Диаграмма статусов измерений выборки интервальных данных

4 сентября 2022 г.

Оглавление

Введение

Методический материал, в перспективе — учебное пособие.

В теоретической части, §2.1, материал заимствован из книги [3], раздел «Выбросы и их выявление».

В исследовательской части, §3, приведены результаты численных экспериментов с модельными и реальными выборками данных.

В заключении представлены выводы и предложения по дальнейшим исследованиям.

Глава 1

Понятия интервального анализа

Понятия, обозначения, методы, программное обеспечение интервального анализа

Глава 2

Понятия, обозначения, специфичные для анализа данных с интервальной неопределённостью

2.1 Выбросы и их выявление

Материал раздела заимствован из книги [3], раздел «Выбросы и их выявление».

2.1.1 Общие идеи выявления выбросов

Понятие «выброс» в статистике и анализе данных, как правило, определяется нечётко и неформально. Объясняется это тем, что основания для признания измерения выбросом лежат за пределами формальной математической постановки задачи анализа данных и требуют привлечения внешних по отношению к ней знаний из предметной области и истории происхождения данных, специфичных в каждом конкретном случае. Тем не менее, главный объединяющий смысл различных определений — указание на нарушение измерением-выбросом некоторой однородности (согласованности, непротиворечивости), ожидаемой

для большинства наблюдений выборки по отношению к заданной математической модели. Подчеркнём эту особую роль модели и неабсолютный характер понятия «выброс», вкупе означающие, что статус измерения в одной и той же выборке может меняться в зависимости от вида модели, рассматриваемой на конкретном этапе анализа данных. Поэтому, строго говоря, утверждения вида «измерение x_i является выбросом в выборке X» всякий раз должны сопровождаться оговоркой — «относительно такой-то модели», если это явно не следует из контекста.

Интервальный подход даёт естественный формальный индикатор согласованности данных, модели и априорной информации — непустоту информационного множества , соответствующего задаче. Пустота информационного множества свидетельствует о возможном наличии тех или иных противоречий между данными и моделью. Поиск причин появления противоречий, а также выбор путей их преодоления — процесс творческий и неформальный, большей частью опирающийся на прикладные соображения и экспертные знания о моделируемом явлении или процессе и условиях получения данных. Формальные приёмы и математические методы, задействованные в этом процессе, выполняют важную, но подчиненную роль. Они используются для получения информации о данных и модели, позволяющей выдвигать гипотезы о причинах противоречий, вырабатывать способы коррекции данных или модели и оценивать обеспечиваемые ею результаты. Иными словами, математические методы отвечают на вопрос «как устроены данные?», в то время как ответы на вопросы «почему так устроены данные?» и «что делать?» может дать только содержательный анализ моделируемого явления.

Причинами возникновения противоречий в задаче анализа данных могут служить как некорректность измерений (вследствие нарушений условий их проведения, регистрации, сбоев при передаче, некорректной оценки уровня неопределённости, нештатного поведения моделируемой системы и т.п.), так и некорректность модели (вид модели не соответствует моделируемому явлению и т.п.). При использовании формальных методов выявления выбросов следует иметь в виду, что выбросы могут оказаться наиболее существенной частью выборки, проливающей свет на то, как собирались данные или каково истинное поведение изучаемой системы или процесса, не укладывающееся в исходные предположения. Учитывая, что предметом интервального анализа часто становятся малые выборки, обычная тактика удаления «подозрительных» измерений должна использоваться с особой осторожностью.

Обозначения (подлежат обсуждению)

$s_i = (x_i, \boldsymbol{y}_i)$ — наблюдение, состоящее из значения входной переменной $x \in \mathbb{R}^m$ и интервального измерения $\boldsymbol{y}_i$ выходной переменной $y \in \mathbb{R}$.

$\mathrm{S}_n = \{s_i\}_{i=1,\ldots,n} = \{(x_i, \boldsymbol{y}_i)\}_{i=1,\ldots,n}$ — выборка из n наблюдений.

$y(x) = f(x, \beta)$ — модель с параметрами $\beta \in \mathbb{R}^{m+1}$, например, линейная $y(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m$.

$\Upsilon(x)$ — коридор совместных зависимостей.

$\Upsilon(x; \mathrm{S}_n)$ — коридор совместных зависимостей, построенных по выборке S_n.

$\Omega_i = \Omega(s_i) = \{\beta \mid f(x_i, \beta) \subset \boldsymbol{y}_i\}$ — информационное множество наблюдения $s_i = (x_i, \boldsymbol{y}_i)$.

$\Omega = \Omega(\mathrm{S}_n) = \cap_{i=1}^n \Omega_i$ — информационное множество задачи построения модели $y(x) = f(x, \beta)$ по выборке S_n.

2.1.2 Статус измерений (неокончательный вариант)

Классификация измерений. Проедём классификация измерений интервальной выборки. О влиянии некоторого интервального измерения $s = (x, \boldsymbol{y})$ на модель, построенную по выборке S_n, можно судить на основе того, в каком взаимоотношении находятся информационные множества $\Omega(s)$ и $\Omega(\mathrm{S}_n)$. Такая характеризация полезна как для «новых» измерений ($s \notin \mathrm{S}_n$), так и для измерений, уже входящих в выборку ($s \in \mathrm{S}_n$).

Измерения, добавление которых к выборке не приводит к модификации модели ($\Omega(\mathrm{S}_n) = \Omega(\mathrm{S}_n \cup s)$), именуются (согласно [1, 2]) *внутренними*, изменяющие же модель ($\Omega(\mathrm{S}_n) \supset \Omega(\mathrm{S}_n \cup s)$) — *внешними*. В каждом из этих классов измерений дополнительно выделяют специальные подклассы — *граничные* измерения и *выбросы* соответственно (Рис. 2.1).

Граничными называют измерения, определяющие какой-либо фрагмент границы информационного множества. Очевидно, это свойство имеет смысл рассматривать для наблюдений, принадлежащих выборке S_n, по которой сконструирована модель и её информационное множество $\Omega(\mathrm{S}_n)$. Подмножество всех граничных наблюдений в S_n играет особую роль, поскольку оно является минимальной подвыборкой, полностью определяющей модель. Удаление негранычных наблюдений из выборки не изменяет модель.

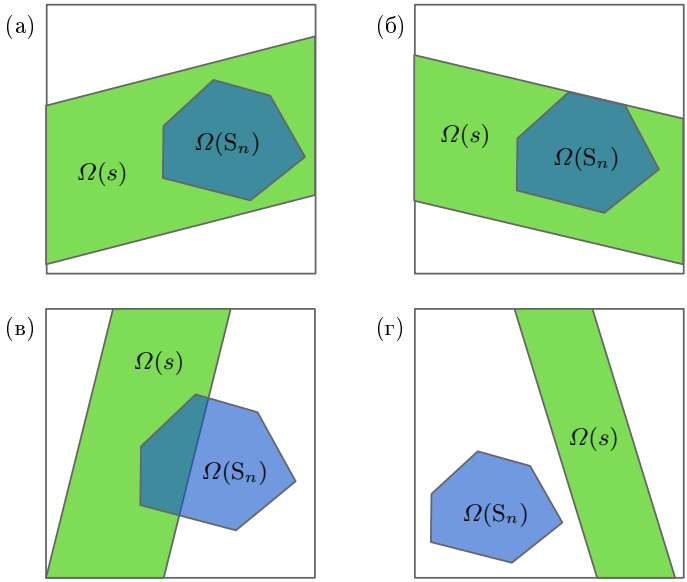

Рис. 2.1. Информационные множества, построенные по выборке S_n и наблюдению s с различными статусами: (а) *внутреннее*, (б) *граничное*, (в) *внешнее*, (г) *выброс*.

