УДК 519.25

В.И.Денисов, Б.Ю.Лемешко

ОПТИМАЛЬНОЕ ГРУППИРОВАНИЕ ПРИ ОБРАБОТКЕ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Статистическая обработка экспериментальных наблюдений в большинстве случаев сопровождается группированием данных. Это вызвано, во-первых, необходимостью сжатия больших объемов накапливаемой информации, вследствие чего упрощается математическая обработка результатов; во-вторых, довольно часто при испытаниях на надежность вообще отсутствует понятие индивидуального измерения, а регистрируется лишь число наблюдений, попавших в определенный интервал; в-третьих, группирование может быть вызвано применяемым методом статистической обработки данных.

В настоящей работе обсуждаются методы группирования регистрируемых данных, применяемые при исследовании законов распределения случайных величин. К группированию наблюдений прибегают при построении гистограмм, оценивании параметров распределений, проверке соответствия между гипотетическим и фактически наблюденным распределениями по критерию χ^2 Пирсона или его модификациям.

До последнего времени в практике статистических вычислений при группировании область определения случайной величины разбивают на интервалы, в основном, одним из двух способов [I] : на интервалы равной длины с последующим объединением смежных, если в них попадает малое число наблюдений, и на интервалы равной вероятности. Во втором случае разбиение связано с видом распределения. Равновероятному группированию соответствует ситуация, когда выборку раз-

бивают на группы так, чтобы в каждой оказалось одинаковое количество наблюдений.

Любое группирование приводит к потере информации, понимаемой в самом общем смысле, по сравнению с исходной негруппированной выборкой. А это отрицательно сказывается на качестве результатов статистических выводов.

Оценки параметров распределений по группированным данным находят обычно методом максимального правдоподобия, методом минимума χ^2 или одним из родственных ему методов. При некоторых условиях [2-6] эти оценки существуют, единственны и обладают свойствами состоятельности и асимптотической эффективности. При этом асимптотическая дисперсионная матрица оценки вектора параметров θ распределения определяется соотношением

$$D(\hat{\theta}) = N^{-1} J^{-1}(\hat{\theta}),$$

где N – объем выборки;

$J(\theta)$ – информационная матрица Фишера по группированным данным.

Так как $J(\theta)$ – положительно определенная матрица, то с ростом потерь информации, вызванным группированием, ухудшается дисперсионная матрица оценок параметров закона распределения.

Рассмотрим, как связана информационная матрица $J(\theta)$ с критерием χ^2 Пирсона. Статистика

$$\chi^2 = N \sum_{i=1}^{k} (n_i / N - P_i(\theta))^2 / P_i(\theta),$$

где k – число интервалов группирования;

n_i – число наблюдений, попавших в i-й интервал;

$P_i(\theta)$ – гипотетическая вероятность попадания наблюдения в i-й интервал,

подчиняется распределению χ^2 с k -1 степенью свободы, если верна нулевая гипотеза, и подчиняется нецентральному распределе-

нию χ^2 с тем же числом степеней свободы и параметром нецентральности

$$\lambda = N \sum_{i=1}^{k} (P_i(\theta_1) - P_i(\theta))^2 / P_i(\theta),$$

если верна конкурирующая гипотеза и выборка соответствует распределению того же типа, но с параметром θ_1. Пусть $\theta_1 = \theta + \delta\theta$. Разложим $P_i(\theta_1)$ в ряд Тейлора, пренебрегая членами высшего порядка

$$\lambda \approx N \sum_{i=1}^{k} \frac{(P_i(\theta) + \nabla P_i^T(\theta)\delta\theta - P_i(\theta))^2}{P_i(\theta)} = N \sum_{i=1}^{k} \frac{\delta\theta^T \nabla P_i(\theta) \nabla P_i^T(\theta) \delta\theta}{P_i(\theta)} =$$

$$= N\delta\theta^T \left(\sum_{i=1}^{k} \frac{\nabla P_i(\theta) \cdot \nabla P_i^T(\theta)}{P_i(\theta)} \right) \delta\theta = N\delta\theta^T \cdot J(\theta) \cdot \delta\theta .$$

Отсюда ясно, что с ростом потерь информации при малых $\delta\theta$, то есть близких альтернативных гипотезах, уменьшается параметр нецентральности λ, а следовательно, и мощность критерия χ^2.

