Для цитирования: Аналитика и контроль. 2023. Т.27, № 1. С. 62-64
DOI:10.15826/analitika.2023.27.1.007

Дополнение к комментариям А.Л. Померанцева к статье И.Г. Зенкевича «Вычисление средних хронологических значений – незаслуженно забытый способ статистической обработки»

И.Г. Зенкевич

ФГАОУ ВО «Санкт-Петербургский государственный университет», Институт химии, Университетский просп., 26, Санкт-Петербург 198504, Российская Федерация

Адрес для переписки: Зенкевич Игорь Георгиевич, E-mail: izenkevich@yandex.ru

Поступило в редакцию 27 февраля 2023 г., после доработки 31 марта 2023 г.

For citation: *Analitika i kontrol'* [Analytics and Control], 2023, vol. 27, no. 1, pp. 62-64
DOI:10.15826/analitika.2023.27.1.007

Supplement to A.L. Pomerantsev's comments on I.G. Zenkevich's article «Calculation of average chronological values – an undeservedly neglected method of statistical data processing»

Igor G. Zenkevich

St. Petersburg State University, Institute for Chemistry, 26, Universitetskii Av., St. Petersburg, 198504, Russian Federation

Corresponding author: Igor G. Zenkevich, E-mail: izenkevich@yandex.ru

Submitted 27 February 2023, received in revised form 31 March 2023

Алексей Леонидович – один из известнейших специалистов России в области хемометрики и обработки данных. Это придает особую значимость его комментариям [1] относительно возможности использования средних хронологических значений при статистической обработке данных и тем интереснее обсуждение дискуссионных моментов рассматриваемой проблемы. Дело в том, что в этих комментариях он вольно или невольно затронул вопросы, выходящие за границы задачи вычисления и применения хронологических средних. Прежде всего, это рекомендуемая им в качестве робастной оценки средних значений медиана и, главное, разброса – медиана абсолютного отклонения (MAD). Основное предназначение таких оценок – хорошо известная и часто обсуждаемая в хемометрике проблема нивелирования влияния возможных выбросов на получаемые результаты. Из такой формулировки задачи следует и используемая терминология, как то: ошибки в исходных данных, «испорченные» данные, данные, загрязненные выбросами и т.д. Однако здесь сразу же следует отметить, что в обсуждаемой статье нет ни слова о выбросах. ***Речь идет о малых выборках, отличающихся повышенным разбросом данных, причем без отбраковки каких-либо значений как ненадежных***. Задачей является более корректная оценка стандартных отклонений (со средними значениями проблем практически нет). Допускаю, что такая формулировка несколько нетипична для «традиционной» хемометрики.

Чтобы не увеличивать объем приведенного ниже текста, указанные в нем ссылки соответствуют работам, цитированным в «исходной» статье.

Дальнейшее обсуждение я хотел бы начать именно с приведенного в комментариях [1], полагаю, выдающегося примера. Если для оценки среднего выбрать медиану, то для робастной оценки разброса исходных данных следует использовать медиану абсолютного отклонения (s). При этом сами медианы используют достаточно редко, а уж про MAD и говорить нечего. Тем не менее, в базе [2]

характеристика средних межлабораторных значений газохроматографических индексов удерживания на стандартных неподвижных фазах имеет следующий вид: “median ± deviation”. Если вычисление медиан не вызывает вопросов, то никаких комментариев о деталях вычисления “deviation” в описании этой базы нет. Если бы речь шла об используемой иногда величине называемой «размах», то его значения значительно превышают приведенные в базе [2] оценки. Самое вероятное, что значения “deviation” представляют собой именно медианы абсолютных отклонений, а это делает справочные данные базы [2] устойчивыми к возможным аномальным значениям индексов удерживания.

В то же время именно это создает серьезные проблемы практического использования хроматографических данных [2]. При сравнении экспериментальных значений индексов со справочными значениями важнейшей проблемой являются допустимые отклонения от средних. При обычной форме представления данных $\langle x \rangle \pm s_x$ («среднее значение ± стандартное отклонение») в интервал $\pm s_x$ попадает около 68 % всех значений, в интервал $\pm 2s_x$ – 95 % , а в интервал $\pm 3s_x$ – уже около 99.7 %. Подобные оценки вероятности для величин (s) или неизвестны, либо сложны для практического использования, либо труднодоступны.