Среди внешних измерений особым образом выделяют *выбросы* (промахи). Построение модели по выборке, пополненной таким наблюдением, приводит не просто к уменьшению информационного множества, а к его пустоте ($\Omega(\mathrm{S}_n \cup s) = \varnothing$), то есть к «разрушению» модели.

Существует экономичный способ определения статуса измерения, не требующий явного перестроения модели для выборки, расширенной анализируемым измерением. Анализ взаимоотношений информационных множеств $\Omega(\mathrm{S}_n)$ и $\Omega(\mathrm{S}_n \cup s)$ или $\Omega(\mathrm{S}_n)$ и $\Omega(s)$ можно заменить выяснением отношений интервала неопределённости $\boldsymbol{y}$ анализируемого измерения $s = (x, \boldsymbol{y})$ и интервального прогнозного значения рассматриваемой модели в той же точке $\Upsilon(x; \mathrm{S}_n)$. На Рис. 2.2 анализируемые измерения показаны чёрными линиями, а соответствующие им интервалы прогнозов — широкими цветными линиями (в данном случае их ширина не имеет содержательного смысла, а лишь упрощает восприятие наложенных друг на друга интервалов).

Внутреннее интервальное измерение $s = (x, \boldsymbol{y})$ полностью со-

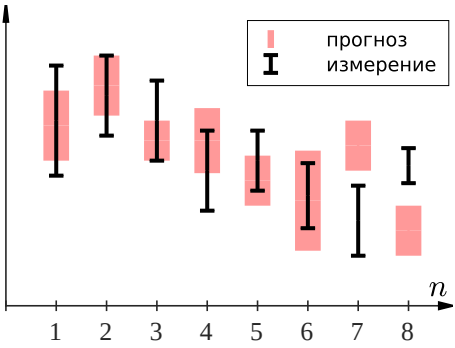

Рис. 2.2. Интервальные наблюдения с различными статусами: *внутреннее* ($n = 1, \ldots, 3$), *граничные* ($n = 2, 3$), *внешние* ($n = 4, \ldots, 8$), *строго внешнее* ($n = 6$), *выбросы* ($n = 7, 8$).

держит в себе прогнозный интервал, оцененный с помощью модели $\Upsilon(x; \mathrm{S}_n)$, или, иными словами, пересечение двух этих интервалов совпадает с прогнозным:

$$\boldsymbol{y} \cap \Upsilon(x; \mathrm{S}_n) = \Upsilon(x; \mathrm{S}_n). \tag{2.1}$$

Будучи перестроенной по выборке, пополненной подобным измерением, модель не претерпит изменений, поскольку соответствующее ей информационное множество окажется внутри ограничения, порожденного добавленным внутренним измерением, а, следовательно, пересечение с ним не изменится. Коридор совместных зависимостей при этом также сохранит прежний вид.

Если внешнее интервальное измерение и соответствующий ему интервал прогноза имеют непустое пересечение, то результирующий интервал сужается по сравнению с прогнозным:

$$\boldsymbol{y} \cap \Upsilon(x; \mathrm{S}_n) \subset \Upsilon(x; \mathrm{S}_n). \tag{2.2}$$

Это означает, что добавление *внешнего* измерения в модель уменьшит информационное множество задачи и коридор совместных зависимостей. Получение пустого множества в пересечении свидетельствует о том, что измерение, возможно, является выбросом по отношению к используемой модели. В некоторых ситуациях, когда требуется более высокий «уровень подозрительности», предпринимать меры можно не

при строгой пустоте информационного множества, а уже при некотором неестественно малом его размере [22].

Взаимные отношения интервалов анализируемого наблюдения и прогнозного интервала рассматриваемой модели. Взаимные отношения интервалов анализируемого наблюдения $(x, \boldsymbol{y})$ и прогнозного интервала рассматриваемой модели $\Upsilon(x)$ удобно характеризовать в специальных терминах. Введём понятия *размаха* (плечо, англ. — high leverage)

$$\ell(x, \boldsymbol{y}) = \frac{\operatorname{rad} \Upsilon(x)}{\operatorname{rad} \boldsymbol{y}} \tag{2.3}$$

и *относительного остатка* (относительное остаточное отклонение, относительное смещение, англ. — relative residual)

$$r(x, \boldsymbol{y}) = \frac{\operatorname{mid} \boldsymbol{y} - \operatorname{mid} \Upsilon(x)}{\operatorname{rad} \boldsymbol{y}}. \tag{2.4}$$

Обе величины являются относительными, поскольку нормируются на величину неопределённости наблюдения y. Размах наблюдения косвенно характеризует положение наблюдения в пространстве независимых переменных x_i. Наблюдения с размахом выше единицы лежат за пределами «области определения» зависимости, образованной наблюдениями выборки, по которой построена зависимость. Остаток характеризует смещение наблюдения по откликовой переменной y относительно коридора совместных зависимостей.

Наблюдения с большими значениями размаха и остатка при их включении в выборку, по которой построен коридор совместных зависимостей, могут существенно повлиять на его вид.

Набор неравенств для классификации измерений. Размах и остаток позволяют установить статус наблюдения, проверив некоторые простые неравенства.

Так для внутренних наблюдений, содержащих в себе прогнозный интервал модели, выполняется нестрогое неравенство

$$|r(x, \boldsymbol{y})| \leq 1 - \ell(x, \boldsymbol{y}), \tag{2.5}$$

а точное равенство в нём является характеристическим условием для граничных наблюдений.

Выбросы — наблюдения, не пересекающиеся с коридором совместных зависимостей, а потому они удовлетворяют неравенству

$$|r(x, \boldsymbol{y})| > 1 + \ell(x, \boldsymbol{y}). \tag{2.6}$$

Интервальные измерения, у которых величина неопределённости меньше, чем ширина прогнозного интервала, то есть

$$\ell(x, \boldsymbol{y}) > 1, \tag{2.7}$$

могут оказывать очень сильное влияние на модель и потому называются *строго внешними*. В [1] и последующих работах О.Е. Родионовой и А.Л. Померанцева, предложивших эту классификацию статусов наблюдений, для обозначения строго внешнего наблюдения используется термин «абсолютно внешнее наблюдение», который неудачен из-за пересечения смысла с общематематическими понятиями «абсолютная величина», «абсолютная погрешность», «абсолютно непрерывный» и т.п.

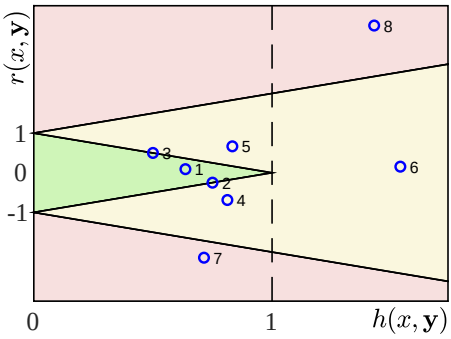

Рис. 2.3. Диаграмма статусов для интервальных наблюдений, показанных на Рис. 2.2. Зоны наблюдений с различными статусами обозначены цветами: зелёный — внутренние наблюдения, жёлтый — внешние, красный — выбросы.
Изменить обозначение h на l

Диаграмма статусов измерений интервальной выборки. Удобным инструментом анализа ролей наблюдений и их влияния на уже имеющуюся модель является *диаграмма статусов*, пример которой

приведен на Рис. 2.3. Наблюдения на диаграмме статусов представляются точками на плоскости, координаты которых задаются остатком и размахом. Неравенства (2.5)–(2.7) на этой плоскости задают границы областей, соответствующих различным статусам наблюдений. Зона внутренних наблюдений выделена зелёным цветом. Наблюдения, размещенные на границе зелёной зоны, являются граничными для информационного множества задачи. Зона внешних наблюдений — жёлтая. Правее вертикали $\ell(x, \boldsymbol{y}) = 1$ лежат строго внешние наблюдения. Выбросы локализуются в красной зоне.

