УДК 614.001.89

Е.Т. Володарский, Д.А. Паляничко

Национальный технический университет Украины «КПИ», Киев

РОБАСТНОЕ ОЦЕНИВАНИЕ ТОЧНОСТНЫХ ХАРАКТЕРИСТИК РЕЗУЛЬТАТОВ НАБЛЮДЕНИЙ

Приводится робастный метод обработки экспериментальных данных при ограниченном объеме выборки, устойчивый к, так называемым, аномальным результатам. Использование данного подхода позволяет на основании полной информации, полученной при проведении исследования, получать статистически надежные и достоверные результаты. Приводится числовой пример применения робастного метода.

Ключевые слова: *выброс, аномальное значение, медиана, робастное среднее, интерквартильный размах, СКО, корректирующий коэффициент.*

Введение

На практике, наличие в выборках даже небольшого числа резко выделяющихся наблюдений (аномальных, экстремальных) способно кардинально изменить результат статистического исследования, и значения, полученные в конечном итоге, становятся недостоверными, а в некоторых случаях вообще перестанут нести какой-либо здравый смысл. Для того, чтобы избежать подобной ситуации, традиционно применяют статистические критерии, которые позволяют выделить, а затем и исключить аномальные данные, которые называют выбросами. Данный подход правомочен и эффективен для выборок большого объема. Однако для выборок малого объема, что имеет место при проведении экспериментальных исследований дорогостоящих, уникальных объектов или при исследованиях с разрушением, без восстановления, необходимо с некоторой осторожностью пользоваться данными приемами. Это обусловлено тем, что при выборках малого объема применяемые статистические критерии теряют чувствительность к аномальным значениям. Кроме того, исключение любого результата из имеющихся экспериментальных данных снижает статистическую надежность получаемой оценки. Например, отношение выборочного значения СКО к его математическому ожиданию при числе наблюдений n=4 составляет 42%, а при n=3 оно становится равным 52%. Как видно, исключение одного результата из имеющихся данных при малых объемах выборки приводит к уменьшению статистической надежности примерно на 10%.

При аттестации характеристик точности методики испытаний проводят совместный эксперимент, для участия в котором привлекаются лаборатории, имеющие соответствующий профессиональную подготовку. При совместном эксперименте существует допущение, что все привлекаемые лаборатории имеют одинаковую повторяемость. На практике же часто оказывается, что некоторые лаборатории имеют худшую повторяемость и в этом есть объективные причины. В соответствии с [1] для установления показателей прецизионности – повторяемости и воспроизводимости, предварительно применяется критерий Граббса для исключения, так называемых, выбросов. Исключение результатов, полученных этими лабораториями, во-первых, несколько идеализирует реальные условия испытаний, а во-вторых, увеличивает неопределенность получаемого результата.

Вся статистическая обработка и принимаемые на ее основании решения базируются на предположении о нормальности распределения [2]. Это в основном обосновано тем, что имеется хорошо разработанная теория статистических выводов. Однако в ряде практических задач нет достаточно объема исходных данных для построения параметрических моделей, адекватных экспериментальным данным. Еще в 60-тые годы прошлого столетия выдающиеся ученые в области матстатистики на основании результатов детальных исследований установили, что данные, которые обрабатываются на основании теории о нормальности распределения, как правило, в среднем около 10% содержат грубые промахи (от 1 до 20%) как в явной, так и в скрытой форме. Пуанкаре указывал на глубокую веру в универсальность нормального закона: математики думают, что физики наблюдают его на опыте; физики же – что математики способны доказать теоретически, что нормальный закон должен выполняться (в соответствии с центральной предельной теоремой). Все теоретические предпосылки базируются на возможности проведения опытов (осуществления наблюдений) при одних и тех же неизменных условиях. На практике же имеет место пространственно-временная изменчивость условий проведения наблюдений, в том числе и изменчивость самого объекта исследования. Таким образом, можно согласиться с Тьюки [3], что нормальность это миф и нормальности распределения не было и никогда не будет. Особенно данное утверждение соответствует при малом объе-

ме. Причинами появления выбросов в результатах исследования могут быть ошибки, сбои средств измерительной техники (СИТ), применяемых при проведении испытаний, несоблюдение правил проведения эксперимента, ошибки и промахи при оформлении результатов исследования, внешние факторы и многое другое. Ввиду условности закона распределения, вид которого фактически является 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 подчиняются другому закону, а не предполагаемому.

