Оптимальные оценки параметров сдвига и масштаба по выборочным квантилям для больших выборок

Б.Ю. Лемешко

Кафедра прикладной математики, НГТУ, Новосибирск, Россия

Аннотация. Рассмотрено применение асимптотически оптимального группирования данных, минимизирующего потери фишеровской информации. Использование решений задачи асимптотически оптимального группирования максимизирует мощность критериев согласия при близких альтернативах, гарантирует минимум асимптотической дисперсии оценок по группированным данным. На основе таблиц оптимального группирования получены простые оптимальные оценки параметров сдвига и масштаба в виде линейной формы квантилей, соответствующих оптимальному группированию. Полученные для ряда распределений коэффициенты линейных форм сведены в таблицы.

Опираясь на асимптотическое распределение k выборочных квантилей, Огавой выведено асимптотическое распределение выборочных квантилей для функции плотности, зависящей только от параметра расположения μ и от параметра рассеяния σ и получены линейные несмещенные оценки параметров сдвига и масштаба по методу наименьших квадратов, в основе которых лежат значения выборочных квантилей [1].

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

Выражения для оценок параметров μ и σ, приводимые в [1], можно представить в виде линейных форм, зависящих от выборочных квантилей. Так оценка параметра μ при известном σ будет иметь вид

$$\tilde{\mu} = \alpha_0 \sigma + \sum_{i=1}^{k-1} \alpha_i x_{(i)}, \qquad (1)$$

где

$$\alpha_0 = -\frac{K_3}{K_1};$$

$$\alpha_1 = \alpha_1' / K_1 = \left(\frac{f_1^2}{P_0} + \frac{f_1^2 - f_1 f_2}{P_1} \right) \Big/ K_1;$$

$$\alpha_i = \alpha_i' / K_1 = \left(\frac{f_i^2 - f_i f_{i-1}}{P_{i-1}} + \frac{f_i^2 - f_i f_{i+1}}{P_i} \right) \Big/ K_1, \quad i = \overline{2,(k-2)};$$

$$\alpha_{k-1} = \alpha_{k-1}' / K_1 = \left(\frac{f_{k-1}^2 - f_{k-1} f_{k-2}}{P_{k-2}} + \frac{f_{k-1}^2}{P_{k-1}} \right) \Big/ K_1;$$

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

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

$f_i = f(t_i)$, $t = (x-\mu)/\sigma$, $P_i = F(t_i) - F(t_{i-1})$. Здесь $x_{(i)}$ - выборочная квантиль, такая, что $F\left(\frac{x_{(i)}-\mu}{\sigma}\right) = F(t_i)$ и t_i - квантиль распределения с нулевым параметром сдвига и единичным масштабным. Через $J_\Gamma(\cdot)$ - обозначено количество информации Фишера о соответствующем параметре. Информационная матрица Фишера о векторе параметров будет иметь вид

$$M_\Gamma(\mu,\sigma) = \sum_{i=1}^{k} \frac{\nabla P_i \nabla^{\mathrm{T}} P_i}{P_i} = \begin{bmatrix} J_\Gamma(\mu) & J_\Gamma(\mu,\sigma) \\ J_\Gamma(\mu,\sigma) & J_\Gamma(\sigma) \end{bmatrix},$$

где ∇P - вектор-градиент по параметрам.

Оценку σ при известном μ можно представить в виде

$$\tilde{\sigma} = \beta_0 \mu + \sum_{i=1}^{k-1} \beta_i x_{(i)}, \qquad (2)$$

где

$$\beta_0 = -\frac{K_3}{K_2};$$

$$\beta_1 = \beta_1' / K_2 = \left(\frac{t_1 f_1^2}{P_0} + \frac{t_1 f_1^2 - t_2 f_1 f_2}{P_1} \right) \Big/ K_2;$$

$$\beta_i = \beta_i' / K_2 = \left(\frac{t_i f_i^2 - t_{i-1} f_i f_{i-1}}{P_{i-1}} + \frac{t_i f_i^2 - t_{i+1} f_i f_{i+1}}{P_i} \right) \Big/ K_2, \quad i = \overline{2,(k-2)};$$

$$\beta_{k-1} = \beta_{k-1}' / K_2 = \left(\frac{t_{k-1} f_{k-1}^2 - t_{k-2} f_{k-1} f_{k-2}}{P_{k-2}} + \frac{t_{k-1} f_{k-1}^2}{P_{k-1}} \right) \Big/ K_2;$$