Примечательно, что характеризация наблюдений в терминах размахов и остатков не зависит от размерности входной переменной x и позволяет поддержать анализ статусов наблюдений визуальными инструментами даже в случаях, когда явное отображение информационного множества задачи и коридора совместных зависимостей затруднительно. По своему назначению диаграмма статусов интервальных наблюдений является содержательным аналогом широко используемого в классическом регрессионном анализе графика влияния (английский термин — influence plot), который также служит для оценки степени однородности (похожести) наблюдений и их потенциальной влиятельности на конструируемую зависимость.

2.1.3 Варьирование величины неопределённости измерений

Один из приёмов выявления выбросов в задаче построения зависимости по интервальным наблюдениям основан на интерпретации выбросов как наблюдений с недооценённой величиной неопределённости [**?**, 25]. Закономерным шагом в этом случае становится поиск некоторой минимальной коррекции величин неопределённости интервальных наблюдений, необходимой для обеспечения совместности задачи построения зависимости. Если величину коррекции каждого интервального наблюдения $\boldsymbol{y}_i = [\mathring{y}_i - \epsilon_i, \mathring{y}_i + \epsilon_i]$ выборки S_n выражать коэффициентом его уширения $w_i \geq 1$, а общее изменение выборки характеризовать суммой этих коэффициентов, то минимальная коррекция выборки в виде вектора коэффициентов $w^* = (w_1^*, \ldots, w_n^*)$, необходимая для совместности задачи построения зависимости $y = f(x, \beta)$ может быть найдена решением задачи условной оптимизации

$$\text{найти} \quad \min_{w,\beta} \ \Sigma_{i=1}^{n} w_i \tag{2.8}$$

при ограничениях

$$\begin{cases} \mathring{y}_i - w_i\,\epsilon_i \le f(x_i,\beta) \le \mathring{y}_i + w_i\,\epsilon_i, \\ w_i \ge 1, \end{cases} \qquad i = 1,\dots,n. \tag{2.9}$$

Результирующие значения коэффициентов w_i^*, строго превосходящие единицу, указывают на наблюдения, которые требуют уширения интервалов неопределённости для обеспечения совместности данных и модели. Именно такие наблюдения заслуживают внимания при анализе данных на выбросы. Значительное количество подобных наблюдений может говорить либо о неверно выбранной структуре зависимости, либо о том, что величины неопределённости измерений занижены во многих наблюдениях (например, в результате неверной оценки точности измерительного прибора).

Следует отметить значительную гибкость языка неравенств. Он даёт возможность переформулировать и расширять систему ограничений (2.9) для учёта специфики данных и задачи при поиске допустимой коррекции данных, приводящей к разрешению исходных противоречий. Например, если имеются основания считать, что величина неопределённости некоторой группы наблюдений одинакова и при коррекции должна увеличиваться синхронно, то система ограничений (2.9) может быть пополнена равенствами вида

$$w_{i_1} = w_{i_2} = \dots = w_{i_K},$$

где $i_1,\dots,i_K$ — номера наблюдений группы. В случае, когда в надёжности каких-либо наблюдений исследователь уверен полностью, при решении задачи (2.8)–(2.9) соответствующие им величины w_i можно положить равными единице, т.е. запретить варьировать их неопределённость.

Задача поиска коэффициентов масштабирования величины неопределённости (2.8)–(2.9) сформулирована для распространённого случая уравновешенных интервалов погрешности и подразумевает синхронную подвижность верхней и нижней границ интервалов неопределённости измерений $\boldsymbol{y}_i$ при сохранении базовых значений интервалов $\mathring{y}_i$ неподвижными. При необходимости постановка задачи легко обобщается. Например, если интервалы наблюдений не уравновешены относительно базовых значений (то есть $\boldsymbol{y}_i = [\mathring{y}_i - \epsilon_i^-,\ \mathring{y}_i + \epsilon_i^+]$ и $\epsilon^- \ne \epsilon^+$), то границы интервальных измерений можно варьировать независимо,

масштабируя величины неопределённости ϵ_i^- и ϵ_i^+ с помощью отдельных коэффициентов w_i^- и w_i^+:

$$\text{найти} \quad \min_{w^-,\, w^+,\, \beta} \Sigma_{i=1}^n (w_i^- + w_i^+) \tag{2.10}$$

при ограничениях

$$\begin{cases} \mathring{y}_i - w_i^- \, \epsilon_i^- \le f(x_i, \beta) \le \mathring{y}_i + w_i^+ \, \epsilon_i^+, \\ w_i^- \ge 1, \qquad\qquad i = 1, \ldots, n. \\ w_i^+ \ge 1, \end{cases} \tag{2.11}$$

Для линейной по параметрам β зависимости $y = f(x, \beta)$ задача (2.8)–(2.9) представляет собою задачу линейного программирования, для решения которой широко доступны хорошие и апробированные программы в составе библиотек на различных языках программирования, в виде стандартных процедур систем компьютерной математики, а также в виде интерактивных подсистем электронных таблиц.

Идея варьирования величины неопределённости интервальных измерений оформилась в 80-е годы XX века (Н.М. Оскорбин [44] и др.), и далее неоднократно переоткрывалась различными исследователями.

Удобное программное обеспечение на языке Octave в виде набора jupyter-блокнотов разработано С.И.Жилиным [33]. Это программное обеспечение и будет использоваться нами для численных экспериментов.

Глава 3

Применение диаграммы статуса измерений

Проведём ряд численных экспериментов по применению диаграммы статуса измерений выборки интревальных данных. Предметом изучения будет случай несовместных выборок, который редко рассматривается в литературе.

Для выясненения основных закономерностей проблемы расссмотрим базовый случай анализа данных — модель постоянной величины. Будем следовать терминологии, введённой в §2.1.2 и представленной графически на Рис. 2.2.

3.1 Применение диаграммы статуса измерений для несовместных выборок

3.1.1 Задача измерения постоянной величины

Приведйм определение задачи измерения постоянной величины и используемые оценки, следуя [3].

Постоянная величина — это величина, которая в рассматриваемом процессе сохраняет свое значение неизменным. В отличие от фундаментальных констант, постоянные величины не обязательно остаются таковыми продолжительное время.

Пусть имеется выборка измерений некоторой величины,

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n, \tag{3.1}$$

или, кратко, $\{\,\boldsymbol{x}_k\}_{k=1}^n$, где k — номер измерения, $\boldsymbol{x}_k$ — интервальный результат измерения, полученный, к примеру, какой-либо из процедур, описанных в предыдущих параграфах. Таким образом, согласно терминологии интервального анализа, рассматриваемая выборка — это вектор интервалов, или интервальный вектор $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$. Число n — размерность вектора данных — будем, как обычно, называть *длиной выборки* (или объёмом выборки). По интервальным результатам измерений или наблюдений требуется построить оценку для интересующей нас величины.

