Научный вестник НГТУ. - 1996. - № 2

ОБРАБОТКА ИНФОРМАЦИИ

УДК 519.2 (75.8)

Вопросы обработки выборок одномерных случайных величин*

Б.Ю. ЛЕМЕШКО, С.Н. ПОСТОВАЛОВ

Рассматриваются вопросы оценивания параметров и проверки гипотез по критериям согласия при различном представлении выборочных данных. Показывается влияние асимптотически оптимального группирования данных на свойства оценок и мощность критериев согласия и влияние группирования данных на устойчивость оценок. Подчеркивается эффективность параметрических методов отбраковки аномальных наблюдений при использовании робастных оценок по группированным данным. Рассматриваются вопросы оценивания смесей усеченных и неусеченных распределений.

ВВЕДЕНИЕ

С задачами статистического анализа одномерных наблюдений случайных величин приходится сталкиваться на каждом шагу. Достаточно хорошо развитый математический аппарат позволяет многим исследователям и практикам использовать известные методы оценивания и процедуры проверки статистических гипотез, задумываясь над справедливостью выводов и эффективностью оценок лишь при наличии нештатных ситуаций. К таким ситуациям можно отнести необычное представление исходных наблюдений, засорение выборки грубыми ошибками измерений, усечение области определения случайной величины, содержание в выборке наблюдений, принадлежащих различным законам распределений. В этих случаях привычные методы становятся неустойчивыми, выводы неоднозначными, не внушающими доверия. Не является секретом и факт, что неоправдано часто результаты анализа стремятся подогнать к нормальному закону распределения. Желание это понятно, так как снимаются многие проблемы и всё становится просто. На самом же деле довольно редко с достаточной степенью уверенности оказывается возможным использование нормального закона в качестве модели, описывающей реальные наблюдения.

В данной работе изложение материала ведется на базе результатов, в том числе оригинальных, вошедших в реализованную версию программной системы "Статистический анализ одномерных наблюдений случайных величин" [1] и используемых в разрабатываемой объектно-ориентированной версии системы. Основное отличие новой версии связано с расширением класса моделей, используемых для описания наблюдаемых величин, усеченными законами распределений и смесями законов распределений.

---

* *Статья получена 27 мая 1996 г.*

ПРЕДСТАВЛЕНИЕ ВЫБОРОЧНЫХ ДАННЫХ

В программной системе все задачи статистического анализа данных рассматриваются с точки зрения *наиболее общего представления экспериментальных наблюдений* в виде *частично группированных выборок* [2,3], частными случаями которых являются негруппированные, группированные и цензурированные выборки. Это обеспечивает проведение анализа реальных данных независимо от того, в каком виде они регистрировались и в каком виде представлены для обработки. Выборка является *негруппированной*, если выборочные значения представляют собой индивидуальные значения наблюдений из области определения случайной величины. Выборка является *группированной*, если область определения случайной величины разбита на k непересекающихся интервалов граничными точками

$$x_0 < x_1 < \ldots < x_{k-1} < x_k ,$$

где x_0, x_k - нижняя и верхняя грани области определения случайной величины X, и зафиксированы количества наблюдений n_i, попавших в i -й интервал значений. Выборка является *частично группированной*, если имеющаяся в нашем распоряжении информация связана с множеством непересекающихся интервалов, которые делят область определения случайной величины так, что каждый интервал принадлежит к одному из двух типов:

а) i -й интервал принадлежит к первому типу, если число n_i известно, но индивидуальные значения x_{ij}, $j = \overline{1, n_i}$ неизвестны;

б) i -й интервал принадлежит ко второму типу, если известно не только число n_i, но и все индивидуальные значения x_{ij}, $j = \overline{1, n_i}$.

Область определения случайной величины в этом случае можно представить в виде $X_{(1)} \cup X_{(2)}$, где $X_{(1)}$ - множество интервалов первого типа, а $X_{(2)}$ - множество интервалов второго типа.

ОЦЕНИВАНИЕ ПАРАМЕТРОВ

Наиболее часто на практике используются следующие методы вычисления оценок: метод моментов, метод минимума χ^2, модифицированный метод минимума χ^2, метод максимального правдоподобия, методы, минимизирующие такие статистики, как расстояние Хеллингера, дивергенция Кульбака-Лейблера, мера расхождения Холдейна.

Все эти методы при соответствующих условиях регулярности, как показал С.Р. Рао, дают состоятельные и асимптотически эффективные оценки [4]. Однако имеются и различия между этими методами, возникающие при учете введённой С.Р. Рао эффективности второго порядка [5]. Им показано, что асимптотическая дисперсия оценки определяется соотношением

$$D(\theta) = \frac{1}{NJ(\theta)} + \frac{\psi(\theta)}{N^2} + o\left(\frac{1}{N^2}\right),$$

где $J(\theta)$ - информационное количество Фишера о параметре, а величина $\psi(\theta)$ определяется методом оценивания и неотрицательна. При учёте эффективности второго порядка, связанной со вторым слагаемым, метод максимального правдоподобия оказывается наилучшим. Кроме того, метод макси-

мального правдоподобия является наиболее универсальным по отношению к форме представления выборочных данных (структуре выборки), по которым оцениваются параметры. Метод моментов требует преобразования группированных данных к негруппированным, только после чего оцениваются параметры с использованием при необходимости (или возможности) поправок на группирование. Напротив, метод минимума χ^2 и родственные с ним используют только группированные данные: если в распоряжении исследователя имеются индивидуальные наблюдения, выборку следует преобразовывать в полностью группированную. Метод максимального правдоподобия в отличие от других позволяет определять оценки максимального правдоподобия (ОМП) параметров по негруппированным, частично группированным и группированным данным, т.е. дает возможность исследователю самому определять, в каком виде регистрировать и в каком виде хранить экспериментальную информацию в зависимости от характеристик приборов, регистрирующих наблюдения, и объема экспериментальной информации.

Основным методом оценивания параметров распределений, заложенным в системе, является метод максимального правдоподобия. Оценки параметров распределений находятся в результате максимизации функции правдоподобия по частично группированной выборке, которая имеет вид:

$$L(\theta) = \prod_{(1)} P_i^{n_i}(\theta) \prod_{(2)} \prod_{j=1}^{n_i} f(x_{ij}, \theta),$$

где $f(x,\theta)$ - функция плотности случайной величины; $P_i(\theta) = \int_{x_{i-1}}^{x_i} f(x,\theta)dx$ - вероятность попадания наблюдения в i -й интервал значений; (1) и (2) означают, что умножение осуществляется по интервалам с группированными и негруппированными данными соответственно.

В общем случае частично группированной выборки ОМП определяются только с использованием численных методов. Лишь в частных случаях негруппированных наблюдений ОМП получаются в виде конкретных формул. Условия существования и единственности ОМП по группированным и частично группированным выборкам рассматривались в [2, 3, 6, 7].

АСИМПТОТИЧЕСКИ ОПТИМАЛЬНОЕ ГРУППИРОВАНИЕ ДАННЫХ

Всякая группировка данных по сравнению с негруппированной выборкой ведет *к потере информации*, понимаемой в общем широком смысле. Асимптотическая дисперсионная матрица ОМП по *группированным* наблюдениям определяется соотношением

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

где

$$M_I(\hat{\theta}) = \sum_{i=1}^{k} \frac{\nabla P_i(\hat{\theta}) \nabla^T P_i(\hat{\theta})}{P_i(\hat{\theta})}$$

- информационная матрица Фишера по группированным данным. Элементы информационной матрицы зависят от граничных точек интервалов, так как

$P_i(\theta) = \int_{x_{i-1}}^{x_i} f(x,\theta)dx$. В случае, когда функция плотности распределения определяется скалярным параметром или осуществляется оценивание только одного параметра при известных остальных, целью задачи асимптотически оптимального группирования является минимизация асимптотической дисперсии ОМП по группированным данным. И эта задача сводится к максимизации информационного количества Фишера о параметре по группированной выборке, т.е. к решению задачи

$$\max_{x_0<x_1<\ldots<x_{k-1}<x_k} \sum_{i=1}^{k} \left(\frac{\partial \ln P_i(\theta)}{\partial \theta} \right)^2 P_i(\theta).$$

При оценивании вектора параметров мы имеем дело с информационной матрицей. В этом случае в качестве критериев оптимальности могут быть выбраны различные функционалы от асимптотической дисперсионной матрицы, но наиболее естественно минимизировать обобщенную асимптотическую дисперсию, т.е. решать задачу

$$\max_{x_0<x_1<\ldots<x_{k-1}<x_k} \det M_I(\theta).$$

Следует отметить, что информационная матрица Фишера зависит от неизвестных оцениваемых параметров. Однако для широкого ряда распределений при решении задач асимптотически оптимального группирования удалось получить граничные точки интервалов в виде, *инвариантном относительно параметров распределений,* и на их основе сформировать таблицы асимптотически оптимального группирования. Ранее в литературе были представлены только фрагмент таблицы для экспоненциального закона распределения и отдельно таблицы для математического ожидания и среднего квадратичного отклонения нормального распределения [2,8]. Эти результаты были уточнены и в дальнейшем в совокупности были получены таблицы для распределений экспоненциального, полунормального, Рэлея, Максвелла, модуля многомерного нормального вектора, Парето, Эрланга, Лапласа, нормального, логарифмически-нормальных (*ln* и *lg*), Коши, Вейбулла, распределений минимального значения и максимального значения, двойного показательного, гамма-распределения. В общей сложности сформировано 54 таблицы оптимальных граничных точек и вероятностей попадания в соответствующие интервалы. Полностью таблицы приведены в [3].