Так как матрица $J(\theta)$ зависит от граничных точек интервалов, то можно подобрать их так, чтобы, задаваясь определенным критерием, минимизировать потери информации от группирования и тем самым максимизировать мощность критерия χ^2 при близких альтернативных гипотезах и соответствующим образом минимизировать асимптотическую дисперсию оценок параметров.

Необходимо отметить, что в общем случае равномерное и равновероятное разбиения весьма далеки от оптимального.

В таблицах представлены асимптотически оптимальные граничные точки, максимизирующие определитель информационной матрицы Фишера по группированным данным. В последних колонках таблиц показаны значения отношения определителя информационной матрицы Фишера по группированным наблюдениям к определителю информационной матрицы по негруппированной выборке. Для сравнения

заметим, что при равновероятном группировании значение этого отношения для экспоненциального распределения и числе интервалов K $K = 10$ составляет всего 0.8928.

В табл.I приведены оптимальные граничные точки для экспоненциального распределения с плотностью распределения $f(x) = \theta \exp(-\theta x)$ распределения Рэлея - $f(x) = (x/\theta^2)\exp(-x^2/2\theta^2)$, распределения Парето - $f(x) = \theta \alpha^{\theta} x^{-(\theta+1)}$, в табл.2 оптимальные граничные точки для распределения Максвелла с плотностью $f(x) = (2x^2/\theta^3\sqrt{2\pi})\exp(-x^2/2\theta^2)$, в табл.3 оптимальные граничные точки для нормального распределения, в табл.4 для распределения Вейбулла с плотностью $f(x) = \frac{\theta}{\theta_1}\left(\frac{x}{\theta_1}\right)^{\theta-1}\exp\left(-\left(\frac{x}{\theta_1}\right)^{\theta}\right)$ и для распределения экстремального значения с плотностью $f(x) = \frac{1}{\sigma}\exp\left\{\pm\frac{\mu-x}{\sigma} - \exp\left[\pm\frac{\mu-x}{\sigma}\right]\right\}$, наибольшего значения со знаком (+), наименьшего значения со знаком (-). Для распределения Коши с плотностью $f(x) = \theta/\pi\left(\theta^2 + (x-\theta_1)^2\right)$ оптимальное группирование совпало с разбиением на интервалы равной вероятности. Поэтому таблица оптимальных граничных точек для него не приводится.

На рисунке в качестве иллюстрации преимуществ оптимального группирования при проверке гипотез о согласии даны функции мощности критерия χ^2 для экспоненциального распределения при 10 интервалах группирования, объема выборки $N = 1000$ и уровня значимости $\alpha = 0.05$. Оптимальное группирование на несколько процентов увеличивает мощность критерия.

Если необходимо проверить согласие выборки с некоторым распределением, то следует непосредственно пользоваться таблицами I-4. Для оценивания параметров распределений процедура использования таблиц оптимального группирования несколько усложняется. Для этого, во-первых, можно воспользоваться априорными сведениями о параметрах и выбрать граничные точки, опираясь на них, а, во-вторых, лучше и проще всего сгруппировать выборку так, чтобы

Таблица I

Таблица I

Оптимальные граничные точки для экспоненциального распределения в виде $t_i = \theta \cdot x_i$, для распределения Рэлея в виде $t_i = (x_i/\theta)^2/2$, для распределения Парето в виде $t_i = \theta \ln(x_i/\alpha)$ и соответствующие значения относительной асимптотической информации A