Найдём внешнюю и внутреннюю оценку $\boldsymbol{X} = \{\,\boldsymbol{x}_k\}_{k=1}^n$ согласно формулам из раздела «Обработка ненакрывающих выборок» книги [3].

Уточнение пересечением здесь уже невозможно, и информационное множество для истинного значения величины имеет смысл взять в виде объединения всех интервалов выборки, т. е. как

$$\bigcup_{1\le k\le n} \boldsymbol{x}_k. \tag{3.2}$$

Это множество может не быть единым интервалом на вещественной оси (подобное часто случается, к примеру, если выборка несовместна). В интервальном анализе такой объект называется *мультиинтервалом*.

В случае мультиинтервальности объединения всех интервалов выборки, следует воспользоваться вместо объединения обобщающей его операцией «$\vee$», т. е. взятием максимума по включению, и вместо (3.2) взять информационный интервал в виде

$$\boldsymbol{J} \;=\; \bigvee_{1\le k\le n} \boldsymbol{x}_k \;=\; \Big[\min_{1\le k\le n} \underline{\boldsymbol{x}}_k,\, \max_{1\le k\le n} \overline{\boldsymbol{x}}_k\Big]. \tag{3.3}$$

Точечной оценкой измеряемой величины может служить середина полученного интервала, т. е.

$$x_{\mathrm{c}} \;=\; \operatorname{mid} \boldsymbol{J} \;=\; \tfrac{1}{2}\Big(\min_{1\le k\le n} \underline{\boldsymbol{x}}_k + \max_{1\le k\le n} \overline{\boldsymbol{x}}_k\Big). \tag{3.4}$$

Другой возможный сценарий обработки данных ненакрывающей выборки может состоять в том, что вместо пересечения интерваль-

ных измерений мы используем обобщающую её операцию «∧», т. е. взятие минимума всех интервальных результатов измерений относительно упорядочения по включению:

$$\boldsymbol{I} \;=\; \bigwedge_{1\le k\le n} \boldsymbol{x}_k \;=\; \Big[\max_{1\le k\le n} \underline{\boldsymbol{x}}_k,\, \min_{1\le k\le n} \overline{\boldsymbol{x}}_k\Big]. \tag{3.5}$$

Здесь по существу требуется использование полной интервальной арифметики Каухера, так как интервал (3.5) может оказаться неправильным. Соответственно, точечной оценкой измеряемой величины целесообразно взять

$$x_{\mathrm{c}} \;=\; \operatorname{mid} \boldsymbol{I} \;=\; \tfrac{1}{2}\Big(\max_{1\le k\le n} \underline{\boldsymbol{x}}_k + \min_{1\le k\le n} \overline{\boldsymbol{x}}_k\Big), \tag{3.6}$$

т. е. середину интервала, который получается как минимум по включению всех интервалов выборки. Если выборка совместна, то (3.6) совпадает с (3.4). Если же выборка несовместна, то результатом (3.5) является неправильный интервал $\boldsymbol{I}$, $\operatorname{rad} \boldsymbol{I} < 0$. Соответственно, информационное множество результатов измерений по обрабатываемой выборке пусто.

Но даже когда интервал (3.5) неправилен, его середина (3.6) — это точка, обладающая определёнными условиями оптимальности. Она первой появляется в непустом пересечении интервалов выборки, если мы станем равномерно уширять их, увеличивая неопределённость измерений (см. §2.1.3).

Важно, что приведённые конструкции удовлетворяют *принципу соответствия*. Этот факт обеспечивает, в частности, единый подход при переходе от ненакрывающей выборки к накрывающей и наоборот, например, при применении процедуры варьирования неопределённости.

3.1.2 Обработка модельной выборки

Пример 1 (Обработка модельной выборки) Рассмотрим пример измерения постоянной величины.

Пусть имеется выборка

$$\boldsymbol{X} = [\ [1,2],[3,4],[2,3],[1,3],[2,4],[1,4]\ ]\,. \tag{3.7}$$

Диаграмма рассеяния выборки $\boldsymbol{X}$ приведена на Рис. 3.1.

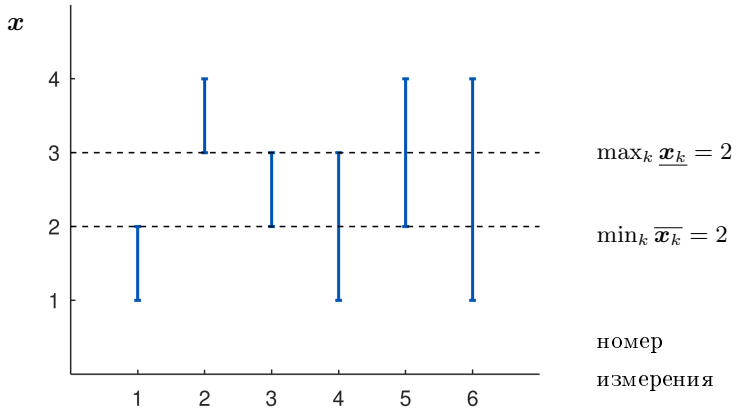

Рис. 3.1. Диаграмма рассеяния интервальной выборки (3.7).

По (3.3) получим максимум по включению

$$\boldsymbol{J} = \bigvee_{1\le k\le n} \boldsymbol{x}_k = \left[\min_{1\le k\le n} \underline{\boldsymbol{x}}_k, \max_{1\le k\le n} \overline{\boldsymbol{x}}_k\right] = [1,4]. \tag{3.8}$$

и по (3.5) — минимум по включению

$$\boldsymbol{I} = \bigwedge_{1\le k\le n} \boldsymbol{x}_k = \left[\max_{1\le k\le n} \underline{\boldsymbol{x}}_k, \min_{1\le k\le n} \overline{\boldsymbol{x}}_k\right] = [3,2]. \tag{3.9}$$

Интервал (3.9) — неправильный. связи с этим, диаграмму статуса измерений Рис. 2.3 следует доработать.

Например, её можно представить как на Рис. 3.2. Поскольку интервал (3.9) — неправильный, его радиус отрицателен. Поэтому ось $l(x, \boldsymbol{y})$ следует продлить в область отрицательных значений, при этом все значения размаха $l(x, \boldsymbol{y})$ (2.3) для несовместной выбоки отрицательны.

Разберём статус различных измерений выборки (3.7) согласно классификации, представленной в §2.1.2.

Как видно из Рис. 3.1 и Рис. 3.2, замеры (3)-(6) образуют совместную подвыборку. Устранение любого из этих замеров не повлияет на информационное множество $\boldsymbol{I}$.

При этом статусы измерений (3)-(6) различны. Замер (6) — *внутренними*, не оказывает влияния на размеры информационного множе-

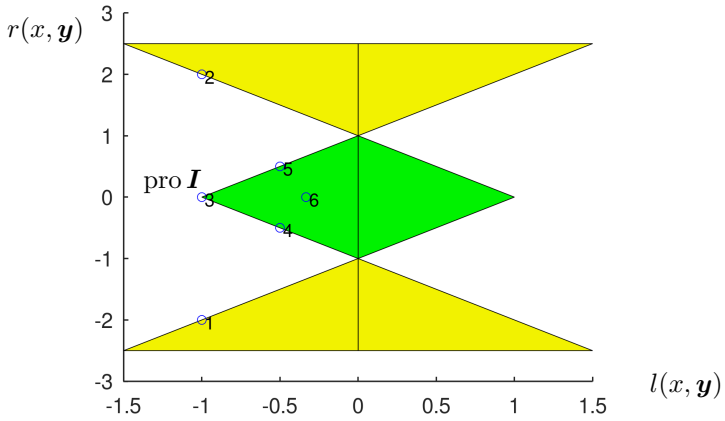

Рис. 3.2. Диаграмма статуса измерений ненакрывающей интервальной выборки (3.7).