Для некоторых распределений граничные точки интервалов не могут быть выражены в виде, инвариантном относительно параметров распределений, т.е. они остаются функциями этих параметров. Это касается, например, таких законов, как гамма- и бета-распределения и ряда других. В этом случае формирование таблиц асимптотически оптимального группирования теряет смысл, невозможно также использовать асимптотически оптимальное группирование при оценивании параметров. Однако для задач проверки гипотез о согласии, как будет показано ниже, решение задачи асимптотически оптимального группирования остается чрезвычайно актуальным, так как возможно ее решение при конкретных значениях параметров непосредственно в процессе проверки гипотез.

В качестве примера в табл. 1-3 приведены значения асимтотически оптимальных граничных точек в виде $t_i = \left(x_i/\theta_1\right)^{\theta_0}$ для распределения Вейбулла с функцией плотности

$$f(x)=\frac{\theta_0(x-\theta_2)^{\theta_0-1}}{\theta_1^{\theta_0}}\exp\left\{-\left(\frac{x-\theta_2}{\theta_1}\right)^{\theta_0}\right\}.$$

При формировании табл. 1 максимизировалось количество информации Фишера по группированным данным о параметре θ_1, при формировании табл. 2 - о параметре θ_0, при формировании табл. 3 максимизировался определитель информационной матрицы Фишера по группированным данным. О качестве группирования (о потерях информации, связанных с группированием) можно судить по величине относительной асимптотической информации A, представляющей собой отношение количества информации Фишера по группированным данным к количеству информации по негруппированным наблюдениям (или отношение соответствующих определителей). Из табл. 1 видно, что при использовании асимптотически оптимального группирования и разбиении выборки на 10 интервалов потери информации от группирования о параметре θ_1 составляют всего около 2 %.

Таблицы асимптотически оптимального группирования могут эффективно использоваться в случае больших выборок при вычислении оценок параметров [3, 9]. Суть этих оценок заключается в следующем. Опираясь на таблицы вероятностей попадания в интервал, соответствующих асимптотически оптимальному группированию, находят приближенные значения граничных точек x_i так, чтобы количества наблюдений n_i, попавших в каждый интервал, было пропорционально оптимальной частоте, т.е. $n_i = nP_i$, где P_i берется из соответствующей таблицы. Из таблицы оптимальных граничных точек при заданном числе интервалов берутся значения t_i, которые связаны с x_i вполне определенной зависимостью вида $t_i = \varphi(x_i, \theta)$. Отсюда можно вычислить значение параметра $\theta = \varphi^{-1}(t_i, x_i)$, а затем усреднить его по всем граничным точкам. Например, для распределения Вейбулла оценка параметра θ_1 будет иметь вид

$$\hat{\theta}_1=\frac{1}{k-1}\sum_{i=1}^{k-1}\frac{x_i}{t_i^{1/\theta_0}},$$

где t_i - берутся из табл. 1. Оценка параметра θ_0 -

$$\hat{\theta}_0=\frac{1}{k-1}\sum_{i=1}^{k-1}\frac{\ln t_i}{\ln(x_i/\theta_1)},$$

где t_i - берутся из табл. 2. Одновременно оценки двух параметров -

$$\hat{\theta}_0=\frac{1}{k-2}\sum_{i=2}^{k-1}\frac{\ln t_{i-1}-\ln t_i}{\ln x_{i-1}-\ln x_i};$$

$$\hat{\theta}_1=\frac{1}{k-2}\sum_{i=2}^{k-1}\exp\left\{\frac{\ln t_{i-1}\ln x_i-\ln t_i\ln x_{i-1}}{\ln t_{i-1}-\ln t_i}\right\},$$

где t_i - берутся из табл. 3. Полная сводка формул такого вида для вычисления оценок параметров распределений приведена в [3].

Следует отметить, что оценки с оптимальным выбором порядковых статистик, рассматриваемые в ряде работ, ссылки на которые даны в [10], также

опираются на решение задачи оптимального группирования и близки к предлагаемым.

КРИТЕРИИ СОГЛАСИЯ

Наиболее часто при проверке гипотез о согласии пользуются критериями отношения правдоподобия, χ^2 Пирсона, Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса.

При проверке гипотез о согласии для найденного значения соответствующей статистики S^* вычисляется вероятность

$$p = P\{S > S^*\} = \int_{S^*}^{\infty} g(s)ds,$$

где $g(s)$ - плотность распределения статистики при условии истинности нулевой гипотезы. При заданном уровне значимости α гипотеза о согласии не отвергается, если $p > \alpha$.

Таблица 1

Оптимальные граничные точки интервалов группирования в виде $t_i = \left(x_i/\theta_1\right)^{\theta_0}$ для оценивания масштабного параметра θ_1 распределения Вейбулла и для проверки гипотез о нем 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	t_8	t_9	A
2	1.5936									0.6476
3	1.0176	2.6112								0.8203
4	0.7541	1.7716	3.3652							0.8910
5	0.6004	1.3545	2.3720	3.9657						0.9269
6	0.4993	1.0997	1.8538	2.8714	4.4650					0.9606
7	0.4276	0.9269	1.5273	2.2813	3.2989	4.8925				0.9606
8	0.3739	0.8015	1.3008	1.9012	2.6553	3.6729	5.2665			0.9693
9	0.3323	0.7063	1.1338	1.6331	2.2336	2.9876	4.0052	5.5988		0.9754
10	0.2990	0.6314	1.0053	1.4329	1.9322	2.5326	3.2866	4.3042	5.8979	0.9798

Критерии χ^2 Пирсона и отношения правдоподобия предусматривают обязательное группирование выборки, если она негруппирована. Напротив, непараметрические критерии Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса, рассматривающие различные меры близости теоретической и эмпирической функций распределения [11], предполагают, что исходные наблюдения негруппированы. В случае, если исходная выборка группирована или частично группирована, даже только цензурирована, при использовании непараметрических критериев возникают сложности, связанные с тем, что статистики этих критериев в такой ситуации не могут быть вычислены, так как выражения для соответствующих статистик предусматривают, что известны все индивидуальные значения наблюдений. Возможный выход может заключаться в следующем. Для статистики соответствующего критерия находятся оцен-

ки сверху и снизу ($\underline{S^*}$ и $\overline{S^*}$), и на основании верхней и нижней границ вероятности согласия ($p_{\max} = P\left\{S > \underline{S^*}\right\}$ и $p_{\min} = P\left\{S > \overline{S^*}\right\}$), которые позволяют оценить степень согласия теоретического и эмпирического законов распределения, делаются статистические выводы. Такой подход предлагается и рассматривается в [12, 13].

Таблица 2

Оптимальные граничные точки интервалов группирования в виде $t_i = \left(x_i/\theta_1\right)^{\theta_0}$ для оценивания основного параметра θ_0 распределения Вейбулла и для проверки гипотез о нем 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	t_8	t_9	A
2	3.4903									0.3282
3	0.1418	3.2891								0.6518
4	0.1505	2.6936	4.5643							0.7481
5	0.0516	0.2486	2.6173	4.4970						0.8235
6	0.0535	0.2580	2.3339	3.6005	5.3984					0.8639
7	0.0244	0.1154	0.3260	2.2878	3.5602	5.3523				0.8936
8	0.0251	0.1181	0.3342	2.1205	3.1036	4.2984	6.0540			0.9141
9	0.0136	0.0639	0.1731	0.3843	2.0935	3.0803	4.2767	6.0333		0.9288
10	0.0137	0.0649	0.1760	0.3917	1.9766	2.7906	3.7069	4.8673	6.6006	0.9408

В развиваемой программной системе реализованы все 6 упомянутых выше критериев. Использование совокупности критериев даёт возможность принимать более обоснованное решение, а при противоречивости выводов по отдельным критериям, что бывает достаточно часто - формировать компромиссный критерий и с учетом его делать окончательный вывод о предпочтительности выбора того или иного закона для описания наблюдаемой случайной величины.

