УДК 614.001.89(045)

Е. Т. ВОЛОДАРСКИЙ[1], Л. А. КОШЕВАЯ[2]

1. Національний технічний університет України «Київський політехнічний інститут», Київ
2. Національний авіаційний університет, Київ

ПРИМЕНЕНИЕ РОБАСТНЫХ МЕТОДОВ ПРИ ОЦЕНИВАНИИ РЕЗУЛЬТАТОВ ЭКСПЕРИМЕНТАЛЬНЫХ ИССЛЕДОВАНИЙ

Аннотация. Показано влияние выбросов, которые могут возникать при проведении экспериментальных исследований, на точность получаемых оценок. Рассмотрены робастные методы оценивания, при реализации которых учитываются все без исключения результаты измерений. Учет всех данных позволяет «подобрать» соответствующий нормальный закон распределения. Это обеспечивает статистическую надежность оценок, особенно для выборок малого объема. Приводятся числовые данные, подтверждающие эффективность робастных методов.
Ключевые слова: устойчивость, аномальность, выбросы, интерквартильный интервал, статистическая надежность оценок
Анотація. Показано вплив викидів, які можуть виникати при проведенні експериментальних досліджень, на точність одержуваних оцінок. Розглянуто робастні методи оцінювання, при реалізації яких враховуються всі без винятку результати вимірювань. Урахування усіх даних дозволяє «підібрати» відповідний нормальний закон розподілу. Це забезпечує статистичну надійність оцінок, особливо для вибірок малого обсягу. Наводяться числові дані, що підтверджують ефективність робастних методів.
Ключові слова: стійкість, аномальність, викиди, інтерквартильний інтервал, статистична надійність оцінок.
Abstract. Shows the effect of emissions which may arise in experimental studies, on accuracy the resulting estimates. Considered robust estimation methods, implementation of which takes into account all measurement results without exception. Accounting for all data allows you to choose the the corresponding normal distribution. This provides a reliable of statistical estimates, especially for small volume samples. Are given numerical data to confirm the efficiency of robust methods.
Keywords: robustness, anomalousness, emissions, interquartile range, the statistical reliability of the estimates

Введение

При статистической обработке результатов экспериментальных исследо-ваний всегда исходят из некоторых предположений об исследуемой ситуации. Даже в простейших случаях делаются явные или неявные предположения о случайности и независимости полученных опытных дан-ных, о виде тех или иных распределений в изучаемой модели, например, о виде исход-ных распределений для некоторых неизвест-ных параметров. Однако, наиболее распрост-раненные процедуры статистического анали-за, которые оптимальны в предположении о нормальности распределения, весьма чувст-вительны к довольно малым отклонениям от этих предположений. На практике в боль-шинстве случаев мы имеем дело с реальными распределениями, несколько отличающимися от идеальных.

Для описания данной ситуации существует несколько подходов. Наиболее широко используемым является подход, предложенный Тьюки [1]. Он предполагает, что имеется большой набор данных n, среди которых имеются «хорошие» и «плохие» случайно перемешанные наблюдения x_i некоторой величины μ. Каждому «хорошему» и каждому «плохому» из всех наблюдений соответствуют вероятности (1 - ε) и ε, где ε – малое число. В первом случае наблюдения x_i имеют нормальное распределение $N(\mu, \sigma^2)$, а во втором – $N(\mu, 9\sigma^2)$. Иначе говоря, все наблюдения имеют одно и то же среднее, а случайная погрешность измерения для некоторых из них в 3 раза больше, чем у остальных.

Постановка проблемы

Предполагая, что все величины x_i являются независимыми и имеют одно и то же распределение, можно дать следующее эквивалентное описание приведенной ситуации:

$$F(x) = (1-\varepsilon)\cdot \Phi\left(\frac{x-\mu}{\sigma}\right) + \varepsilon \cdot \Phi\left(\frac{x-\mu}{3\sigma}\right), \tag{1}$$

где $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}} dy$.