Тем не менее, параметрический подход с предположением, что закон распределения известен (должен быть нормальным) настолько глубоко вошел в практику статистической обработки данных, что нецелесообразно от него отказываться.

Целью данной статьи является презентация подхода к повышению достоверности результатов испытаний при выборках малого объема.

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, в общем случае неизвестными, причинами.

Модель засорения характеризуется растянутыми «хвостами» плотностей вероятности. В схеме с засорением имеется средняя устойчивая часть распределения, которая обусловлена обычными малыми составляющими погрешности, и растянутые хвосты, характеризующиеся относительно редкими выбросами.

Использование такой схемы позволяет, с одной стороны, сохранить удобное вероятностное предположение об однородности гипотетической генеральной совокупности, по которой строятся все вероятностные оценки, а, с другой - ввести требуемое представление о возможности появления погрешностей высокого уровня.

Имеется истинное значение параметра. Из-за влияния случайных величин определяем некоторую его оценку с погрешностью. На практике погрешности по отношению к истинному значению располагаются несимметрично. Чем меньше объем, тем больше несимметричность. Выборочное среднее минимизирует сумму квадратов отклонений результатов от истинного значения. Однако это и является источником чувствительности оценки к выбросам, которые существенно увеличивают эту сумму. Как уже отмечалось, имеется некоторая центральная часть распределения экспериментальных данных, которая соответствует предположению о распределении генеральной совокупности. Для данных, которые находятся в этой части распределения целесообразно проводить осреднение, т.е. использовать метод наименьших квадратов МНК. Модульный критерий, предложенный Лапласом, является более устойчивым чем МНК к выбросам – дает наилучший результат при наиболее неблагоприятном распределении. Поэтому, исходя из вышеприведенных соображений, при построении робастных методов делается «симбиоз» - для некоторой центральной группы берется метод наименьших квадратов , а, начиная с некоторого предела, для уменьшения влияния выбросов, но с сохранением данных, применяется модульный критерий.

Чтобы уменьшить чувствительность к выбросам в [4] предложено минимизировать функционал

$$\sum_{i=1}^{n} \rho(\varepsilon)(x_i - \mu),$$

где $\rho(\varepsilon)$ - функция потерь, которая позволяет менее «строго» подходить к отбраковке выбросов, которые отстоят от центра распределения на значение ε больших, чем $c\sigma$. Константа *c* регулирует степень робастности, значение этой константы зависит от степени «засорения». Так при «засорении» 1% $c = 2$, а при «засорении» 5% $c = 1{,}4$. Обычно выбирают значение $c = 1{,}5$.

В соответствии с выбранным критерием необходимо провести модификацию имеющихся данных, а именно:

$$x_i^* = \begin{cases} x_i \text{ при } |x_i - \hat{\mu}| \le c\sigma, \\ \hat{\mu} - c\sigma \text{ при } x_i < \hat{\mu} - c\sigma, \\ \hat{\mu} + c\sigma \text{ при } x_i > \hat{\mu} + c\sigma \end{cases} \qquad (1)$$

где $x_i^* = \text{med}\{x_i\}$, предварительно x_i ранжированы в порядке возрастания.

В качестве первоначальной оценки центра распределения $\hat{\mu}$, устойчивой к выбросам, берется выборочная медиана $\text{med}\{x_i\}$.

Проведенные исследования показали [5], что наилучшими свойствами с точки зрения устойчивости к выбросам обладает середина интервала, находящаяся между выборочными квартилями.

В качестве устойчивого берется интерквартильный интервал (размах) (interquartile range) – разность между значениями третьего $p = 3/4$ и первого $p = 1/4$ квартилей.

Интерквартильный интервал является характеристикой разброса распределения. Площадь под кривой распределения плотности вероятности на этом интервале составляет 50%. В предположении о

возможном законе распределения длина интервала однозначно соответствует дисперсии этого распределения. Для взаимосвязи интерквартильного интервала с дисперсией вводится абсолютное медианное отклонение (АМО), которое как раз и является оценкой масштаба – перехода от «полного» распределения к «усеченному»:

$$\mathrm{AMO}_n = \mathrm{med}\left\{\left|x_i - M_n\right|\right\}$$

где $M_n = \mathrm{med}\{x_i\}$, а индекс n соответствует числу элементов в выборке.