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

При одновременном оценивании μ и σ оценки будут выглядеть следующим образом

$$\tilde{\mu} = \sum_{i=1}^{k-1} \gamma_i x_{(i)}, \qquad (3)$$

$$\tilde{\sigma} = \sum_{i=1}^{k-1} \upsilon_i x_{(i)}. \qquad (4)$$

где $\gamma_i = \left(\alpha_i' K_1 - \beta_i' K_3\right)/\Delta$, $\upsilon_i = \left(-\alpha_i' K_3 + \beta_i' K_1\right)/\Delta$, $\Delta = K_1 K_2 - K_3^2$.

Все эти оценки асимптотически эффективны и их асимптотические дисперсии определяются количеством информации Фишера по группированным данным, а в случае векторного параметра информационной матрицей по группированным данным.

Коэффициенты α_i, β_i, γ_i, υ_i зависят от граничных точек t_i (квантилей стандартизованного распределения). Очевидно, что так как рассматриваемые оценки асимптотически эффективны, то использование квантилей (граничных точек интервалов), соответствующих асимптотически оптимальному группированию, обеспечит оптимальные свойства этих оценок: минимум асимптотической дисперсии, а в случае оценивания сразу двух параметров - минимум обобщенной асмптотической дисперсии. Опираясь на построенную нами совокупность таблиц асимптотически оптимального группирования [2], значения коэффициентов α_i, β_i, γ_i, υ_i вычислены и сведены в соответствующие таблицы. Если в случае *больших* выборок мы будем выбирать $x_{(i)}$ таким образом, чтобы $n_i \approx nP_i$, где P_i - соответствует вероятности попадания в интервал при асимптотически оптимальном группировании [2], и использовать соответственно формулы (1), (2), (3) и (4) с полученными коэффициентами, то будем получать оптимальные оценки.

Таблицы коэффициентов для формул вида (1-4) сформированы для нормального распределения, для логистического распределения с функцией плотности

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

для распределения Коши с плотностью

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

для распределения наименьшего экстремального значения с плотностью

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

для 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 параметра, наборам коэффициентов α_i, β_i и паре γ_i, υ_i соответствуют свои таблицы асимптотически оптимального группирования. В частности, для нормального распределения полученные значенья коэффициентов γ_i, υ_i представлены в табл. 1 - 2. При определении $x_{(i)}$ вероятности P_i должны выбираться из табл. 3 [2].

Для распределений экспоненциального с плотностью

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

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

Таблица 1

Нормальное распределение. Неизвестны оба параметра.
Коэффициенты для оценивания параметра сдвига μ.

k	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7
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	0.029871
9	0.021547	0.068108	0.148605	0.261739	0.261739	0.148605	0.068108
10	0.016187	0.050213	0.107748	0.196679	0.258345	0.196679	0.107748
11	0.012568	0.037666	0.080500	0.145524	0.223741	0.223741	0.145524
12	0.002356	0.078666	0.032450	0.099837	0.188002	0.197378	0.188002
13	0.008056	0.023716	0.048181	0.086640	0.138225	0.195182	0.195182
14	0.006737	0.018173	0.039623	0.068299	0.109776	0.161527	0.191732
15	0.005076	0.015581	0.032157	0.055371	0.088028	0.130918	0.172869

Продолженние табл. 1

k	γ_8	γ_9	γ_{10}	γ_{11}	γ_{12}	γ_{13}	γ_{14}
3							
4							
5							
6							
7							
8							
9	0.021547						
10	0.050213	0.016187					
11	0.080500	0.037666	0.012568				
12	0.099837	0.032450	0.078666	0.002356			
13	0.138225	0.086640	0.048181	0.023716	0.008056		
14	0.161527	0.109776	0.068299	0.039623	0.018173	0.006737	
15	0.172869	0.130918	0.088028	0.055371	0.032157	0.015581	0.005076