МОЩНОСТЬ КРИТЕРИЕВ СОГЛАСИЯ

Способность критерия различать альтернативы определяется такой его характеристикой как мощность. Чем выше мощность критерия, тем лучше он различает соответствующие альтернативы. Наиболее ценной является способность критерия отличать близкие гипотезы.

Для критериев χ^2 Пирсона и отношения правдоподобия из совокупности критериев, включенных в программную систему, известны асимптотические распределения статистик при истинности альтернативных гипотез. Это значит, что могут быть построены функции мощности этих критериев.

Таблица 3

Оптимальные граничные точки интервалов группирования в виде $t_i = \left(x_i/\theta_1\right)^{\theta_0}$ для одновременного оценивания двух параметров распределения Вейбулла 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	t_8
3	0.2731	2.6067						
4	0.2109	1.3979	3.4137					
5	0.1044	0.5123	1.9590	3.8606				
6	0.0772	0.3649	1.2269	2.5726	4.4096			
7	0.0501	0.2318	0.6758	1.7192	2.9922	4.7959		
8	0.0377	0.1740	0.4837	1.1904	2.2041	3.4285	5.2049	
9	0.0275	0.1269	0.3431	0.7829	1.6027	2.5713	3.7667	5.5273
10	0.0213	0.0988	0.2638	0.5770	1.1805	1.9932	2.9269	4.1024
11	0.0165	0.0771	0.2046	0.4359	0.8560	1.5344	2.3192	3.2319
12	0.0123	0.0618	0.1638	0.3434	0.6517	1.1789	1.8570	2.6163
13	0.0106	0.0500	0.1326	0.2754	0.5106	0.9030	1.4807	2.1401
14	0.0087	0.0412	0.1094	0.2261	0.3126	0.7116	1.1798	1.7608
15	0.0072	0.0344	0.0913	0.1881	0.3394	0.5734	0.9387	1.4426

Окончание табл. 3

k	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}	A
3							0.4079
4							0.5572
5							0.6836
6							0.7571
7							0.8109
8							0.8480
9							0.8756
10	5.8478						0.8963
11	4.3930	6.1270					0.9123
12	3.5103	4.6589	6.3853				0.9248
13	2.8810	3.7623	4.9016	6.6208			0.9349
14	2.4019	3.1286	3.9997	5.1314	6.8444		0.9431
15	2.0116	2.6381	3.3538	4.2169	5.3425	7.0506	0.9498

Непараметрические критерии еще называют критериями свободными от распределения. Такие критерии не могут быть очень чувствительными [4]. Использование свободного от распределения критерия вместо наиболее эффективного параметрического приводит к некоторой потере эффективности или мощности [14]. Критерий χ^2 асимптотически свободен от распределения в случае простой гипотезы и в случае сложной гипотезы, если в последнем случае используются ОМП по группированным данным [15].

Точные зависимости мощностей непараметрических критериев неизвестны. В ряде источников приводятся лишь различные оценки, на основании которых можно построить оценки функций мощности от конкретных альтернатив. В [15, 16] даны ссылки на результаты Мэсси, из которых следует, что при больших объемах выборки критерии типа Колмогорова-Смирнова оказываются значительно лучшими, чем критерий χ^2. С другой стороны, эксперименты Дурбина с умеренными выборками, на которые приводится ссылка в [16], не позволяют утверждать, что всегда критерий Колмогорова обладает большей мощностью, чем критерий χ^2.

Аналогичное сравнение критериев ω^2 и χ^2, проводившееся Кацем, Кифером и Вольфовицем, упоминается в [16]. И здесь из приведенных результатов видно, что для достижения той же минимальной мощности критерию χ^2 требуется больший объем выборки, чем критерию ω^2.

Опыт эксплуатации программной системы [1], в которой возможна параллельная проверка согласия по всем 6 критериям, приводит к необходимости отметить следующий факт. Если по данной выборке оцениваются параметры распределения и затем осуществляется проверка гипотез о согласии, то критерии Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса по сравнению с критериями χ^2 Пирсона и отношения правдоподобия *практически всегда* отличаются повышенной вероятностью согласия вида $P\{S > S^*\}$, где S^* - вычисленное значение соответствующей статистики. Это может свидетельствовать о большей чувствительности критериев χ^2 Пирсона и отношения правдоподобия к близким альтернативам.

Очевидно, что мощность такого критерия, как χ^2 Пирсона, существенно зависит от способа группирования. Естественно, что желательно так проводить группирование данных, чтобы максимизировать мощность. Фишеровская информация служит мерой внутренней близости распределений случайных величин, и этот внутренний характер связан с мощностью различения между близкими значениями параметра [4]. Статистика редуцирует выборочные данные, и поэтому мощность различения с помощью статистики не больше, чем с помощью всей выборки. А это значит, если нужно выбирать между несколькими статистиками, следует предпочесть ту, для которой потери фишеровской информации минимальны.

Статистика критерия согласия χ^2 Пирсона вычисляется в соответствии с соотношением

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

и в пределе подчиняется χ^2-распределению с $k-1$ степенью свободы, если верна нулевая гипотеза, и подчиняется нецентральному χ^2-распределению с тем же числом степеней свободы и параметром нецентральности

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

если верна конкурирующая гипотеза и выборка соответствует распределению того же типа, но с параметром θ_1 (в общем случае векторным). Разлагая $P_i(\theta_1)$ в ряд Тейлора при малых $\Delta\theta = \theta_1 - \theta$ и пренебрегая членами высшего порядка, получаем

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

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

Мощность критерия χ^2 Пирсона является неубывающей функцией от λ. Матрица потерь информации, вызванных группированием данных, $\Delta M = M(\theta) - M_\Gamma(\theta)$, где $M(\theta)$ - информационная матрица Фишера по негруппированным наблюдениям, является неотрицательно определённой, и, следовательно, $\delta\theta^T\Delta M\delta\theta \geq 0$. А так как $\delta\theta^T M_\Gamma(\theta)\delta\theta = \delta\theta^T M(\theta)\delta\theta - \delta\theta^T\Delta M\delta\theta$, то очевидно, что с ростом потерь информации падает и мощность критерия при близких альтернативных гипотезах. Эти потери можно уменьшить, подбирая граничные точки так, чтобы $M_\Gamma(\theta)$ стремилась к информационной матрице по негруппированным данным $M(\theta)$, т.е. в данном случае приходим к той же самой задаче асимптотически оптимального группирования, что и при оценивании параметров.

Аналогичные результаты справедливы для критерия отношения правдоподобия и ряда других [3].

Асимптотически оптимальное группирование минимизирует потери информации, связанные с группированием, и, следовательно, гарантирует *максимальную мощность различения близких альтернатив* для соответствующих критериев. А в критериях согласия особенно важна именно способность различать близкие гипотезы.

К сожалению, на практике наиболее часто, применяя критерий χ^2 Пирсона, используют интервалы равной длины или, в лучшем случае, интервалы равной вероятности. Естественно, что в такой ситуации мощность критерия обычно далека от максимально возможной. В частности, теряется чувствительность на краях. На рис. 1 и 2 для сравнения приведены построенные предельные функции мощности критерия χ^2 Пирсона при проверке гипотез об основном параметре θ_0 распределения Вейбулла при использовании асимптотически оптимального группирования и при разбиении области определения случайной величины на равновероятные интервалы.

РОБАСТНОСТЬ ОЦЕНОК

В статистике под робастностью понимают *нечувствительность к малым отклонениям от предположений* [17].

При решении задач статистического анализа и, в частности, при вычислении оценок параметров распределений чрезвычайно важное значение приобретает проблема наличия в выборке грубых ошибок измерений или факт принадлежности выборки другому закону распределения. Присутствие единственного аномального наблюдения обычно приводит к резкому изменению оценки. Аналогично, если для описания выборочных данных исполь-

зуется некоторый закон распределения, в то время как на самом деле выборка принадлежит существенно отличающемуся от него, найденные оценки определяют закон, не согласующийся с выборочными данными.