Для того, чтобы установить взаимосвязь между параметрами «усеченного» и предполагаемого генерального распределению совокупности, т.е. выполнить условие масштабирования, необходимо осуществить пересчет СКО (σ), воспользовавшись корректирующим коэффициентом, который определим из исходной плотности распределения (рис. 1):

$$f(x) = \frac{c}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}.$$

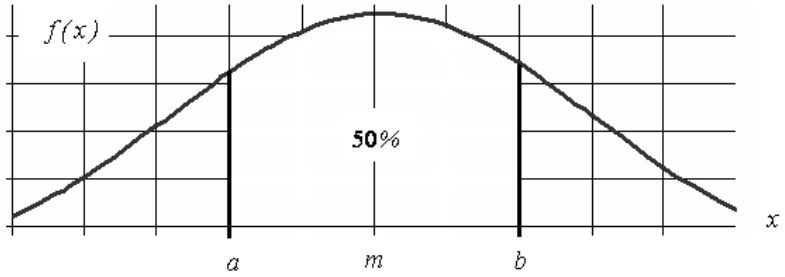

Рис. 1. Определение интерквартильного интервала;
a и b – ординаты первого и третьего квартиля

При «неусеченном» распределении *c*=1, при интерквартильном «усеченном» распределении:

$$c = \frac{0,5}{\hat{O}\left(\frac{b-m}{\sigma}\right) - \hat{O}\left(\frac{a-m}{\sigma}\right)},$$

где Φ(z) – интегральная нормированная функция нормального распределения. Воспользовавшись таблицами для нормального распределения [6], найдем

$$c = \frac{1}{0,6745} = 1,4826 \approx 1,483.$$

Следовательно изменение масштаба при переходе от «неусеченного» распределения к «усеченному» составляет 1,483.

Таким образом оценка СКО s^*, для выборки объема *n* находится на основании нормированного интерквартильного размаха будет:

$$s^* = 1,483 \cdot \mathrm{AMO}_n \qquad (2)$$

и, по сравнению с СКО, которое определяется на основании известного выражения

$$s = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}, \qquad (3)$$

будет устойчивой, даже если выборка содержит до 50% аномальных результатов.

Следовательно, для реализации критерия устойчивости к выбросам на основании выбранной функции потерь, необходимо определить медиану исходного массива данных, а затем, воспользовавшись выражением (2), вычислить s^*, которое вместе с выбранным c = 1,5, позволит установить границу перехода от метода наименьших квадратов МНК к методу наименьшего модуля МНМ $\varphi = cs^*$. В качестве оценки центра распределения берется медиана исходного ряда, которая более устойчива к выбросам, чем среднее значение.

Применив условия (1), получим модифицированный ряд x_i^*, для которого вычисляется среднее значение:

$$\bar{x}^* = \frac{1}{n}\sum_{i=1}^{n} x_i^*, \qquad (4)$$

а также, в соответствии с выражением (2), после подставки x_i^* и $\bar{x}_i^*$, значение СКО $s(x^*)$.

Вычисленное значение $s(x^*)$ используется для вычисления «нового» устойчивого значения СКО

$$s^* = 1,134 \cdot s(x^*),$$

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

Найденное значение s^* используется для вычисления нового значения границы φ и процедура продолжается в соответствии с ранее рассмотренным.

По соотношению между значениями $\bar{x}^*$, вычисляем 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.

Пример

Рассмотрим пример, данные для которого взяты из [1]. Проводились совместные лабораторные

исследования с целью определения характеристик точности методики испытаний. Для этого было привлечено 9 лабораторий. По результатам проведенных исследований были вычислены средние значения в лабораториях, которые представлены следующим образом: 24,140; 20,155; 19,500; 20,300; 20,705; 17,570; 20,100; 20,940; 21,185.

Таблица 1

Применение метода к средним значениям в элементах (% креозота)

Итерация	$0^{1)}$	1	2	3	4
φ		1,424	1,478	1,514	1,539
$x^* - \varphi$		18,876	18,909	18,893	18,872
$x^* + \varphi$		21,724	21,865	21,921	21,950
x_1^*	17,570	18,876	18,909	18,893	18,872
x_2^*	19,500	19,500	19,500	19,500	19,500
x_3^*	20,100	20,100	20,100	20,100	20,100
x_4^*	20,155	20,155	20,155	20,155	20,155
x_5^*	20,300	20,300	20,300	20,300	20,300
x_6^*	20,705	20,705	20,705	20,705	20,705
x_7^*	20,940	20,940	20,940	20,940	20,940
x_8^*	21,185	21,185	21,185	21,185	21,185
x_9^*	24,140	21,724	21,865	21,921	21,950
Среднее	20,511	20,387	20,407	20,411	20,412
Стандартное отклонение	1,727	0,869	0,890	0,905	0,916
Новые x^*	$20,300^{2)}$	20,387	20,407	20,411	20,412
Новые s^*	$0.949^{2)}$	0,985	1,009	1,026	1,039

После ранжирования по возрастанию, соответствующие результаты (нулевая итерация) $x_i^*, i = (1...9)$ занесены в табл. 1, колонка «0».

