8	P_9	P_{10}	P_{11}	P_{12}	P_{13}	P_{14}	P_{15}	A
-	-	-	-	-	-	-	-	0,4065
-	-	-	-	-	-	-	-	0,5527
-	-	-	-	-	-	-	-	0,6826
-	-	-	-	-	-	-	-	0,7557
-	-	-	-	-	-	-	-	0,8103
0,0141	-	-	-	-	-	-	-	0,8474
0,0422	0,0102	-	-	-	-	-	-	0,8753
0,0748	0,0317	0,0077	-	-	-	-	-	0,8960
0,1074	0,0567	0,0243	0,0059	-	-	-	-	0,9121
0,1392	0,0829	0,0442	0,0190	0,0047	-	-	-	0,9247
0,1670	0,1085	0,0652	0,0352	0,0152	0,0037	-	-	0,9348
0,1850	0,1327	0,0862	0,0524	0,0283	0,0124	0,0030	-	0,9430
0,1838	0,1532	0,1066	0,0698	0,0427	0,0232	0,0101	0,0025	0,9498

Таблица 3

Коэффициенты для параметра сдвига нормального распределения
в случае одновременного оценивания двух параметров

k	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6
3	0,500000	0,500000	-	-	-	-
4	0,224374	0,551252	0,224374	-	-	-
5	0,108579	0,391421	0,391421	0,108579	-	-
6	0,067815	0,234061	0,396249	0,234061	0,067815	-
7	0,043180	0,141936	0,314884	0,314884	0,141936	0,043180
8	0,029871	0,096902	0,216939	0,312575	0,216939	0,096902
9	0,021547	0,068108	0,148605	0,261739	0,261739	0,148605
10	0,016187	0,050213	0,107748	0,196679	0,258345	0,196679
12	0,002356	0,078666	0,032450	0,099837	0,188002	0,197378
13	0,008056	0,023716	0,048181	0,086640	0,138225	0,195182
14	0,006737	0,018173	0,039623	0,068299	0,109776	0,161527
15	0,005076	0,015581	0,032157	0,055371	0,088028	0,130918

γ_7	γ_8	γ_9	γ_{10}	γ_{11}	γ_{12}	γ_{13}	γ_{14}
-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-
0,029871	-	-	-	-	-	-	-
0,068108	0,021547	-	-	-	-	-	-
0,107748	0,050213	0,016187	-	-	-	-	-
0,188002	0,099837	0,032450	0,078666	0,002356	-	-	-
0,195182	0,138225	0,086640	0,048181	0,023716	0,008056	-	-
0,191732	0,161527	0,109776	0,068299	0,039623	0,018173	0,006737	-
0,172869	0,172869	0,130918	0,088028	0,055371	0,032157	0,015581	0,005076

Таблица 4

Коэффициенты для параметра масштаба нормального распределения в случае одновременного оценивания двух параметров

k	ν_1	ν_2	ν_3	ν_4	ν_5	ν_6
3	-0,450207	0,450207	-	-	-	-
4	-0,361428	0	0,361428	-	-	-
5	-0,201360	-0,229872	0,229872	0,201360	-	-
6	-0,140732	-0,235892	0	0,235892	0,140732	-
7	-0,095717	-0,186279	-0,136715	0,136715	0,186279	0,095717
8	-0,070411	-0,147147	-0,166972	0	0,166972	0,147147
9	-0,052747	-0,114684	-0,153492	-0,090860	0,090860	0,153492
10	-0,040995	-0,091463	-0,132388	-0,123812	0	0,123812
11	-0,032533	-0,073373	-0,111849	-0,124976	-0,064980	0,064980
12	-0,019239	-0,098971	-0,069069	-0,112065	-0,080404	0
13	-0,021688	-0,050048	-0,079521	-0,102566	-0,102357	-0,048835
15	-0,014676	-0,035496	-0,058652	-0,079499	-0,092263	-0,085051

ν_7	ν_8	ν_9	ν_{10}	ν_{11}	ν_{12}	ν_{13}	ν_{14}
-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-
0,070411	-	-	-	-	-	-	-
0,114684	0,052747	-	-	-	-	-	-
0,132388	0,091463	0,040995	-	-	-	-	-
0,124976	0,111849	0,073373	0,032533	-	-	-	-
0,080404	0,112065	0,069069	0,098971	0,019239	-	-	-
0,048835	0,102357	0,102566	0,079521	0,050048	0,021688	0,018214	-
-0,038353	0,038353	0,085051	0,092263	0,079499	0,058652	0,035496	0,014676