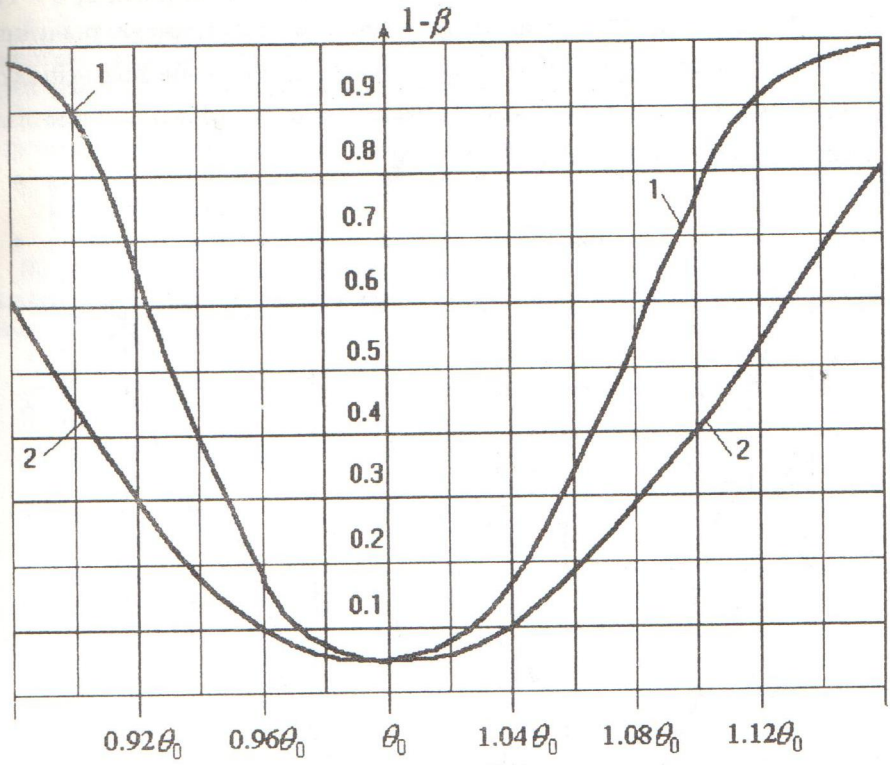

Рис. 1. Функции мощности критерия χ^2 при проверке гипотез о параметре θ_0 распределения Вейбулла: уровень значимости α =0.05; объем выборки n =100; число интервалов k =3; 1 - при оптимальном группировании; 2 - при равновероятном

Естественно желание, чтобы найденные оценки были как можно менее чувствительны к аномальным наблюдениям. Так как в противном случае, прежде чем переходить к оцениванию, приходится использовать процедуры исключения грубых ошибок измерений, что выливается в не совсем простую задачу. В данном случае следует подчеркнуть достоинство оценок, использующих группирование исходных выборочных данных, так как очевидно, что они менее чувствительны к случайным выбросам. *Группирование выборки позволяет резко снизить влияние аномальных наблюдений*, а иногда и совсем исключить влияние случайных выбросов. Оценки, определяемые по предварительно сгруппированным данным, оказываются устойчивыми и в ситуациях, когда наши предположения о наблюдаемом законе распределения сильно отличаются от действительного.

Посмотрим, как поведут себя оценки максимального правдоподобия параметров распределения Лапласа с плотностью $f(x)=\frac{\theta_0}{2}e^{-\theta_0|x-\theta_1|}$, когда на самом деле выборка принадлежит распределению Коши с плотностью $f(x)=\frac{\theta_0}{\pi[\theta_0^2+(x-\theta_1)^2]}$. Распределение Коши - это распределение с "тяжё-

лыми" хвостами. Смоделируем выборку по закону Коши с параметрами $\theta_0 = 1,\ \theta_1 = 0$. На рис. 3 представлены результаты моделирования и анализа. На этом и последующих рисунках $\theta_0 = t[0],\ \theta_1 = t[1]$. На рисунках 3, 6 - 11 отражаются результаты проверки гипотез о согласии: вычисленные значения S^* соответствующих статистик S и вероятности превышения полученного значения статистики при истинности нулевой гипотезы $P\{S > S^*\}$. Гипотеза о согласии не отвергается, если $P\{S > S^*\} > \alpha$.

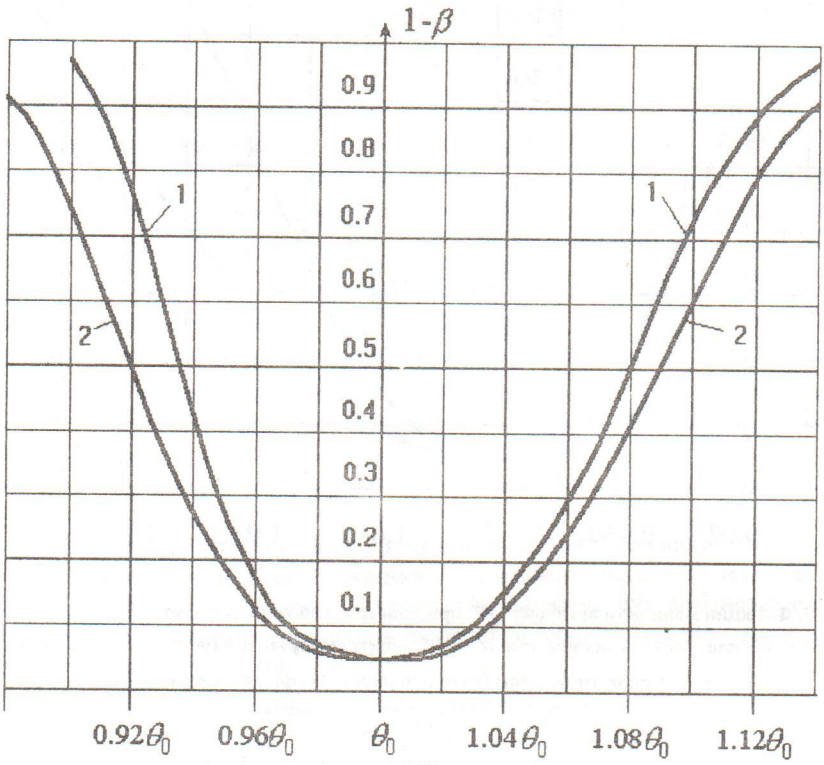

Рис. 2. Функции мощности критерия χ^2 при проверке гипотез о θ_0 параметре распределения Вейбулла: уровень значимости α =0.05; объем выборки n =1000; число интервалов k =10; 1 - при оптимальном группировании; 2 - при равновероятном

Предполагая, что на самом деле выборка принадлежит распределению Лапласа, оценим его параметры. Результаты оценивания и проверки гипотез о согласии даны на рис. 4, где *1* - эмпирическая, а *2* - теоретическая функции распределения. Далее после предварительного группирования выборки, разбив её на интервалы равной частоты (равной вероятности), найдены робастные оценки параметров распределения Лапласа. В данном случае выборка разбивалась на 14 интервалов. Результаты оценивания и анализа показаны на рис. 5. Качественная картина рис. 4 и 5 позволяет увидеть насколько ближе во втором случае теоретический и эмпирический законы распределения. Еще более впечатляющая картина получается при оценивании в аналогичной ситуации параметров нормального распределения [1].

ИСКЛЮЧЕНИЕ АНОМАЛЬНЫХ ИЗМЕРЕНИЙ

При вычислении оценок параметров распределений присутствие единственного аномального наблюдения может приводить к оценкам, которые совершенно не согласуются с выборочными данными.

В борьбе с грубыми погрешностями измерений, если они не были обнаружены в процессе измерений, используют два подхода:

- исключение резко выделяющихся аномальных измерений из дальнейшей обработки;
- использование робастных методов обработки.

Во втором случае обычно подразумевается и дальнейшее выделение аномальных наблюдений, так как именно они могут нести необходимую исследователю информацию.