На основании этих данных вычисляются исходные значения:

- среднее для p = 9 лаборатории

$$x_0^* = \frac{1}{p}\sum_{i=1}^{p} x_{i(0)} = 20{,}511;$$

- стандартное отклонение

$$s_0^* = \sqrt{\frac{1}{p-1}\sum_{i=1}^{p}(x_{i(0)}^* - x_0^*)^2} = 1{,}727,$$

где x_i^* берется из столбца «0» таблицы 1.

В качестве исходной (грубой) оценки центра распределения для совокупности берется исходная медиана $\text{med}x_{i(0)}^* = x_5^* = 20{,}300$ и в соответствии с выражением вычисляется:

$$s_1^* = 1{,}483 \cdot \text{med}\left\{\left|x_{i(0)}^* - x_0^*\right|\right\},$$

где ряд $\left|x_{i(0)}^* - x_0^*\right|$ представляем значениями: 2,73; 0,8; 0,2; 0,145; 0; 0,405; 0,64; 0,889; 3,84.

Находим $s_1^* = 1{,}483 \cdot 0{,}640 = 0{,}949$, а затем пороговое значение $\varphi_1 = 1{,}5 \cdot s_1^* = 1{,}5 \cdot 0{,}949 = 1{,}424$, которое используется для вычисления границ перехода от МНК к МНМ

$$x_{(0)}^* - \varphi_1 \text{ и } x_{(0)}^* + \varphi_1 .$$

Результаты вычисления

$$x_{(0)}^* - \varphi_1 = 20{,}300 - 1{,}424 = 18{,}876$$

$$x_{(0)}^* + \varphi_1 = 20{,}300 + 1{,}424 = 21.724$$

заносим в столбец «1» (первый шаг итерации) таблицы 1.

Данным $x_{i(0)}^*$, которые выходят за пороговые значения, присваиваются пороговые значения, а остальные остаются без изменений.

Так как выполняются неравенства $x_{1(0)}^* < (x_0^* - \varphi_1)$, и $x_{9(0)}^* > (x_0^* + \varphi_1)$, то присваивается $x_{1(1)}^* = 18{,}876$ $x_{9(1)}^* = 21{,}724$.

В колонке «1» таблицы 1 представлены значения $x_{i(1)}$, которые используются для вычисления на первом шаге итерации.

Проведя вычисления, аналогичные проделанному выше, получим:

- среднее значение $x_1^* = \frac{1}{p}\sum_{i=1}^{p} x_{i(1)} = 30{,}387;$

- стандартное отклонение $s_1^* = 0{,}869,$

для вычисления которого были использованы значения $x_{i(1)}^*$, взятые из столбца «1» таблицы 1 и использовалось в качестве центра распределения $x_1^* = 30{,}387$.

Затем определяем новое значение

$$s_2^* = 1{,}134\sqrt{\frac{1}{p-1}\sum_{i=1}^{p}(x_{i(1)}^* - x_1^*)^2} = 0{,}985,$$

где $\text{med}\left\{\left|x_{i(1)}^* - x_1^*\right|\right\} = 0{,}664.$

Верхняя и нижняя границы соответственно равны $x_{(1)}^* - \varphi_2 = 18{,}909$ и $x_{(1)}^* + \varphi_2 = 21{,}865$, где $\varphi_2 = 1{,}5 \cdot s_2^* = 1{,}5 \cdot 0{,}985 = 1{,}478.$

Тогда на втором шаге итерации будем иметь модифицированные значения:

$$x_{1(2)}^* = x_{(1)}^* - \varphi_2 = 18{,}909,$$

$$x_{9(2)}^* = x_{(1)}^* + \varphi_2 = 21{,}865.$$

Занесем модифицированные данные $x^*_{1(2)}$ и $x^*_{9(2)}$ во «2» столбец и вместе с немодифицированными данными $x^*_{i(0)}$ (i=2…8) используем на втором шаге итерации.