соотношения (12) и (13) (см. табл. 3, 4)

$$\tilde{\mu} = 0,021547(\hat{x}_{(1)} + \hat{x}_{(8)}) + 0,068108(\hat{x}_{(2)} + \hat{x}_{(7)}) + 0,148605(\hat{x}_{(3)} + \hat{x}_{(6)}) + 0,261739(\hat{x}_{(4)} + \hat{x}_{(5)}),$$

$$\widetilde{\sigma} = 0,052747(\hat{x}_{(8)} - \hat{x}_{(1)}) + 0,114684(\hat{x}_{(7)} - \hat{x}_{(2)}) + 0,153492(\hat{x}_{(6)} - \hat{x}_{(3)}) + 0,090860(\hat{x}_{(5)} - \hat{x}_{(4)}).$$

Если мы оцениваем оба параметра, для определения $\hat{x}_{(i)}$ вероятности P_i выбираются из табл. 2. И при объеме выборки в 1000 наблюдений в качестве $\hat{x}_{(i)}$, $i = \overline{1,8}$, можно взять средние значения между следующими парами членов вариационного ряда: $X_{(10)} - X_{(11)}$, $X_{(52)} - X_{(53)}$, $X_{(153)} - X_{(154)}$, $X_{(350)} - X_{(351)}$, $X_{(649)} - X_{(650)}$, $X_{(846)} - X_{(847)}$, $X_{(947)} - X_{(948)}$, $X_{(989)} - X_{(990)}$.

3. Точность оценивания квантилей и *L*-оценок. Оптимальные L-оценки параметров сдвига и масштаба являются асимптотически эффективными. На практике же мы имеем дело с выборками ограниченного объема. Понятно, что и точность оценивания квантилей $\hat{x}_{(i)}$, и точность вычисления L-оценок зависят от объема выборки n. В качестве основного возражения против использования L-оценок обычно выдвигают возможную значительную неточность в определении выборочных квантилей $\hat{x}_{(i)}$, которая должна отражаться на точности L-оценок. Методами статистического моделирования мы исследовали законы распределения выборочных квантилей и получаемых L-оценок в зависимости от конкретных объемов выборок n и числа используемых квантилей для различных законов распределений. На рис. 1 для случая оценивания масштабного параметра σ экспоненциального закона при использовании 5 квантилей (число интервалов $k = 6$) приведены центрированные плотности выборочных квантилей $\hat{x}_{(1)} \div \hat{x}_{(5)}$ и L-оценок $\widetilde{\sigma}$, построенных по этим выборочным квантилям (центрированные относительно истинных значений квантилей $x_{(i)}$ и параметра $\sigma = \sigma_0$) при объемах выборок $n = 1000$. Экспоненциальный закон моделировался с масштабным параметром $\sigma = 1$. Значения асимптотически оптимальных квантилей $x_{(i)}$, $i = \overline{1,5}$, для данной ситуации соответственно равны [11, 18, 20] $0,4993$; $1,0997$; $1,8538$; $2,8714$; $4,4650$. Значение L-оценки $\widetilde{\sigma}$ определялось по формуле

$$\widetilde{\sigma} = -0,81488\mu + 0,347021\hat{x}_{(1)} + 0,232423\hat{x}_{(2)} + 0,140462\hat{x}_{(3)} + 0,071103\hat{x}_{(4)} + 0,023870\hat{x}_{(5)}.$$

Для построения приведенных на рис. 1 законов распределения формировались выборки оценок из $N = 2000$ значений, каждое из которых находилось по выборке объема n случайных величин, распределенных по экспоненциальному закону. Для большей точности параметры найденных моделей законов распределения оценок усреднялись по 100 таким экспериментам. Для сравнения на рис. 1 построена также плотность асимптотического распределения оценки максимального правдоподобия (ОМП) $\widehat{\sigma}$ по точечной выборке. Сравнивая плотность асимптотически эффективной ОМП по точечной выборке с плотностью L-оценки, мы видим, что последние мало уступают ОМП. Это естественно, так как в данном случае при асимптотически оптимальном группировании сохраняется $94,76\,\%$ информации Фишера о параметре масштаба σ. Следовательно, стандартное отклонение предельного распределения $f(\widetilde{\sigma})$ превышает стандартное отклонение распределения $f(\widehat{\sigma})$ не более чем на $2,73\,\%$.