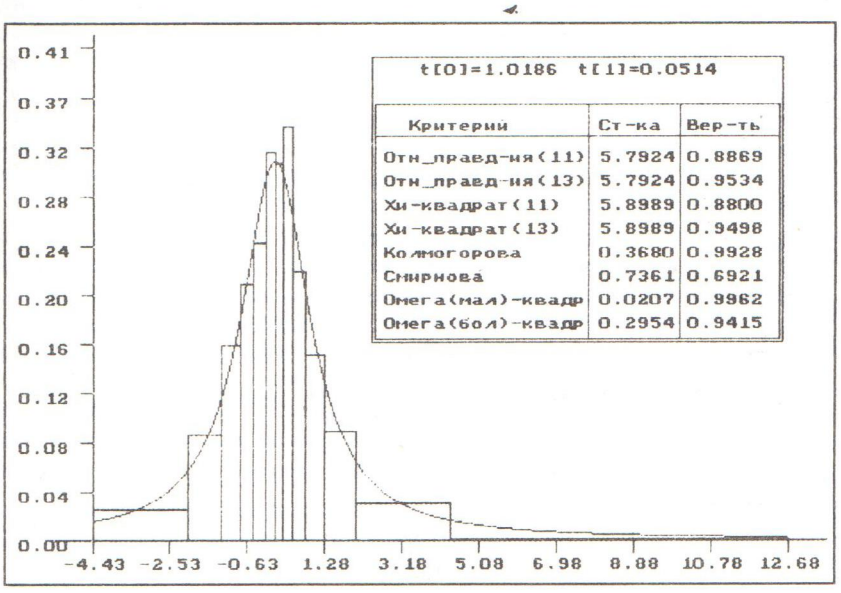

Рис. 3. Результаты оценивания параметров распределения Коши и проверки согласия с исходной выборкой

Параметрическая процедура отбраковки грубых ошибок измерений в одномерной ситуации выглядит следующим образом [18]. Рассматривается ситуация, когда $x_1, x_2, \ldots, x_n$ числа. Резко выделяется одно наблюдение, для определенности $x_{\max}$. При нулевой гипотезе H_0 наблюдения $x_1, x_2, \ldots, x_n$ рассматриваются как реализация независимых одинаково распределенных случайных величин $X_1, X_2, \ldots, X_n$ с функцией распределения $F(x)$. При альтернативной гипотезе H_1 случайные величины $X_1, X_2, \ldots, X_n$ также независимы, $X_1, X_2, \ldots, X_{n-1}$ имеют распределение $F(x)$, а X_n - распределение $G(x)$, которое "существенно сдвинуто вправо" относительно $F(x)$, например $G(x) = F(x - A)$, где A достаточно велико. Если $x_{\max} \le d$, то принимается гипотеза H_0, в противном случае - гипотеза H_1. При справедливости гипотезы H_0 $P\{\max\limits_{1 \le i \le n} X_i \le d\} = [F(d)]^n = 1 - \alpha$, и критическое значение $d = d(\alpha, n)$

определяется из уравнения $F(d) = \sqrt[n]{1-\alpha}$. Если рассматриваем принадлежность к выборке $x_{\min}$, то гипотеза H_0 принимается при $x_{\min} \geq d_1$. Если справедлива гипотеза H_0, то $P\{\min\limits_{1 \leq i \leq n} X_i \geq d_1\} = [1 - F(d_1)]^n = 1 - \alpha$. Тогда значение $d_1 = d_1(\alpha, n)$ определяется из уравнения $1 - F(d_1) = \sqrt[n]{1-\alpha}$.

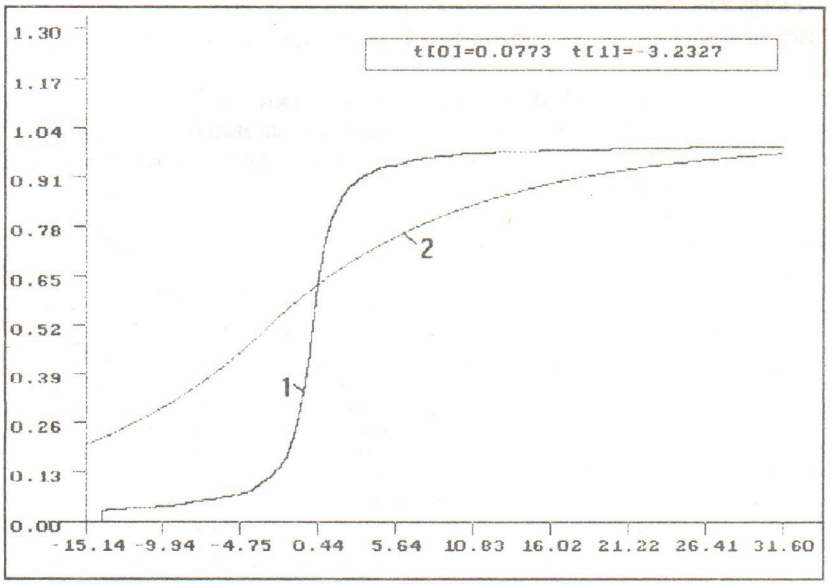

Рис. 4. Результаты оценивания параметров распределения Лапласа и проверки согласия с исходной выборкой

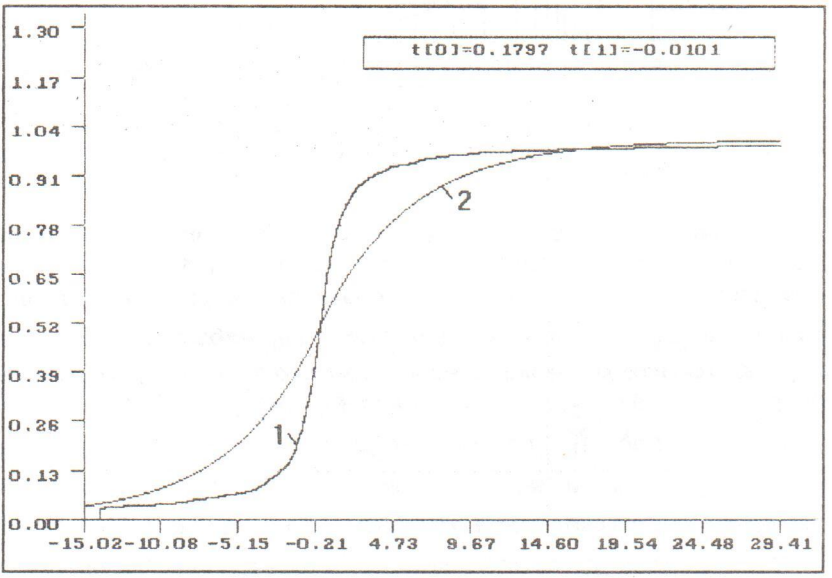

Рис. 5. Результаты оценивания параметров распределения Лапласа по группированной выборке и проверки согласия с исходной

Неустойчивость такой процедуры отбраковки в основном связана с возможно неточной идентификацией закона $F(x)$ из-за влияния грубых ошибок на оценки параметров и последующего неверного выбора закона из заданного класса с помощью критериев согласия.

В задаче отбраковки аномальных наблюдений на разных этапах её решения к статистическим процедурам оценивания и проверки гипотез предъявляются прямо противоположные требования. На этапе идентификации закона распределения и при оценивании его параметров методы должны быть как можно менее чувствительны к наличию аномальных ошибок измерений. Наоборот, на последующем этапе исключения аномальных измерений критерий должен улавливать их наличие и позволять их отсекать. Отметим, что все критерии согласия не чувствительны к аномальным наблюдениям и не могут использоваться для отбраковки.

Таким образом, при идентификации (при оценивании параметров распределений) мы должны использовать робастные алгоритмы (устойчивые к наличию аномальных наблюдений). В этой связи на первом этапе рекомендуется использовать оценки по группированным данным. Как уже говорилось, их важным достоинством является малая чувствительность к случайным выбросам. Причем для большей устойчивости оценок следует осуществлять разбиение выборки на интервалы равной вероятности (равночастотные интервалы). В то же время необходимо учитывать, что в общем случае в выборке, сгруппированной с использованием асимптотически оптимального группирования содержится существенно больше информации о параметрах распределения, чем при равновероятном группировании. Это значит, что потери информации от группирования при асимптотически оптимальном группировании могут быть меньше, чем мешающая информация, связанная с засорением выборки. Поэтому желательно на первом этапе находить не одну, а две оценки по группированным данным с использованием как оптимального, так и равновероятного группирования, и остановиться на той оценке, которая дает лучшие результаты (лучшее согласие). Это не отрицает и применения разбиения выборки на интервалы равной длины. Правильный выбор закона $F(x)$ и определение робастных оценок его параметров гарантируют успех на последующем этапе отбраковки в соответствии с изложенной выше параметрической процедурой.