ства и может быть исключен из выборки. Тоже самое относилось бы ко всем замерам внутри зелёной области диаграммы Рис. 3.2.

Замеры (4)-(5) являются *граничными*, но с разных сторон информационного множества, что отражается в равных по абсолютной величине, но разных по знаку значениях относительных остатков $r(x, \boldsymbol{y})$ (2.4).

Замер (3) имеет выделенное положение. Он равен правильной проекции минимума по включению (3.9)

$$\boldsymbol{x}_3 = \operatorname{pro} \boldsymbol{I}. \tag{3.10}$$

Как мы увидим далее, «угловое» расположение на диаграмме Рис. 3.2 такого представителя выборки является характерным.

Замеры (1)-(2) являются *внешними* и *граничными*, и как и пара (4)-(5), имеют равные по абсолютной величине, но разные по знаку значения относительных остатков $r(x, \boldsymbol{y})$. Они расположены на границе, определяемой неравенством (2.6), за которой находятся выбросы.

Роль замеров (1)-(2) и (4)-(5), а также (3) состоит в формировании информационного множества задачи определения постоянной величины. Малые возмущения этих замеров могут изменить величину оценки (3.5).

■

Пример 2 (Регуляризация модельной выборки) Продолжим работу с выборкой (3.7)

$$\boldsymbol{X} = [\ [1,2],[3,4],[2,3],[1,3],[2,4],[1,4]\ ]\,.$$

Поставим задачу добиться для $\boldsymbol{X}$ свойства накрытия.

Воспользуемся техникой, представленной в §2.1.3. Поставим задачу линейного программирования, описываемую условиями (2.8) и (2.9). В отсутствие содержательных гипотез, будем считать все измерения равнонадёжными.

Тем самым исходим из того, что величина неопределённости некоторой группы наблюдений одинакова и при коррекции должна увеличиваться синхронно, то система ограничений (2.9) может быть пополнена равенствами вида

$$w_{i_1} = w_{i_2} = \cdots = w_{i_K},$$

где $i_1, \ldots, i_K$ — номера наблюдений группы.

Используем программное обеспечение С.И.Жилина [33] и найдём минимальный коэффициент $w = 2$, при котором модифицированная выборка $\boldsymbol{X}^1$ становится совместной.

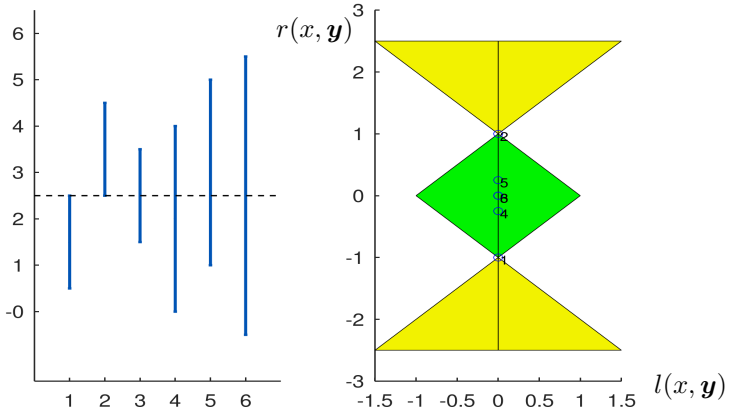

Рис. 3.3. Диаграммы рассеяния и статуса измерений модифицированная выборки $\boldsymbol{X}^1$.

Информационное множество $\boldsymbol{I}$ состоит из одной точки $I = 2.5$. Соответственно, rad $\boldsymbol{I} = 0$ и на диаграмме статуса измерений все замеры имеют значения размаха $l(x, \boldsymbol{y}) = 0$.

На Рис. 3.3 представлена диаграммы рассеяния и статуса измерений регуляризованной выборки $\boldsymbol{X}^1$.

$$\boldsymbol{X}^1 = [\ [0.5, 2.5], [2.5, 4.5], [1.5, 3.5], [0, 4], [1, 5], [-0.5, 5.5]\ ]\,.$$

Разберём статус различных измерений выборки $\boldsymbol{X}^1$. Все замеры являются *внутренними*.

При этом замеры (1)-(2) являются *граничными*, имеют равные по абсолютной величине, но разные по знаку значения относительных остатков $r(x, \boldsymbol{y}) = \pm 1$. Они расположены в точках, определяемых неравенствами (2.5) и (2.6), за которой находятся выбросы.

■

Пример 3 (Работа с совместной модельной выборкой) Продолжим работу с выборкой (3.7) и выберем теперь коэффициент $w = 4$ таким, что информационое множество стало правильным интервалом.

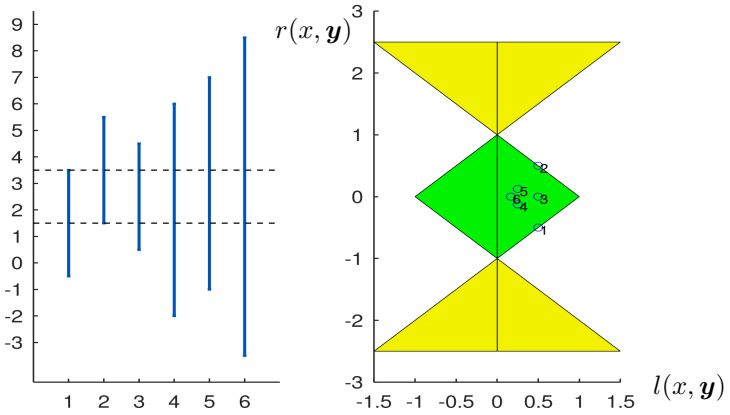

Рис. 3.4. Диаграммы рассеяния и статуса измерений модифицированная выборка $\boldsymbol{X}^2$.

На Рис. 3.4 представлена диаграммы рассеяния и статуса измерений регуляризованной выборки $\boldsymbol{X}^2$ после синхронного расширения интервалов.

$$\boldsymbol{X}^2 = [\ [-0.5, 3.5], [1.5, 5.5], [0.5, 4.5], [-2, 6], [-1, 7], [-3.5, 8.5]\ ].$$

Информационное множество по$\boldsymbol{I} = [1.5, 3.5]$ — правильный интервал. На диаграмме статуса измерений все замеры имеют положительные значения размаха $l(x, \boldsymbol{y})$. Тем самым, диаграмме статуса измерений Рис. 3.4 приобрела вид, введённый в практику авторами [1], [2].

Разберём статус различных измерений выборки $\boldsymbol{X}^2$. Как видно из Рис. 3.4, замеры (3)-(6) являются *внутренними* и образуют совместную подвыборку. Устранение любого из этих замеров не повлияет на информационное множество $\boldsymbol{I}$.

Замеры (1)-(2) являются *внутренними* и *граничными*, имеют равные по абсолютной величине, но разные по знаку значения относительных остатков $r(x, \boldsymbol{y})$. Они расположены на границе, определяемой неравенством (2.5). ■

Заметим, что все измерения при увеличении их радиусов стремятся к «угловой» точке области внутренних измерений. Придадим этому факту предельную форму, взяв интервальную выборку из одинаковых элементов.