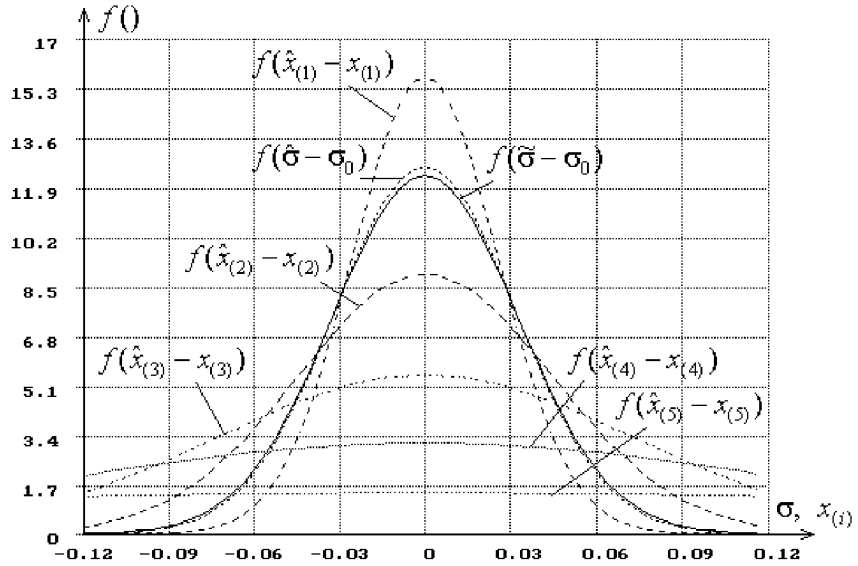

Рис. 1. Центрированные распределения выборочных асимптотически оптимальных квантилей и L-оценок масштабного параметра экспоненциального распределения при объемах выборок $n = 1000$

Рис. 1 наглядно демонстрирует, что несмотря на относительно невысокую точность оценивания квантилей $x_{(5)}$, $x_{(4)}$ и $x_{(3)}$ мы имеем достаточно высокую точность оценивания параметра масштаба σ наблюдаемого экспоненциального закона. При этом очевидно, что эти оценки немногим уступают ОМП по точечным (негруппированным) наблюдениям, имея существенное преимущество в робастности.

4. Точность *L*-оценок в зависимости от объема выборок. Характер изменения точности L-оценок с ростом объема выборок n при фиксированном числе используемых квантилей показывают рис. 2 и 3. На рисунках приведены соответственно плотности оценок $\tilde{\mu}$ и $\widetilde{\sigma}$ параметров нормального закона, центрированные относительно истинных значений параметров μ_0 и σ_0, для случая $k = 3$ в зависимости от n, когда при построении L-оценок используются всего две выборочные квантили, соответствующие асимптотически оптимальному группированию, и одновременно оцениваются оба параметра. Выборки нормального закона объема n генерировались с параметрами $\mu_0 = 0$ и $\sigma_0 = 1$.

О сравнительной точности оценивания можно судить по значениям среднеквадратичного отклонения закона, описывающего распределение соответствующих оценок при конкретных объемах выборок. Значения среднеквадратичного отклонения характеризуют рассеяние оценок. Например, в табл. 5 для различных объемов выборок представлены значения среднеквадратичных отклонений (СКО) для ОМП по точечным выборкам $\hat{\mu}$ и $\widehat{\sigma}$ и для L-оценок $\tilde{\mu}$ и $\widetilde{\sigma}$ параметров сдвига и масштаба логистического закона при $k = 5$. Характеристики рассеяния для ОМП по группированным наблюдениям в данном случае совпадают с характеристиками рассеяния L-оценок. В то же время следует отметить, что в общем случае ОМП по группированным наблюдениям все-таки несколько точнее. Исследования при конечных объемах выборок распределений оценок параметров сдвига и масштаба, рассматриваемых в данной работе законов, показали, что всегда выполняются неравенства $D[\hat{\mu}_\Gamma] \le D[\tilde{\mu}]$ и $D[\widehat{\sigma}_\Gamma] \le D[\widetilde{\sigma}]$. Однако если это преимущество и оказывается за ОМП по группированным наблюдениям, то оно незначительно.