Таблица 2

Нормальное распределение. Неизвестны оба параметра.
Коэффициенты для оценивания масштабного параметра σ.

k	υ_1	υ_2	υ_3	υ_4	υ_5	υ_6	υ_7
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	0.070411
9	-0.052747	-0.114684	-0.153492	-0.090860	0.090860	0.153492	0.114684
10	-0.040995	-0.091463	-0.132388	-0.123812	0	0.123812	0.132388
11	-0.032533	-0.073373	-0.111849	-0.124976	-0.064980	0.064980	0.124976
12	-0.019239	-0.098971	-0.069069	-0.112065	-0.080404	0	0.080404
13	-0.021688	-0.050048	-0.079521	-0.102566	-0.102357	-0.048835	0.048835
14	-0.018214	-0.041087	-0.068460	-0.090418	-0.099244	-0.075434	0
15	-0.014676	-0.035496	-0.058652	-0.079499	-0.092263	-0.085051	-0.038353

Продолжение табл. 2

k	υ_8	υ_9	υ_{10}	υ_{11}	υ_{12}	υ_{13}	υ_{14}
3							
4							
5							
6							
7							
8							
9	0.052747						
10	0.091463	0.040995					
11	0.111849	0.073373	0.032533				
12	0.112065	0.069069	0.098971	0.019239			
13	0.102357	0.102566	0.079521	0.050048	0.021688		
15	0.075434	0.099244	0.090418	0.068460	0.041087	0.018214	
15	0.038353	0.085051	0.092263	0.079499	0.058652	0.035496	0.014676

Таблица 3

Вероятности (частоты) 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	P_8
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	0.0141
9	0.0102	0.0422	0.1009	0.1976	0.2982	0.1976	0.1009	0.0422
10	0.0077	0.0317	0.0748	0.1438	0.2420	0.2420	0.1438	0.0748
11	0.0059	0.0243	0.0567	0.1074	0.1823	0.2468	0.1823	0.1074
12	0.0047	0.0190	0.0442	0.0829	0.1392	0.2100	0.2100	0.1392
13	0.0037	0.0152	0.0352	0.0652	0.1085	0.1670	0.2104	0.1670
14	0.0030	0.0124	0.0283	0.0524	0.0862	0.1327	0.1850	0.1850
15	0.0025	0.0101	0.0232	0.0427	0.0698	0.1066	0 1532	0.1838

Продолжение табл. 3

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
							0.8474
0.0102							0.8753
0.0317	0.0077						0.8960
0.0567	0.0243	0.0059					0.9121
0.0829	0.0442	0.0190	0.0047				0.9247
0.1085	0.0652	0.0352	0.0152	0.0037			0.9348
0.1327	0.0862	0.0524	0.0283	0.0124	0.0030		0.9430
0.1532	0.1066	0.0698	0.0427	0.0232	0.0101	0.0025	0.9498

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

частными случаями которого являются полунормальное распределение - $m=1$, Рэлея - $m=2$ и Максвелла - $m=3$ таблицы коэффициентов α_i, β_i, γ_i, υ_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 количества информации о μ для построения асимптотически оптимальных граничных точек относительно этого параметра.

Значения $x_{(i)}$, фигурирующие в формулах (1-4), следует выбирать из условия

$$X_{\left(\left[nP^i\right]\right)}\le x_{(i)}\le X_{\left(\left[nP^i\right]+1\right)},$$

где $X_{(i)}$ - члены вариационного ряда $X_{(1)}\le X_{(2)}\le\ldots\le X_{(n)}$, построенного по исходной выборке, $P^i=\sum_{j=1}^{i}P_j$, $[\cdot]$ - означает целую часть числа, а P_j - выбираются из соответствующей строки таблицы оптимальных вероятностей. Это могут быть средние значения между соответствующими соседними членами вариационного ряда.

