

Тема X-1. Обработка и анализ данных с интервальной неопределённостью.

ФТИ им. А.Ф.Иоффе

a_bazhenov@inbox.ru

16.03.2023

Интервал — замкнутый отрезок вещественной оси, а **интервальная неопределенность** — состояние неполного знания об интересующей нас величине, когда известна лишь ее принадлежность **некоторому интервалу**.

Интервальный анализ — отрасль математического знания, исследующая задачи с интервальными неопределенностями и методы их решения.

Поиск множества, удовлетворяющего постановке задачи.

Интервалом $[a, b]$ вещественной оси R называется множество всех чисел, расположенных между заданными числами включая их самих, т.е.

$$[a, b] := \{x \in \mathbb{R} \mid a \leq x \leq b\}$$

При этом a и b называются концами интервала.

«... В большинстве случаев некорректно говорить о «решении интервальных уравнений» (систем уравнений, неравенств и т. п.) вообще.

Правильнее вести речь о решении тех или иных **постановок задач**, связанных с интервальными уравнениями (системами уравнений, неравенств и т. п.). В свою очередь, формулировка постановки интервальной задачи подразумевает указание, по крайней мере, **множества решений задачи и способа его оценивания**».

С.П.Шарый. Конечномерный интервальный анализ, 2022

Обработка и анализ данных с интервальной неопределённостью.

Общий план

- Общие понятия
- Обработка постоянной величины
- Задача восстановления зависимостей
- Задача классификации данных

Теория:

А.Н. БАЖЕНОВ, С.И. Жилин, С.И. Кумков, С.П. ШАРЫЙ.
Обработка и анализ данных с интервальной неопределённостью.
РХД. Серия «Интервальный анализ и его приложения». Ижевск. 2022.
с.270.

Общие понятия.

Отношения между интервалами.

Интервалы являются множествами, составленными из вещественных чисел, и неудивительно, что большую роль для них играют теоретико-множественные отношения и операции (объединение, пересечение и др.). Особенно важно отношение включения одного интервала в другой:

$$\underline{a} \subseteq \underline{b} \text{ равносильно тому, что } \underline{a} \geq \underline{b} \text{ и } \overline{a} \leq \overline{b}. \quad (1)$$

Отношение включения является частичным порядком и превращает множество интервалов в частично упорядоченное множество, важную и хорошо изученную математическую структуру.

Отношения между интервалами.

Помимо порядка по включению на множестве интервалов огромную роль играют также другие отношения, которые обобщают хорошо известный порядок « \leq » на вещественной оси \mathbb{R} .

Фундаментальным фактом является то, что порядок « \leq » между вещественными числами может быть обобщен на интервалы многими осмысленными способами (и даже бесконечно большим числом способов). Значительная часть получающихся при этом отношений на \mathbb{IR} не являются полноценными порядками.

Отношения между интервалами.

Помимо порядка по включению на множестве интервалов огромную роль играют также другие отношения, которые обобщают хорошо известный порядок « \leq » на вещественной оси \mathbb{R} .

Фундаментальным фактом является то, что порядок « \leq » между вещественными числами может быть обобщен на интервалы многими осмысленными способами (и даже бесконечно большим числом способов). Значительная часть получающихся при этом отношений на \mathbb{IR} не являются полноценными порядками.

Отношения между интервалами.

Важную роль играет следующее упорядочение

Definition

Для интервалов $\mathbf{a}, \mathbf{b} \in \mathbb{IR}$ условимся считать, что \mathbf{a} не превосходит \mathbf{b} и писать « $\mathbf{a} \leq \mathbf{b}$ » тогда и только тогда, когда $\underline{\mathbf{a}} \leq \underline{\mathbf{b}}$ и $\overline{\mathbf{a}} \leq \overline{\mathbf{b}}$.

Интервал называется *неотрицательным*, т. е. « ≥ 0 », если неотрицательны оба его конца. Интервал называется *неположительным*, т. е. « ≤ 0 », если неположительны оба его конца.

Теоретико-множественные операции между интервалами.

Если интервалы ***a*** и ***b*** имеют непустое пересечение, т.е. $\mathbf{a} \cap \mathbf{b} \neq \emptyset$, то можно дать простые выражения для результатов теоретико-множественных операций пересечения и объединения через концы этих интервалов

$$\mathbf{a} \cap \mathbf{b} = [\max\{\underline{\mathbf{a}}, \underline{\mathbf{b}}\}, \min\{\overline{\mathbf{a}}, \overline{\mathbf{b}}\}], \quad \mathbf{a} \cup \mathbf{b} = [\min\{\underline{\mathbf{a}}, \underline{\mathbf{b}}\}, \max\{\overline{\mathbf{a}}, \overline{\mathbf{b}}\}]. \quad (2)$$

Если же $\mathbf{a} \cap \mathbf{b} = \emptyset$, т.е. интервалы ***a*** и ***b*** не имеют общих точек, то эти равенства уже неверны.

Теоретико-множественные операции между интервалами.

Обобщением операций пересечения и объединения являются операции взятия минимума и максимума относительно включения « \subseteq »:

$$\mathbf{a} \wedge \mathbf{b} = [\max\{\underline{\mathbf{a}}, \underline{\mathbf{b}}\}, \min\{\overline{\mathbf{a}}, \overline{\mathbf{b}}\}], \quad \mathbf{a} \vee \mathbf{b} = [\min\{\underline{\mathbf{a}}, \underline{\mathbf{b}}\}, \max\{\overline{\mathbf{a}}, \overline{\mathbf{b}}\}]. \quad (3)$$

Они также понадобятся нам при обработке интервальных измерений.

Первая из этих операций, « \wedge », не всегда выполнима во множестве обычных интервалов, но это затруднение преодолевается путём расширения множества интервалов специальными элементами — неправильными интервалами.

Definition

Измерением (замером, наблюдением) будем называть измеренное значение величины.

По способу получения результата измерения все процессы измерения разделяются на *прямые, косвенные и совокупные*.

- Погрешности квантования
- Неопределённость измерения нуля
- Агрегирование результатов многократных наблюдений

Агрегирование результатов многократных наблюдений.

Во многих практических ситуациях измерение интересующей нас величины выполняется для надёжности многократно. Тем не менее, повторные измерения над одними и теми же явлениями не показывают разумное (в пределах точности измерений) совпадение результатов.

Приняв все необходимые меры предосторожности, обеспечив постоянные условия измерения, мы всё равно не получаем разумно согласующихся друг с другом результатов.

Скажем, в промышленности, как бы тщательно ни был отрегулирован измерительный прибор, колебания в его показаниях не могут быть уменьшены ниже некоторого предела.

Агрегирование результатов многократных наблюдений.

В этих условиях результатом серии повторяющихся измерений можно взять интервал от минимального до максимального из полученных результатов, т. е. агрегировать (объединить) результаты отдельных измерений.

Математически, если результаты повторных измерений величины равны x_1, x_2, \dots, x_n , то интервальным результатом следует взять

$$x = \left[\min_{1 \leq i \leq n} x_i, \max_{1 \leq i \leq n} x_i \right].$$

Будем называть этот способ получения интервального результата измерения *агрегированием*.

Агрегирование результатов многократных наблюдений.

Используя операции взятия *интервальной оболочки* множества и *максимума по включению* этот результат можно записать следующим равносильным образом:

$$\mathbf{x} = \square\{x_1, x_2, \dots, x_n\}$$

или

$$\mathbf{x} = \bigvee_{1 \leq i \leq n} x_i.$$

Эти представления хороши тем, что могут быть обобщены на более сложные случаи.

Модель погрешности наблюдения.

Интервалы в результатах измерений могут возникать различным способом. Они могут получаться сразу, в виде готовых интервалов, но могут возникать в результате коррекции точечных результатов.

Один из распространённых способов получения интервальных результатов в первичных измерениях — это «обинтерваливание» точечных значений, когда к точечному *базовому значению* \hat{x} прибавляется *интервал погрешности* ϵ :

$$x = \hat{x} + \epsilon \quad (4)$$

Модель погрешности наблюдения.

Интервал погрешности, вообще говоря, может быть произвольным, но если он уравновешен, то есть

$$\epsilon = [-\epsilon, \epsilon],$$

то это можно трактовать, как отсутствие систематических погрешностей в прямом измерении.

На практике концы интервалов, представляющие результаты измерений, сами могут быть известны неточно, так что возникает необходимость работы с интервалами, имеющими интервальные концы.

Такие объекты известны в интервальном анализе и называются *твинами* (по английски *twın*, как сокращение фразы *twice interval*, «двойной интервал»).

