Робастные алгоритмы оценивания и параметрические методы отбраковки аномальных наблюдений

Б.Ю. Лемешко, С.Н. Постовалов

Кафедра прикладной математики, НГТУ, Новосибирск, Россия

Аннотация. Показана устойчивость к аномальным измерениям оценок максимального правдоподобия параметров законов распределения при использовании предварительного группирования данных. Показано, что применение асимптотически оптимального группирования данных при проверке гипотез по критериям χ^2 Пирсона и отношения правдоподобия делает их более чувствительными к наличию аномальных наблюдений. Проиллюстрирована эффективность отбраковки грубых ошибок измерений параметрическими методами, когда на первом этапе идентификации закона используются робастные алгоритмы.

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.

Параметрическая процедура отбраковки грубых ошибок измерений в одномерном случае выглядит следующим образом. 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 нулевой гипотезы $P\{\max\limits_{1\le i\le n} X_i \le d\} = [F(d)]^n = 1-\alpha$, 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 $x_{\min} \ge d_1$. При справедливости гипотезы H_0 $P\{\min\limits_{1\le i\le n} X_i \ge d_1\} = [1 - F(d_1)]^n = 1-\alpha$. И значение $d_1 = d_1(\alpha, n)$ определяется из уравнения $1 - F(d_1) = \sqrt[n]{1-\alpha}$.

Неустойчивость такой процедуры отбраковки может быть связана с возможным неточным определением закона $F(x)$ и трудностью различения близких законов распределения с помощью критериев согласия.

Приводимые далее примеры и выводы получены на основании результатов, которые были использованы при создании программной системы [1], дополненной параметрической процедурой отбраковки аномальных наблюдений.

В данной работе отметим два достоинства методов, реализованных в программной системе. Во-первых, свойства получаемых оценок, использующих группирование исходных выборочных данных. Очевидно, что они менее чувствительны к случайным выбросам. *Группирование выборки позволяет резко снизить влияние аномальных наблюдений, а иногда и практически исключить влияние грубых ошибок измерений.* Во-вторых, использование асимптотически оптимального группирования в критериях отношения правдоподобия и χ^2 Пирсона [2]. Мощности критериев отношения правдоподобия и χ^2 Пирсона пропорциональны количеству информации Фишера о параметрах распределения в группированной выборке. Асимптотически оптимальное группирование минимизирует потери информации, связанные с группированием и следовательно гарантирует *максимальную мощность различения близких альтернатив* для этих критериев.

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 методы должны быть как можно менее чувствительны к наличию аномальных ошибок измерений. Наоборот, на последующем этапе отбраковки критерий должен улавливать их наличие и позволять отсекать.

Таким образом, при идентификации (при оценивании параметров распределений) мы должны использовать робастные алгоритмы (устойчивые к наличию аномальных наблюдений), а на последующем этапе отбраковки желательна максимальная мощность критерия для различения близких альтернатив (чувствительность к грубым ошибкам). В этой связи мы рекомендуем на первом этапе использовать оценки по группированным данным, причем для большей устойчивости оценок осуществлять разбиение выборки на интервалы равной вероятности (равночастотные интервалы), а на втором этапе при проверке согласия - разбиение на асимптотически оптимальные интервалы.

Продемонстрируем сказанное на конкретном примере. Была смоделирована выборка объёмом 1000 наблюдений в соответствии с распределением Вейбулла с плотностью

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

При моделировании были заданы параметры: $\theta_0=2,\ \theta_1=1,\ \theta_2=0.$ В процессе регистрации 8 наблюдений "подверглись" сильным искажениям.

На рис.1-2 приведены результаты статистического анализа полученной выборки. Здесь и в дальнейшем $t[0]=\theta_0,\ t[1]=\theta_1,\ t[2]=\theta_2$. На рисунках отражены результаты проверки гипотез о согласии: вычисленные значения S^* соответствующих статистик S и вероятности превышения полученного значения статистики при истинности нулевой гипотезы $P\{S>S^*\}$. Проверка гипотез о согласии осуществляется по ряду критериев: отношения правдоподобия, χ^2 Пирсона, Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса. Гипотеза о согласии по соответствующему критерию не отвергается, если $P\{S>S^*\}>\alpha$. В данном случае получили закон распределения Вейбулла с параметрами $\theta_0=1.4433,\ \theta_1=1.0613,\ \theta_2=0.$ Как видим из рис. 1, согласие по всем критериям отвергается: наличие аномальных наблюдений сыграло свою роль.

