УДК 519.222

Д.А. Паляничко,
В.И. Згуря, канд. техн. наук,
Р.Г. Папуша

ПРИМЕНЕНИЕ РОБАСТНЫХ МЕТОДОВ АНАЛИЗА ДАННЫХ ПРИ ОГРАНИЧЕННОМ ЧИСЛЕ НАБЛЮДЕНИЙ

Измерительная информация малого объема характеризуется асимметрией распределения, что приводит к ситуации, когда необходимо принимать решения, которые могут снижать статистическую надежность результатов. Применение робастных методов анализа данных помогает избежать подобной ситуации, при этом позволяет восстановить возможную генеральную совокупность.

Ключевые слова: *асимметрия, выброс, робастность, СКО, среднее значение, медиана.*

D.A. Palianychko,
V.I. Zgurya, PhD.,
R.G. Papusha

APPLICATION OF THE ROBUST DATA ANALYSIS METHODS WITH LIMITED NUMBER OF OBSERVATIONS

Measuring information of a small size distribution is characterized by asymmetry, which leads to a situation when you need to make decisions that can reduce the statistical reliability of the results. The use of a robust data analysis methods helps to avoid such situation, while allowing recovering a possible general population.

Keywords*: asymmetry, outlier, robust, standard error, mean value, median.*

Д.О. Паляничко,
В.І. Згуря, канд. техн. наук,
Р.Г. Папуша

ЗАСТОСУВАННЯ РОБАСТНИХ МЕТОДІВ АНАЛІЗУ ДАНИХ ПРИ ОБМЕЖЕНІЙ КІЛЬКОСТІ СПОСТЕРЕЖЕНЬ

Вимірювальна інформація малого обсягу характеризується асиметрією розподілу, що призводить до ситуації, коли необхідно приймати рішення, які можуть знижувати статистичну надійність результатів. Застосування робастних методів аналізу даних допомагає уникнути подібної ситуації, при цьому дає змогу відновити можливу генеральну сукупність.

Ключові слова*: асиметрія, викид, робастність, СКВ, середнє значення, медіана.*

Введение. Процесс измерения и обработка полученной при этом информации являются неотъемлемыми этапами процесса измерения, потому что в зависимости от того, какие данные получает специалист, принимается решение о том, какие методы будут использованы для обработки полученных данных. Ограниченный объем данных, чаще всего, является результатом дороговизны процесса получения измерительной информации, применения методов с разрушением объекта или же проведения уникальных исследований. В той или иной мере приходится работать с «неполной» информацией. Процесс усугубляется еще и проявлением аномальных значений, которые вносят свои негативные коррективы в общую картину.

© Паляничко Д.А., Згуря В.И.,
Папуша Р.Г., 2012

Выброс (аномальное значение) — значениевыборки, которое резко отличается от всего набора данных и по решению специалиста по статистике может быть исключено из дальнейшей обработки. Данное решение «облагородит» выборку и сделает ее однородной, но данный подход правомочен, когда мы имеем дело с неограниченным объемом информации (на практике, при числе значений больше 30). При этом могут быть использованы параметрические методы, к примеру, метод усеченного среднего. Когда же число значений не превосходит 10 (а в некоторых случаях и 5), то отбрасывание даже одного значения приводит к потере статистической надежности.

В настоящее время для выявления выбросов рекомендуется применять положение [2], в котором для определения аномальных значений предлагается использовать критерий Граббса, предписывающий проверять наи-

большее либо же наименьшее из значений результатов испытаний на наличие выбросов и квазивыбросов. Данный критерий хорошо себя зарекомендовал, но для малых объемов выборки [3] существует такое понятие, как асимметрия распределения, ввиду которого значения из одной генеральной совокупности могут классифицироваться как выбросы. Другими словами, так называемые «плохие» значения будут на самом деле «хорошими» и наоборот.

Вся статистическая обработка и принимаемые на ее основании решения базируются на предположении о нормальности распределения [4]. Это в основном обосновано тем, что имеется хорошо разработанная теория статистических выводов и подходов. Однако в ряде практических задач нет достаточного объема исходных данных для построения параметрических моделей, адекватных экспериментальным данным. Ввиду условности закона распределения фактически являющегося предполагаемой моделью, которой должны соответствовать экспериментальные данные, сама по себе реальная выборка может иметь некоторые расхождения с идеалом (особенно при малых объемах) – содержать некоторые значения, которые подчиняющиеся другому закону, а не предполагаемому.

В связи с этим предлагается использовать робастные методы обработки данных, которые устойчивы к выбросам и «не привязаны» к форме закона распределения, при этом нет необходимости уменьшать объем выборки, тем самым снижая статистическую надежность.