Твины были введены в научный оборот в начале 80-х годов XX века в работах испанских исследователей.

Развёрнутый анализ дан в диссертации В.М.Нестерова. Твинные арифметики и их применение в методах и алгоритмах двустороннего интервального оценивания. – Санкт-Петербург, 1999.

<http://www.nsc.ru/interval/Library/InteDiss/Nesterov-disser-1999.pdf>

Твин, как «интервал интервалов» или интервал с интервальными концами, можно представить как

$$X = [a, b] = [\underline{a}, \bar{a}], [\underline{b}, \bar{b}]. \quad (5)$$

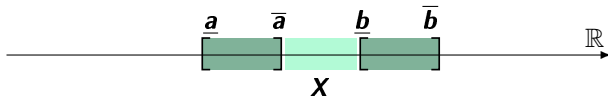


Рис.: Твины на вещественной оси.

На рисунке твин X представлен в графической форме. Концы твина, т.е. интервалы \underline{a} и \underline{b} , даны более тёмной заливкой, чем остальная часть твина.

Твин является множеством всех интервалов, больших или равных $[\underline{a}, \bar{a}]$ и меньших или равных $[\underline{b}, \bar{b}]$, и точное определение зависит от смысла, который вкладывается в понятия «больше или равно», «меньше или равно».

Поскольку интервалы могут быть упорядочены различными способами, то существуют различные виды твинов. Двум основным частичным порядкам на \mathbb{IR} и \mathbb{KR} , « \subseteq » и « \leq », соответствуют два основных типа твинов. Разработаны различные операции с твинами, а также способы оценок значений функций от них.

Измерение температуры термометром сопротивления.

В повседневной лабораторной и промышленной практике широко применяются термометры сопротивления.

Один из типов таких датчиков, платиновый термометр Pt100, имеет номинальное сопротивление 100 Ом при температуре 0°C и систематическую погрешность

$$\Delta t = \pm 0.35 \text{ }^{\circ}\text{C}.$$

Пусть измеряемая температура находится в диапазоне $[19.5, 20.5] ^\circ\text{C}$, которую представим как интервал \mathbf{t} :

$$\mathbf{t} = [19.5, 20.5] ^\circ\text{C}. \quad (6)$$

Аналогично рассмотренному выше примеру, представим границы $\underline{\mathbf{t}}$, $\overline{\mathbf{t}}$ интервала \mathbf{t} как интервалы. С учётом систематической погрешности твин температур \mathbf{T} , даваемый датчиком, составит

$$\mathbf{T} = \left[[19.15, 19.85], [20.15, 20.85] \right] ^\circ\text{C}. \quad (7)$$

Графическое представление твина \mathbf{T} (7) дано на рисунке 2.

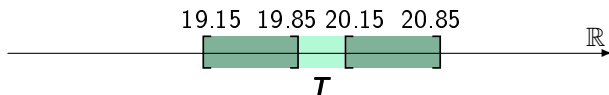


Рис.: Температура как твин.

Форма записи температуры в виде твина T (7) выразительно и полно представляет информацию об измеряемых данных. В случае, если концы интервала в выражении (6) могут меняться независимо, возможны различные ситуации. В частности, может реализоваться ситуация, подобная рассмотренной выше для твина R_2 . Также может оказаться, что значения температур для левого конца будут выше, чем для правого.

Мультиинтервалы.

В ряде разделов науки и техники имеют место ситуации, когда исследуемая величина содержится в неодносвязной области.

Мультиинтервал — это объединение конечного числа несвязных интервалов числовой оси (Рис. 3).



Рис.: Мультиинтервал в \mathbb{R} .

Мультиинтервалы.

Между мультиинтервалами также могут быть определены арифметические операции «по представителям», аналогично тому, как это делается на множестве интервалов.

Мультиинтервальная арифметика применяется редко ввиду серьёзных ограничений, которые возникают при алгебраических операциях с мультиинтервальными величинами и вычислительных сложностей. Тем не менее, сама по себе идея мультиинтервалов содержательна и полностью отменить её не стоит.

Ряд научных и технических примеров возникновения мультиинтервалов приводится в материале А.Н.Баженов.

Естественнонаучные и технические применения интервального анализа: учебное пособие.

<https://elib.spbstu.ru/dl/5/tr/2021/tr21-169.pdf/info>.

Рассмотрим задачу калибровки временной шкалы прибора. Для этого на прибор подаётся гармонический сигнал. В силу того, что на промышленно выпускаемых генераторах положительный и отрицательный фронт имеет разную длительность, необходимо различать эти части временной шкалы.

На рисунке 4 черным цветом показан гармонический сигнал и выделены соответственно красным и синим цветом области положительной и отрицательной производной сигнала. Эти области образуют мультиинтервалы. Они преобразуются при изменении калибровочного сигнала.

При изменении частоты составляющие мультиинтервалов расширяются или сужаются. При изменении фазы происходит их сдвиг.

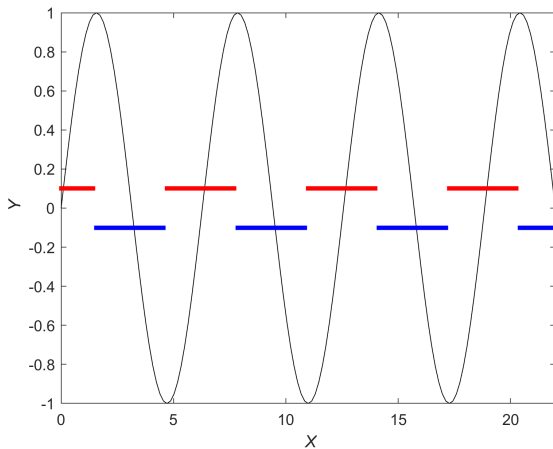


Рис.: Мультиинтервалы фаз гармонического сигнала.

На практике измерения и наблюдения, как правило, подвержены неизбежным внешним влияниям, выполняющие их средства измерений и приборы не вполне точны и т. п., что в целом приводит к отличию измеренного значения от истинного (идеального) значения физической величины.

По отношению к неточным измерениям иногда используют термин «зашумлённые» (зашумлённые данные и т. п.), особенно, когда проводится целая серия таких измерений или наблюдений. Чтобы количественно охарактеризовать неточности измерений, вводится понятие *погрешности*.

Погрешность измерений — вещественная арифметика

Погрешность измерения — это отклонение результата измерения от истинного значения измеряемой величины. Математически погрешность равна алгебраической разности измеренного значения и истинного значения величины.

Если это истинное значение x^* и результат измерения \tilde{x} — вещественные числа, то погрешностью является разность $\tilde{x} - x^*$.

Погрешность измерений — интервальная арифметика

Если истинное значение и результат измерения — интервалы x^* и \tilde{x} соответственно, то погрешность Δ определяется как *алгебраическая разность*

$$\Delta = \tilde{x} \ominus x^* \quad (8)$$

в полной интервальной арифметике Каухера.

Напомним, что обычное интервальное вычитание, которое обозначается традиционным знаком « $-$ » и является интервальным расширением вычитания, не является операцией, алгебраически обратной сложению и для нашей цели непригодно.

Формула (8) справедлива и в том случае, когда истинное значение величины x^* — точечное, а результат её измерения \tilde{x} интервальный. При этом в (8) полагаем $x^* = [x^*, x^*]$.

Расстояние на множестве интервалов.

Расстояние между интервалами \mathbf{a} и \mathbf{b} из \mathbb{IR} или \mathbb{KR} определяется как

$$\text{dist}(\mathbf{a}, \mathbf{b}) = \max\{|\underline{\mathbf{a}} - \underline{\mathbf{b}}|, |\overline{\mathbf{a}} - \overline{\mathbf{b}}|\}. \quad (9)$$

Оно обладает всеми свойствами абстрактного расстояния (метрики) и ещё некоторыми хорошими свойствами в связи с интервальными арифметическими операциями. Кроме того, легко убедиться, что

$$\text{dist}(\mathbf{a}, \mathbf{b}) = |\mathbf{a} \ominus \mathbf{b}|.$$

Эта формула является полным аналогом расстояния между точками вещественной оси, как модуля их разности, т. е. $|a - b|$.

Расстояние на множестве интервалов.

Рассмотрим интервал $[3, 5]$ и точку 3.6 внутри него. Расстояние от этой точки, отождествляемой с вырожденным интервалом $[3.6, 3.6]$, до данного интервала равно

$$\text{dist}(3.6, [3, 5]) = \max\{|3.6 - 3|, |3.6 - 5|\} = 1.4.$$

Рассмотрим дуальный интервал к интервалу $[3, 5]$. Это интервал $\text{dual}[3, 5] = [5, 3]$. Расстояние его до исходного интервала равно

$$\text{dist}([3, 5], [5, 3]) = 2.$$

Расстояние важно для определения отклонения интервалов друг от друга и, как следствие, для определения погрешности интервальных измерений.

