Для цитирования: Аналитика и контроль. 2023. Т.27, № 1. С. 59-61
DOI:10.15826/analitika.2023.27.1.006

Комментарии к статье И.Г. Зенкевича «Вычисление средних хронологических значений – незаслуженно забытый способ статистической обработки»

А.Л. Померанцев

Федеральный исследовательский центр химической физики РАН (ФИЦ ХФ РАН)
Российская Федерация, 119991, Москва, ул. Косыгина, 4

Адрес для переписки: Померанцев Алексей Леонидович, E-mail: forecast@chph.ras.ru

Поступила в редакцию 18 февраля 2023 г., после доработки - 04 марта 2023 г.

For citation: *Analitika i kontrol'* [Analytics and Control], 2023, vol. 27, no. 1, pp. 59-61
DOI:10.15826/analitika.2023.27.1.006

Comments on the article by I.G. Zenkevich «Calculation of average chronological values – an undeservedly neglected method of statistical data processing»

A.L. Pomerantsev

Federal Research Center for Chemical Physics of the Russian Academy of Sciences (FITC HF RAS)
4 Kosygin Str., Moscow, 119991, Russian Federation

Corresponding author: Alexey L. Pomerantsev, E-mail: : forecast@chph.ras.ru

Submitted 18 February 2023, received in revised form 4 March 2023

Игорь Георгиевич известный специалист в хроматографии и им опубликовано много превосходных работ в этой области. Моя же область интересов – анализ данных, в том числе и получаемых в аналитической химии. Поэтому я не мог не обратить внимания на публикацию, в которой он предлагает неизвестный мне подход к статистической обработке.

Проблема, обсуждаемая в статье И.Г. Зенкевича, имеет важное практическое и теоретическое значение. Действительно, каждый аналитик сталкивался в своей практике с данными (измерениями), при анализе которых получается ошибка, фатально влияющая на сделанные или предполагаемые выводы. Чтобы не быть голословным, приведу пример.

В таблице 1приведены некоторые условные данные, которые могут являться, например, значениями концентраций (площади или высоты пиков), найденными в эксперименте. Первая колонка – это метки, вторая колонка – значения, записанные в лабораторном журнале, а последняя колонка – величины, введенные в таблицу Excel. Внимательный читатель заметит (курсив ему в помощь), что при переписывании значений была совершена ошибка – не так поставлена запятая в образце 3. К чему это привело, видно из последних двух рядов: среднее значение выросло в 2 раза, а среднеквадратичное отклонение более чем в 5 раз.

А теперь представим, что эти данные должны доказать, что предлагаемый новый метод лучше чем общепризнанный. И вот они показывают ровно обратное. Значит надо искать и исправлять ошибку. Все просто в этом иллюстративном примере, а как

Таблица 1

Данные и оценки

Table 1

Data and estimates

Имя	В тетради	В компьютере
образец 1	0,0254	0,0254
образец 2	0,0359	0,0359
образец 3	*0,0156*	*0,1560*
образец 4	0,0202	0,0202
образец 5	0,0112	0,0112
образец 6	0,0322	0,0322
среднее арифметическое	0,023	0,047
среднеквадратичное отклонение	0,010	0,054

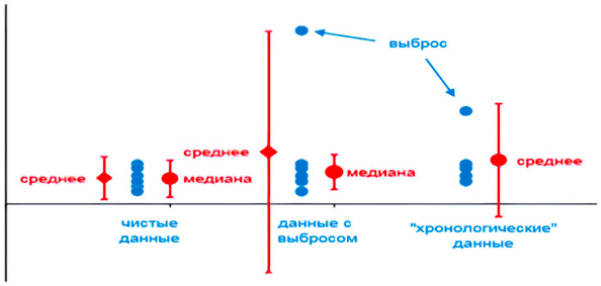

Рис. 1. Оценки положения и размаха в данных без выброса, с выбросом и «хронологических»

быть в случае, когда данных многие тысячи? А если это не ошибка переписывания, а более сложная причина: неверно подготовленный образец, сбой в работе прибора? Как тогда найти и удалить этот выброс в данных?

Оказывается можно его не удалять, но применять робастные (устойчивые) методы оценивания, которые слабо чувствительны к выбросам.

Посмотрим на таблицу 2 и рисунок 1, в которых показаны классические и робастные оценки среднего и разброса, вычисленные по данным, приведенным в табл. 1.