Ну а теперь можно вернуться к обсуждению средних хронологических значений, прежде всего, областей их применения. Действительно, изначально они были предложены исключительно для временных рядов. Однако никто ведь не ограничивает, например, средние геометрические значения только имеющими отношение к геометрии величинами, так что их использование, например, для количественного хроматографического анализа способом двойного внутреннего стандарта [3, 4] не вызывает удивления. Суть состоит в том, что применение средних хронологических величин вполне можно распространить и на иные типы переменных, но это требует их предварительного ранжирования по возрастанию. В результате получаем последовательности, которые не имеют отношения к временным рядам, но в которых, как и в таких рядах, сразу выявляется корреляция ближайших значений. Каждое значение x_i (кроме первого и последнего) может быть оценено линейной интерполяцией соседних с ним значений, $x_i \approx (x_{i-1} + x_{i+1})$. Этим приемом, в частности, воспользовался Д.И. Менделеев, с достаточно хорошей точностью оценив некоторые физико-химические характеристики еще не открытого германия (экасилиций) средними значениями свойств ***соседних*** известных элементов по группе (Si – Sn) и по периоду. Заметим, что эта операция, как и создание самой Периодической Системы, потребовала предварительного ранжирования совокупности атомных масс известных элементов по их возрастанию. Без такого ранжирования задачу, полагаю, решить бы не удалось.

Таким образом, нельзя не признать, что ранжирование случайных значений изменяет некоторые характеристики выборок данных. Должен заметить, что, видимо по этой причине, не все специалисты согласны с применением такой операции, так что, при необходимости, дискуссию по этому вопросу можно продолжить.

Комментарии [1] закономерно начинаются с таблицы, иллюстрирующей искажение результатов обработки массивов данных случайными (чаще всего немногочисленными, а то и вовсе единичными) ошибками. Конкретно, в последовательности шести значений 0.0156 было заменено ошибочным 0.1560, что привело к увеличению среднего арифметического вдвое, а его стандартного отклонения – более чем в пять раз. Кроме того, в комментариях [1] весьма примечателен рисунок, иллюстрирующий средние значения, их стандартные отклонения и медианы не искаженного выбросом массива данных (1), массива данных с выбросом (2) и такого же массива, преобразованного к «хронологическому» виду (3). Подробные комментарии к этому рисунку приведены далее.

Вот в этом и состоит главная особенность применения средних хронологических величин (это отнюдь не предложенный Зенкевичем метод, а использование известной, но незаслуженно игнорируемой разновидности средних значений). В рассмотренном в комментариях [1] примере предполагается, что одно значение ошибочно (выброс), а «традиционная» задача статистической обработки данных состоит в получении устойчивых к выбросам робастных оценок среднего и разброса. Однако следующее важное положение стоит повторить еще раз: ***применение средних хронологических значений не ориентировано на массивы, содержащие значения, которые могли бы быть классифицированы как выбросы***. Речь идет о малых выборках данных, отличающихся их повышенным разбросом. Рассмотрим один из примеров, приведенных в табл. 2.

Площади пиков кумола в серии из пяти дозирований его раствора в газовый хроматограф равны 199941, 186049, 207540, 189010 и 182562 мВ´мс. Основная причина наблюдаемого разброса – недостаточная практика студентов в выполнении этой операции и малый объем проб (1 мкл). За счет неодинакового времени пребывания иглы шприца в нагретом испарителе к этому объему добавляется некоторое неучтенное количество раствора из иглы (0.7 мкл), так что реальный объем проб теоретически может варьировать от 1.0 до 1.7 мкл. Однако эта серия не содержит значений, которые следовало бы классифицировать как ошибочные; все они соответствуют анализируемому образцу, причем причина разброса известна. Обычное среднее арифметическое значение и его стандартное отклонение составляют 193020 ± 10402, среднее хронологическое и его стандартное отклонение – 192513 ± 8241, а медиана выборки и соответству-

ющая ему медиана абсолютного отклонения – 189010 ± 8690. Различие среднего хронологического и среднего арифметического значений составляет всего (-0.3 %), тогда как медианы и среднего арифметического – (-2.1 %), то есть в семь раз больше. Как в случае обычных, так и хронологических средних в интервалы $<A_{отн}> \pm 2s_A$ должны попадать приблизительно 95% значений исходных выборок, как и в нашем случае (в обоих случаях выпадает единственное значение 207540). То же самое наблюдается, если же мы примем диапазон допустимых отклонений от медианы равным удвоенной медиане абсолютного отклонения 2·8690 = 17380. Таким образом, в рассматриваемом примере существенных различий в способах усреднения нет.