Абсолютной погрешностью измерения назовём модуль (абсолютное значение) погрешности.

Для интервальных измерений абсолютная погрешность равна модулю интервала разности $\tilde{x} \ominus x$, и, как легко видеть, она равна расстоянию (9) между измеренным и истинным значениями величины.

Рассмотрим для примера ситуацию, когда истинное значение измеряемой величины, скажем, массы какого-то груза, является интервалом $[3, 4]$ кг, а её измерение дало интервал $[3, 5]$ кг. Тогда его погрешность равна

$$[3, 5] \text{ кг} \ominus [3, 4] \text{ кг} = [0, 1] \text{ кг}. \quad (10)$$

Пример

Если в результате измерения мы получим вещественное значение 3.8 кг, которое отождествляется с интервалом $[3.8, 3.8]$ кг, то его погрешность

$$[3.8, 3.8] \text{ кг} \ominus [3, 4] \text{ кг} = [0.8, -0.2] \text{ кг} \quad (11)$$

— неправильный интервал.

Может показаться, что он бессмыслен с физической точки зрения, но это поспешный вывод. Ситуация здесь совершенно аналогична, например, тому, как при измерении положительных физических величин (массы, плотности, давления и т. п.) мы получаем отрицательную погрешность, если измеренное значение приближает истинное значение снизу.

Абсолютная погрешность измерения равна 1 в случае (10) и 0.8 в случае (11).

Накрывающие и ненакрывающие измерения

Если результат измерения — точечная величина, то для неё возможны только два исхода проведения измерения: либо она получается равной истинному значению интересующей нас физической величины, либо не равной ей. Как говорят математики и программисты, исход измерения является «булевозначным», «да» или «нет».

При этом ясно, что в случае измерения непрерывных физических величин равенство является исключительным событием и почти никогда не достигается. Если же оно по каким-то причинам произошло, то является неустойчивым к сколь угодно малым возмущениям или же погрешностям в вычислительных алгоритмах.

Принципиально другая ситуация возникает, если результат измерения может быть интервалом.

Интервал по своей сути является двусторонней «вилкой» значений, и принадлежность ей истинного значения — это уже не исключительное событие. Оно, как правило, устойчиво к возмущениям и погрешностям обработки. Как следствие, для теории обработки интервальных данных фундаментальный характер имеют следующие определения:

Накрывающие и ненакрывающие измерения

Definition

Накрывающее измерение (накрывающий замер) — это интервальная оценка неизвестной истинной величины, гарантированно ее содержащая.

Измерение, не являющееся накрывающим, будем называть *ненакрывающим* (Рис. 5 и Рис. 6).

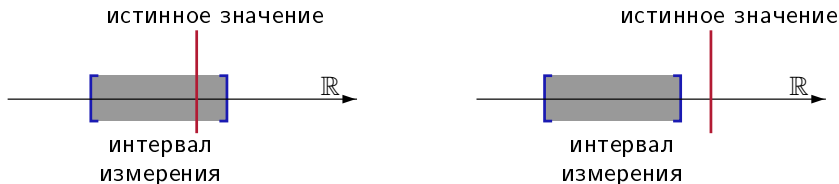


Рис.: Накрывающее (слева) и ненакрывающее (справа)

измерения точечного истинного значения некоторой физической величины.

Накрывающие и ненакрывающие выборки

Definition

Накрывающая выборка — совокупность накрывающих измерений, т. е. выборка, в которой все измерения (наблюдения) являются накрывающими. Напротив, выборка называется *ненакрывающей*, если хотя бы одно из входящих в неё измерений — ненакрывающее.

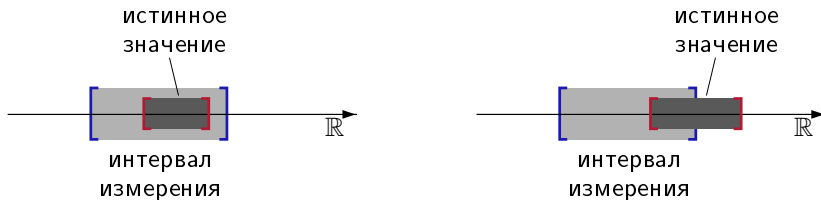


Рис.: Накрывающее (слева) и ненакрывающее (справа)

измерения интервального истинного значения некоторой физической величины.

Неформально говоря, *информационное множество* — это множество параметров задачи, которые совместны с данными измерений в рамках выбранной модели их обработки.

Аналогом «информационного множества» может отчасти служить понятие *доверительного интервала* оцениваемой случайной величины в традиционной вероятностной статистике.

В определение доверительного интервала входит дополнительный параметр — *уровень статистической значимости*, без которого понятие становится бессодержательным из-за неограниченности носителей большинства вероятностных распределений, но смысл доверительного интервала примерно соответствует «информационному множеству».

Далее для обозначения различных информационных множеств мы будем использовать прописную греческую букву

$$\Omega$$

(«омега»), добавляя к ней при необходимости параметры, обозначающие контекст задачи.

Так как информационное множество может быть достаточно произвольным множеством в пространстве параметров и не обязательно является интервалом, интервальным вектором или интервальной матрицей, мы не выделяем его символ жирным шрифтом.

Принцип соответствия в методологии науки — это утверждение, что любая новая научная теория должна включать старую теорию и её результаты как частный предельный случай.

Мы будем использовать принцип соответствия, как инструмент проверки «разумности» и адекватности наших конструкций, понятий и методов обработки данных с интервальными неопределённостями, который позволяет отсекаать заведомо «неразумные».

Выбросами или *промахами* в метрологии называются такие измерения, результаты которых не приносят информацию об исследуемом объекте в рамках его принятой модели.

Другое популярное определение выбросов или промахов состоит в том, что это результаты измерений, которые для данных условий резко отличаются от остальных результатов общей выборки.

Что считать выбросом (промахом) в случае интервальных результатов измерений? Прежде всего, не стоит связывать выбросы со свойством измерений быть накрывающими или ненакрывающими.

Более точно, из того, что интервальное измерение не является накрывающим, не следует, что оно представляет выброс или промах.

Отождествление выбросов (промахов) со свойством ненакрывания противоречит принципу соответствия, сформулированному в предыдущем параграфе.

В самом деле, при стремлении ширины интервальных измерений к нулю они переходят в точечные измерения, которые, как правило, всегда ненакрывающие. Тем не менее, различие для них выбросов (промахов) от этого не исчезает.

Измерение физической величины (константы).

Физическая величина взята в качестве примера. Данные могут быть любой природы: из наук о Земле, биологии, науках об обществе, экономики, etc.

Измерение физической величины — пример.

Проведём рассмотрение обработки данных физического эксперимента по измерению константы. В качестве источника данных будем использовать публикацию [2], представляющую результаты измерения циркулярной поляризации гамма-кванта в реакции захвата поляризованного нейтрона протоном.

Приведём часть данных таблицы 1 из публикации [2].

В таблице 1 основные данные измерения содержатся в столбцах Peak — средние значения и std Peak — оценки ошибки. В столбцах BG и std BG приведены данные, которые можно использовать для коррекции систематических ошибок. В первом столбце дан условный номер эксперимента.

Исходные данные. Величина $\delta \times 10^5$.

Номер замера	Peak	std Peak	BG	std BG
1	-4.4	2.7	4.2	6.7
2	-3.4	1.9	-3.2	4.8
3	-6.9	2.4	12.1	9
4	-1.2	2.4	12.4	7.2
5	-1.0	2.7	9.4	5.1
6	-10.8	3.5	1	12.4
7	-10.2	2.8	-0.6	6.1
8	-6.3	2	3.9	4.3
9	-10.4	4.1	10.3	10
10	0.6	3.4	-4.8	10.6
11	-1.8	2	4.6	4.2
12	-6.6	2.1	-5.7	4.6
13	-4.9	2.1	13	3
14	-6.0	2.4	8.4	4.6
15	-4.0	2.7	10.6	5.5

Таблица: Данные таблицы 1 для величины $\delta \times 10^5$ [2].

Представление данных.

В первую очередь представим данные таким образом, чтобы применить понятия статистики данных с интервальной неопределённостью.

Согласно терминологии интервального анализа, рассматриваемая выборка — это вектор интервалов, или интервальный вектор $\mathbf{x} = (x_1, x_2, \dots, x_n)$.