Таблица 5

Объем выборки n	ОМП по точечной выборке		L-оценки	
	СКО $\hat{\mu}$	СКО $\widehat{\sigma}$	СКО $\tilde{\mu}$	СКО $\widetilde{\sigma}$
100	0,0947	0,0833	0,0997	0,0927
300	0,0550	0,0482	0,0577	0,0541
500	0,0426	0,0373	0,0446	0,0420
1000	0,0301	0,0264	0,0315	0,0297
2000	0,0214	0,0187	0,0224	0,210

5. Точность *L*-оценок в зависимости от числа используемых квантилей. Характер изменения точности L-оценок с ростом числа используемых квантилей при фиксированном объеме выборки n показывают рис. 4 и 5. На этих рисунках приведены центрированные относительно истинных значений параметров μ_0 и σ_0 плотности оценок $\tilde{\mu}$ и $\widetilde{\sigma}$ параметров нормального закона при объеме выборки $n = 500$ и различном числе $k - 1$ используемых выборочных квантилей для случая одновременного оценивания двух параметров. Выборки нормального закона, как и в предыдущем случае, генерировались с параметрами $\mu_0 = 0$ и $\sigma_0 = 1$. Для сравнения на рисунках представлены центрированные распределения ОМП $\hat{\mu}$ и $\widehat{\sigma}$, полученные также в результате моделирования. Сохраняемое различие в законах распределения ОМП и L-оценок при $k = 9$ связано с величиной относительной асимптотической информации о параметрах закона $\det \mathbf{J}_\Gamma(\tilde{\mu}, \widetilde{\sigma}) / \det \mathbf{J}(\hat{\mu}, \widehat{\sigma})$. Эта величина определяет часть информации, сохраняющейся при группировании выборки (при переходе к выборочным квантилям) и составляющую в данном случае величину $0,8753$.

6. Распределения статистики χ^2 Пирсона при использовании *L*-оценок. При анализе наблюдений случайных величин оценивание параметров модели наблюдаемого закона всегда оказывается лишь первым этапом. Следующим этапом является проверка адекватности построенной модели наблюдаемым данным. Проверка адекватности найденной теоретической модели закона распределения наблюдаемому эмпирическому распределению осуществляется с использованием критериев согласия. Если мы проверяем согласие по той же выборке, по которой оценивали и параметры, то имеем дело с проверкой сложной гипотезы. В этом случае *предельное распределение статистики* любого критерия согласия (касается ли это критериев типа χ^2 или непараметрических критериев типа Колмогорова и типа ω^2 Мизеса) *зависит от применяемого метода оценивания* параметров. И для того, чтобы воспользоваться каким-либо критерием согласия, вычислив L-оценки, необходимо знать (предельное) распределение статистики этого критерия, соответствующее данной проверяемой сложной гипотезе.

В частности, при справедливости сложной проверяемой гипотезы H_0 предельным распределением $G(X_n^2|H_0)$ статистики критерия согласия Пирсона

$$X_n^2 = n \sum_{i=1}^{k} \frac{(n_i/n - P_i(\theta))^2}{P_i(\theta)},$$

где n — объем выборки, n_i — количество наблюдений, попавших в i-й интервал, $P_i(\theta) = \int\limits_{x_{(i-1)}}^{x_{(i)}} f_0(x, \theta)\, dx$ — вероятность попадания наблюдения в интер-

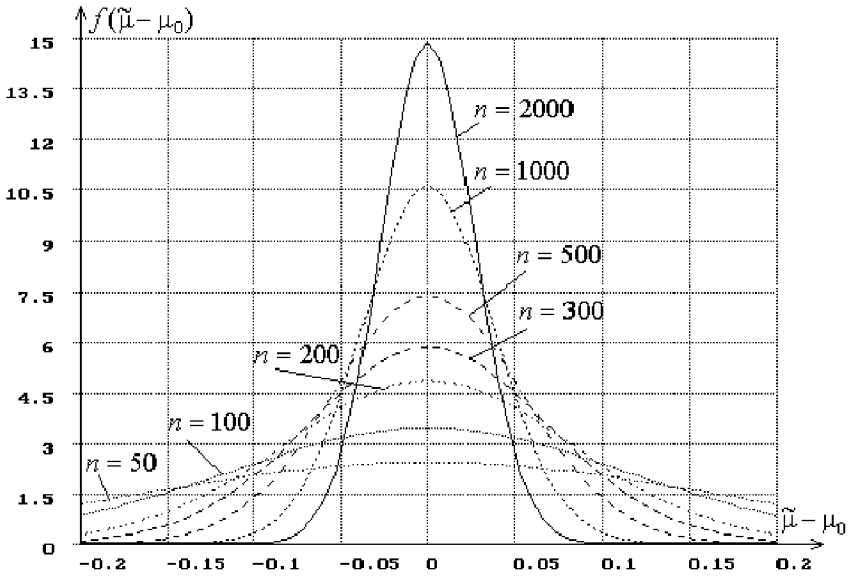

Рис. 2. Плотности распределения L-оценок $\tilde{\mu}$ при $k = 3$ в зависимости от n