Иначе выглядит пример с относительными оптическими плотностями $A_{отн}$ пропиофенона в серии из восьми определений (табл. 1), которые составляют 2.87, 2.89, 2.84, 2.42, 2.53, 2.11, 2.72 и 2.91. Причина разброса в данном случае – искажение значений $A_{отн}$ примесями, которые не удается разделить с пиком целевого компонента. Обычное среднее значение и его стандартное отклонение составляют 2.66 ± 0.29, среднее хронологическое и его стандартное отклонение – 2.68 ± 0.23, а медиана выборки и соответствующая ему медиана абсолютного отклонения – 2.78 ± 0.12. Различие среднего хронологического и среднего арифметического значений составляет всего (+0.8 %), тогда как медианы и среднего арифметического – (+4.5 %), то есть и в данном случае в шесть раз больше. Как в случае обычных, так и хронологических средних в интервалы $<A_{отн}> \pm 2s_A$ должны попадать около 95 % значений исходных выборок, что и наблюдается в нашем случае (в обоих случаях выпадает минимальное значение 2.11). Если же мы примем диапазон допустимых отклонений от медианы равным удвоенной медиане абсолютного отклонения 2·0.12 = 0.24, то мы «потеряем» сразу три значения: 2.11, 2.42 и 2.53.

Однако особенно показателен в комментариях [1] рисунок, иллюстрирующий среднее значения и медиану не искаженных выбросом данных (слева), медианы данных с выбросом (в середине) и среднего хронологического значения данных с выбросом (справа). Если находиться в рамках концепции обязательного исключения выброса, то значение медианы не подвержено его влиянию и практически совпадает со средним арифметическим. Однако ***если исходная задача иная, а именно не выявлять выбросы в массиве данных, а более корректно оценить как среднее значение, так и стандартное отклонение***, то предпочтение все-таки следует отдать хронологическим средним. На том же рисунке в комментариях [1] отчетливо видно, что их стандартное отклонение приблизительно вдвое меньше, чем в случае обычной статистической обработки. При этом важно, что соответствующая выбросу точка попадает в допустимый интервал отклонений от среднего, что полностью отвечает формулировке исходной задачи (не потерять данные). При вычислении медианы такая точка полностью исключена из рассмотрения.

Таким образом, автор статьи не склонен противопоставлять робастные оценки медианы и соответствующей медианы абсолютного отклонения и вычисление средних хронологических значений. ***Задачи этих способов статистической обработки данных различны***: если в первом случае это получение надежных средних значений с полным пренебрежением возможными выбросами, то во втором – корректный учет всех значений исходного массива данных, даже если они по объективным (экспериментальным) причинам характеризуются повышенным разбросом, но необходимости классификации некоторых из таких данных как выбросов нет.

ЛИТЕРАТУРА

1. Померанцев А.Л. Комментарии к статье И.Г. Зенкевича «Вычисление средних хронологических значений – незаслуженно забытый способ статистической обработки» // Аналитика и контроль. 2023.Т. 27, № 1. С. 59-61.
2. The NIST 17 Mass Spectral Library (NIST17/2017/EPA/NIH). Software/Data Version (NIST17); NIST Standard Reference Database, Number 69, June 2017. National Institute of Standards and Technology, Gaithersburg, MD 20899. [Электронный ресурс]: http://webbook.nist.gov (дата обращения: 01. 2023 г.).
3. Вигдергауз М.С., Краузе И.М. Развитие методов количественной интерпретации хроматограмм сложных смесей // Ж. аналит. химии. 1986. Т. 41, № 11. С. 2064-2074.
4. Zenkevich I.G., Makarov E.D. Chromatographic quantitation at losses of analyte during sample preparation. Application of the modified method of double internal standard // J. Chromatogr. A. 2007. V. 1150. P. 117-123.

REFERENCES

1. Pomerantsev A.L. [Comments on the article by I.G. Zenkevich «Calculation of average chronological values – an undeservedly neglected method of statistical data processing»]. *Analytika i Kontrol' [Analytics and Control],* 2023, vol. 27, pp. 59-61. doi: 10.15826/analitika.2023.27.1.00X
2. The NIST 17 Mass Spectral Library (NIST17/2017/EPA/NIH). Software/Data Version (NIST17); NIST Standard Reference Database, Number 69, June 2017. National Institute of Standards and Technology, Gaithersburg, MD 20899: Available at: http://webbook.nist.gov (Accessed January 01. 2023).
3. Vigdergauz M.S., Krauze I.M. Development of the methods of quantitative interpretation of chromatograms of complex mixtures. *J. Analyt. Chem. (Rus.).* 1986. vol. 41. no. 11. pp. 2064-2074.
4. Zenkevich I.G., Makarov E.D. Chromatographic quantitation at losses of analyte during sample preparation. Application of the modified method of double internal standard. *J. Chromatogr. A,* 2007, vol. 1150, pp. 117-123. doi: 10.1016/j.chroma.2006.08.083.