Пример 4 (Интервальная выборка из одинаковых элементов) Рассмотрим интервальную выборку из одинаковых элементов.

$$\boldsymbol{X}^{eq} = [\ [2, 3], [2, 3], [2, 3], [2, 3], [2, 3], [2, 3]\ ].$$

На Рис. 3.5 представлены диаграммы рассеяния и статуса измерений выборки $\boldsymbol{X}^{eq}$.

Информационное множество $\boldsymbol{I} = [2, 3]$ совпадает со всеми интервалами замеров выборки.

$$\boldsymbol{x}_i = \operatorname{pro} \boldsymbol{I}, \quad i = 1, 2, \ldots, 6.$$

На диаграмме статуса измерений все замеры имеют значения размаха и относительных остатков.

$$l(x, \boldsymbol{y}) = 1, \quad r(x, \boldsymbol{y}) = 0. \tag{3.11}$$

Такое расположение на диаграмме следует сравнить со случаем положения интервала $\boldsymbol{x}_3$ исходной выборки $\boldsymbol{X}$ (3.10). Он был равен равен правильной проекции минимума по включению (3.9)

$$\boldsymbol{x}_3 = \operatorname{pro} \boldsymbol{I}.$$

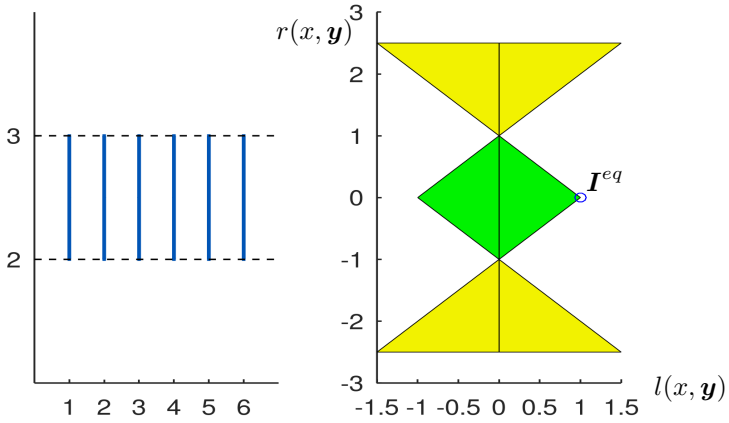

Рис. 3.5. Диаграммы рассеяния и статуса измерений модифицированная выборка $\boldsymbol{X}^2$.

Таким образом, «угловое» расположение элементов, равных $\boldsymbol{I}$ или $\operatorname{pro} \boldsymbol{I}$ на диаграммах Рис. 3.2 и Рис. 3.5 является характерным предельным случаем.

Если даже выборка сначала состояла из разных замеров, при увеличении радиусов, все положения её замеров будут стремиться к точке (3.11). ■

3.1.3 Диаграмма статусов измерений интервальной выборки. Обобщение на случай несовместных выборок.

На основании рассмотренных в §3.1.2 примеров сделаем предложение по обобщению вида диаграмме статуса измерений, введённого в практику Померанцевым А.Л. и Родионовой О.Е. [1], [2].

При рассмотрении интервальных выборкок, в том числе несовместных, будем использовать арифметику Каухера. В таком случае информационное множество задачи измерения постоянной величины $\boldsymbol{I}$ в форме (3.5) может быть неправильным интервалом. В таком случае его радиус в формуле (2.3) для размаха следует полагать отрицательным.

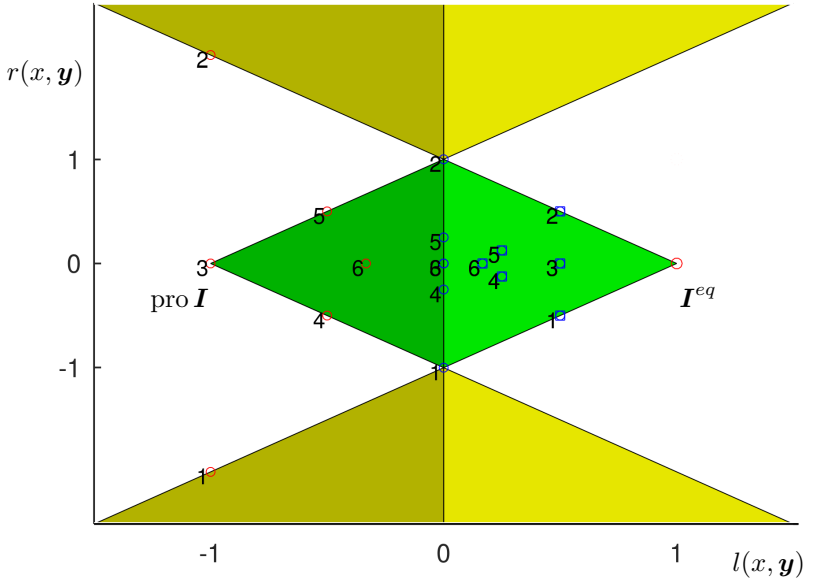

Рис. 3.6. Диаграмма статуса измерений ненакрывающей интервальной выборки (3.7) и её последовательных трансформаций при увеличении радиусов измерений.

На Рис. 3.6 представлена диаграмма статуса измерений ненакрывающей интервальной выборки (3.7) и её последовательных трансформаций при увеличении радиусов измерений. Красными кружками даны измерения для исходной выборки $\boldsymbol{X}$, синими кружками — для регуляризованной выборки $\boldsymbol{X}^1$, синими квадратами — для совместной выборки $\boldsymbol{X}^2$. Также показаны положения информационных множеств pro $\boldsymbol{I}$ и $\boldsymbol{I}^{eq}$.

При последовательном увеличении радиусов измерений все измерения в конечном счёте попадают в точку с координатами (3.11), соот-

ветсвующую выборке одинаковых измерений $\boldsymbol{I}^{eq}$.

На Рис. 3.6 представлена диаграмма статуса измерений ненакрывающей интервальной выборки (3.7) и «траектории» статусов измерений при увеличении их радиусов. Измерение $\boldsymbol{x}_6$ исключено из выборки, из-за сходства его траектории с $\boldsymbol{x}_3$.

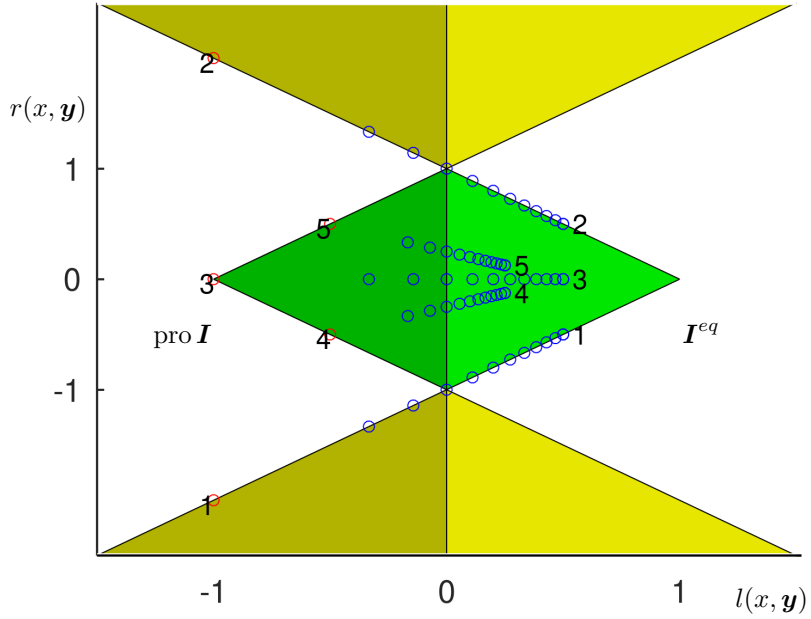

Рис. 3.7. Диаграмма статуса измерений ненакрывающей интервальной выборки (3.7) и траектории статусов измерений при увеличении их радиусов.