Классические оценки – среднее арифметическое *m* и среднеквадратичное отклонение *s* – вычисляются привычным образом, так как показано в уравнениях

$$m = \frac{1}{N}\sum_{n=1}^{N} x_n \qquad s = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(x_n - m)^2} \qquad (1)$$

С их робастными аналогами сложнее, но не намного. Для робастной оценки среднего обычно используется медиана – величина μ, относительно которой все исходные данные лежат с двух сторон, больше и меньше μ, поровну. Для вычисления медианы в Excel есть встроенная функция, которая так и называется – MEDIAN.

Для робастной оценки разброса используется медиана абсолютного отклонения MAD (Median of Absolute Deviation), которая определяется по формуле, приведенной в уравнении (2).

$$\mu = \text{median}(\mathbf{X}) \qquad \sigma = \text{median}\left(\left|\mathbf{X} - \mu\right|\right) \qquad (2)$$

Для вычисления MAD в Excel можно использовать простую формулу

$$\{=\text{MEDIAN}(\text{ABS}(\mathbf{X} - \text{MEDIAN}(\mathbf{X}))\} \qquad (3)$$

Обратите внимание на фигурные скобки в уравнении (3). Вводить их не нужно, и они показаны здесь только для того, чтобы напомнить – это формула массива, ввод которой завершается нажатием двух клавиш Shift+Enter одновременно. И уже после этого появятся скобки. Конечно, **X** – это массив данных, который указывается целиком, например B3:B8 или с помощью его имени в Excel.

Посмотрим теперь на результаты оценивания в таблице 2. Видно, что для «загрязненных» данных робастные оценки значительно лучше классических – они ближе к тем, которые были получены на «чистых» данных. Это означает, что они менее чувствительны к выбросам.

Обратимся теперь к методу, который предложен в работе И.Г. Зенкевича. Как указано в статье, источник «среднего хронологического» – это временные ряды, т.е. последовательность наблюдений, полученных в последовательные моменты времени. Главное, в чем временной ряд отличается от выборки случайных значений – это наличие корреляций между ближайшими значениями, тогда как в обычной практике измеренные значения принимаются независимыми. Если же говорить о самой формуле «среднего хронологического», то ее применение ограничивается коротким списком эконометрических задач учета запасов и движения средств. Приведем пример, иллюстрирующий такой подход.

Предположим, некто имел на своем счету 100 рублей и не пополнял его весь год вплоть до декабря, когда он положил на счет еще 100 000 рублей. Очевидно, в надежде на процентный доход за год. Если для расчета среднего годового вклада использовать арифметическое среднее, то мы получим (100 100+100x11)/12 = 8 433,33 руб. А вот хронологическое среднее даст совсем другой результат (0,5x(100 100+100)+100x10)/11 = 4 645,45 руб. Выгода банка – очевидна.

Но вернемся к нашей теме и посмотрим, является ли предлагаемое «хронологическое среднее» робастной оценкой.

Из табл. 2 мы видим, что «хронологический» подход плохо справляется с задачей робастного оценивания – величины среднего и разброса значительно отличаются от своих аналогов, найден-

Таблица 2

Классические и робастные оценки

Table 2

Classical and robust estimates

Оценка	В тетради	В компьютере
среднее арифметическое	0,023	0,047
медиана	0,023	0,029
среднее хронологическое	0,023	0,039
среднеквадратичное отклонение	0,010	0,054
медиана абсолютного отклонения	0,008	0,008
среднеквадратичное хронологическое	0,006	0,025
среднее арифметическое	0,023	0,047
среднеквадратичное отклонение	0,010	0,054

ных для неиспорченных данных. Таким образом, хронологическая оценка не является робастной, поскольку она чувствительна к выбросам.

Есть критерий, определяющий, является ли оценка робастной и в какой степени. Он называется «предел устойчивости». Эта величина определяется тем, сколько (в процентах) выпадающих значений (выбросов) можно добавить к исходной, свободной от выбросов выборке до тех пор, пока оценка останется устойчивой. Известно, что оценки, основанные на медиане, обладают очень высоким пределом устойчивости, который равен 50%. Это означает, что, например, к выборке из 6 элементов можно добавить 3 выброса и оценка все еще будет устойчивой. Для хронологического подхода предел устойчивости равен 0.

Таким образом, для данных загрязненных выбросами лучше использовать традиционные робастные оценки, например медиану.