МИНИСТЕРСТВО ВЫСШЕГО И СРЕДНЕГО СПЕЦИАЛЬНОГО ОБРАЗОВАНИЯ
РСФСР

НОВОСИБИРСКИЙ ЭЛЕКТРОТЕХНИЧЕСКИЙ ИНСТИТУТ

В.И.ДЕНИСОВ, Г.Г.ЗАЧЕПА, Б.Ю.ЛЕМЕШКО

№ 3338 - 75 Деп.

УДК 519.24

ОБ АСИМПТОТИЧЕСКИ ОПТИМАЛЬНОМ ГРУППИРОВАНИИ ПРИ ОЦЕНИВАНИИ ПАРАМЕТРОВ ПО ГРУППИРОВАННЫМ ДАННЫМ

Новосибирск - 1975

При оценивании параметров непрерывных распределений методом максимума функции правдоподобия по группированным наблюдениям значительный интерес представляет асимптотическое поведение оценок. Дело в том, что при малых объемах выборки оценки параметров по группированным наблюдениям малопривлекательны. Оценки не всегда существуют и единственны, получение оценок сопряжено с решением сложных трансцендентных уравнений, оценки, вообще говоря, оказываются смещенными, судить об их эффективности затруднительно из-за невозможности получить аналитическое выражение для дисперсии и т.п. В то же время, как показал Куллдорф [I] и Бодин [2], выполнение определенных условий достаточно для того, чтобы при больших объемах выборки оценки параметров непрерывных распределений по группированным наблюдениям существовали почти наверное, были асимптотически состоятельны и асимптотически эффективны по Вальду.

Напомним определение состоятельности и асимптотической эффективности оценок параметров по группированным наблюдениям [2].

Пусть случайная величина X распределена на множестве $\mathcal{X}$ с плотностью $f(x;\theta)$; $x_1 < x_2 < \cdots < x_{k-1}$ - граничные точки разбиения множества $\mathcal{X}$ на k интервалов $R_i = (x_{i-1}, x_i]$, $i = 1,2,\ldots,k$, где $x_0 = \inf \mathcal{X}$, $x_k = \sup \mathcal{X}$; N_i, $i = 1,\ldots,k$ - групповые частоты, показывающие, сколько выборочных значений случайной величины X попало в R_i; $n = \sum_{i=1}^{k} N_i$ - объем выборки и, наконец $\hat{\theta}_n$ - оценка максимального правдоподобия

параметра θ по группированной выборке

$$N_1, \ldots, N_K ; x_1, \ldots, x_{K-1} .$$

Определение I. Оценка максимального правдоподобия $\hat{\theta}_n$ называется состоятельной, если по любым $\varepsilon > 0$ и $\delta > 0$ можно найти такое $n_0 = n_0(\varepsilon, \delta)$, что для $n > n_0$ имеет место соотношение для вероятности

$$p(|\hat{\theta}_n - \theta| > \delta) < \varepsilon .$$

Определение 2. Оценка максимального правдоподобия $\hat{\theta}_n$ называется асимптотически эффективной, если

а) $\lim\limits_{n \to \infty} E[\sqrt{n}(\hat{\theta}_n - \theta)] = 0 ;$

б) $\lim\limits_{n \to \infty} E[n(\hat{\theta}_n - \theta)^2] = \left[\sum\limits_{i=1}^{K} \left(\frac{\partial}{\partial \theta} \ln p_i\right)^2 \cdot p_i \right]^{-1} = j^{-1} ,$

где $p_i = \int\limits_{R_i} f(t; \theta) dt ;$

в) распределение случайной величины $\sqrt{n}(\hat{\theta}_n - \theta)$ при $n \to \infty$ стремится к нормальному распределению с параметрами 0 и j^{-1} .

Дисперсия предельного нормального распределения асимптотически эффективной оценки $\hat{\theta}_n$ зависит от граничных точек группирования. Поэтому за счет подходящего выбора граничных точек можно сделать эту дисперсию наименьшей и, следовательно, информационное количество Фишера j - наибольшим. Так возникла задача оптимального группирования при оценивании параметров распределения по группированным наблюдениям, по-видимому, впервые поставленная в таком аспекте Куллдорфом [I].

