

Тема X-2. Обработка и анализ данных с интервальной неопределённостью.

А.Н. Баженов

ФТИ им. А.Ф.Иоффе

a_bazhenov@inbox.ru

17.03.2022

Общий план

- Общие понятия
- **Обработка константы (физической величины)**
- Задача восстановления зависимостей

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. 2021. с.200.

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. 1 и Рис. 2).

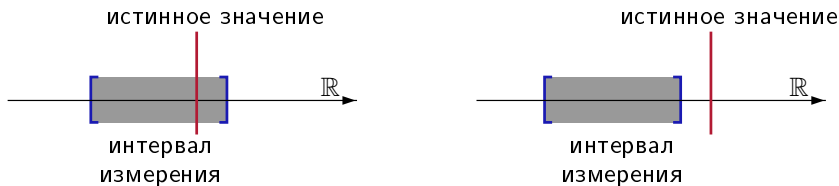


Рис.: Накрывающее (слева) и ненакрывающее (справа)

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.

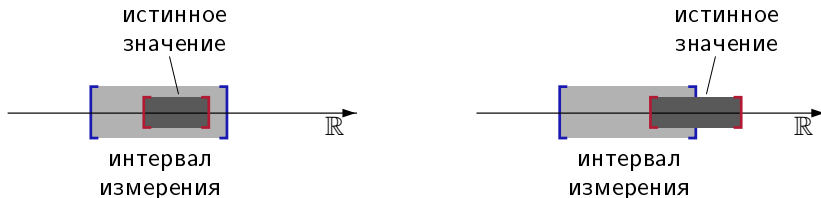


Рис.: Накрывающее (слева) и ненакрывающее (справа)

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.

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
1	-4.4	2.7
2	-3.4	1.9
3	-6.9	2.4
4	-1.2	2.4
5	-1.0	2.7
6	-10.8	3.5
7	-10.2	2.8
8	-6.3	2
9	-10.4	4.1
10	0.6	3.4
11	-1.8	2
12	-6.6	2.1
13	-4.9	2.1
14	-6.0	2.4
15	-4.0	2.7

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. 7, который по статистической традиции мы будем называть *диаграммой рассеяния*.

Диаграмма рассеяния интервальных измерений.

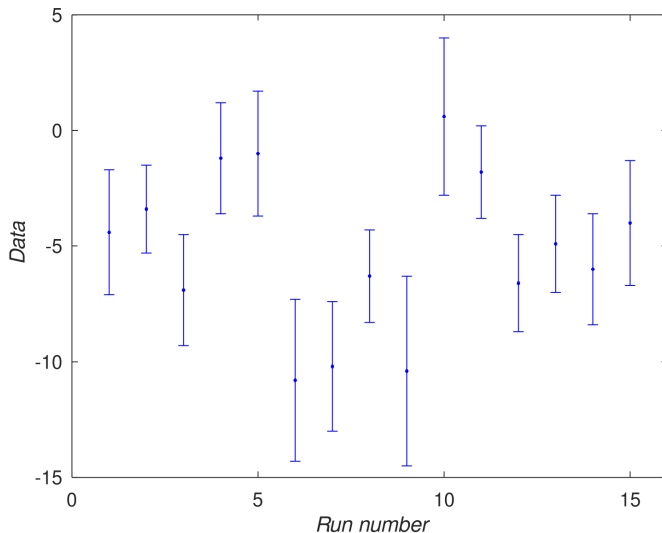


Рис.: Диаграмма рассеяния интервальных измерений [2]

Диаграмма рассеяния интервальных измерений.

Из таблицы 1 и Рис. 7 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.).

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

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

Это и будет информационный интервал I оценки измеряемой физической величины (см. Рис. 4). Явные выражения для его левой (нижней) и правой (верхней) границ даются следующими формулами:

$$\underline{I} = \max_{k=1, \dots, n} \underline{x}_k, \quad \bar{I} = \min_{k=1, \dots, n} \bar{x}_k. \quad (2)$$

