ПРОГРАММНАЯ СИСТЕМА СТАТИСТИЧЕСКОГО АНАЛИЗА "ОЦЕНИВАНИЕ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЙ"

Б.Ю. Лемешко

Россия, Новосибирск, НГТУ

Программная система предназначена для статистической обработки экспериментальных данных. Она позволяет находить оценки максимального правдоподобия параметров для 26 наиболее часто используемых в приложениях распределений: экспоненциального, полунормального, Рэлея, Максвелла, модуля многомерного нормального вектора, Парето, Эрланга, Лапласа, нормального, логарифмически нормальных (ln и lg), Коши, Вейбулла, Накагами, распределения минимального значения, распределения максимального значения, двойного показательного, гамма-распределения, логистического, бета-распределения 1-го рода, стандартного бета-распределения 2-го рода, бета-распределения 2-го рода, распределений Sb-Джонсона, Sl-Джонсона и Su-Джонсона, семейства экспоненциальных распределений.

По ряду возможностей система не имеет аналогов среди программного обеспечения задач статистического анализа. **Во-первых** исходная выборка может быть негруппированной, группированной или частично группированной. Простейшими случаями частично группированных выборок являются *цензурированные* выборки. **Во-вторых**, при проверке гипотез по критериям χ^2 Пирсона или отношения правдоподобия используются полученные таблицы асимптотически оптимального группирования данных [1], обеспечивающие максимальную мощность критерия при близких альтернативных гипотезах. **В-третьих**, очевидным достоинством оценок, использующих группирование исходных выборочных данных является то, что они менее чувствительны к случайным выбросам. *Группирование выборки позволяет резко снизить влияние аномальных наблюдений, а иногда и совсем исключить влияние случайных выбросов*, что бывает чрезвычайно важно, например, при передаче информации по каналам связи. **В-четвертых**, применение системы при обработке результатов даст возможность использовать оптимальное группирование при проведении экспериментов и регистрации их результатов, что позволяет резко сократить объем хранимых и передаваемых по каналам связи данных без существенной потери информации о законе распределения наблюдаемой случайной величины. И в-пятых. Проверка гипотез о согласии осуществляется по ряду критериев: χ^2-Пирсона, отношения правдоподобия, Колмогорова, Смирнова, ω^2 и Ω^2 Мизеса. При использовании общепринятой методики проверки гипотез по критериям согласия, когда гипотеза о согласии с данным распределением не отвергается, если вычисленное значение статистики не превышает критического, соответствующего заданному уровню значимости α, обычно оказывается, что нет причин отказаться от целого ряда распределений. В описываемом программном обеспечении при проверке гипотез о согласии для каждой используемой статистики S_i, $i=\overline{1,m}$,

вычисляются вероятности вида $P\{S_i > S_i^*\} = \int\limits_{S_i^*}^{\infty} g_i(s)ds$, где S_i^* - найденное по выборке значение соответствующей статистики, $g_i(s)$ - функция плотности распределения статистики S_i при условии, что гипотеза H_0 является истинной. Гипотеза о согласии не отвергается, если $P\{S_i > S_i^*\} > \alpha$. Обычно вывод о наиболее подходящем распределении можно сделать однозначно. Однако довольно часто для различных, но близких законов распределения выводы по различным критериям указывают на предпочтительность различных законов. Т.е. при общем "согласии" по всем критериям один критерий указывает на предпочтительность одного закона, второй - на предпочтительность некоторого другого и т.д. Это означает, что решения задачи выбора распределения по различным критериям не совпадают. Такая "несогласованность" объясняется различием мер, используемых в критериях. В этом случае решается *многокритериальная задача выбора* распределения.

Предусмотрена возможность вычисления вероятностей различных событий для найденных законов. Встроенные датчики позволяют имитировать выборки в соответствии с включенными в систему законами распределений.

Чтобы показать, насколько эффективно можно бороться с влиянием аномальных измерений, используя группирование данных, приведем пример с нормальным распределением. Была смоделирована выборка по нормальному закону с нулевым математическим ожиданием и единичной дисперсией, состоящая из 500 наблюдений. На рис. 1 представлены результаты оценивания параметров по этой выборке и проверки гипотез о согласии. Оценка математического ожидания равна 0.01469 и среднего квадратического отклонения - 0.99.

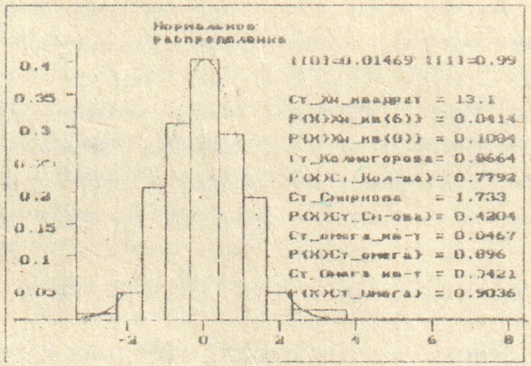

Рис. 1. Результаты анализа на исходной выборке

Затем в данной выборке увеличили первое наблюдение на 20 и снова провели соответствующий анализ. Его результаты приведены на рис. 2 и 3. Как и следовало ожидать, наиболее существенно изменилась оценка среднеквадратичного отклонения. Согласие по всем критериям отвергается.