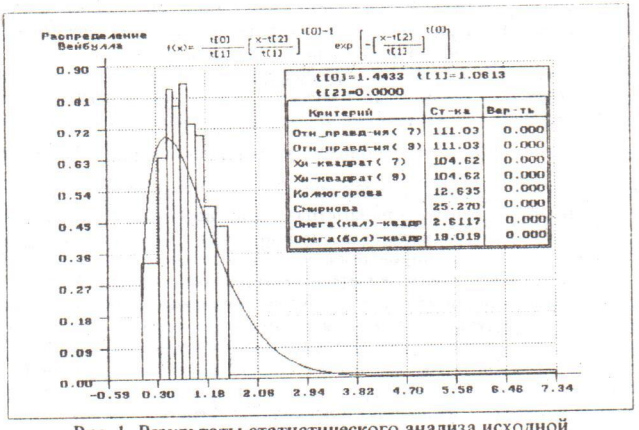

Рис. 1. Результаты статистического анализа исходной выборки по негруппированным данным

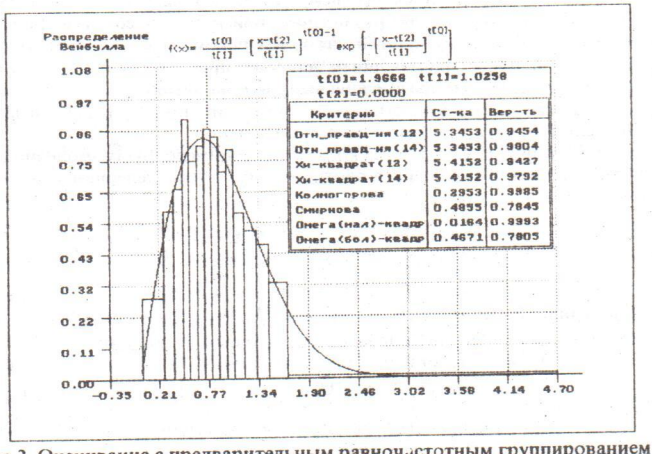

Рис.2. Оценивание с предварительным равночастотным группированием и проверкой гипотез о согласии с разбиением на равночастотные интервалы.

На рис. 2 представлены результаты статистического анализа, когда перед оцениванием выборка была разбита на интервалы равной частоты, затем по получившейся группированной выборке были найдены оценки параметров $\theta_0 = 1.9668$, $\theta_1 = 1.0258$, $\theta_2 = 0$. При проверке гипотез о согласии исходная выборка разбивалась на интервалы равной вероятности. Как видим, результаты проверки гипотез о согласии по всем критериям очень хорошие.

Отличие результатов на рис. 3 определяется тем, что при проверке гипотез о согласии исходная выборка разбивалась на асимптотически оптимальные

интервалы. В данном случае критерии отношения правдоподобия и χ^2 Пирсона оказываются более чувствительными, чем остальные, *улавливают наличие аномальных измерений*. Гипотезы о согласии при $\alpha \geq 0.0027$ по этим критериям должны быть отвергнуты.

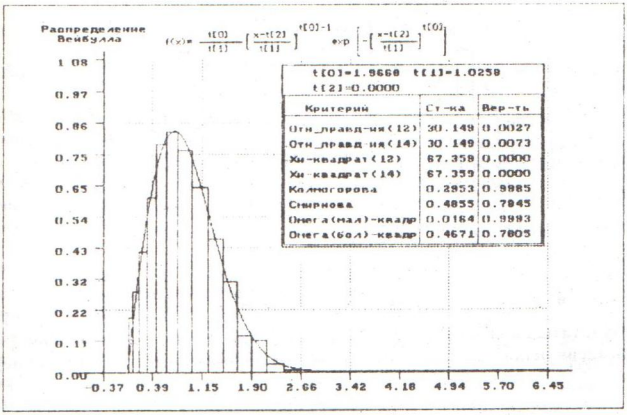

Рис. 3. Оценивание с предварительным равночастотным группированием и проверкой гипотез о согласии с разбиением на асимптотически оптимальные интервалы