Для того, чтобы придать данным таблицы 1 необходимую форму, примем, что в качестве элементов \mathbf{x} будут выступать данные

$$\text{mid } x_k = \text{Peak}(k), \quad \text{rad} x_k = \text{std Peak}(k), \quad k = 1, 2, \dots, 15.$$

Для наглядного представления выборки часто рисуют образующие её интервалы в виде графика, изображённого на Рис. 11, который по статистической традиции мы будем называть *диаграммой рассеяния*.

Диаграмма рассеяния интервальных измерений.

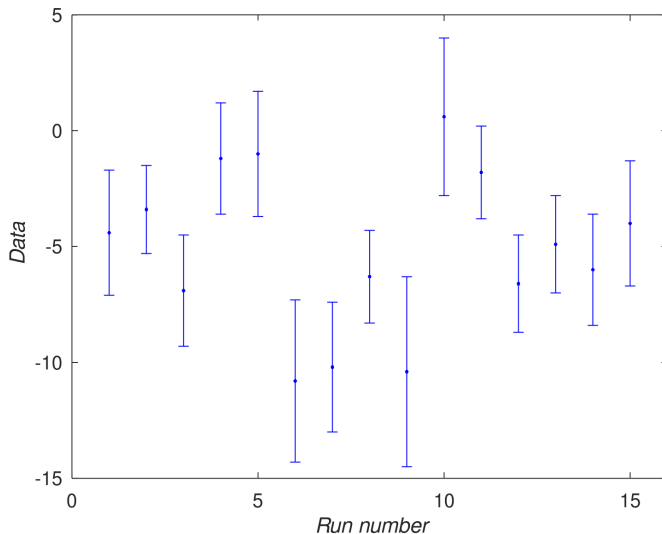


Рис.: Диаграмма рассеяния интервальных измерений [2].

Диаграмма рассеяния интервальных измерений.

Из таблицы 1 и Рис. 11 видно, что элементы выборки *неравноширинные*, поскольку величина неопределённости $\text{rad}x_k$ меняется в зависимости от измерения выборки, $k = 1, \dots, n$.

Информационное множество.

Информационным множеством в случае оценивания единичной физической величины по выборке интервальных данных будет также интервал, который называют *информационным интервалом*.

Неформально говоря, это интервал, содержащий значения оцениваемой величины, которые «совместны» с измерениями выборки («согласуются» с данными этих измерений).

Конкретный смысл, вкладываемый в понятия «совместные» или «согласующиеся», будет различен для разных ситуаций. В частности, он зависит от того, является ли выборка интервальных данных накрывающей или нет.

Важным внутренним свойством интервальной выборки, характеризующим согласование её данных между собой, является понятие совместности.

Definition

Выборка $\{x_k\}_{k=1}^n$ называется *совместной*, если пересечение всех интервалов составляющих её измерений непусто, т.е.

$$\bigcap_{1 \leq k \leq n} x_k \neq \emptyset.$$

В противном случае, если пересечение всех интервалов x_k , $k = 1, \dots, n$, является пустым, то выборка называется *несовместной*.

Свойство совместности характеризует саму выборку и, строго говоря, не связано напрямую с её свойством быть накрывающей выборкой, т. е. с включением ею истинного значения измеряемой величины.

Выборка может быть совместной, но ненакрывающей. Но если выборка накрывающая, то она обязана быть совместной.

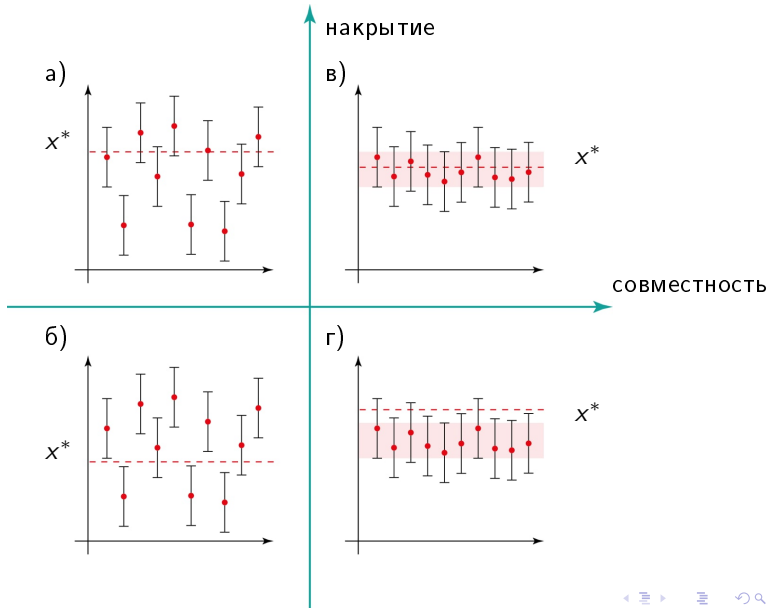
Эквивалентная формулировка этого свойства: если выборка несовместна, то она и ненакрывающая.

Основываясь на этих соображениях, в практической обработке результатов измерений трудный анализ накрытия выборкой истинного значения часто заменяют анализом её совместности, так как это удобнее и нагляднее (хотя и не вполне строго).

Если обрабатываемая выборка несовместна, то это может вызываться следующими причинами:

- (а) неверно заданным значением неопределённости измерений $\text{rad}x_k$ для каких-то $k \in \{1, 2, \dots, n\}$, которое занижено в сравнении с фактическим значением неопределённости;
- (б) наличием в этой выборке выбросов (промахов), т. е. сбойных измерений;
- (в) невыполнением условий на измеряемую физическую величину (её непостоянство и т. п.).

Диаграмма совместность—накрытие



Обработка накрывающей выборки

Если истинное значение величины содержится во всех интервалах измерений выборки $\{x_k\}_{k=1}^n$, то оно должно принадлежать также пересечению этих интервалов. Следовательно, уточнённым интервалом принадлежности истинного значения можно взять

$$I = \bigcap_{1 \leq k \leq n} x_k. \quad (12)$$

Это и будет информационный интервал I оценки измеряемой физической величины (см. Рис. 9). Явные выражения для его левой (нижней) и правой (верхней) границ даются следующими формулами:

$$\underline{I} = \max_{k=1, \dots, n} \underline{x}_k, \quad \bar{I} = \min_{k=1, \dots, n} \bar{x}_k. \quad (13)$$

Обработка накрывающей выборки

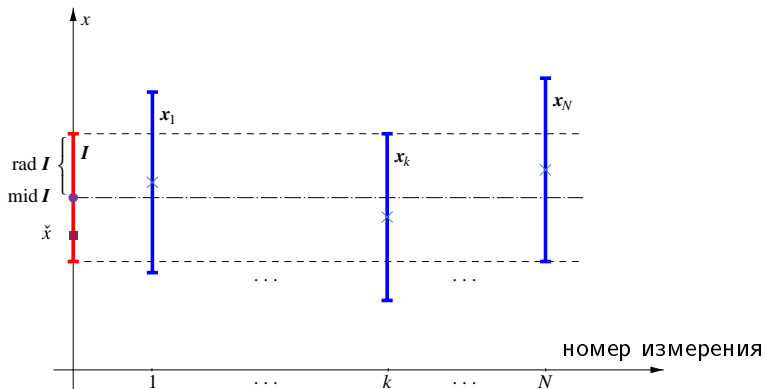


Рис.: Обработка накрывающей выборки интервальных измерений величины.

Предел совместности выборки

В силу сделанного допущения о том, что выборка покрывает истинное значение величины, имеем $\underline{I} \leq \bar{I}$.

При этом интересен предельный случай совместной выборки, когда

$$\underline{I} = \bar{I} = x^*.$$

Тогда выборка совместна, но мы, образно говоря, находимся на пределе её совместности, и информационный интервал I вырождается при этом в точку.

Уточнение априрным интервалом

Если известен некоторый априорный интервал возможных значений оцениваемой физической величины $I_{\text{апр}} = [L_{\text{апр}}, \bar{I}_{\text{апр}}]$, который должен гарантированно содержать её, то границы результирующего интервала (12) могут быть уточнены пересечением

$$I = I \cap I_{\text{апр}}. \quad (14)$$

Отметим, что априорный интервал $I_{\text{апр}}$ может задавать одностороннее ограничение, если он имеет вид $[L_{\text{апр}}, +\infty]$ или $[-\infty, \bar{I}_{\text{апр}}]$, т. е. является полубесконечным интервалом из арифметики Кахана.