3.1.4 Диаграмма статусов измерений интервальной выборки. Присутствие выбросов. (не сделано)

Рассмотрим теперь пример с добавлением к выборке дополнительного измерения, попадающего по критерию (2.6) в область выбросов.

Литература

[1] Померанцев А.Л., Родионова О.Е. Построение многомерной градуировки методом простого интервального оценивания // Журнал Аналитической Химии. – 2006. – Т. 61, №10. – С. 1032–1047.

[2] Родионова О.Е. Интервальный подход к анализу больших массивов физико-химических данных. – Диссертация ... доктора физико-математических наук. – Институт физической химии им. Н.Н. Семёнова РАН: Москва, 2007.

[3] А.Н. Баженов, С.И. Жилин, С.И. Кумков, С.П. Шарый. Обработка и анализ данных с интервальной неопределённостью. 2022

[4] А.Н. Баженов. Интервальный анализ. Основы теории и учебные примеры: учебное пособие. Санкт-Петербург, 2020.
https://elib.spbstu.ru/dl/2/s20-76.pdf/info

[5] А.Н. Баженов. Естественнонаучные и технические применения интервального анализа. Санкт-Петербург, 2021.
https://elib.spbstu.ru/dl/5/tr/2021/tr21-169.pdf/info

[6] Шарый С.П. Конечномерный интервальный анализ. – ФИЦ ИВТ: Новосибирск, 2021. Электронная книга, доступная на `http://www.nsc.ru/interval/Library/InteBooks/SSharyBook.pdf`

[7] А.Н. Баженов. Введение в анализ данных с интервальной неопределённостью: учебное пособие. Санкт-Петербург, 2022.

[8] А.Н. Баженов, А.Ю. Тельнова. Обобщение коэффициента Жаккара для анализа данных с интервальной неопределенностью.
Измерительная техника. 2022

[9] A.N. Bazhenov, O.M Skrekel. A revision of data processing of the circular polarization of the γ-quanta in the $np \to d\gamma$ reactions with polarized neutrons.
NIM A (to be submitted)

[10] Шарый С.П. Сильная согласованность в задаче восстановления зависимостей при интервальной неопределенности данных // Вычислительные технологии. – 2017. – Т. 2, №2. – С. 150–172.

[11] Nguyen H.T., Kreinovich V., Wu B., Xiang G. Computing Statistics under Interval and Fuzzy Uncertainty. Applications to Computer Science and Engineering. – Springer, Berlin-Heidelberg, 2012.

[12] Tukey J.W. The Future of Data Analysis // Annals of Mathematical Statistics – 1962. – Vol. 33, Issue 1. – P. 1–67

[13] Коэффициент Жаккара
https://en.wikipedia.org/wiki/Jaccard_index
Jaccard P. Distribution de la flore alpine dans le Bassin des Dranses et dans quelques regions voisines // Bull. Soc. Vaudoise sci. Natur. 1901. V. 37. Bd. 140. S. 241—272.

[14] Б.И. Сёмкин. О СВЯЗИ МЕЖДУ СРЕДНИМИ ЗНАЧЕНИЯМИ ДВУХ МЕР ВКЛЮЧЕНИЯ И МЕРАМИ СХОДСТВА. Бюллетень Ботанического сада-института ДВО РАН, 2009. Вып. 3. С. 91-101

[15] Ю.А. Шрейдер. Равенство, сходство, порядок: Популярное введение в теорию бинарных отношений. С примерами из математической лингвистики № 248. Изд. 2 URSS. 2021. 256 с. ISBN 978-5-9710-8453-2.

[16] Kearfott R. B., Nakao M. T., Neumaier A., Rump S. M., Shary S. P., van Hentenryck P. Standardized notation in interval analysis // Вычислительные технологии. 2010. Т. 15. № 1. С. 7–13.

[17] IEEE Std 1788-2015 – IEEE standard for interval arithmetic
https://standards.ieee.org/standard/1788-2015.html

[18] Шарый С.П. Метод максимума согласования для восстановления зависимостей по данным с интервальной неопределённостью // Известия Академии Наук. Теория и системы управления. – 2017. – №6. – С. 3–19.

[19] Шарый С. П. Выявление выбросов в методе максимума согласования при анализе интервальных данных // Сборник трудов Всероссийской конференции по математике с международным участием «МАК-2018». – Барнаул : Изд-во АлтГУ, 2018. – С. 215–218.
Доступна на http://elibrary.asu.ru/handle/asu/6303

[20] Шарый С.П. О мере вариабельности оценки параметров в статистике интервальных данных // Вычислительные технологии. – 2019. – Т. 24, №5. – С. 90–108.

[21] S.P. Shary Numerical computation of formal solutions to interval linear systems of equations. https://arxiv.org/abs/1903.10272v1

[22] Dbouk H., Schon S., Neuman I. Kreinovich V. When can we be sure that measurement results are consistent: 1-D interval case and

beyond // Technical Report: UTEP-CS-20-67, University of Texas at El Paso, July 2020. https://scholarworks.utep.edu/cgi/viewcontent.cgi?article=2457&context=cs_techrep

[23] Шарый С.П. Задача восстановления зависимостей по данным с интервальной неопределённостью // Заводская лаборатория. Диагностика материалов. – 2020. – Т. 86, №1. – С. 62–74. DOI: 10.26896/1028-6861-2020-86-1-62-74

[24] Zhilin, S.I. On fitting empirical data under interval error // Reliable Computing. – 2005. – Vol. 11. – P. 433–442. DOI: 10.1007/s11155-005-0050-3

[25] Zhilin S.I. Simple method for outlier detection in fitting experimental data under interval error // Chemometrics and Intelligent Laboratory Systems. – 2007. – Vol. 88, No. 1. – P. 60-68 DOI: 10.1016/j.chemolab.2006.10.004

[26] Кумков С.И. Обработка экспериментальных данных ионной проводимости расплавленного электролита методами интервального анализа // Расплавы. – 2010. – №3. – С. 79–89.

[27] Kumkov, S.I., Mikushina, Yu. V. Interval approach to identification of catalytic process parameters // Reliable Computing. – 2013. – Vol. 19. – P. 197–214.

[28] В.И. Левин Сравнение интервальных чисел и оптимизация систем с интервальными параметрами, Автомат. и телемех., 2004, № 4, 133–142; Autom. Remote Control, 65:4 (2004), 625–633

[29] Kabir, S., Wagner, C., Havens, T. C., Anderson, D. T. and Aickelin, U. Novel Similarity Measure for Interval-Valued Data Based on Overlapping Ratio. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2017), IEEE.
https://doi.org/10.1109/FUZZIEEE.2017.8015623.

[30] T. Wilkin and G. Beliakov, "The Mode of Interval-Valued Data,"2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2019, pp. 1-6,
doi: 10.1109/FUZZ-IEEE.2019.8858850.