Для отбраковки аномальных наблюдений зададимся уровнем значимости $\alpha = 0.1$ и при объёме выборки $n = 1000$ и векторе параметров $\theta^T = [1.9668,\ 1.0258,\ 0]$ распределения Вейбулла найдем $d = F^{-1}(\sqrt[n]{1-\alpha}) \approx$ ≈ 3.1721 (в систему встроена возможность вычисления различных вероятностей для законов распределения). Далее, мы должны исключить те наблюдения, которые превышают величину 3.1721. Таких наблюдений оказалось 8.

На рис.4 отражены результаты статистического анализа выборки после исключения из неё аномальных наблюдений. При проверке согласия использовано асимптотически оптимальное группирование. Как видим, согласие по всем критериям очень хорошее.

В довершение картины на рис.5 приведены результаты проверки согласия найденного закона (после отбраковки грубых ошибок измерений) с исходной выборкой, содержащей ошибки измерений, с применением асимптотически оптимального группирования. Заметно, что параметрические критерии "не замечают" присутствия грубых ошибок в выборке.

Выводы

1. При решении задачи отбраковки на этапе идентификации закона распределения следует использовать робастные алгоритмы, устойчивые к наличию аномальных наблюдений. Высокую устойчивость к присутствию в выборке грубых искажений или принадлежности выборки к другому закону распределения проявляют оценки максимального правдоподобия по *группированной* выборке. Обычно такие оценки наиболее устойчивы при разбиении области определения случайной величины на *интервалы равной вероятности*. В некоторых случаях более устойчивы оценки с использованием оптимального группирования.

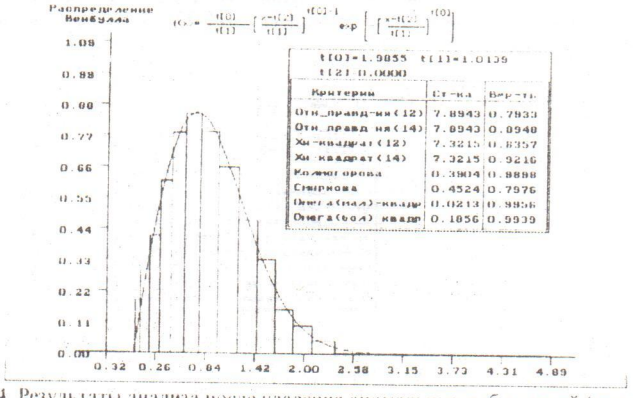

Рис.4. Результаты анализа после удаления аномальных наблюдений (при проверке согласия использовано асимптотически оптимальное группирование)

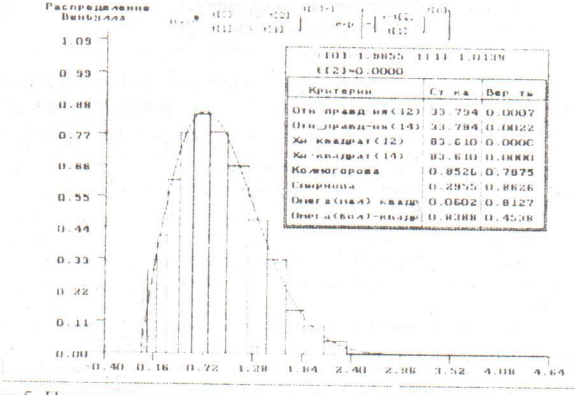

Рис.5. Проверка согласия с исходной выборкой (при проверке согласия использовано асимптотически оптимальное группирование)

2. При использования робастных оценок параметрический метод отбраковки грубых ошибок позволяет эффективно исключать аномальные наблюдений.

Литература

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

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. - 316 с.