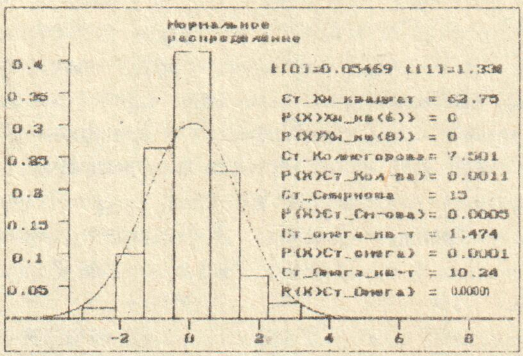

Рис. 2. Результаты анализа при наличии "аномального" наблюдения

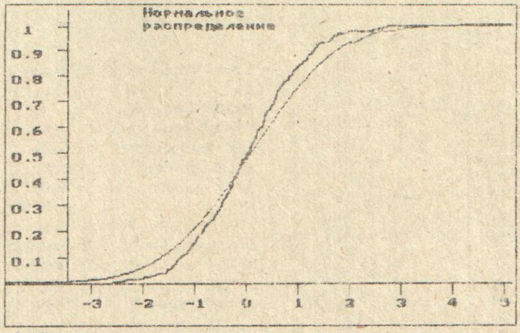

Рис. 3. Теоретическая и эмпирическая функции распределения при наличии "аномального" наблюдения

А далее осуществили группирование выборки с "аномальным" наблюдением, провели оценивание по группированной выборке и проверили гипотезы о согласии. Результаты представлены на рис. 4.

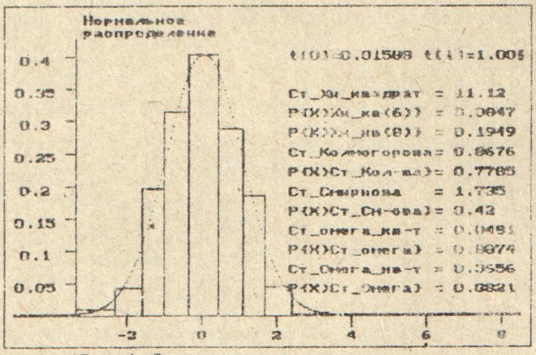

Рис. 4. Результаты робастного оценивания

Как видим, "случайная" ошибка в данных практически не повлияла на оценки параметров. Посмотрим, как поведут себя оценки максимального правдоподобия параметров нормального распределения, когда на самом деле выборка принадлежит другому распределению, например, распределению Коши. Распределение Коши - это распределение с "тяжёлыми" хвостами. Моделировалась выборка по закону Коши с параметрами $\theta_0 = 1$, $\theta_1 = 0$. На рис. 5. представлены результаты моделирования.

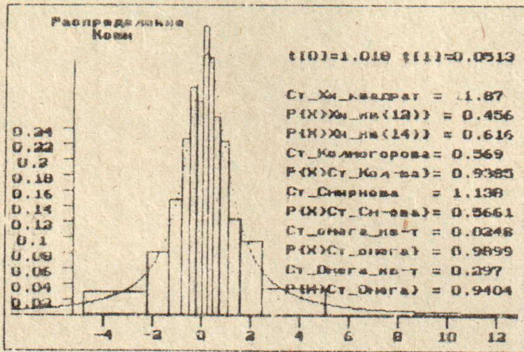

Рис. 5. Результаты моделирования распределения Коши

Предполагая, что на самом деле выборка принадлежит нормальному распределению, оценили его параметры. Результаты оценивания и проверки гипотез о согласии представлены на рис. 6, где пологая кривая это теоретическая функция распределения, а ступенчатая - эмпирическая.

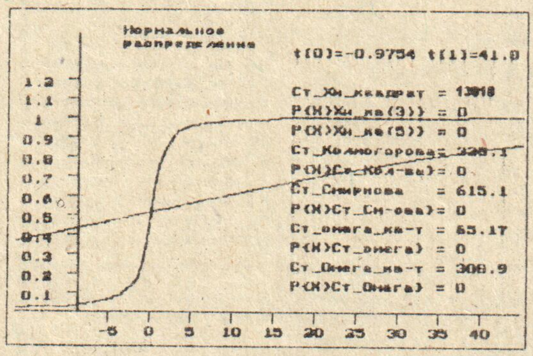

Рис. 6. Эмпирическая и теоретическая функции распределения нормального закона при вычислении оценок по негруппированной выборке

Далее нашли робастные оценки. Для этого предварительно сгруппировали выборку, разбив её на интервалы равной частоты (равной вероятности), а затем по группированной выборке оценили параметры нормального распределения. В данном случае выборка разбивалась на 6 интервалов. Результаты оценивания и анализа представлены на рис. 7.

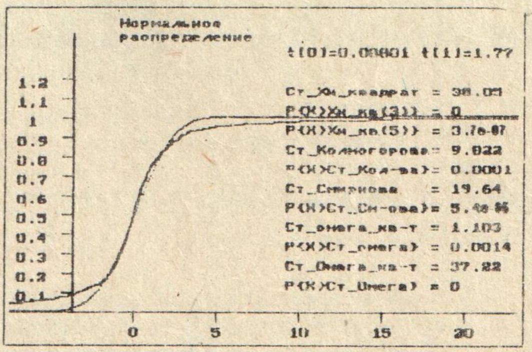

Рис. 7. Эмпирическая и теоретическая функции распределения нормального закона при вычислении оценок по группированной выборке

Естественно, что хотя о согласии найденного нормального закона с выборочными данными говорить не приходится, так как моделировался всё-таки закон распределения Коши, преимущество оценок, найденных по сгруппированной выборке, очевидно. Из рисунка видна близость эмпирической и теоретической функций распределения в середине и расхождение на хвостах.

Программная система успешно используется в Новосибирском государственном техническом университете в научных исследованиях и в учебном процессе, поставлена в ряд ВУЗов России и стран СНГ.

Литература

1. В.И.Денисов, Б.Ю. Лемешко, Е.Б. Цой. Оптимальное группирование, оценка параметров и планирование регрессионных экспериментов. Новосибирск, 1993, 347 с.