Продемонстрируем сказанное на конкретном примере исключения аномальных измерений из исходной негруппированной выборки. Выборка объёмом 1000 наблюдений была смоделирована в соответствии с логистическим распределением с плотностью

$$f(x) = \frac{\pi}{\theta_1\sqrt{3}}\exp\left\{-\frac{\pi(x-\theta_0)}{\theta_1\sqrt{3}}\right\}\Bigg/\left[1+\exp\left\{-\frac{\pi(x-\theta_0)}{\theta_1\sqrt{3}}\right\}\right]^2 .$$

При моделировании были заданы параметры: $\theta_0 = 1$, $\theta_1 = 1$. В процессе регистрации три наблюдения "подверглись" сильным искажениям.

На рис. 6 приведены результаты статистического анализа полученной выборки. В данном случае получили логистический закон распределения с параметрами $\theta_0 = 1.0133$, $\theta_1 = 1.1508$. Как видим из рис. 6, согласие по всем критериям плохое: наличие аномальных наблюдений сыграло свою роль.

А на рис. 7 представлены результаты статистического анализа, когда перед оцениванием выборка была разбита на интервалы, а затем по группированной выборке были найдены оценки параметров распределения $\theta_0 = 1.0066$, $\theta_1 = 1.0058$, после чего проверены гипотезы о согласии исходной выборки с полученным законом распределения. Как видим, результаты проверки гипотез о согласии по всем критериям очень хорошие. Это является

косвенным подтверждением того, что все попытки выделения с помощью критериев согласия грубых ошибок измерений заведомо обречены на неудачу.

Для отбраковки аномальных наблюдений зададимся уровнем значимости $\alpha = 0.1$, при объёме выборки $n = 1000$ и найденном векторе параметров $\theta^T = [1.0066,\ 1.0058]$ логистического распределения определяем критическое значение $d_1 = -4.101$ из условия $F(d_1) \approx 0.0001 = \alpha/n$ и $d = 6.114$ из условия $F(d) \approx 0.9999 = 1 - \alpha/n$ (в систему встроена возможность вычисления различных вероятностей для законов распределения). При анализе минимального наблюдения мы должны исключить его, если оно оказывается левее d_1, а при анализе максимального элемента выборки - правее d. Последовательно используя эту процедуру, были исключены все три грубые "ошибки" измерений.

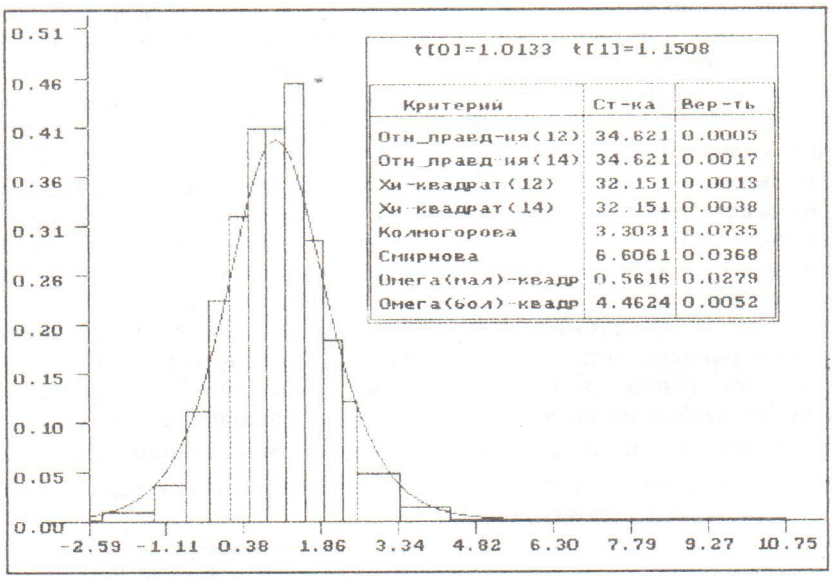

Рис. 6. Результаты оценивания параметров логистического распределения и проверки согласия с "исходной" выборкой

ОЦЕНИВАНИЕ ПАРАМЕТРОВ УСЕЧЕННЫХ РАСПРЕДЕЛЕНИЙ

Множество законов распределения, включенных в программную систему [1], охватывает 26 наиболее часто используемых в приложениях распределений: экспоненциальное, полунормальное, Рэлея, Максвелла, модуля многомерного нормального вектора, Парето, Эрланга, Лапласа, нормальное, логарифмически-нормальные (*ln* и *lg*), Коши, Вейбулла, Накагами, распределение минимального значения, распределение максимального значения, двойное показательное, гамма-распределение, логистическое, бета-распределение 1-го рода, стандартное бета-распределение 2-го рода, бета-распределение 2-го рода, распределения *Sb*-Джонсона, *Sl*-Джонсона и *Su*-Джонсона, экспоненциальное семейство распределений.

Понятно, что это не покрывает всего множества законов реальных случайных величин. Особенно часто оказываются неудачными попытки описания набором вышеуказанных распределений наблюдений, принадлежащих

усеченным законам распределения или смесям законов. Использование усечённых законов распределения и смесей различных законов существенно расширяет множество вероятностных моделей, которые могут применяться для описания реальных данных. Именно поэтому реализуемая версия программной системы дополняется возможностями, позволяющими решать задачи статистического анализа, описывая наблюдаемые величины, в том числе и смесями усеченных и неусеченных законов распределения.

Появление реальных случайных величин, подчиняющихся усеченным законам, обычно связано с существующими физическими ограничениями на область их определения. В случае двустороннего усечения закона с функцией плотности $f(x)$ имеем закон с функцией плотности

$$f_{yc}(x) = \begin{cases} 0, & x < a, \\ \dfrac{f(x)}{F(b) - F(a)}, & x \in [a, b], \\ 0, & x > b. \end{cases}$$

где a и b - параметры усечения.

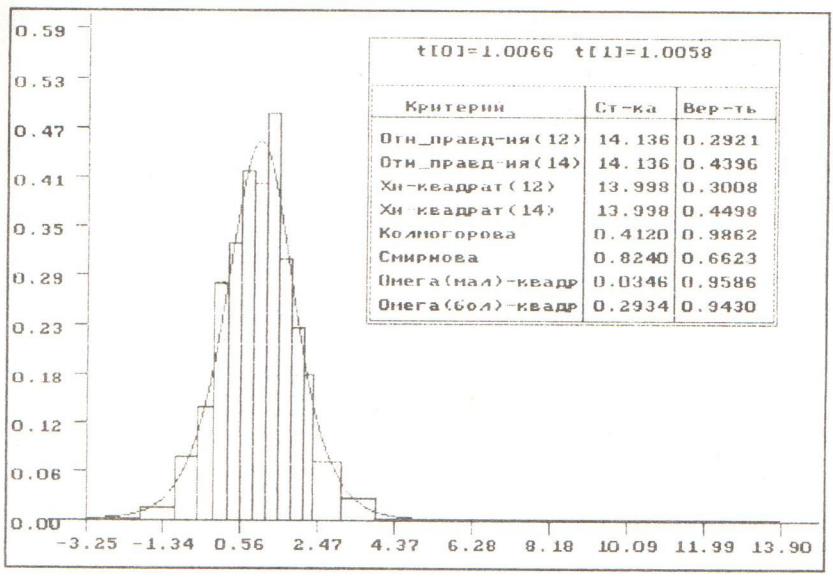

Рис. 7. Результаты оценивания параметров логистического распределения по сгруппированной выборке и проверки согласия с "исходной" выборкой

Основные сложности при анализе частично группированных выборок усеченных величин носят вычислительный характер. Для усечённых законов распределения решение задачи асимптотически оптимального группирования [2, 3], имеющее принципиально важное значение для качества статистических выводов, оказывается зависящим от параметров усечения. Это означает, что таблицы асимптотически оптимального группирования для использования в критериях согласия могут формироваться либо для конкретных значений параметров усечения, либо, что более предпочтительно, соответствующая задача должна решаться непосредственно в ходе проверки гипотезы о согласии. Свои проблемы возникают при оценивании параметров усечения. Эти параметры могут определяться исследователем исходя из

области определения наблюдаемой физической величины, либо оцениваться с использованием порядковых статистик.

Применение неусеченного закона распределения для описания усеченных наблюдений, даже если оно правомочно, обычно даёт неудовлетворительные результаты. Например, на рис. 8 приведены результаты использования нормального закона для описания смоделированной усеченной слева нормальной выборки. Согласие полученного нормального распределения с такой выборкой оказалось очень плохое, зато использование усеченного нормального распределения (рис. 9) дало очень хорошие результаты.