K	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	A
2	I.5936										0.6476
3	I.0I76	2.6II2									0.8203
4	0.754I	I.77I6	3.3652								0.89I0
5	0.6004	I.3545	2.3720	3.9657							0.9269
6	0.4993	I.0997	I.8538	2.87I4	4.4650						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	0.63I4	I.0053	I.4329	I.9322	2.5326	3.2866	4.3042	5.8979		0.9798
II	0.27I6	0.5695	0.90I4	I.2746	I.70I5	2.I989	2.7955	3.5429	4.5480	6.II76	0.9832

Таблица 2

Оптимальные граничные точки для распределения Максвелла в виде $t_i = x_i/\theta$ и соответствующие значения относительной асимптотической информации A

k	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	A
2	2.045I									0.645I
3	I.6762	2.5366								0.8I79
4	I.4689	2.I292	2.8402							0.8892
5	I.3292	I.8879	2.422I	3.0583						0.9254
6	I.226I	I.7205	2.I667	2.6379	3.2274					0.9464
7	I.I458	I.5947	I.9859	2.3759	2.808I	3.3649				0.9596
8	I.0807	I.4952	I.848I	2.I879	2.543I	2.9480	3.4803			0.9685
9	I.0267	I.4I38	I.7377	2.0423	2.3499	2.6803	3.0652	3.5789		0.9747
I0	0.9798	I.3447	I.6460	I.9252	2.2003	2.4857	2.7984	3.I668	3.6625	0.9792

Таблица 3

Оптимальные граничные точки для

Таблица 3

Оптимальные граничные точки для нормального распределения в виде $t_i(x_i - \mu)/\sigma$ и соответствующие значения относительной асимптотической информации A

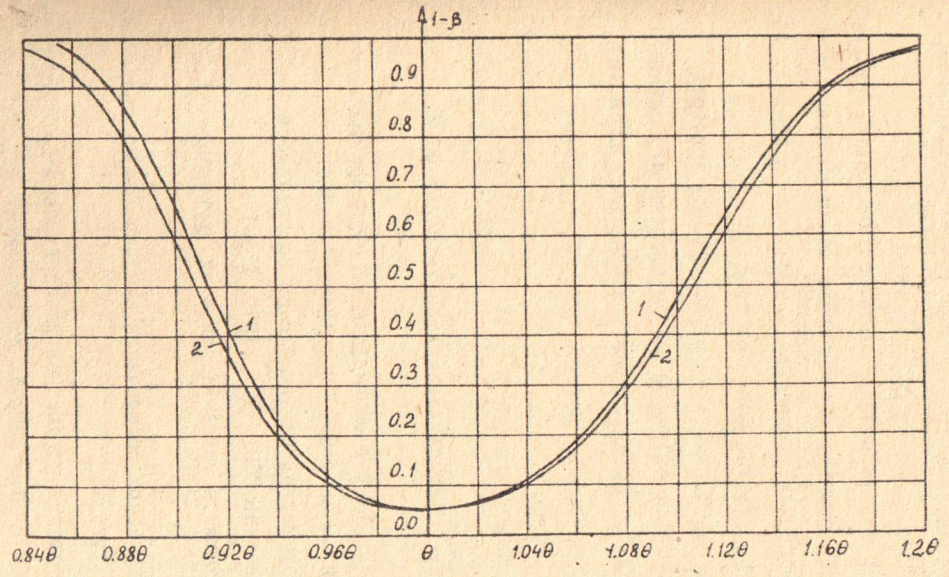
k	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	A
3	-I.III	I.III											0.4065
4	-I.383	0.0	I.383										0.5527
5	-I.696	-0.689	0.689	I.696									0.6827
6	-I.882	-0.997	0.0	0.997	I.882								0.7557
7	-2.060	-I.265	-0.492	0.492	I.265	2.060							0.8I03
8	-2.I95	-I.455	-0.786	0.0	0.786	I.455	2.I95						0.8474
9	-2.3I9	-I.622	-I.022	-0.383	0.383	I.022	I.622	2.3I9					0.8753
I0	-2.423	-I.758	-I.205	-0.650	0.0	0.650	I.205	I.758	2.423				0.8960
II	-2.5I7	-I.878	-I.360	-0.862	-0.3I4	0.3I4	0.862	I.360	I.878	2.5I7			0.9I2I
I2	-2.599	-I.983	-I.49I	-I.033	-0.553	0.0	0.553	I.033	I.49I	I.983	2.599		0.9247
I3	-2.675	-2.076	-I.607	-I.I78	-0.746	-0.267	0.267	0.746	I.I78	I.607	2.076	2.675	0.9348