Результат вычислений занесем в соответствующий столбец. Затем в дальнейшем реализуем алгоритм, аналогичный рассмотренному ранее. Как видно из результатов, представленных в табл. 1, среднее значения, полученные на четвертом $x^*_4 = 20{,}411$ и на пятом $x^*_5 = 20{,}412$, практически не отличаются. При этом робастная оценка стандартного отклонения средних значений, полученных в лабораториях, s^*_y будет соответствовать $s^*_4 = 1{,}026$ и $s^*_5 = 1{,}039$.

Если использовать традиционный подход, т.е. выявить выбросы с применением критерия Граббса, то исключив выбросы

$$x^*_{1(0)} = 17{,}570 \text{ и } x^*_{9(0)} = 24{,}140$$

из дальнейшей обработки, для p = 7 среднее значение m = 20,412, что свидетельствует о совпадении центров распределения выборочных значений, т.е. получим одинаковые оценки характеристики точности исследуемой методики испытаний. Однако применение робастного метода дает значение $s^* = 1{,}039$, которое соответствует действительному рассеиванию исходных величин. Исключение двух выбросов (средние значения результатов, полученных «первой» и «девятой» лабораториями) из результатов совместных экспериментальных испытаний дает значение s = 0,501. Но при этом «уходят» от реальной экспериментальной ситуации, идеализируя ее, предполагая, что экспериментальные данные в обязательном порядке должны принадлежать генеральной совокупности, имеющей нормальное распределение. Так как при этом были учтены результаты p=7 лабораторий, то найденная оценка СКО будет иметь меньшую статистическую надежность, Различие между найденными оценками СКО может быть использовано для анализа организации испытательного процесса в двух вышеупомянутых лабораториях.

Выводы

Проведенное исследование показало целесообразность проведения статистической обработки данных в выборках малого объема, содержащих результаты, которые, в предположении нормального распределения генеральной совокупности, относят к выбросам, с использованием робастных методов

Список литературы

1. ДСТУ ГОСТ ISO 5725:2005 Точность (правильность и прецизионность) методов и результатов измерения.

2. Сархан А.Е. Введение в теорию порядковых статистик / А.Е. Сархан, Б.Г. Гринберг. – М.: Статистика, 1970. – 414 с.

3. Тьюки Дж. Анализ результатов наблюдений / Дж. Тьюки. – М.: Мир, 1981. – 702 с.

4. Хьюбер Дж.П. Робастность в статистике: пер. с англ. / Дж.П. Хьюбер. – М.: Мир, 1984. – 304 с.

5. Analyst Robust statistics. – How Not to Reject Outliers. – December 1989. – Vol. 114.

6. Большев Л.Н. Таблицы математической статистики / Л.Н. Большев, Н.В Смирнов. – М.: Наука. Главная редакция физико-математической литературы, 1983. – 416 с.

Поступила в редколлегию 15.11.2011

Рецензент: д-р техн. наук, проф. И.П. Захаров, Харьковский национальный университет радиоэлектроники, Харьков.

РОБАСТНЕ ОЦІНЮВАННЯ ТОЧНОСТНИХ ХАРАКТЕРИСТИК РЕЗУЛЬТАТІВ СПОСТЕРЕЖЕНЬ

Є.Т. Володарський, Д.А. Палянічко

Наводиться робастний метод обробки експериментальних даних при обмеженому об'ємі вибірки, стійкий до так званих, аномальних результатів. Використання даного підходу дозволяє на підставі повної інформації, отриманої при проведенні дослідження, отримувати статистично надійні і достовірні результати. Наводиться числовий приклад застосування робастного методу.

Ключові слова: *викид, аномальне значення, медіана, робастне середнє, інтерквартільний розмах, СКВ, коректуючий коефіцієнт.*

ROBAST EVALUATION OF DESCRIPTIONS ACCURACY OF SUPERVISIONS RESULTS

E.T. Volodarskiy, D.A. Palyanichko

A robust method over of processing of experimental data is brought at the limited sample size, steady to anomalous results. The use of this approach allows on the basis of complete information, got during the leadthrough of research, to get reliable and reliable results statistically. A numerical example of application of robust method is made.

Keywords: *troop landing, anomalous value, median, robust middle, interquartile scope, root-mean-square error, correcting a coefficient.*