На практике часто необходимо работать не с интервалами интересующей нас величины — (12) или (14), а с некоторой точечной оценкой \check{x} . Все точки информационного интервала вполне равноценны друг другу, так что эту точечную оценку \check{x} можно выбирать достаточно произвольно (см. Рис. 9). Тем не менее, имеет смысл взять из интервала некоторое точечное значение, которое представляет его наилучшим образом.

В качестве такой величины можно использовать, к примеру, его *центральную оценку* x_c ,

$$x_c = \text{mid } I = \frac{1}{2} (\underline{I} + \overline{I}). \quad (15)$$

Напомним, что середина интервала обладает определённой оптимальностью, являясь точкой, которая наименее удалёна от других точек этого интервала.

Обработка ненакрывающей выборки

Если выборка — ненакрывающая, так что некоторые из её измерений не содержат истинного значения измеряемой величины, то приведённые в предыдущем параграфе рассуждения и приёмы частично теряют свой смысл.

Поскольку кроме информации, представленной выборкой, в нашем распоряжении ничего нет, то следует бережно относиться ко всем измерениям и считать, что каждое из них несёт существенную информацию об измеряемой величине, которая не должна быть потеряна.

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

$$\bigcup_{1 \leq k \leq n} x_k. \quad (16)$$

Обработка ненакрывающей выборки

Это множество может не быть единым интервалом на вещественной оси (подобное часто случается, к примеру, если выборка несовместна). Разумно тогда воспользоваться вместо объединения обобщающей его операцией « \vee » (см. (3)), т.е. взятием максимума по включению, и вместо (16) взять информационный интервал в виде

$$\mathbf{J} = \bigvee_{1 \leq k \leq n} \mathbf{x}_k = \left[\min_{1 \leq k \leq n} \underline{\mathbf{x}}_k, \max_{1 \leq k \leq n} \bar{\mathbf{x}}_k \right]. \quad (17)$$

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

$$x_c = \text{mid } \mathbf{J} = \frac{1}{2} \left(\min_{1 \leq k \leq n} \underline{\mathbf{x}}_k + \max_{1 \leq k \leq n} \bar{\mathbf{x}}_k \right). \quad (18)$$

Обработка ненакрывающей выборки

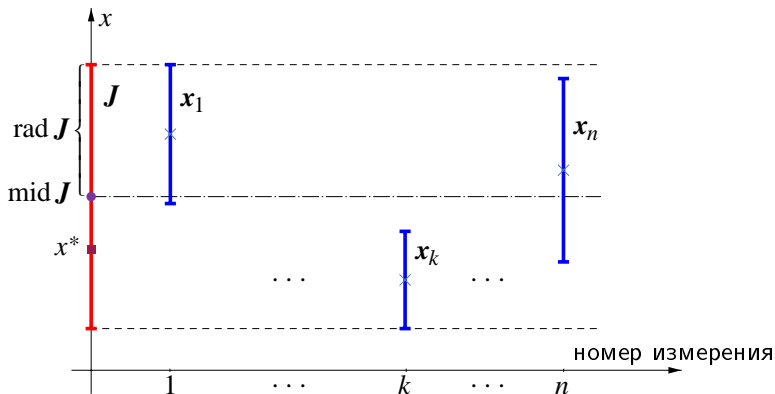


Рис.: Обработка ненакрывающей выборки интервальных измерений величины.

Уточнение априорным интервалом

Как и ранее, нам может быть известен некоторый априорный интервал возможных значений оцениваемой физической величины

$J_{\text{апр}} = [\underline{J}_{\text{апр}}, \overline{J}_{\text{апр}}]$, который должен гарантированно содержать её. Его могут задавать внешние физические (химические, биологические, экономические и т. п.) условия или ограничения.

Тогда границы результирующего интервала (17) могут быть уточнены пересечением

$$J = J \cap J_{\text{апр}}. \quad (19)$$

В данной ситуации это уточнение имеет даже бóльший смысл, чем в случае накрывающей выборки.

Взятие минимума по включению

Другой возможный сценарий обработки данных ненакрывающей выборки может состоять в том, что вместо пересечения интервальных измерений мы используем обобщающую её операцию « \wedge », т. е. взятие минимума всех интервальных результатов измерений относительно упорядочения по включению:

$$I = \bigwedge_{1 \leq k \leq n} x_k = \left[\max_{1 \leq k \leq n} \underline{x}_k, \min_{1 \leq k \leq n} \bar{x}_k \right]. \quad (20)$$

Здесь по существу требуется использование полной интервальной арифметики Каухера, так как интервал (20) может оказаться неправильным.

Точечная оценка ненакрывающей выборки

Соответственно, точечной оценкой измеряемой величины целесообразно взять

$$x_c = \text{mid } I = \frac{1}{2} \left(\max_{1 \leq k \leq n} \underline{x}_k + \min_{1 \leq k \leq n} \bar{x}_k \right), \quad (21)$$

т. е. середину интервала, который получается как минимум по включению всех интервалов выборки (см. (3)).

Если выборка совместна, то (21) совпадает с (15). Если же выборка несовместна, то результатом (20) является неправильный интервал I , $\text{rad } I < 0$. Соответственно, информационное множество результатов измерений по обрабатываемой выборке пусто.

Но даже когда интервал (20) неправилен, его середина (21) — это точка, обладающая определёнными условиями оптимальности. Она первой появляется в непустом пересечении интервалов выборки, если мы станем равномерно уширять их, увеличивая неопределённость измерений.

В самом деле, пусть радиусы всех интервалов выборки увеличились на s , $s \geq 0$, тогда как середины остались неизменными. Вместо радиусов $\text{rad} \mathbf{x}_k$ мы получили $\text{rad} \mathbf{x}_k + s$, $k = 1, 2, \dots, n$. Кроме того, все нижние концы интервальных измерений стали теперь $\underline{\mathbf{x}}_k - s$, а верхние концы — $\bar{\mathbf{x}}_k + s$, $k = 1, 2, \dots, n$.

Как следствие, $\max_{1 \leq k \leq n} \underline{\mathbf{x}}_k$ уменьшается на s , а $\min_{1 \leq k \leq n} \bar{\mathbf{x}}_k$ увеличивается на s , а радиус получающегося интервала (20) теперь равен $\text{rad} \mathbf{I} + s$.

Как следствие, $\max_{1 \leq k \leq n} \underline{x}_k$ уменьшается на s , а $\min_{1 \leq k \leq n} \bar{x}_k$ увеличивается на s , а радиус получающегося интервала (20) теперь равен $\text{rad}I + s$.

Поэтому, если взять s таким, чтобы $s \geq |\text{rad}I|$, то получившийся интервал станет правильным, и точка x_c будет лежать в нём.

Можно также сказать, что в точке (21) минимизируется равномерное уширение интервалов данных рассматриваемой выборки, необходимое для достижения её совместности.

«Средняя» оценка ненакрывающей выборки

Наконец, если выборка интервальных измерений — ненакрывающая, то иногда имеет смысл взять среднее арифметическое образующих её интервалов, т. е.

$$K = \frac{1}{n} \sum_{k=1}^n x_k.$$

Его середина может служить точечной оценкой измеряемой величины.

Нетрудно убедиться в том, что все три рассмотренных выше приёма обработки ненакрывающей выборки при стремлении ширины интервальных данных к нулю переходят в осмысленные методы оценивания физической величины по точечным данным.

В частности, она полагается равной среднему арифметическому измерений выборки в третьем случае. То есть, эти методы удовлетворяют «принципу соответствия».

Пример выборки данных [2].

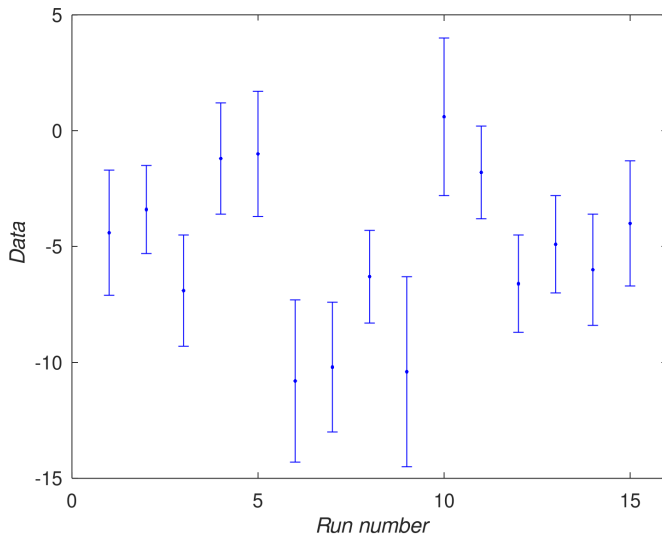


Рис.: Диаграмма рассеяния интервальных измерений [2].