Пример. Для нормального распределения при $k=9$ соотношения (3-4) принимают вид (см. табл. 1-2) -

$$\tilde{\mu}=0.021547(x_{(1)}+x_{(8)})+0.068108(x_{(2)}+x_{(7)})+$$
$$+0.148605(x_{(3)}+x_{(6)})+0.261739(x_{(4)}+x_{(5)}),$$
$$\tilde{\sigma}=0.052747(-x_{(1)}+x_{(8)})+0.114684(-x_{(2)}+x_{(7)})+$$
$$+0.153492(-x_{(3)}+x_{(6)})+0.090860(-x_{(4)}+x_{(5)}).$$

Для смоделированной нормальной выборки объемом 1000 наблюдений оценки максимального правдоподобия (ОМП) по негруппированным данным оказались равны $\hat{\mu}=-0.01672934$, $\hat{\sigma}=0.9849848$, оценки по формулам (3) и (4) при $k=9$ - $\tilde{\mu}=-0.018158$, $\tilde{\sigma}=0.975442$. При определении $x_{(i)}$ вероятности P_i выбирались из табл. 3 и в качестве $x_{(i)}$ принимались средние значения между следующими парами членов вариационного ряда $X_{(10)}-X_{(11)}$, $X_{(52)}-X_{(53)}$, $X_{(153)}-X_{(154)}$, $X_{(350)}-X_{(351)}$, $X_{(649)}-X_{(650)}$, $X_{(846)}-X_{(847)}$, $X_{(947)}-X_{(948)}$, $X_{(989)}-X_{(990)}$. Для сравнения вычисленные ОМП по группированным данным с использованием тех же самых граничных точек равны $\hat{\mu}_\Gamma=-0.01680601$ и $\hat{\sigma}_\Gamma=0.9766322$. Естественно, что при проверке критериев согласия полученных теоретических законов с исходной выборкой наилучшие результаты получены для ОМП. Для проверки согласия применялись критерии отношения правдоподобия, χ^2 Пирсона, Колмогорова, Смирнова, ω^2 и

Ω^2 Мизеса [3]. Заслуживает внимания факт, что по всем этим критериям оценки $\tilde{\mu}$ и $\tilde{\sigma}$ оказались *предпочтительнее* оценок $\hat{\mu}_\Gamma$ и $\tilde{\sigma}_\Gamma$. Результаты статистического анализа для оценок $\tilde{\mu}$ и $\tilde{\sigma}$ представлены на рис. 1. На этом рисунке приведены вычисленные значения статистик соответствующих критериев S^* и вероятности вида $P\{S > S^*\}$, позволяющие судить о степени согласия. Гипотезы о согласии не отвергаются, если задаваемый уровень значимости $\alpha < P\{S > S^*\}$.

Аналогичные результаты характерны для оценок параметров сдвига и масштаба для законов распределений Коши, логистического, наибольшего и наименьшего экстремальных значений.

Следует отметить, что рассмотренные оценки, как и все оценки по группированным данным являются робастными. Они устойчивы к наличию аномальных ошибок измерений, к малым отклонениям от исходных предположений о виде наблюдаемого закона распределения.

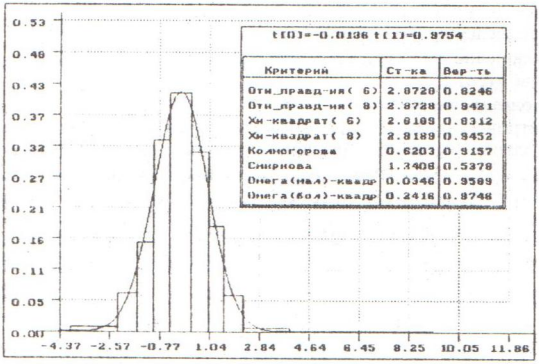

Рис. 1. Результаты статистического анализа для нормального закона с параметрами $\tilde{\mu}$ и $\tilde{\sigma}$.

Литература

1. Сархан А.Е., Гринберг Б.Г. Введение в теорию порядковых статистик. - М.: Статистика, 1970. - 414 с.
2. Денисов В.И., Лемешко Б.Ю., Цой Е.Б. Оптимальное группирование, оценка параметров и планирование регрессионных экспериментов. В 2-х ч. / Новосиб. гос. техн. ун-т. - Новосибирск, 1993. - 347 с.
3. Лемешко Б.Ю. Статистический анализ одномерных наблюдений случайных величин: Программная система / Новосиб. гос. техн. ун-т. - Новосибирск, 1995. - 125 с.