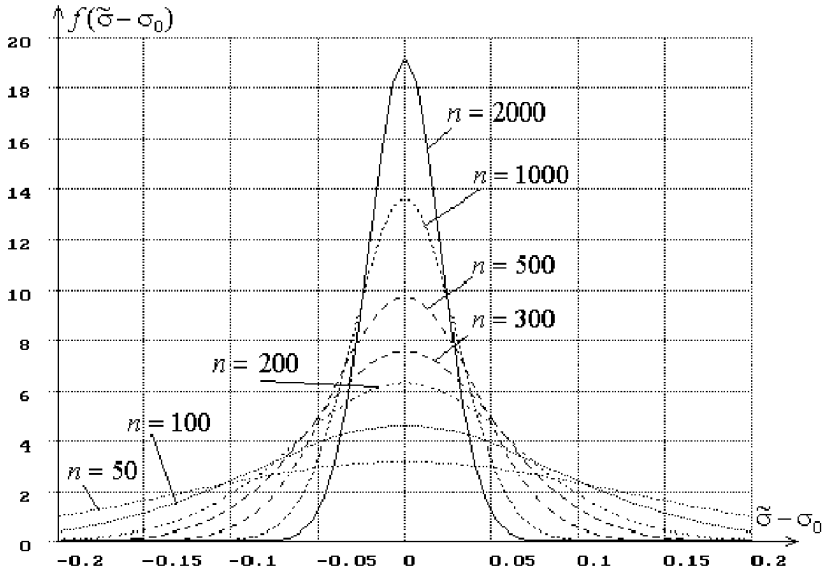

Рис. 3. Плотности распределения L-оценок $\tilde{\sigma}$ при $k = 3$ в зависимости от n

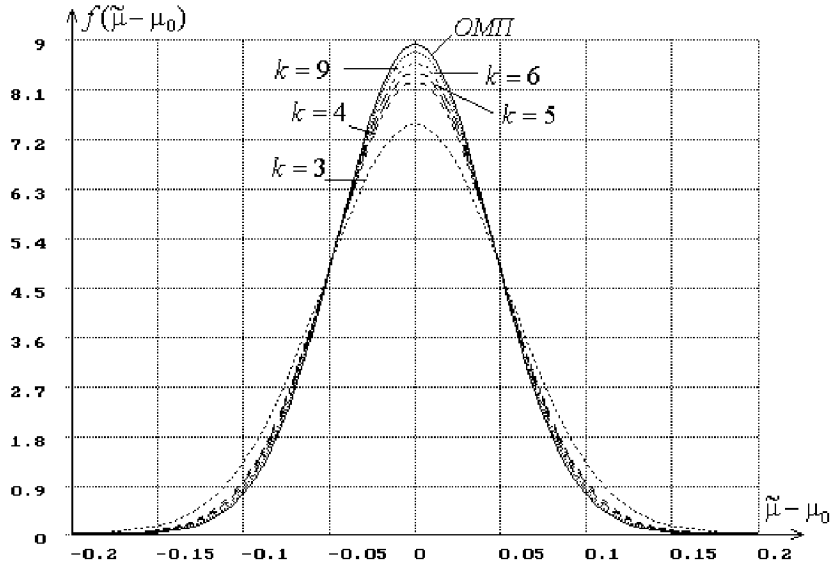

Рис. 4. Плотности распределения L-оценок $\tilde{\mu}$ при $n = 500$ в зависимости от k

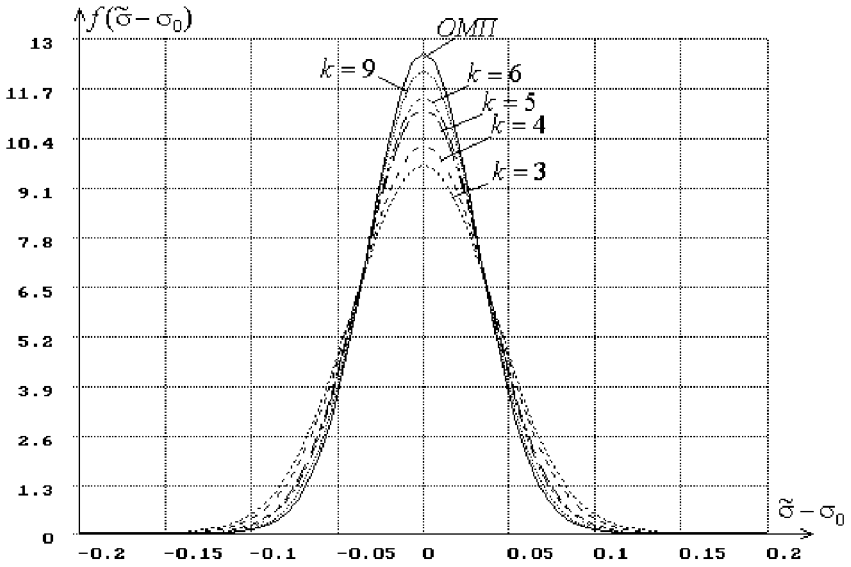

Рис. 5. Плотности распределения L-оценок $\tilde{\sigma}$ при $n = 500$ в зависимости от k