Целью данной статьи является презентация подхода, направленного на повышение достоверности результатов испытаний.

Основная часть. Робастность в статистике предоставляет подходы, направленные на снижение влияния выбросов и других отклонений в исследуемой величине от моделей, используемых в классических методах статистики. Под робастностью понимают нечувствительность к различным отклонениям и неоднородностям в выборке, связанными с теми или иными причинами.

Рассмотрим алгоритм итерационного робастного метода.

Обозначим индексом p общее число данных, расположенных в порядке возрастания: $x_1, x_{2,\ldots}, x_{i,\ldots}, x_p$.

Обозначим робастные среднее и стандартное отклонения этих данных соответственно x^* и s^*.

Рассчитаем [5] первоначальные значения для x^* и s^* в виде:

$$x^* = \text{медиана от } x_i \ (i = 1,2,\ldots,p),$$

$$s^* = 1{,}483 * \text{медиана от } |x_i - x^*| (i = 1,2,\ldots,p),$$

где 1,483 поправочный коэффициент для приведения к нормальному закону распределения.

Обновим значения x^* и s^* по следующему алгоритму:

Рассчитываем пороговое значение:

$$\varphi = 1.5 s^*,$$

где 1,5 – степень робастности и выбирается из диапазона (1…2), около 1 % выбросов предполагают c=2,0 и около 5 % c=1,4. Значение c=1,5 используется в подавляющем большинстве случаев. [6]

Для каждого значения $x_i \ (i = 1,2,\ldots,p)$ необходимо провести модификацию имеющихся данных, а именно:

$$x_i^* = \begin{cases} x^* - \varphi, \text{при } x_i < x^* - \varphi, \\ x^* + \varphi, \text{при } x_i > x^* + \varphi, \\ x_i - \text{в остальных случаях.} \end{cases}$$

Новые робастные значения среднего и СКО x^* и s^* рассчитываем по формулам

$$x^* = \sum_{i=1}^{p} x_i^* / p, \quad (1)$$

$$s^* = 1.134 \sqrt{\sum_{i=1}^{p} (x_i^* - x^*)^2 / (p-1)}, \quad (2)$$

где коэффициент 1,134 соответствует c=1,5 для нормального распределения [7].

По соотношению между значениями x^* и s^*, вычисляемыми на текущем и предыдущем этапе итерации, определяем сходимость алгоритма. Процедура повторяется до тех пор, пока x^* и s^* от одного расчета до следующего станут минимальными.

Моделирующий эксперимент. Была проанализирована выборка из четырех значений, взятых из генеральной нормально распределенной совокупности с центром $\mu_X = 79,75$ и $\sigma = 8,15$.

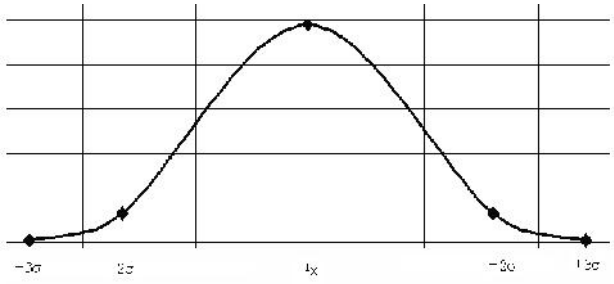

Рис. 1. Графическое изображение распределения генеральной совокупности

По правилу «трех сигм» в нормально распределенных данных отклонение значения от его математического ожидания не должно превышать $\pm 3\sigma$ (т.е. вероятность такой ситуации $P = 0{,}01$). Рассмотрим значения, которые отстоят от центра распределения не более чем на $\pm 2\sigma$, что соответствует вероятности P=0,05, т.е. выборка значений из этой генеральной совокупности не будет содержать выбросов и квазивыбросов.

На рис. 2 и 3 изображены выборки, в которых по критерию Граббса отстоящее значение признано выбросом. Проблема использования критерия Граббса при малом числе наблюдений, в основном, состоит в том, что, к примеру, для нашего случая, исключив одно значение из дальнейшей обработки мы потеряем 25 % информации и существенно занизим СКО по сравнению с реальным.