Задача оптимального группирования является задачей оптимизации для нелинейной функции цели при ограничениях вида $0 < x_1 < x_2 < \cdots < x_{K-1} < +\infty$. Она решена для оценивания по группированным наблюдениям:

- математического ожидания нормального распределения при известной его дисперсии;

- среднего квадратического отклонения нормального распределения при известном его математическом ожидании;

- масштабного параметра показательного распределения.

В работе [I] приведены таблицы оптимального группирования для указанных задач оценивания.

В предлагаемой статье излагаются результаты численного исследования оптимального группирования для оценивания основного параметра распределения Вейбулла. В качестве метода оптимизации был взят метод сопряженных градиентов с использованием квадратичной интерполяции при поиске максимума в выбранном направлении. Для приведения задачи оптимизации к безусловной была использована функция штрафа.

Разработанный нами алгоритм аппробирован на задаче оптимального группирования для оценивания по группированным наблюдениям масштабного параметра показательного распределения. У Куллдорфа [I] таблица оптимального группирования составлена для $K = 2,3,4,5,6$. В ходе тестирования алгоритма нами были пересчитаны результаты Куллдорфа для $K = 4,5,6$ и затем таблица была продолжена для $K = 7,8,9,10$ и 11.
Одновременно с этим для каждого значения K вычислялась относительная асимптотическая эффективность A как мера

11250.

Таблица I

Значения граничных точек θx_i оптимального группирования для различных K и соответствующие им значения относительной асимптотической эффективности A (по Куллдорфу)

K	θx_1	θx_2	θx_3	θx_4	θx_5	A
2	I.5936	-	-	-	-	0,6476
3	I.0I76	2.6II2	-	-	-	0.8203
4	0,7540	I,77I6	3,3652	-	-	0,89I0
5	0,6004	I,3545	2,3720	3.9657	-	0,9269
6	0,4994	I,0997	I,8538	2,87I4	4,4650	0,9476

Таблица 2

Значения граничных точек θx_i оптимального группирования для различных K и соответствующие им значения относительной асиптотической эффективности A (по нашим исследованиям.)

K	θx_1	θx_2	θx_3	θx_4	θx_5	θx_6	θx_7	θx_8	θx_9	θx_{10}	A
4	0.7540	I.77I6	3.3652	-	-	-	-	-	-	-	0.89I0
5	0.6004	I.3545	2.3720	3.9657	-	-	-	-	-	-	0.9269
6	0.4993	I.0998	2.85I0	2.87I7	4.4654	-	-	-	-	-	0.9476
7	0.4276	0.9269	I.5273	2.28I3	3.2989	4.8925	-	-	-	-	0.9606
8	0.3739	0.80I5	I.3008	I.90I2	2.6553	3.6729	5.2665	-	-	-	0.9693
9	0.3323	0.7063	I.I338	I.633I	2.2336	2.9876	4.0052	5.5988	-	-	0.9754
I0	0.2990	0.63II	I.0050	I.4326	I,9320	2.5324	3.2863	4.3042	5.8979	-	0.9798
II	0.27I9	0.5709	0,9032	I.2772	I.7047	2.204I	2.8045	3.5585	4.576I	6.I697	0.9832

асимптотической эффективности оценивания по группированным наблюдениям по сравнению с оцениванием по негруппированным наблюдениям. На стр.5,6 приведены для сравнения обе таблицы.

Сравнение контрольных результатов таблицы 2 с соответствующими значениями таблицы I показывает, что при $K = 4,5$ мы имеем полное совпадение, а при $K = 6$ – незначительное отклонение в третьем десятичном знаке для значений θx_i при полном совпадении для A . Это обстоятельство означает, что примененный нами алгоритм численного исследования гарантирует достаточную точность вычислений.

Аппробированный указанным образом алгоритм был использован при решении задачи оптимального группирования для оценивания основного параметра распределения Вейбулла по группированным наблюдениям.