вал, θ — вектор параметров закона с плотностью $f_0(x,\theta)$, относительно которого проверяется гипотеза, $x_{(i)}$ — граничные точки интервалов, является χ^2_{k-m-1}-распределение в том случае, если m компонентов вектора параметров закона оцениваются по этой же выборке в результате минимизации этой же статистики. Статистика X^2_n подчиняется χ^2_{k-m-1}-распределению и в том случае, если используются ОМП по группированным наблюдениям (см. [20, с. 563–567; 21, с. 460–470; 22]). Последнее подтверждают и наши исследования методами статистического моделирования, которые показали хорошее согласие получаемых эмпирических распределений статистики X^2_n с χ^2_{k-m-1}-распределениями при проверке сложных гипотез с использованием ОМП по группированным наблюдениям (при конечных объемах выборок).

Начиная исследование распределений статистики X^2_n при проверке сложных гипотез с использованием L-оценок, мы надеялись на справедливость наших предположений о том, что и в данном случае предельными распределениями статистики являются χ^2_{k-m-1}-распределения. Действительно, статистическое моделирование распределений статистики X^2_n с использованием L-оценок (для различных наблюдаемых законов; при различном числе используемых квантилей, которое соответствует числу интервалов группирования при вычислении статистики; при различном числе оцениваемых параметров) и последующий анализ показали очень хорошее согласие получаемых эмпирических распределений статистики с соответствующими χ^2_{k-m-1}-распределениями.

Например, на рис. 6 представлены эмпирическая функция распределения статистики X^2_n при проверке согласия с экспоненциальным законом распределения в случае использования L-оценок масштабного параметра этого закона при объеме выборок $n = 500$ и числе интервалов $k = 5$ и функция χ^2_3-распределения ($k-m-1=3$). Эмпирическая функция распределения построена по выборке из $N = 2000$ смоделированных значений статистики X^2_n. На рис. 7 приведена аналогичная картина, соответствующая проверке согласия с нормальным законом распределения при использовании L-оценок параметров сдвига и масштаба, также при объеме выборок $n = 500$ и числе интервалов $k = 5$. В этом случае число степеней свободы предельного χ^2-распределения $k - m - 1 = 2$. Как видим, на приводимых рисунках эмпирические функции распределений статистики визуально практически совпадают с теоретическими χ^2-распределениями. Проверка гипотез о согласии с χ^2-распределениями по критериям (χ^2 Пирсона, отношения правдоподобия, Колмогорова, ω^2 и Ω^2 Мизеса), реализованным в [23], подтвердила очень хорошее согласие.

χ^2_r-Распределения являются частным случаем гамма-распределения с плотностью

$$f(x) = \frac{1}{\sigma^\theta \Gamma(\theta)} x^{\theta-1} e^{-x/\sigma},$$

в котором параметр формы $\theta = r/2$ и параметр масштаба $\sigma = 2$. Наилучшей моделью для эмпирических распределений статистики X^2_n, получаемых в результате моделирования, оказались гамма-распределения. При повторении испытаний, указанных в предыдущем абзаце, была получена серия из 10 эмпирических распределений, каждое из которых было сглажено гамма-распределением, параметры которого оценивались по выборке значений статистик. Средние значения параметров гамма-распределения по серии из 10 экспериментов, соответствующих проверке согласия с экспоненциальным законом, составили: $\theta = 1,02405$; $\sigma = 1,966607$ (вместо положенных для χ^2-распределения значений параметров соответственно 1 и 2). А для рассмотренной выше ситуации проверки согласия с нормальным законом получены параметры гамма-распределения $\theta = 1,51723$, $\sigma = 2,003205$ (вместо $1,5$ и 2). Очевидно, что усреднение по большему числу реализаций приведет нас к соответствующим χ^2-распределениям.