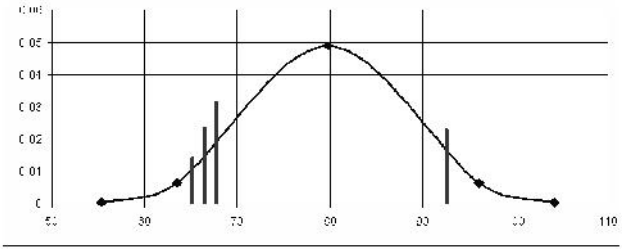

Рис. 2. Графическая иллюстрация 1 выборки с «псевдо выбросом»

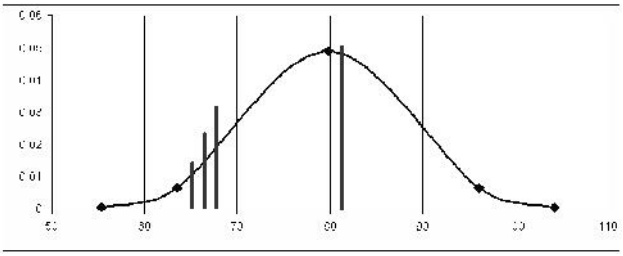

Рис. 3. Графическая иллюстрация 2 выборки с «псевдо выбросом»

При этом могут делаться неверные выводы. В этой связи становится актуальным использование методов, которые бы позволили не отбрасывать «подозрительное» значение и показывали бы реальное распределение данных [1].

Рассмотрим рис. 3. Исходная выборка значений будет иметь вид 75,3; 76; 76,3; 89.

Для начала проверим по критерию Граббса, является ли максимальное значение выбросом:

$$G_{max} = \frac{(x_{max} - \overline{x})}{\hat{s}} = 1.4969 > G_{кр} > 1.496,$$

где $G_{кр}$ – 1 %-ное критическое значение критерия Граббса (для четырех значений равно 1,496); $\overline{x}$, $\hat{s}$ – среднее значение и СКО, рассчитанные по известным формулам:

$$\overline{x} = \frac{1}{p}\sum_{i=1}^{p} x_i;\ \hat{s} = \sqrt{\frac{1}{p-1}\sum_{i=1}^{p}(x_i - \overline{x})^2}.$$

Так как $G_{max} > G_{кр} > 1.496$, то тестируемая позиция 89 признается статистическим выбросом и специалист по статистике должен принять решение об исключении данного значения из дальнейшей обработки. Проводится пересчет оценок среднего и СКО без отброшенного значения.

Таким образом, получаем $\overline{x} = 75.87$ и $\hat{s} = 0.513$. Это значение СКО значительно меньше теоретического ($\sigma = 8.15$), что говорит о негативных последствиях отбраковки результата. К примеру, робастный алгоритм дал результаты: $\overline{x} = 79.15$ и $\hat{s} = 7.46$.

Для дальнейших исследований, максимальное значение исходной выборки (x_{max}) будем уменьшать до значения 82 (это минимальное значение, которое по критерию Граббса определен как выброс). К полученным данным применим робастную процедуру, а результаты занесем в табл. 1.

По результатам эксперимента (табл.) построим графики зависимостей показателей $(\overline{x} - \mu)$ и $(\hat{s}^2_{роб} / \sigma^2)$ от количества шагов алгоритма. В результате было определено, что число шагов напрямую связано с отклонением, чем больше отклонение – тем больше шагов.

Результаты эксперимента

x_{max}	$(\bar{x}-\mu)$	$(\hat{s}^2_{роб}/\sigma^2)$	Количество шагов робастного алгоритма
82	2,35	0,185	13
83	2,1	0,25	15
84	1,85	0,324	16
85	1,6	0,407	17
86	1,35	0,5	18
87	1,1	0,603	19
88	0,85	0,716	20
89	0,6	0,838	21

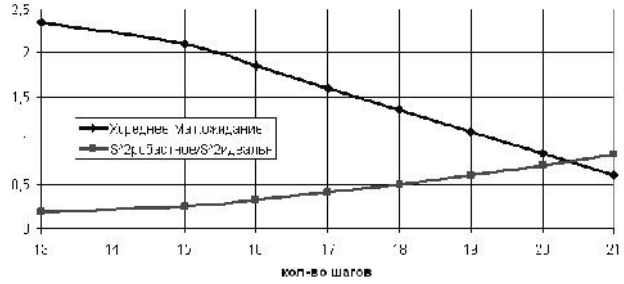

Рис. 4. Зависимость показателей $(\bar{x}-\mu)$ и $(\hat{s}^2_{роб}/\sigma^2)$ от количества итераций робастного алгоритма

Выводы

При работе с выборками малых объемов (меньше 10) главной проблемой для специалиста по статистике является асимметрия (перекос) закона распределения. Из-за этого часто измерительная информация трактуется неверно; так, к примеру, значения, попадающие на границы $(\pm 2\sigma)$, могут классифицироваться как выбросы, или же квазивыбросы. И наоборот, выбросами могут признаваться «хорошие» данные. В связи с этим и возникает проблема оценки данных при малом числе наблюдений.