Пример данных [2].

Информация, представленная выборкой Табл. 1, уникальна, так что следует бережно относиться ко всем измерениям и считать, что каждое из них несёт существенную информацию об измеряемой величине, которая не должна быть потеряна.

Попробуем взять в качестве информационного множества для истинного значения величины объединение всех интервалов выборки, т. е.

$$I_{Uni} = \bigcup_{1 \leq k \leq n} x_k = [-14.5, 4.0]. \quad (22)$$

Пример данных [2].

По существу измеряемая величина является константой неизвестного, но определённого знака. Оценка (16) в данном случае имеет разные знаки концов интервалов и противоречит постановке задачи.

Можно было бы отбросить элементов выборки, имеющие «неправильный» знак, но это представляется недопустимым произволом.

Вместе с тем, середина интервала (16)

$$\text{mid } I_{Uni} = -5.25$$

может быть разумной точечной оценкой, и её будет полезно сравнить с оценками, полученными на основе других подходов.

Пример данных [2].

Продemonстрируем наглядно, что получается в конкретном случае. Будем представлять теперь данные в несколько ином виде, чем на рисунке 11, откладывая номер измерения по вертикальной шкале.

При этом мы будем действовать согласовано с представлением подобных результатов при обработке данных на ресурсе С.И.Жилина [3].

Вычисления проводились в среде Octave в классической интервальной арифметике с использованием стандартной библиотеки `interval` и полной интервальной арифметики с использованием библиотеки `kinterval` [4].

Пример данных [2].

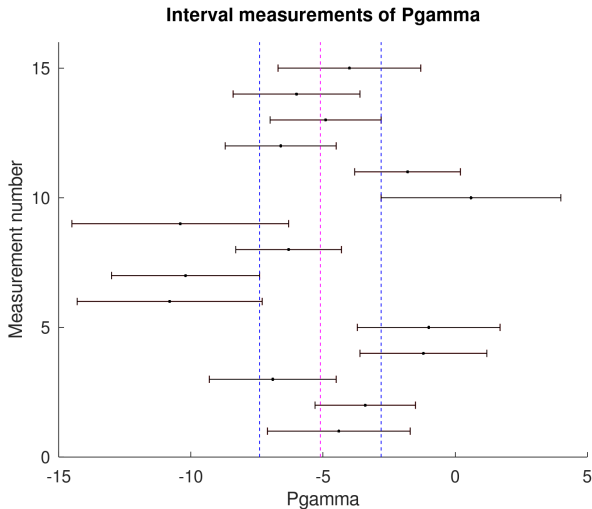


Рис.: Диаграмма рассеяния интервальных измерений величины, полоса минимума по включению (20) и точечная оценка (21).

Пример данных [2].

На Рис. 12 синими вертикальными линиями показаны границы информационного множества, полученные по формуле (20)

$$I = \bigwedge_{1 \leq k \leq n} x_k = \left[\max_{1 \leq k \leq n} \underline{x}_k, \min_{1 \leq k \leq n} \bar{x}_k \right] = [-2.8, -7.4].$$

Также вычислим точечную оценку измеряемой величины по формуле (21)

$$x_c = \text{mid } I = \frac{1}{2} \left(\max_{1 \leq k \leq n} \underline{x}_k + \min_{1 \leq k \leq n} \bar{x}_k \right) = -5.1.$$

На Рис. 12 эта величина показана вертикальной линией цветом magenda. Интервал I — неправильный. Смысл значения x_c прояснён в комментарии после формулы (21) как точки, которая первой появляется в непустом пересечении интервалов выборки, если мы станем равномерно ушивать их.

Пример данных [2].

Наконец, если выборка интервальных измерений — ненакрывающая, то иногда имеет смысл взять среднее арифметическое образующих её интервалов, т. е.

$$\mathbf{K} = \frac{1}{n} \sum_{k=1}^n \mathbf{x}_k = [-7.77, -2.54]. \quad (23)$$

Середина этого интервала

$$\text{mid } \mathbf{K} = -5.15$$

также может служить точечной оценкой измеряемой величины.

Вариабельность оценки — радиус

Рассмотрим теперь характеристики разброса оценок физической величины, полученных по интервальной выборке. Её наиболее естественной мерой, если информационный интервал непуст, является его *радиус* ϱ , т. е.

$$\varrho = \text{rad}I = \frac{1}{2} (\overline{I} - \underline{I}).$$

Фактически, это максимальное отклонение границ информационного интервала от центральной оценки.

При анализе данных имеет также смысл знать отклонения точечных или интервальных измерений выборки от итоговой точечной оценки. Они дают возможность судить о степени разброса измерений относительно полученной оценки, что помогает при анализе «качества» выборки и выявлении выбросов.

Отклонения Δ_k для первичных интервальных измерений рассчитываются как

$$\Delta_k = \text{dist}(\mathbf{x}_k, x_c), \quad k = 1, \dots, n. \quad (24)$$

В некоторых случаях имеет смысл отсчитывать отклонения от базовых точечных измерений, вокруг которых строятся далее интервальные результаты, т. е. рассматривать в качестве отклонений результатов отдельных измерений величины

$$\Delta_k = |\dot{x}_k - x_c|, \quad k = 1, \dots, n. \quad (25)$$

Норма вектора $\Delta = (\Delta_1, \dots, \Delta_n)$ может служить аналогом выборочной дисперсии оценки из традиционной вероятностной статистики.

Приём варьирования неопределённости

Выше мы видели, что величина реальной неопределённости измерения, т. е. радиуса интервала измерения, определяется не просто и подчас неоднозначно. С другой стороны, он сильно влияет на свойства как отдельного измерения, так и выборки интервальных измерений. Совместность выборки и свойство накрытия истинного значения существенно зависят от правильно назначенной величины неопределённости — радиуса интервальных измерений. Наконец, если некоторое Δ является величиной неопределённости интервального измерения или выборки, то и любое Δ' , удовлетворяющее $\Delta' \geq \Delta$, также может служить величиной неопределённости.

Сказанное выше приводит к мысли о том, что при обработке интервальных данных величиной неопределённости можно управлять, виртуально варьируя её, с целью исследования интервальных измерений, их выборок и построения оценок с нужными свойствами.

Приём варьирования неопределённости

Если выборка интервальных измерений несовместна, то, увеличивая одновременно величину неопределённости всех измерений, мы всегда сможем добиться того, чтобы выборка сделалась совместной, т. е. чтобы пересечение интервалов стало непустым, а интервал минимума по включению (20) — правильным.

Кроме того, точка (или точки), которая первой появляется в непустом пересечении интервалов при расширении интервальных измерений, и тем самым требует наименьшего увеличения неопределённости измерений для достижения совместности выборки, является «наименее несовместной». Её разумно брать в качестве оценки величины (или оценки параметров зависимости).

Приём варьирования неопределённости

В конкретной ситуации данных [2], измерения выборки являются существенно неравноширинными. Одновременное изменение величины неопределённости для всех измерений на одно и то же значение может оказаться неразумным.

Пусть задан некоторый положительный весовой вектор $w = (w_1, w_2, \dots, w_n)$, $w_k > 0$, размерность которого равна длине исследуемой выборки, причём изменение величины неопределённости k -го измерения — $\text{rad}x_k$, должно быть пропорциональным w_k , т. е. для любых k и l справедливо

$$\frac{\text{изменение } \text{rad}x_k}{\text{изменение } \text{rad}x_l} = \frac{w_k}{w_l}.$$

Приём варьирования неопределённости

Идея варьирования величины неопределённости интервальных измерений оформилась в 80-е годы XX века (Н.М. Оскорбин [5] и др.), и далее неоднократно переоткрывалась различными исследователями.

Применительно к данным таблицы 1, применение методики приведено на Рис. 13.

Красным цветом даны исходные данные таблицы 1, а чёрным цветом — «расширенные» интервалы данных при выбранном коэффициенте расширения.

Приём варьирования неопределённости

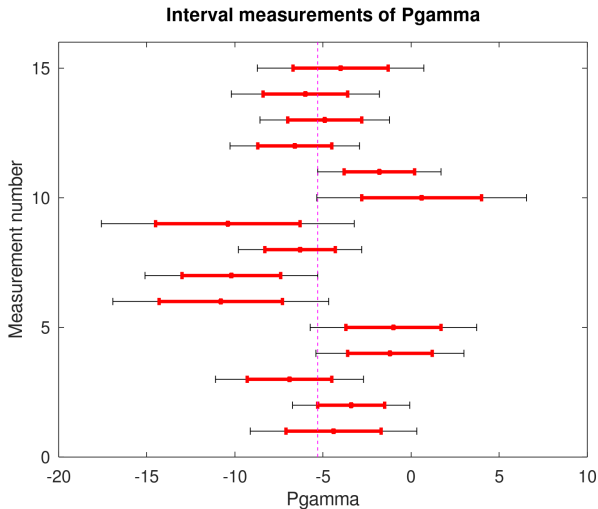


Рис.: Графическое представление интервальных данных и результаты обработки по методике [5].