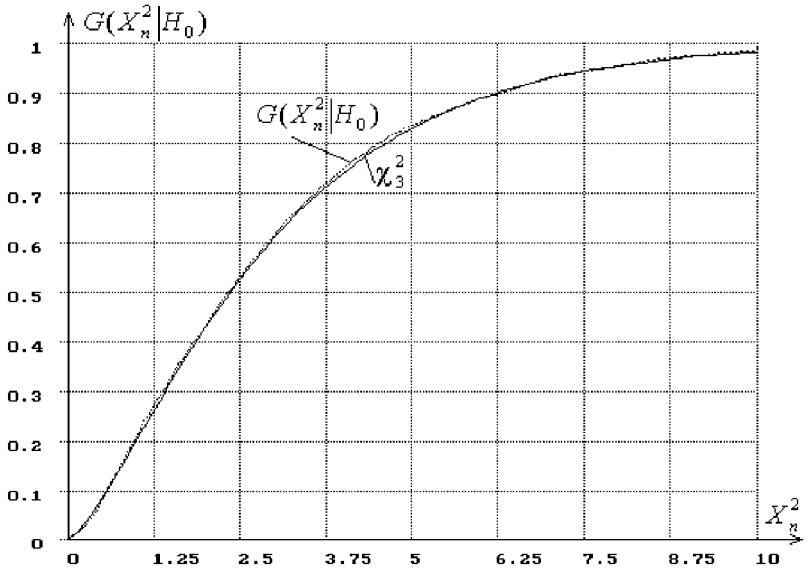

Рис. 6. Распределение статистики X_n^2 с использованием L-оценок параметра экспоненциального распределения при $n = 500$, $k = 5$

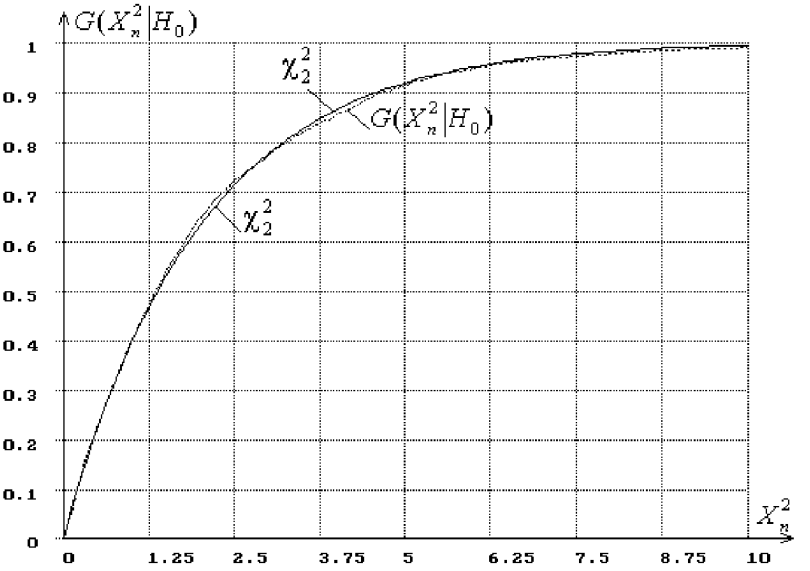

Рис. 7. Распределение статистики X_n^2 с использованием L-оценок параметров сдвига и масштаба нормального распределения при $n = 500$, $k = 5$

Заключение. L-Оценки асимптотически эквивалентны ОМП по группированным наблюдениям, и асимптотические дисперсионные матрицы этих оценок определяются соотношением (14). Но при конечных n разница между свойствами этих оценок все же заметна. Дисперсионные матрицы оценок практически совпадают при $n \geq 2000$, а при меньших объемах выборок преимущество, хотя и незначительное, за ОМП по группированным данным.

Преимущество L-оценок в другом. Определение ОМП по группированным наблюдениям всегда, а ОМП по точечным выборкам за редким исключением (например, экспоненциальный и нормальный законы) связано с проблемами вычислительного характера, так как требуется реализация итерационного процесса для определения максимума функции правдоподобия или решения системы уравнений правдоподобия. По сравнению с этим вычисление L-оценок параметров сдвига и масштаба реализуется элементарно. При этом самой трудоемкой операцией является процедура упорядочивания исходных наблюдений. Применение таблиц вероятностей попадания в интервал, соответствующих асимптотически оптимальному группированию, и формул (10)–(13), опирающихся на вычисленные таблицы коэффициентов, позволяют легко получать оптимальные оценки параметров сдвига и масштаба для больших выборок.

Использование L-оценок не вызывает проблем в применении критериев согласия типа χ^2 Пирсона и отношения правдоподобия, так как распределениями статистик этих критериев являются χ^2_{k-m-1}-распределения. А применение готовых таблиц вероятностей попадания в интервал, соответствующих асимптотически оптимальному группированию, делает элементарной и процедуру вычисления статистики X^2_n.