Статистический критерий Граббса, как показывают результаты приведенного примера, при малых объемах выборки не учитывает асимметрию, соответственно возникают ситуации, когда приходится исключать данные, тем самым беспочвенно занижать реальное СКО и терять в статистической надежности.

Результаты проведенного моделирующего эксперимента показали, что итерационный робастный метод прекрасно работает со значениями, которые имеют значительные отклонения. Оценки среднего и СКО несущественно отличаются от теоретически рассчитанных, в отличие от критерия Граббса, по которому ложно классифицируется максимальное значение как выброс. Процесс обработки результатов не занимает много времени и ресурсов, что свидетельствует об универсальности робастных методов.

Список использованной литературы

1. Володарский Е. Т. Робастное оценивание точностных характеристик результатов наблюдений: «Системи обробки інформації» / Е. Т., Володарский, Д. А. Паляничко. – Харків: // Збірник наук. Праць, видав. Харківського ун-ту Повітряних Сил ім. І. Кожедуба. – 2011. – Вип. 8 (98).

2. ГОСТ ИСО 5725-2:2005 Точность (правильность и прецизионность) методов и результатов измерений. Часть 2. Основной метод определения повторяемости и воспроизводимости стандартного метода измерения (ГОСТ ИСО 5725-2-2003, IDT).

3. Володарський Е. Т. Корректность применения критерия Граббса при анализе результатов испытания с тремя элементами: «Системи обробки інформації» / Е. Т. Володарський, И. А. Харченко, В. И. Згуря, М. Е. Молочков // Збірник наук. Праць, видав. Харківського ун-ту Повітряних Сил ім. І. Кожедуба. –:. Харків: – 2007. – Вип. № 6 (64). – С.20 – 22.

4. Тьюки Дж. Анализ результатов наблюдений / Дж. Тьюки // Пер. с англ. – М.: Мир, 1981. – 325 с.

5. Хьюбер Дж. П. Робастность в статистике / Дж. П. Хьюбер // Пер. с англ. – М.: Мир, 1984. – 308 с., ил.

6. Analyst Robust statistics – How Not to Reject Outliers. December 1989, vol. 114.

7. James N. Miller Statistics and Chemometrics for Analytical Chemistry / James N. Miller, Jane C. Miller – 6th ed. – 2010.

Получено 11.05.2012

References

1. Volodarsky E. T., Palianychko D. A. Robust estimation accuracy characteristics of the observations: "Information processing systems" Scientific Papers. Issue 8 (98) - Kharkov, publisher Kharkiv Air Force University named after Ivan Kozhedub, 2011. [in Ukrainian].

2. ISO 5725-2:1994 Accuracy (trueness and precision) of measurement methods and results -- Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method [in Russian].

3. Volodarsky E.T, Kharchenko I.A, Zgurya V.I., Molochkov M.E. The correctness of Grubbs criterion in the analysis of test results with three elements: "Information processing systems" Scientific Papers. Issue 6 (64) - Kharkov, publisher Kharkiv Air Force University named after Ivan Kozhedub, 2007. - P.20 - 22. [in Russian].

4. Tukey, J. An analysis of observations - first. from English. – Moscow. Mir: 1981. - 325 p. [in Russian].

5. Peter J. Huber Robust statistics - first. from English. – Moscow. Mir: 1984. - 308 p. [in Russian].

6. Analyst Robust statistics – How Not to Reject Outliers. December 1989, vol. 114. [in English].

7. James N. Miller, Jane C. Miller Statistics and Chemometrics for Analytical Chemistry – 6th ed. 2010. [in English].

Паляничко
Денис Александрович,
аспирант каф. автоматизации экспериментальных исследований Нац. технич. ун-та Украины «Киевский политехнический институт;
E-mail: polunychko@ukr.net.

Згуря
Вячеслав Иванович,
канд. техн. наук, ст. науч. сотрудник Госуд. предприятия "Всеукраинский государственный научно-производствен. центр стандартизации, метрологии, сертификации и защиты прав потребителей» (ГП «Укрметртестстандарт»);
E-mail: v.i.zgurya@ukr.net.

Папуша Роман Григорьевич,
начальник отдела стандартизации и системы управления качеством научно-исслед. центра нормативно-правового регулирования Украинского научно-исследоват. ин-та гражданской защиты;
E-mail: roman_p_r_g@ukr.net.