Приём варьирования неопределённости

Вычисления проведены по методике [5] и с использованием кода С.И.Жилина [3]. При этом решается задача линейного программирования, в ходе которой вычисляются 2 параметра: оптимальное положение «центра неопределенности» `oskorbin_center` и коэффициент расширения радиусов замеров.

$$x_{MM} = \text{oskorbin_center} = -5.30, \quad k = 1.75.$$

Здесь в индексе x_{MM} , MM соответствует Minimal Module, функции оптимизации задачи линейного программирования.

Информационное множество представляет точку

$$I_{MM} = \bigcap_{1 \leq k \leq n} x_k = x_{MM}.$$

Содержательным результатом вычислений является уточнение положения наиболее вероятной точечной оценки физической величины [2] и вычисление дополнительной погрешности для каждого элемента выборки, необходимой для достижения совместности данных.

Следует заметить, что значение x_{MM} , полученное варьированием неопределённости, ненамного отличается от полученных ранее оценок.

Это свидетельствует в пользу того, что выборка данных таблицы 1 не обладает какими-то патологическими свойствами. При этом для данных требуется увеличение неопределённости. Таким образом, можно говорить о наличии систематических погрешностей.

Definition

Модой интервальной выборки назовём интервал пересечения её наибольшей совместной подвыборки.

Вход

Интервальная выборка $X = \{x_i\}_{i=1}^n$.

Выход

Мода $\text{mode } X$ выборки X и её ранг μ .

Алгоритм

$I = \bigcap_{i=1}^n x_i$;

IF $I \neq \emptyset$ THEN

$\text{mode } X = I$;

$\mu = n$

ELSE

 объединяем все концы $\underline{x}_1, \bar{x}_1, \underline{x}_2, \bar{x}_2, \dots, \underline{x}_n, \bar{x}_n$

 интервалов выборки в одно множество $C = \{c_i\}_{i=1}^{2n}$;

 упорядочиваем элементы C по возрастанию значений;

 порождаем интервалы $c_i = [c_i, c_{i+1}]$, $i = 1, 2, \dots, 2n - 1$;

 для каждого c_i подсчитываем число μ_i интервалов

 из выборки X , имеющих непустое пересечение с c_i ;

 выбираем из всех c_i интервалы с максимальным

 значением μ_i , т.е. такие c_i , что $\mu_i = \max_i \mu_i$;

$\text{mode } X = \bigcup_i c_i$

$\mu = \mu_i$

END IF

Рис.: Алгоритм для нахождения моды интервальной выборки.

Диаграмма рассеяния интервальных измерений.

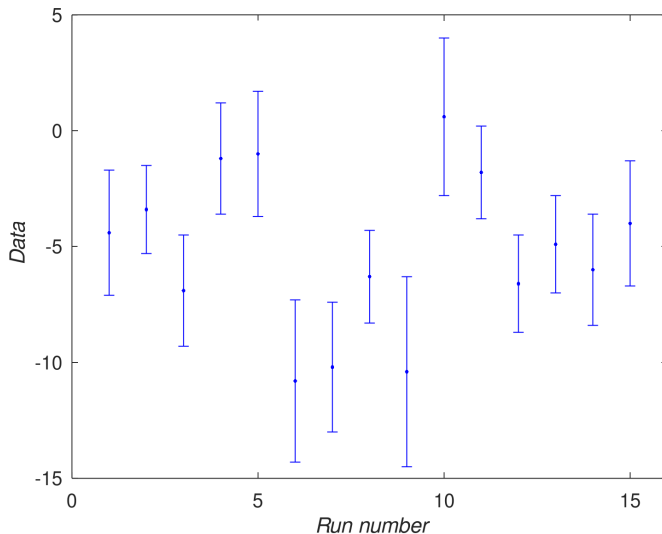


Рис.: Диаграмма рассеяния интервальных измерений [2].

Массив подинтервалов.

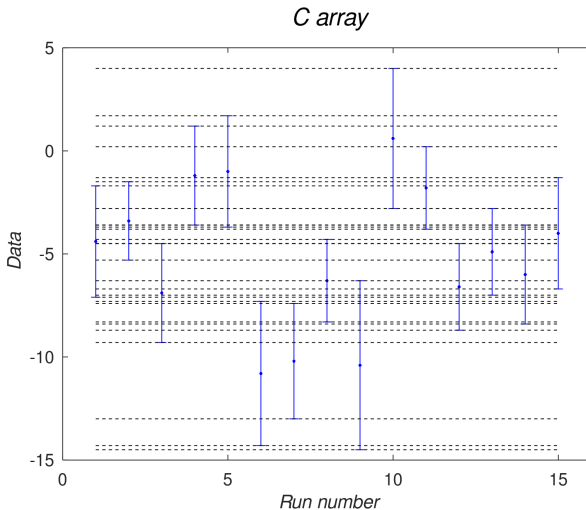


Рис.: Массив *c*.

Массив подинтервалов.

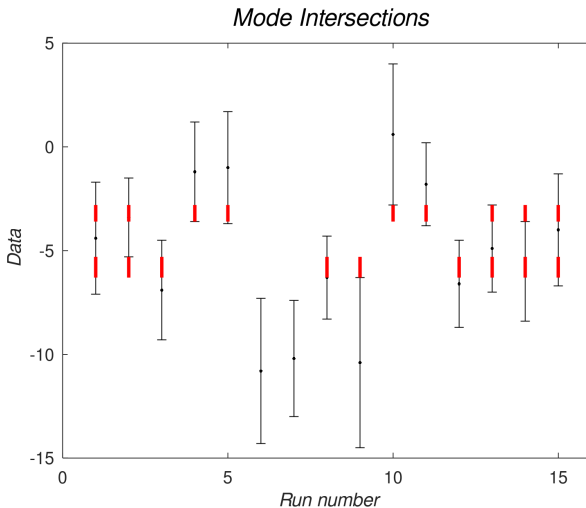


Рис.: Массив c .

Частоты пересечений

Частоты пересечений подинтервалов с исходными интервалами

$$\mu = \{2, 3, 4, 5, 6, 7, 7, 7, 7, 7, 8, 8, \underline{9}, 8, 8, 7, 7, 7, 8, 8, \underline{9}, 8, 8, 7, 6, 5, 4, 3, 2\}$$

Максимум пересечений имеет множество подинтервалов

$$\iota = \{13, 21\}$$

Ранг = 9

Мода — мультиинтервал

$$\text{mode } \mathbf{X} = \bigcup_{\iota} \mathbf{c}_{\iota} = \{ [-6.3001, -5.2999]; \quad [-3.6001, -2.8] \}$$

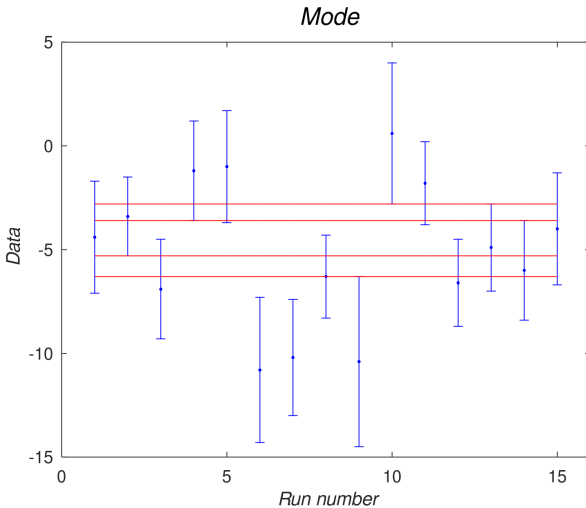


Рис.: Интервальная мода выборки данных таблицы 1.

Диаграмма рассеяния интервальных измерений.