Как и все оценки по группированным данным, L-оценки являются робастными. Они устойчивы к наличию аномальных ошибок измерений, к малым отклонениям от исходных предположений о виде наблюдаемого закона распределения.

Все вышесказанное позволяет настоятельно рекомендовать использование L-оценок в приложениях. Полный состав таблиц, которыми можно воспользоваться при вычислении L-оценок и проверке гипотез о согласии, представлен на указанном выше WEB-сайте.

ЛИТЕРАТУРА

1. *Сархан А. Е., Гринберг Б. Г.* Введение в теорию порядковых статистик. М.: Статистика, 1970.
2. *Лемешко Б. Ю.* Группирование наблюдений как способ получения робастных оценок // Надежность и контроль качества. 1997. № 5. С. 26–35.
3. *Лемешко Б. Ю.* Робастные методы оценивания и отбраковка аномальных измерений // Заводская лаборатория. 1997. Т. 63, № 5. С. 43–49.
4. *Hampel F. R.* The influence curve and its role in robust estimation // J. Amer. Statist. Assoc. 1974. V. 69, N 346. P. 383–393.
5. *Хьюбер П.* Робастность в статистике. М.: Мир, 1984.
6. *Шуленин В. П.* Введение в робастную статистику. Томск: Изд-во Томск. гос. ун-та, 1993.
7. *Mosteller F.* On some useful inefficient statistics // An. Math. Statist. 1946. V. 17. P. 377–407.
8. *Ogawa J.* Contributions to the theory of systematic statistics // Inst. Osaka Math. J. 1951. N 3. P. 175–213.
9. *Лемешко Б. Ю.* Оптимальные оценки параметров сдвига и масштаба по выборочным квантилям для больших выборок // Тр. 3 Междунар. науч.-техн. конф. АПЭП–96. Т. 6. Ч. I. Новосибирск, 1996. С. 37-44.
10. *Куллдорф Г.* Введение в теорию оценивания по группированным и частично группированным выборкам. М.: Наука, 1966.
11. *Денисов В. И., Лемешко Б. Ю., Цой Е. Б.* Оптимальное группирование, оценка параметров и планирование регрессионных экспериментов. В 2 ч. Новосибирск: Изд-во Новосиб. гос. техн. ун-та, 1993.
12. *Eisenberger J., Posner E. C.* Systematic statistics used for data compression in space telemetry // J. Amer. Statist. Assoc. 1965. V. 60. P. 97–133.

13. *Saleh A. K. M. J., Ali M. M.* Asymptotic optimum quantiles for the estimation of the parameters of the negative exponential distribution // An. Math. Statist. 1966. V. 37. P. 143–151.
14. *Gupta S. S., Gnanadesikan M.* Estimation of the parameters of the logistic distribution // Biometrika. 1966. V. 53. P. 565–570.
15. *Bloch D.* A note on the estimation of the location parameter of the Cauchy distribution // J. Amer. Statist. Assoc. 1966. V. 61. P. 852–855.
16. *Hassanein K. M.* Analysis of extreme-value data by sample quantiles for very large samples // J. Amer. Statist. Assoc. 1968. V. 63. P. 877–888.
17. *Sarndal C. E.* Estimation of the parameters of the gamma distribution by sample quantiles // Technometrics. 1964. N 6. P. 405–414.
18. *Денисов В. И., Лемешко Б. Ю., Постовалов С. Н.* Прикладная статистика. Правила проверки согласия опытного распределения с теоретическим. Методические рекомендации. Ч.I. Критерии типа χ^2. Новосибирск: Изд-во Новосиб. гос. техн. ун-та, 1998.
19. *Дэйвид Г.* Порядковые статистики. М.: Наука, 1979.
20. *Кендалл М., Стьюарт А.* Статистические выводы и связи. М.: Наука, 1973.
21. *Крамер Г.* Математические методы статистики. М.: Мир, 1975.
22. *Birch M. W.* A new proof of the Pearson—Fisher theorem // An. Math. Statist. 1964. V. 35. P. 817.
23. *Лемешко Б. Ю.* Статистический анализ одномерных наблюдений случайных величин: Программная система. Новосибирск: Изд-во Новосиб. гос. техн. ун-та, 1995.

г. Новосибирск
Новосибирский гос. техн. университет
e-mail: headrd@fpm.ami.nstu.ru
chim@mail.ru

Статья поступила 11 июля 2001 г.