При этом используют две широко известные оценки разброса – среднее квадратическое отклонение (СКО):

$$s_n = \left[1/n \cdot \sum_i (x_i - \bar{x})^2 \right]^{1/2};$$

и среднее абсолютное отклонение САО:

$$d_n = 1/n \cdot \sum_i |x_i - \bar{x}|.$$

Существуют противоположные точки зрения о преимуществах этих оценок [1]. Главным аргументом в пользу СКО является то, что для нормального распределенных результатов измерений величина s_n примерно на 12% более эффективна, чем d_n. Однако, как следует из исследований Тьюки, асимптотическая эффективность СКО по отношению к САО резко снижается при малейшем отклонении предполагаемой модели разброса от нормальной. Уже при наличии двух «плохих» из тысячи наблюдений (ε = 0,002) преимущество в 12% пропадает, а при наличии пяти «плохих» из тысячи наблюдений (ε = 0,005) асимптотическая эффективность САО становится равной двум и уменьшаясь, достигает единицы при ε = 0,5. Как правило, типичные выборки «хороших» данных, которые встречаются на практике, довольно точно моделируются законом распределения (1), где ε лежит в пределах от 0,01 до 0,1. Таким образом, выражение (1) служит удобным описанием функции распределения с более длинными «хвостами», чем у классического нормального распределения.

Традиционно при статистической обработке экспериментальных данных производится их «редактирование» путем усечения выделяющихся наблюдений по некоторому признаку. Таким примером могут быть, исходя из уровня статистической значимости 5% и 1%, квазивыбросы и выбросы соответственно. Однако, данные после редактирования не будут иметь нормальное распределение, что объясняется статистическими ошибками первого и второго рода при проверке гипотезы на анормальность. Еще сложнее обстоит дело при работе с выборками малого объема, когда экспериментальные данные, полученные из даже нормально распреде-ленной генеральной совокупности, могут быть асимметричными. Причем, чем меньше объем, тем больше вероятность асимметрии. Это приводит к явному искажению оценок центра распределения и особенно дисперсии, что не допустимо. например, при проверке профессионального уровня испытательных лабораторий, когда объем эксперименталь-ных данных из практических соображений ограничен. Более того, при выборках малых объемов каждый результат имеет особый вес [2] – точность оценивания СКО при числе наблюдений $n \leq 10$ может изменяться на десятки процентов. Однако классический параметрический подход при оценивании результатов экспериментальных исследо-ваний с использованием нормального распределения настолько вошел в практику, что отказываться от него нецелесообразно. Эти методы просты и очевидны, для них разработана полная и глубокая теория статистических выводов. Поэтому и возникла проблема приспособления «старой» модели к новым задачам, т.е. нахождения методов оценивания, которые учитывали бы на определенных условиях «грубые ошибки» и позволяли при этом достаточно точно определять оценки параметров, исходя из имеющихся данных. Эти методы получили название робастных.

Основная часть

Под робастностью (устойчивостью) в статистике понимается нечувствительность к различным отклоне-ниям и неоднородностям в выборке, связан-ным, в общем случае, с неизвестными причи-нами. Для робастных методов базовой является не модель случайной величины с нормальным законом распределения, а распределение «смеси» – нормальная модель с «засорениями», характеризирующаяся «растянутыми» хвостами плотности вероят-ности, что соответствует выражению (1). При этом имеет место устойчивая средняя часть распределения, соответствующая пред-полагаемой нормальной модели, и растя-нутые «хвосты», характеризующиеся относи-тельно редкими выбросами (или квази-выбросами для малых выборок). Использо-вание такой модели позволяет, с одной стороны, сохранить удобное традиционное представление об однородности гипотети-ческой генеральной совокупности, на которой строятся все вероятностные оценки, а с другой – ввести требуемое представление о возможности появления больших отклонений на «хвостах» распределения.

Как было установлено ранее, модульный критерий (предложенный Лапласом) является при оценке центра распределения более устойчивым к выбросам, чем метод наименьших квадратов (предложенный Гауссом), т.е. дает меньшее смещение оценки. Поэтому, исходя из выше-приведенных соображений, при построении робастных процедур имеет место «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 [3] предложено минимизировать функционал

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

где ρ – функция потерь, которая позволяет менее «строго» подходить к отбраковке выбросов, которые отстоят от центра распределения на значение, большее, 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:

$$x_i^* = \begin{cases} x_i \ при \ \left|\mathrm{x_i} - \hat{\mu}\right| \le c\sigma, \\ \hat{\mu} - c\sigma \, при \ \mathrm{x_i} < \hat{\mu} - c\sigma, \\ \hat{\mu} + c\sigma \ при \ \mathrm{x_i} > \hat{\mu} + c\sigma \end{cases} \tag{2}$$

где x_i предварительно ранжированы в порядке возрастания.