ОЦЕНИВАНИЕ ПАРАМЕТРОВ СМЕСЕЙ РАСПРЕДЕЛЕНИЙ

Смесь законов распределений может образоваться, например, при объединении выборок, распределенных по разным законам, или когда наблюдаемая величина является следствием различных причин.

При анализе смесей распределений возникает ряд принципиальных моментов. Рассмотрим, что представляет собой область определения параметра смеси на примере смеси двух распределений. Параметр w выбирают таким образом, чтобы функция плотности была неотрицательной:

$$f(x, w) = wf_1(x) + (1 - w) f_2(x) \geq 0, \ \forall w \in \Omega, \forall x \in X .$$

Это условие можно преобразовать к виду:

$$w\left[f_2(x) - f_1(x)\right] \leq f_2(x) .$$

Пусть $X = A \cup B \cup C$,

где

$$A = \left\{x \in X: f_2(x) - f_1(x) < 0\right\};$$

$$B = \left\{x \in X: f_2(x) - f_1(x) > 0\right\};$$

$$C = \left\{x \in X: f_2(x) - f_1(x) = 0\right\}.$$

Тогда $\Omega = \Omega_A \cap \Omega_B \cap \Omega_C$, где $\Omega_C = R$, а

$$\Omega_A = \left\{w \in R: w \geq \frac{f_2(x)}{f_2(x) - f_1(x)}, \forall x \in A\right\};$$

$$\Omega_B = \left\{w \in R: \ w \leq \frac{f_2(x)}{f_2(x) - f_1(x)}, \forall x \in B\right\}.$$

Обозначим

$$a = \max_{x \in A} \frac{f_2(x)}{f_2(x) - f_1(x)} \leq 0;$$

$$b = \min_{x \in B} \frac{f_2(x)}{f_2(x) - f_1(x)} \geq 1$$

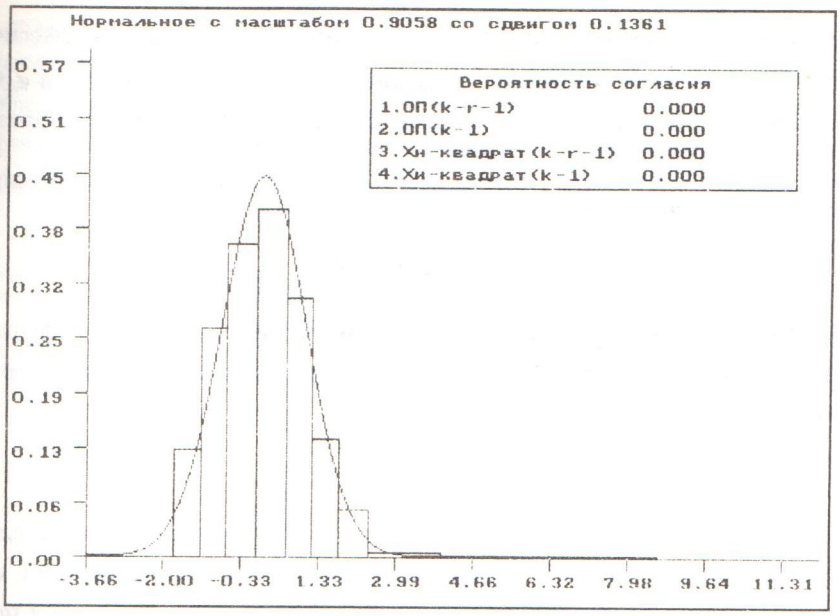

Рис.8. Результаты оценивания параметров нормального распределения по группированным данным

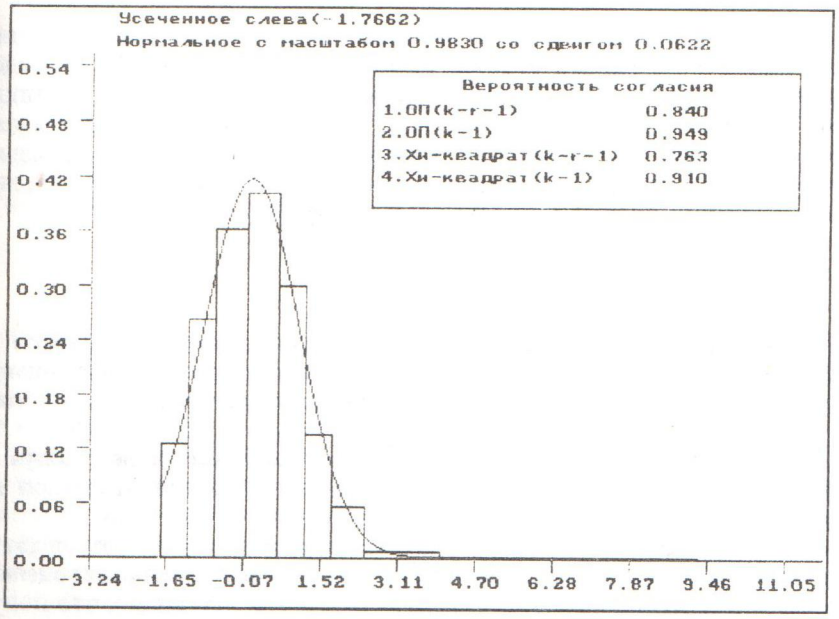

Рис. 9. Результаты оценивания параметров усеченного нормального распределения по группированной выборке

и $a=0$, если $\exists x$: $f_2(x)=0 \wedge f_1(x)\neq 0$, $b=1$, если $\exists x$: $f_1(x)=0 \wedge f_2(x)\neq 0$. В результате получим, что область определения параметра смеси имеет вид: $\Omega=\Omega_A \cap \Omega_B \cap \Omega_C=[a,b]\supseteq[0,1]$

Когда параметр смеси принадлежит интервалу $[0,1]$, мы имеем классическую смесь, получаемую, например, объединением выборок. Если же $w \notin [0,1]$, то одно из распределений входит в смесь со знаком минус и, таким образом, вычитается из другого распределения.

Обобщением смеси двух распределений является смесь из S распределений. Функция распределения в этом случае имеет вид

$$F(x) = \sum_{i=1}^{S} w_i F_i(x, \theta_i),$$

где S - число распределений в смеси; w_i - параметры смеси; F_i - *i-я* функция распределения; θ_i - вектор её параметров. Параметры смеси удовлетворяют условию нормировки:

$$\sum_{i=1}^{S} w_i = 1.$$

Смесь является *идентифицируемой*, если для $\forall w_1, w_2 \in \Omega$: $w_1 \neq w_2$ следует, что $\exists x \in X$: $f(x, w_1) \neq f(x, w_2)$. Очевидно, что смесь неидентифицируема, если $f_1 \equiv f_2$. Численная реализация метода максимального правдоподобия показала, что функция $L(\theta)$ для смеси распределений зачастую является многоэкстремальной, и поэтому получаемые оценки параметров существенно зависят от начального приближения. Качество оценки параметра смеси заметно ухудшается, когда входящие в смесь распределения достаточно близки по форме.

Остаются в силе соображения о робастности оценок по группированным данным параметров смесей законов распределения. На рис. 10 приведены результаты описания смесью двух нормальных распределений группированной выборки разных сортов бобов [19]. Смесь законов распределения может эффективно использоваться для описания ошибок измерительных приборов в тех случаях, когда реальное распределение ошибок прибора двухмодально [20], как, например, на рис. 11, или многомодально.

ЗАКЛЮЧЕНИЕ

1. Оценивание параметров распределений может осуществляться по любой выборке, т.е., как бы не были представлены исходные данные, алгоритмическое и программное обеспечение позволяет находить оценки параметров.

2. Асимптотически оптимальное группирование обеспечивает получение оценок по группированным данным с минимальной асимптотической дисперсией.

3. Таблицы асимптотически оптимального группирования позволяют получать простые оценки, частными случаями которых являются оценки с использованием оптимальных порядковых статистик. Очевидно, что получаемые оценки являются робастными.

4. Применение асимптотически оптимального группирования при проверке согласия по критериям χ^2 Пирсона и отношения правдоподобия обеспечивают им в больших выборках максимальную мощность при близких альтернативах. Это позволяет *рекомендовать обязательное* использование таблиц асимптотически оптимального группирования в процедурах проверки согласия, введение их в стандарты.

5. Есть основания, базирующиеся на экспериментальных исследованиях, для предположений в пользу того, что при асимптотически оптимальном группировании критерии χ^2 Пирсона и отношения правдоподобия являются более мощными при близких альтернативах, чем непараметрические критерии Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса.