Информационное количество Фишера j для оценок основного параметра θ распределения Вейбулла по группированным наблюдениям имеет вид

$$j = \frac{\left[\exp(-x_1^{\theta})\cdot x_1^{\theta}\cdot \ln x_1\right]^2}{1-\exp(-x_1^{\theta})} + \sum_{i=2}^{K-1} \frac{\left[\exp(-x_i^{\theta})\cdot x_i^{\theta}\ln x_i - \exp(-x_{i-1}^{\theta})\cdot x_{i-1}^{\theta}\cdot \ln x_{i-1}\right]^2}{} +$$

$$+ \exp(-x_{K-1}^{\theta})\cdot x_{K-1}^{2\theta}\cdot \ln^2 x_{K-1} .$$

Так как оптимальное группирование ищется при фиксированном θ , то преобразуем j , положив $x_i^{\theta} = t_i$, $i = 1,2,\ldots,K-1$,

$$j = \frac{1}{\theta^2}\left\{\frac{\exp(-2t_1)\cdot t_1^2\cdot \ln^2 t_1}{1-\exp(-t_1)} + \sum_{i=2}^{K-1} \frac{\left[\exp(-t_i)\cdot t_i \ln t_i - \exp(-t_{i-1})\cdot t_{i-1}\cdot \ln t_{i-1}\right]^2}{} +\right.$$

$$\left. + \exp(-t_{K-1})\cdot t_{K-1}^2\cdot \ln^2 t_{K-1} \right\} .$$

Очевидно, максимум $\dot{J}$ по $x_1, \ldots, x_{K-1}$ будет при фиксированном θ совпадать с максимумом $\dot{J}$ по $t_1, \ldots, t_{K-1}$.

Таким образом задача оптимального группирования для распределения Вейбулла свелась к задаче максимизации выражения

$$\frac{e^{-2t_1} \cdot t_1^2 \cdot \ln^2 t_1}{1 - e^{-t_1}} + \sum_{i=2}^{K-1} \frac{\left(e^{-t_i} \cdot t_i \cdot \ln t_i - e^{-t_{i-1}} \cdot t_{i-1} \cdot \ln t_{i-1}\right)^2}{e^{-t_{i-1}} - e^{-t_i}} + e^{-t_{K-1}} \cdot t_{K-1}^2 \cdot \ln^2 t_{K-1}$$

по $t_1 < t_2 < \ldots < t_{K-1}$.

Результаты вычислений на ЭВМ ОДРА-1204 приведены в таблице 3 (стр.9).

Практическое значение Таблицы 2 и Таблицы 3 заключается в том, что по ним могут быть найдены близкие к оптимальным групповые пределы для соответствующих задач оценивания. Дело в том, что зависимость оптимальных групповых пределов от истинного значения параметра θ, который в задачах оценивания неизвестен, заставляет при пользовании Таблицами придавать параметру θ некоторые априорные значения, которые могут заметно отличаться от истинных значений. Естественно, такой шаг приводит к потере относительной асимптотической эффективности. Но, как указывает Куллдорф [I] величина потери может быть рассчитана заранее и притом достаточно элементарным образом. Рассмотрим соображения, которые кладутся в основу предлагаемой процедуры, на примере показательного распределения. Информационные количества Фишера $\dot{J}$ при оценивании масштабного параметра показательного распределения по группированным и негруппированным наблюдениям равны

Таблица 3

Значения граничных точек x_i^θ оптимального группирования для распределения Вейбулла и соответствующие им значения относительной асимптотической эффективности A .

K	x_1^θ	x_2^θ	x_3^θ	x_4^θ	x_5^θ	x_6^θ	x_7^θ	x_8^θ	x_9^θ	A
2	3.4903	-	-	-	-	-	-	-	-	0.3282
3	0.I4I8	3.289I	-	-	-	-	-	-	-	0.6518
4	0.I505	2.6936	4.5643	-	-	-	-	-	-	0.7481
5	0.05I6	0.2486	2.6I73	4.4970	-	-	-	-	-	0.8235
6	0.0534	0.2580	2.3339	3.6005	5.3934	-	-	-	-	0.8639
7	0.0244	0.II54	0.3260	2.2878	3.5602	5.3523	-	-	-	0.8936
8	0.025I	0.II8I	0.3342	2.I205	3.I036	4.2984	6.0540	-	-	0.9141
9	0.0I35	0.0639	0.I73I	0.3843	2.0935	3.0803	4.2767	6.0333	-	0.9288
I0	0.0I37	0.0649	0.I760	0.39I7	I.9766	2.7906	3.7069	4.8673	6.6006	0.9408