Таким образом, при робастной процедуре аномальные («плохие») результаты не исключаются, а модифицируются в соответ-ствии с выражением (2) и тем самым повышается статистическая надежность оценки. В качестве первоначальной оценки центра распределения μ, устойчивой к выбросам, берется выборочная медиана $med\{x_i\}$.

Проведенные исследования показали [4], что наилучшими свойствами с точки зрения устойчивости к выбросам обладает середина интервала, находящаяся между выбороч-ными квартилями. Интерквартильный интервал является характеристикой разброса распределения. Площадь под кривой распределения плотности вероятности на этом интервале составляет 50%. В предположении о нормальном законе распределения длина интервала однозначно соответствует дисперсии этого распределения. Для взаимосвязи интерквартильного интервала с дисперсией вводится абсолют-ное медианное отклонение MAD (*Median Absolute Deviation)*

$$MAD_n = med\left\{\left|x_i - M_n\right|\right\}$$

где $M_n = med\{x_i\}$ медиана выборки, а индекс n соответствует числу элементов в ней. Для того, чтобы оценка MAD_n при нормальном законе распределения была состоятельной, ее следует поделить на величину Φ^{-1} (3/4) = 0,6745. Таким образом, оценка СКО, которая в соотношении (2) используется вместо σ при вычислении граничных значений, по отношению к которым осуществляется модификация данных, определяется из выражения:

$$s^* = 1{,}483 \ MAD_n.$$

После модификации исходных результатов вычисляют среднее значение

$$x^* = \sum_{i=1}^{n} x_i^* / n, \tag{3}$$

и СКО

$$s^* = 1{,}134\sqrt{\sum_{i=1}^{n}(x_i^* - x^*)^2/(n-1)}. \tag{4}$$

Коэффициент 1,134 учитывает тот факт, что анализируются данные для усеченного распределения при c = 1,5, что соответствует 86,6% предполагаемого нормального распре-деления, под которое «подгоняются» исход-ные данные, содержащие и аномальные ре-зультаты. Такая подгонка осуществляется последовательно в несколько этапов, на каждом из которых исходными являются модифицированные в соответствии с (2) данные предыдущего этапа. Исходя из выражений (3) и (4) находят уточненные значения среднего и СКО. Итерационная процедура продолжается до тех пор, пока разница между средними значениями на текущем и предыдущем шагах становится несущественной. При

этом оценка СКО будет находиться между оценками СКО, полученными на основании всех экспери-ментальных данных, и данными с исключён-ными аномальными результатами. Данная оценка будет адекватной. Таким образом, ро-бастная процедура реализует плавный пере-ход от выборки с «плохими» результатами экспериментальных исследований к усечен-ной выборке, подгоняя при этом, имеющиеся данные к центральной части предпола-гаемого нормального распределения.

При проведении совместных экспери-ментальных исследований возникает необхо-димость объединять оценки отдельных СКО для определения, например, оценки повто-ряемости или воспроизводимости [5]. В большинстве практических случаев эти отдельные оценки из-за тех или иных ограничений определяются на основании выборок малого объема, которые, как уже отмечалось, могут иметь асимметрию. В этой связи, в соответствии с критерием Кохрена [4] некоторые из них могут признаваться «выбросом» и поэтому их необходимо исключить при дальнейшей статистической обработке. В такой ситуации также необходимо применять робастные процедуры, которые базируются на всех имеющихся экспериментальных данных, а, следовательно, дают статистически надежную оценку.

Необходимым условием, которое должно при этом выполняться, является несмещен-ность робастной оценки s_j^*, которая на каждом j-м шаге итерации «подгоняется» под СКО предполагаемой генеральной совокупности σ, наилучшим образом соответствующей имеющейся совокупности оценок СКО. Для обеспечения несмещенности вводится согласующий фактор ξ. При этом должно выполняться условие:

$$E\left\{\left(\xi s^*\right)^2\right\} = \sigma^2. \tag{5}$$

Значком * здесь и в дальнейшем будем обозначать робастную оценку. При этом допускается, что робастная оценка СКО s^* является в заданных пределах с некоторой вероятностью, устойчивой к выбросам. С этой целью вводится верхнее максимальное отклонение $\eta\sigma$ от центра распределения, которое задается в виде неравенства

$$P\left\{s^* > \eta\sigma\right\} = \alpha, \tag{6}$$

где σ – СКО нормально распределенной, величины, соответствующее имеющимся экспериментальным данным, η – ограничительный фактор, зависящий от числа элементов в выборке; $P = (1 - \alpha)$ – вероят-ность выполнения ограничения сверху на допустимое отклонение робастной оценки СКО s^* (чаще всего выбирается $\alpha = 0.1$). Иными словами, ограничительный фактор η выбирается как верхняя $(1-\alpha)$ 100% точка распределения СКО σ.