Таблица 4

Оптимальные граничные точки для распределения Вейбулла в виде $t_i = (x_i/\theta_1)^{\beta}$, для распределения наибольшего экстремального значения в виде $t_i = exp((\mu - x_i)/\sigma)$, для распределения наименьшего экстремального значения в виде $t_i = exp(-(\mu - x_i)/\sigma)$ и соответствующие значения относительной асимптотической информации A

K	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	A
3	0.273	2.607											0.4079
4	0.2II	I.398	3.4I4										0.5572
5	0.I04	0.5I2	I.959	3.86I									0.6836
6	0.077	0.365	I.227	2.573	4.4I0								0.757I
7	0.050	0.232	0.676	I.7I9	2.992	4.795							0.8I09
8	0.038	0.I74	0.484	I.I90	2.204	3.429	5.205						0.8480
9	0.028	0.I27	0.343	0.783	I.603	2.57I	3.767	5.527					0.8756
I0	0.02I	0.099	0.264	0.577	I.I8I	I.993	2.927	4.I02	5.848				0.8963
II	0.0I7	0.077	0.205	0.436	0.856	I.534	2.3I9	3.232	4.393	6.I27			0.9I23
I2	0.0I3	0.062	0.I64	0.343	0.652	I.I79	I.857	2.6I6	3.5I0	4.659	6.385		0.9248
I3	0.0II	0.050	0.I33	0.275	0.5II	0.903	I.48I	2.I40	2.88I	3.762	4.902	6.62I	0.9349

I3	0.0II	0.050	0.I33	0.275	0.5II	0.903	I.48I	2.I40	2.88I	3.762	4.902	6.62I	0.9349



Функции мощности критерия χ^2 при уровне значимости $\alpha = 0.05$, объёме выборки $N = 1000$, числе интервалов группирования $k = 10$:

I – при оптимальном группировании;

2 – при равновероятных интервалах

частоты были пропорциональны вероятностям, соответствующим оптимальному группированию. Последнюю процедуру группирования целесообразно использовать при выравнивании эмпирических распределений по определенным законам.

Число интервалом в случае использования оптимального группирования определяется очень просто. Его выбирают таким образом, чтобы произведение объема выборки на вероятность попадания в крайний интервал с наименьшей вероятностью было больше I.

Для проверки гипотез об отдельных параметрах по критерию χ^2 полезно использовать таблицы оптимального группирования, минимизирующие потери информации о соответствующих параметрах, приводимые в [7].

Литература

I. Кендалл М., Стьюарт А. Статистические выводы и связи. М., "Наука", 1973.

2. Куллдорф Г. Введение в теорию оценивания по группированным и частично группированным выборкам. М., "Наука", 1966.

3. Бодин Н.А. Оценка параметров распределения по группированным выборкам. - Тр.Матем.инст.АН СССР, III, 1970, 110-154.

4. Рао С.Р. Линейные статистические методы и их применение. М., "Наука", 1968.

5. Денисов В.И. Математическое обеспечение системы ЭВМ - экспериментатор. М., "Наука", 1977.

6. Лемешко Б.Ю. Об оценивании параметров распределений по группированным наблюдениям. - В сб.: Вопросы кибернетики. М., 1977, вып.30.

7. Губинский А.И., Денисов В.И., Гречко Ю.П., Лемешко Б.Ю., Цой Е.Б. Методические рекомендации по планированию экспериментов и обработке экспериментальных данных при исследовании надежности и качества функционирования СЧТ. Л., ЛЭТИ, 1978.