соответственно

$$n\cdot\left[\frac{e^{-2\theta x_1}\cdot x_1^2}{1-e^{-\theta x_1}}+\sum_{i=2}^{\kappa-1}\frac{\left(e^{-\theta x_i}\cdot x_i-e^{-\theta x_{i-1}}\cdot x_{i-1}\right)^2}{e^{-\theta x_{i-1}}-e^{-\theta x_i}}+e^{-\theta x_{\kappa-1}}\cdot x_{\kappa-1}^2\right]$$

и

$$n/\theta^2.$$

Пусть теперь θ - истинное значение параметра, а $\tilde{\theta}=\lambda\theta$ - гипотетическое его значение, где $\lambda=\frac{\tilde{\theta}}{\theta}$ относительная ошибка прогноза. Оптимальное группирование по прогнозу имеет вид

$$x_i = t_i/\lambda\theta,$$

а соответствующая ему относительная асимптотическая эффективность равна

$$\theta^2\cdot\left[\frac{e^{-2t_1/\lambda}\cdot t_1^2/\lambda^2\theta^2}{1-e^{-t_1/\lambda}}+\sum_{i=2}^{\kappa-1}\frac{\left(e^{-t_i/\lambda}\cdot t_i/\lambda\theta-e^{-t_{i-1}/\lambda}\cdot t_{i-1}/\lambda\theta\right)^2}{e^{-t_{i-1}/\lambda}-e^{-t_i/\lambda}}+\frac{e^{-t_{\kappa-1}/\lambda}\cdot t_{\kappa-1}^2}{\lambda^2\theta^2}\right].$$

После сокращения на θ^2 получаем, что относительная асимптотическая эффективность по прогнозируемому оптимальному группированию имеет значение

$$\tilde{A}=\frac{1}{\lambda^2}\left[\frac{e^{-\frac{2t_1}{\lambda}}\cdot t_1^2}{1-e^{-t_1/\lambda}}+\sum_{i=2}^{\kappa-1}\frac{\left(e^{-\frac{t_i}{\lambda}}\cdot t_i-e^{-\frac{t_{i-1}}{\lambda}}\cdot t_{i-1}\right)^2}{e^{-t_{i-1}/\lambda}-e^{-t_i/\lambda}}+e^{-\frac{t_{\kappa-1}}{\lambda}}\cdot t_{\kappa-1}^2\right] \quad (I)$$

Придавая κ определенное значение и беря из соответствующей строки Таблицы 2 t_i и A, можно для различных λ вычислить по формуле (I) относительную асимптотическую эффективность по прогнозируемому оптимальному группированию и в сравнении с A узнать величину потери. Например, при $\kappa = 3$ имеем: $t_1 = 1.0176$, $t_2 = 2.6112$ и $A = 0.8203$. Поэтому для $\lambda = 0.8$ по формуле (I) находим, что $\tilde{A} = 0.8057$. Следовательно, потеря в эффективности равна

$$A - \tilde{A} = 0.0146,$$

что составляет чуть меньше 1.8%.

Результаты подобных вычислений приведены в табл.4 (стр.I2).

Аналогичным образом может быть выведена формула для вычисления относительной асимптотической эффективности $\tilde{A}$ оценки основного параметра θ распределения Вейбулла по прогнозируемому оптимальному группированию и составлена таблица потерь в относительной асимптотической эффективности вследствие ошибки в прогнозе.

Приведем окончательный результат

$$\tilde{A} = \frac{1}{1.8237 \cdot \lambda^2} \cdot \left[\frac{e^{-2t_1^{1/\lambda}} \cdot t_1^{2/\lambda} \cdot \ln^2 t_1}{1 - e^{-t_1^{1/\lambda}}} + \right.$$

$$+ \sum_{i=2}^{k-1} \frac{\left(e^{-t_i^{1/\lambda}} \cdot t_i^{1/\lambda} \cdot \ln t_i - e^{-t_{i-1}^{1/\lambda}} \cdot t_{i-1}^{1/\lambda} \cdot \ln t_{i-1}\right)^2}{e^{-t_{i-1}^{1/\lambda}} - e^{-t_i^{1/\lambda}}} +$$

$$\left. + e^{-t_{k-1}^{1/\lambda}} \cdot t_{k-1}^{2/\lambda} \cdot \ln^2 t_{k-1} \right].$$

Таблица потерь в относительной асимптотической эффективности вследствие ошибки в прогнозе приведена на стр. I3.