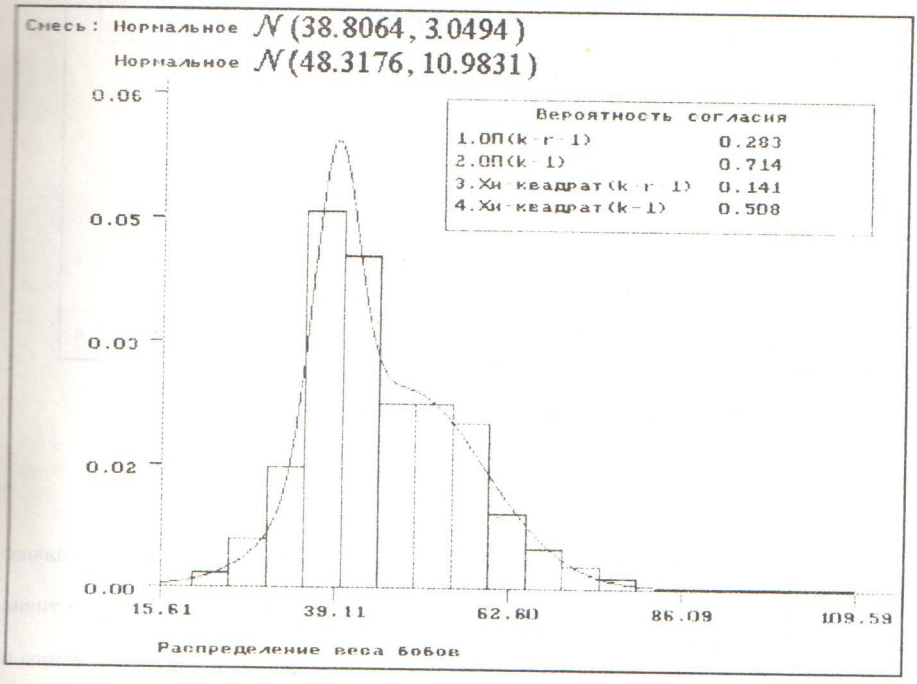

Рис. 10. Результаты оценивания параметров смеси двух нормальных распределений по группированной выборке

6. Предварительное группирование исходной выборки и последующее вычисление ОМП по группированным данным приводит к робастным оценкам, устойчивым как к наличию в исходной выборке аномальных измерений, так и к отклонениям закона распределения выборки от предполагаемого.

7. Предварительное группирование данных для вычисления робастных оценок и асимптотически оптимальное группирование в критериях согласия, используемые на этапе идентификации закона распределения по выборке, содержащей аномальные наблюдения, позволяют реализовать эффективный алгоритм параметрической отбраковки грубых ошибок измерений.

8. Использование для описания законов ошибок измерительных приборов смесей распределений усеченных и неусеченных случайных величин и возможность оценивания параметров смесей по частично группированным данным существенно расширяет класс моделей и множество приложений.

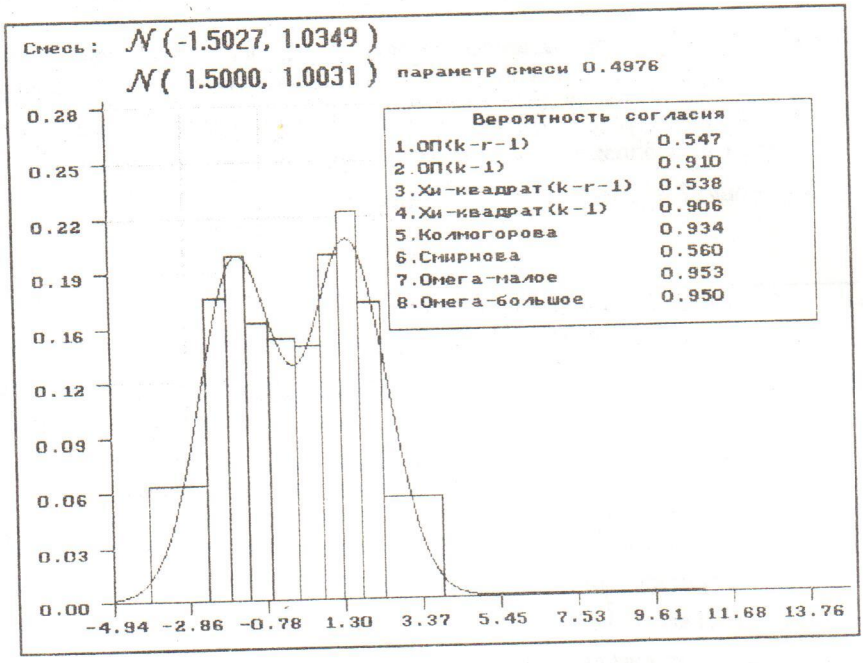

Рис. 11. Двухмодальный закон как смесь двух нормальных распределений

СПИСОК ЛИТЕРАТУРЫ

1. **Лемешко Б.Ю.** Статистический анализ одномерных наблюдений случайных величин: Программная система. - Новосибирск : Изд-во НГТУ, 1995. - 125 с.

2. **Куллдорф Г.** Введение в теорию оценивания по группированным и частично группированным выборкам. - М.: Наука, 1966. - 176 с.

3. **Денисов В.И., Лемешко Б.Ю., Цой Е.Б.** Оптимальное группирование, оценка параметров и планирование регрессионных экспериментов: В 2 ч. - Новосибирск: Изд-во НГТУ, 1993. - 346 с.

4. **Рао С.Р.** Линейные статистические методы и их применения. - М.: Наука, 1968. - 548 с.

5. **Rao C.R.** Criteria of estimation in large samples. - Sankhua, 1962. - Vol. 25. - P. 189-206.

6. **Бодин Н.А.** Оценка параметров распределения по группированным выборкам // Тр. ин-та им В.А. Стеклова АН СССР. - 1970. - Т.111. - С. 110-154.

7. **Лемешко Б.Ю.** Оценивание параметров распределений по группированным наблюдениям // Вопросы кибернетики. - М.,1977. - Вып. 30. - С.80-96.

8. **Cox D.R.** Note on grouping // J. of the American Statistical Assocition. - 1957. - Vol. 52, № 280. - P. 543-547.

9. **Денисов В.И., Лемешко Б.Ю.** Вычисление оценок параметров распределений с использованием таблиц асимптотически оптимального группирования // Применение ЭВМ в оптимальном планировании и проектировании. - Новосибирск, 1981. - С. 3-17.

10. **Дэйвид Г.** Порядковые статистики. -М.: Наука, 1979. - 336 с.

11. **Большев Л.Н., Смирнов Н.В.** Таблицы математической статистики. - М.: Наука,1965.-464 с.

12. **Лемешко Б.Ю., Постовалов С.Н.** Статистический анализ одномерных наблюдений по частично группированным данным // Изв. высших учебных заведений. Физика. - Томск, 1995. - № 9. - С. 39-45.

13. **Лемешко Б.Ю., Постовалов С.Н.** К использованию непараметрических критериев по частично группированным данным // Сб. науч. тр. НГТУ. - Новосибирск, 1995. - № 2. - С. 21-30.

14. **Кокс Д., Хинкли Д.** Теоретическая статистика. - М.: Мир, 1978. -560 с.

15. **Кендалл М.Дж., Стьюарт А.** Статистические выводы и связи. - М.: Наука, 1973. - 903 с.

16. **Тарасенко Ф.П.** Непараметрическая статистика. - Томск: Изд-во Том. ун-та, 1976. - 292 с.

17. **Хьюбер П.** Робастность в статистике. - М.: Мир, 1984. - 303 с.

18. **Орлов А.И.** Неустойчивость параметрических методов отбраковки резко выделяющихся наблюдений // Заводская лаборатория. - 1992. - Т. 58, № 7. - С. 40-42.

19. **Ван дер Варден Б.Л.** Математическая статистика. - М.: Иностранная лит., 1960. - 435 с.

20. **Новицкий П.В., Зограф И.А.** Оценка погрешностей результатов измерений. - Л.: Энергоатомиздат, 1991. - 303 с.

Лемешко Борис Юрьевич, кандидат технических наук, доцент, начальник отдела планирования и управления НИОКР. Основное направление научных исследований - статистический анализ данных. Имеет более 50 публикаций, в том числе 2 монографии.

Постовалов Сергей Николаевич, аспирант кафедры прикладной математики. Направление научных исследований - статистический анализ частично группированных и интервальных наблюдений. Имеет 7 публикаций.