За первоначальную оценку s^* генерального СКО принимается медиана упорядоченных выборочных оценок. На основании этой оценки находится значение предельного отклонения от центра, служащее для модификации текущих оценок СКО на первом шаге приближения:

$$\psi_j = \eta \cdot s_j^*. \tag{7}$$

Для каждого значения рассматриваемой совокупности оценок СКО осуществляют модификацию s_{ij}^* в соответствии с соотношением:

$$s_{ij}^* = \begin{cases} \psi_j, \text{ если } s_{ij} > \psi_j \\ s_{\mathrm{i}\,j} \text{ в остальных случаях} \end{cases} \tag{8}$$

Для первого шага $j = 1$. На последующих шагах итерации приближенная оценка генерального СКО, например, повторя-емости, находится, исходя из выражения:

$$s_{j+1}^* = \xi\sqrt{\sum_{i=1}^{n}\frac{\left(s_{ij}^*\right)^2}{n}},$$

где n – общее число оценок в рассмат-риваемой совокупности. Вычисленное значение s^*_{j+1} используется для определения, в соответствии с (7), нового ограничи-тельного значения ψ_{j+1}, на основании которого, из соотношгния (8) модифици-руются текущие данные.

Итерационная процедура продолжается, пока все исходные значения СКО рассмат-риваемой совокупности не «войдут» в пре-делы текущего ограничительного значения. Ограничительный фактор η, при котором выполняется условие (6), определяется как

$$\eta = \frac{\chi^2_{\nu,\, P=0,1}}{\nu}.$$

Значение согласующего фактора ξ находится, исходя из (5), представив его в виде

$$E\left\{\nu \cdot \left(s^* \cdot \sigma\right)^2\right\} = \frac{\nu}{\xi^2}.$$

Отсюда следует, что

$$\xi = \frac{1}{\sqrt{z + 0{,}1\eta^2}},$$

где значение $z = P\left(\chi^2_\omega \le \nu \cdot \eta^2\right)$ находится из таблицы распределения Пирсона [4].

Выводы

Известно, что реальное распре-деление данных, полученных при экспери-ментальных исследованиях, несколько отли-чается от нормального. Исключение, так называемых. выбросов, приводит к искаже-нию оценок среднего значения результатов наблюдений и, в большей степени, СКО. Особенно это проявляется для выборок малого объема. Применение робастных методов оценивания позволяет при обработ-ке результатов использовать все имеющиеся опытные данные и, тем самым, получить статистически надежные оценки, адекватные реальному распределению опытных данных.

Литература

1) Тьюки Дж. У. Анализ результатов наблюдений. Разведочный анализ. Пер с англ. – М: Советское радио, 1981, 693 с.

2) Guide to the Expression of Uncertainty in Measurement: First edition. ISO, Switzerland. – 1993. – pp.101.

3) Хьюбер Дж. П. Робастность в статистике. Пер. с англ.-М.: Мир, 1984.-304 с.

4) Большев Л.Н., Смирнов Н.В Таблицы математической статистики. - М.: Наука, 1983.- 416 с.

5) ISO 5725, 1-6:1994. Accuracy (trueness and precision) of measurement methods.

Стаття надійшла: 20.10.2014.

Відомості про авторів

Володарський Євген Тимофійович, д.т.н., професор кафедри автоматизації експериментальних досліджень, Національний технічний університет України «Київський політехнічний інститут», просп. Перемоги, 37, Київ 03056, Україна, тел / факс (044) 4549800.

Кошева Лариса Олександрівна, д.т.н., професор кафедри біокібернетики та аерокосмічної медицини, Національний авіаційний університет, просп. Космонавта Комарова, 1, Київ, 03058, Україна, тел / факс (044) 4067442 .