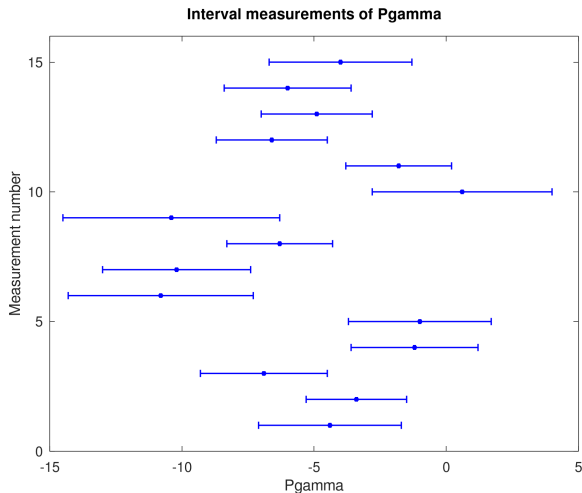
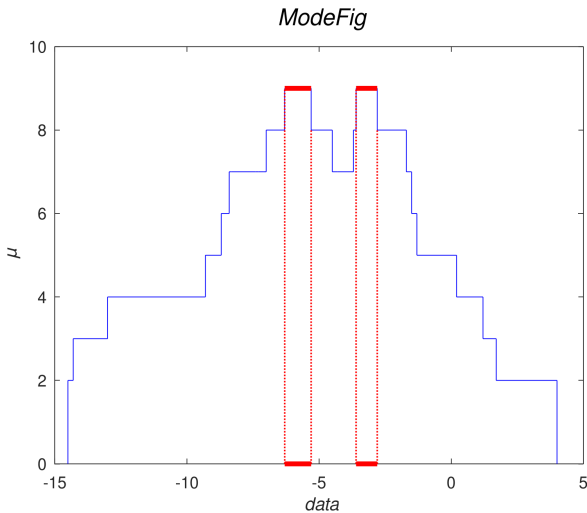


Рис.: Диаграмма рассеяния интервальных измерений [2].

График частоты пересечений подинтервалов с исходными интервалами.



Данные и мода.

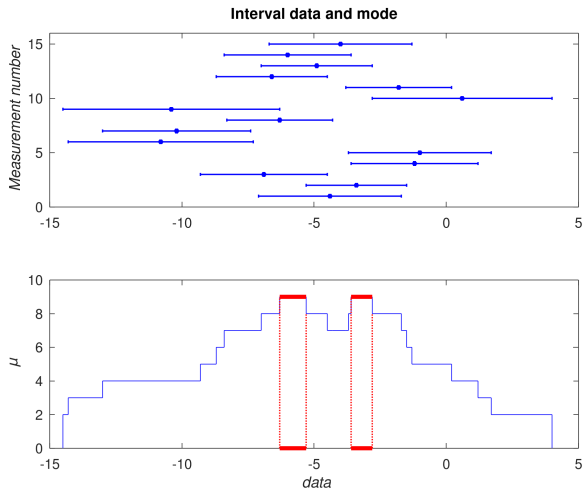


Рис.: Данные и мода.

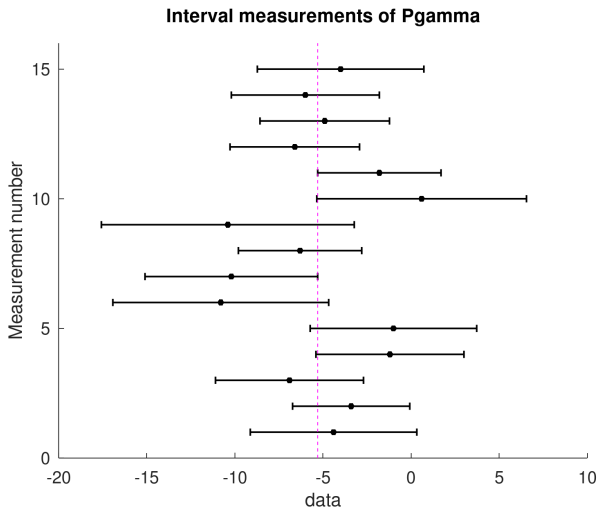


Рис.: Регуляризованные данные.

Частоты пересечений

Частоты пересечений подинтервалов с исходными интервалами

$$\mu = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 14, 15, 14, 14, 13, 12, 11, 10, \\ 9, 8, 7, 6, 5, 4, 3, 2\}$$

Максимум пересечений имеет подинтервал

$$\iota = 15$$

Ранг = 15

Мода — точка

$$\text{mode } \mathbf{X} = \bigcup_{\iota} \mathbf{c}_{\iota} = -5.30$$

Мода выборки с регуляризованными данными.

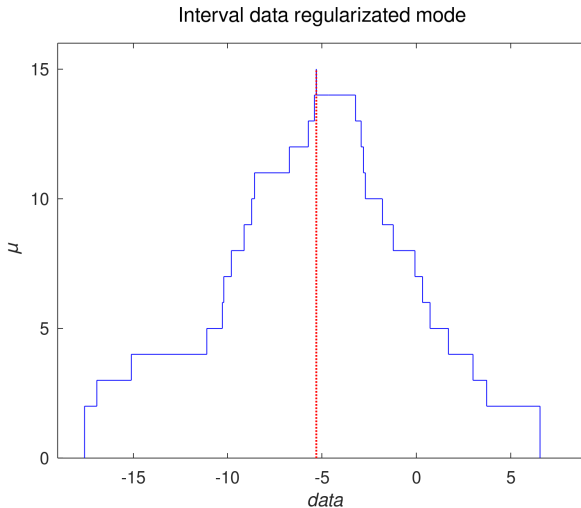


Рис.: Мода выборки с регуляризованными данными.

Регуляризованные данные и мода.

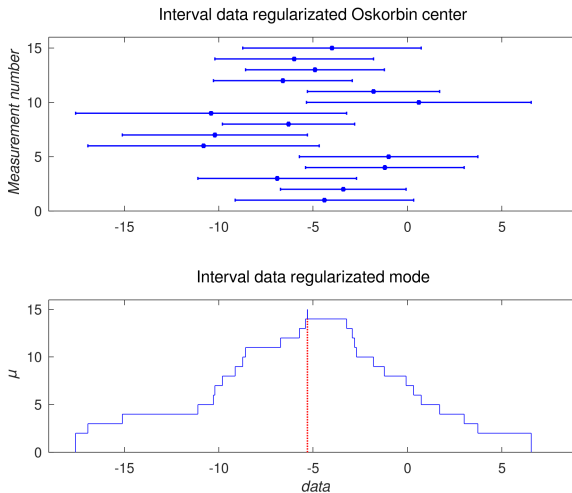
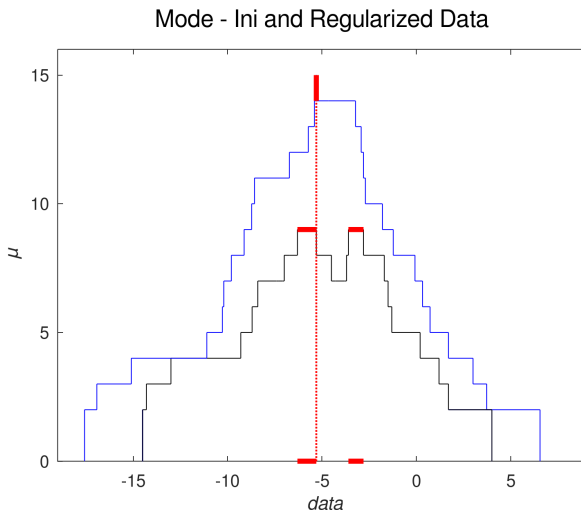







Рис.: Регуляризованные данные и мода.

Моды выборки с исходными и регуляризованными данными.



Материалы 2022-2023

- Обобщение мер совместности в анализе данных
<https://elib.spbstu.ru/dl/5/tr/2022/tr22-142.pdf/info>
- Примеры обработки измерений постоянной величины в анализе данных с интервальной неопределённостью
<https://elib.spbstu.ru/dl/5/tr/2023/tr23-29.pdf/info>

-  А.Н. БАЖЕНОВ, С.И. ЖИЛИН, С.И. КУМКОВ, С.П. ШАРЫЙ. Обработка и анализ данных с интервальной неопределённостью. РХД. Серия «Интервальный анализ и его приложения». Ижевск. 2021. с.200.
-  V.M.LOBASHEV ET AL, Circular polarization of γ -quanta in the $np \rightarrow d\gamma$ reactions with polarized neutrons. Physics Letters B, Volume 289, Issues 1–2, 3 September 1992, Pages 17-21.
-  С.И.ЖИЛИН. Примеры анализа интервальных данных в Octave <https://github.com/szhilin/octave-interval-examples>
-  С.И.ЖИЛИН. Библиотека полной интервальной арифметики `kinterval` в среде Octave. Частное сообщение.
-  ОСКОРБИН Н.М. Некоторые задачи обработки информации в управляемых системах // Синтез и проектирование многоуровневых иерархических систем. Материалы конференции. – Барнаул: Алтайский государственный университет, 1983.