[31] Shaily Kabir, Christian Wagner, Timothy C. Havens, Derek T. Anderson, Uwe Aickelin: Novel similarity measure for interval-valued data based on overlapping ratio. FUZZ-IEEE 2017: 1-6

[32] Shaily Kabir, Christian Wagner, Zack Ellerby Towards Handling Uncertainty-at-Source in AI - A Review and Next Steps for Interval Regression. CoRR abs/2104.07245 (2021)

[33] С.И.Жилин. Примеры анализа интервальных данных в Octave. Сборник jupyter-блокнотов с примерами анализа интервальных данных. `https://github.com/szhilin/octave-interval-examples`

[34] А.Н. Баженов, А.А. Карпова Интервальный анализ для исследователей. РХД, Ижевск. 2022.

[35] А.Н. Баженов, А.Н. Коваль, С.Ю. Толстяков, Е.Е. Мухин, А.М. Дмитриев, Д.С. Самсонов Стенд для термовакуумных механических испытаний. // Приборы и техника эксперимента, 2021, т.1 с.: 151–152.

[36] Н.В. Ермаков, А.Н. Баженов, А.Н. Смирнов, С.Ю. Толстяков. Стенд для испытаний шаговых двигателей. // Приборы и техника эксперимента. 2022.

[37] Hu C., Hu Z.H. On statistics, probability, and entropy of interval-valued datasets // Lesot MJ. et al. (eds) Information Processing and Management of Uncertainty in Knowledge-Based Systems. IPMU 2020. Communications in Computer and Information Science, vol 1239. – Cham: Springer, 2020.

[38] Нестеров В.М. Твинные арифметики и их применение в методах и алгоритмах двустороннего интервального оценивания. дисс. д.ф.-м.н. Санкт-Петербург, Санкт-Петербургский институт информатики и автоматизации РАН, 1999, 234 с.

[39] Робертс Ф.С., Дискретные математические модели с приложениями к социальным, биологическим и экологическим задачам. Изд. "Наука 1986, 496 с.

[40] Шарый С.П. Алгебраический подход к анализу линейных статических систем с интервальной неопределённостью. // Известия РАН. Теория и системы управления. – 1997. – № 3. – С. 51-61.

[41] Шарая И.А. Пакет `IntLinInc2D` для визуализации множеств решений интервальных линейных систем с двумя неизвестными. – Программное обеспечение, доступное на `http://www.nsc.ru/interval/sharaya/`. Описание `http://www.nsc.ru/interval/Programing/MCodes/IntLinInc2D.pdf`

[42] A.N.Bazhenov, L.A.Grigor'eva, V.V.Ivanov, E.A.Kolomensky, V.M.Lobashev , V.A.Nazarenko, A.N.Pirozhkov, Yu.V.Sobolev. Circular polarization of γ-quanta in $np \to d\gamma$ reactions with polarized neutrons Physics Letters B Volume 289, Issues 1–2, 3 September 1992, Pages 17-21

[43] С.И.Жилин. Библиотека полной интервальной арифметики `kinterval` в среде `Octave`. Частное сообщение.

[44] Оскорбин Н.М. Некоторые задачи обработки информации в управляемых системах // Синтез и проектирование многоуровневых иерархиче-

ских систем. Материалы конференции. – Барнаул: Алтайский государственный университет, 1983.

[45] М.З.Шварц, рабочие материалы

[46] Шарый С.П. О мере вариабельности оценки параметров в статистике интервальных данных // Вычислительные технологии. – 2019. – Т. 24, №5. – С. 90–108.

Обозначения

$\Rightarrow$	логическая импликация
$\Leftrightarrow$	логическая равносильность
$\&$	логическая конъюнкция, связка «и»
$\to$	отображение множеств; предельный переход
$\varnothing$	пустое множество
$x \in X$	элемент x принадлежит множеству X
$x \notin X$	элемент x не принадлежит множеству X
$X \cup Y$	объединение множеств X и Y
$X \cap Y$	пересечение множеств X и Y
$\mathrm{JK}(\boldsymbol{x}, \boldsymbol{y})$	мера совместности интервалов $\boldsymbol{x}$ и $\boldsymbol{y}$
$\mathrm{JK}_p(\boldsymbol{x}, \boldsymbol{y})$	мера совместности интервалов $\boldsymbol{x}$ и $\boldsymbol{y}$ положительно определённая
$s_x(\boldsymbol{x}, \boldsymbol{y})$	мера совместности интервалов $\boldsymbol{x}$ и $\boldsymbol{y}$ несимметричная
$X \setminus Y$	разность множеств X и Y
$X \subseteq Y$	множество X есть подмножество множества Y
$X \subset Y$	X есть собственное подмножество множества Y
$X \times Y$	прямое декартово произведение множеств X и Y
$\mathbb{R}$	множество вещественных (действительных) чисел
$\mathbb{IR}$	классическая интервальная арифметика
$\mathbb{KR}$	полная интервальная арифметика Каухера

$\mathbb{R}^n$	множество вещественных n-мерных векторов
$\mathbb{IR}^n$, $\mathbb{KR}^n$	множества n-мерных интервальных векторов
$\mathbb{R}^{m\times n}$	множество вещественных $m \times n$-матриц
$\mathbb{IR}^{m\times n}$, $\mathbb{KR}^{m\times n}$	множества интервальных $m \times n$-матриц
$[a, b]$	интервал с нижним концом a и верхним b
$]a, b[$	открытый интервал с концами a и b
$\underline{\boldsymbol{a}}$, $\inf \boldsymbol{a}$	левый конец интервала $\boldsymbol{a}$
$\overline{\boldsymbol{a}}$, $\sup \boldsymbol{a}$	правый конец интервала $\boldsymbol{a}$
$\operatorname{mid} \boldsymbol{a}$	середина интервала $\boldsymbol{a}$
$\operatorname{wid} \boldsymbol{a}$	ширина интервала $\boldsymbol{a}$
$\operatorname{rad} \boldsymbol{a}$	радиус интервала $\boldsymbol{a}$
$\operatorname{dual} \boldsymbol{a}$	дуальный (двойственный) к $\boldsymbol{a}$ интервал
$\operatorname{pro} \boldsymbol{a}$	правильная проекция интервала $\boldsymbol{a}$
$\big[\,[\underline{\boldsymbol{a}}, \overline{\boldsymbol{a}}], [\underline{\boldsymbol{b}}, \overline{\boldsymbol{b}}]\,\big]$	интервал с интервальными концами, твин
$\big[\,[\underline{\boldsymbol{X}}^{in}, \overline{\boldsymbol{X}}^{in}], [\underline{\boldsymbol{X}}^{out}, \overline{\boldsymbol{X}}^{out}]\,\big]$	твин в форме Нестерова
$\lvert\boldsymbol{a}\rvert$	абсолютное значение (модуль) интервала $\boldsymbol{a}$
$\square X$	интервальная оболочка множества $X \subseteq \mathbb{R}^n$
$\wedge$	операция минимума по включению
$\vee$	операция максимума по включению
dist	расстояние (метрика) на множестве интервалов
Dist	векторнозначное расстояние (мультиметрика)
$\operatorname{ran}(f, X)$	область значений функции f на множестве X
min, max	операции взятия минимума и максимума
$\sum$	символ суммы нескольких слагаемых
$\lVert\cdot\rVert$	векторная или матричная норма
x^*	истинное значение измеряемой величины
$\mathring{x}$	базовое измеренное значение величины
Ω	информационное множество задачи
Ξ	множество решений интервальной системы

Предметный указатель