Таблица 4

Потери в относительной асимптотической эффективности оценки масштабного параметра показательного распределения вследствие оптимального группирования по ошибочному прогнозу

Ошибка прогноза	Потеря в относительной асимптотической эффективности									
	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$	$K = 7$	$K = 8$	$K = 9$	$K = 10$	$K = 11$
– 80%	96.6%	80.4%	61.9%	47.3%	36.7%	29.0%	23.4%	19.2 %	16.0 %	13.6 %
– 60%	53.5%	29.8%	18.8%	12.9%	9.4%	7.1%	5.6%	4.5%	3.7 %	3.1 %
– 40%	17.7%	9.3%	5.7%	3.9%	2.8%	2.1%	1.6 %	1.3%	1.1%	0.9%
– 20%	3.2%	1.8%	1.1%	0.8%	0.6%	0.4%	0.3%	0.3%	0.2%	0.2%
+ 20%	1.8%	1.1%	0.8%	0.6 %	0.4%	0.3%	0.3%	0.2 %	0.2%	0.17%
+ 40%	5.7 %	3.7 %	2.7 %	2.0%	1.6 %	1.3%	1.0 %	0.9%	0.8%	0.6 %
+ 60%	10.3%	7.0%	5.2%	4.0%	3.2%	2.6%	2.2%	1.9 %	1.6 %	1.4 %
+ 80%	15.0%	10.6%	8.0%	6.3%	5.2%	4.3%	3.7 %	3.2 %	2.8%	2.5%

Таблица 5

Потери в относительной асимптотической эффективности оценки основного параметра распределения Вейбулла вследствие оптимального группирования по ошибочному прогнозу

Ошибка прогноза	Потеря в относительной асимптотической эффективности									
	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$	$K = 7$	$K = 8$	$K = 9$	$K = 10$	$K = 11$
- 80%	100%	99.5%	99.4%	96.9%	96.7%	92.9%	92.4%	—	—	-
- 60%	100%	84.9%	85.3%	74.2%	70.2%	63.3%	56.1%	51.5%	44.3%	-
- 40%	84.8%	53.9%	38.8%	31.2%	23.2%	20.1%	15.6%	14.2%	11.2%	-
- 20%	20.7%	11.2%	7.8%	6.4%	4.5%	4.0%	3.0%	2.8%	2.2%	-
- 10%	4.3%	2.3%	1.7%	1.4%	1.0%	0.9%	0.7%	0.7%	0.5%	-
+ 10%	2.8%	1.5%	1.3%	1.1%	0.9%	0.8%	0.6%	0.6%	0.5%	-
+ 20%	9.1%	4.9%	4.4%	3.6%	3.1%	2.8%	2.4%	2.2%	1.9%	-
+ 40%	24.3%	13.8%	12.8%	10.8%	9.7%	8.7%	7.9%	7.4%	6.7%	-
+ 60%	38.3%	22.7%	21.8%	18.5%	17.3%	15.7%	14.6%	13.7%	12.7%	-
+ 80%	49.6%	30.6%	29.9%	26.4%	24.5%	22.3%	21.1%	19.8%	18.8%	-

I. Г.Куллдорф. Введение в теорию оценивания по группированным и частично группированным выборкам, "Наука", М., I966.

2. Н.А.Бодин. Оценка параметров распределения по группированным выборкам. Труды Матем. ин-та им. В.А.Стеклова АН СССР, т.Ш, "Наука", Л., I970.

Печатается в соответствии с решением Учёного Совета факультета Автоматизированных систем управления Новосибирского электротехнического Института от8 октября I975 года.

В печать 30.10.75.
Тир 1 . Цена 71 коп. Зак 4250.

Производственно-издательский комбинат ВИНИТИ
Люберцы, Октябрьский пр. 403.