Обработка накрывающей выборки

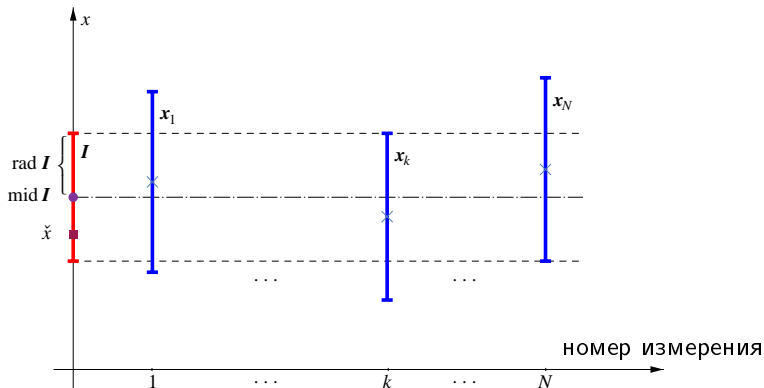


Рис.: Обработка накрывающей выборки интервальных измерений величины.

Пример - обработка накрывающей выборки

Выберем из данных Табл. 1 накрывающую подвыборку. Эти данные приводятся в Табл. 2.

Номер замера	Peak	std Peak
1	-4.4	2.7
2	-3.4	1.9
3	-6.9	2.4
8	-6.3	2
12	-6.6	2.1
13	-4.9	2.1
14	-6.0	2.4
15	-4.0	2.7

Таблица: Накрывающая подвыборка данных таблицы 1.

Пример - обработка накрывающей выборки

Диаграмма рассеяния выборки Табл. 2 приводятся на Рис. 5. Также на этом рисунке приведены оценки границы информационного множества (2).

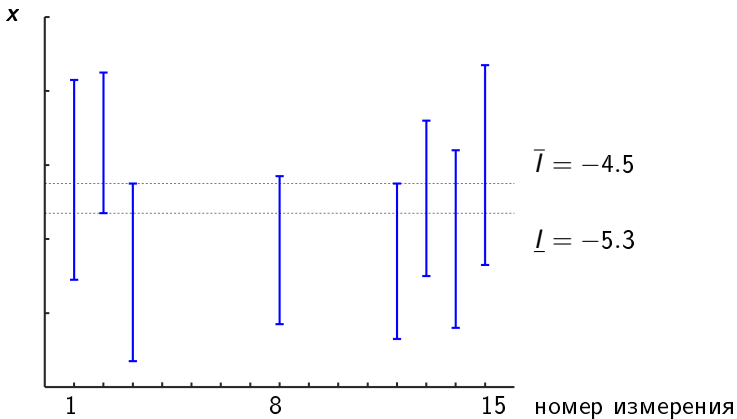


Рис.: Пример обработки накрывающей выборки интервальных измерений

Пример - обработка накрывающей выборки

Численно оценки границ информационного множества (2) в единицах 10^{-5} составляют

$$\underline{I} = \max_{1 \leq k \leq n} \underline{x}_k = -5.3,$$

$$\bar{I} = \min_{1 \leq k \leq n} \bar{x}_k = -4.5.$$

Центральная оценка (4) в единицах 10^{-5} равна

$$x_c = \text{mid } I = \frac{1}{2} (\underline{I} + \bar{I}) = -4.9.$$

В дальнейшем мы сможем сравнить полученные оценки с другими способами оценивания по полной ненакрывающей выборке Табл. 1.

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 (1) могут быть уточнены пересечением

$$I = I \cap I_{\text{апр}}. \quad (3)$$

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 — (1) или (3), 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. 4). 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 ,

$$x_c = \text{mid } I = \frac{1}{2} (\underline{I} + \overline{I}). \quad (4)$$

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

$$\bigcup_{1 \leq k \leq n} x_k. \quad (5)$$

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. (??)), т. е. взятием максимума по включению, и вместо (5) взять информационный интервал в виде

$$\mathbf{J} = \bigvee_{1 \leq k \leq n} \mathbf{x}_k = \left[\min_{1 \leq k \leq n} \underline{\mathbf{x}}_k, \max_{1 \leq k \leq n} \bar{\mathbf{x}}_k \right]. \quad (6)$$

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

$$x_c = \text{mid } \mathbf{J} = \frac{1}{2} \left(\min_{1 \leq k \leq n} \underline{\mathbf{x}}_k + \max_{1 \leq k \leq n} \bar{\mathbf{x}}_k \right). \quad (7)$$

Обработка ненакрывающей выборки

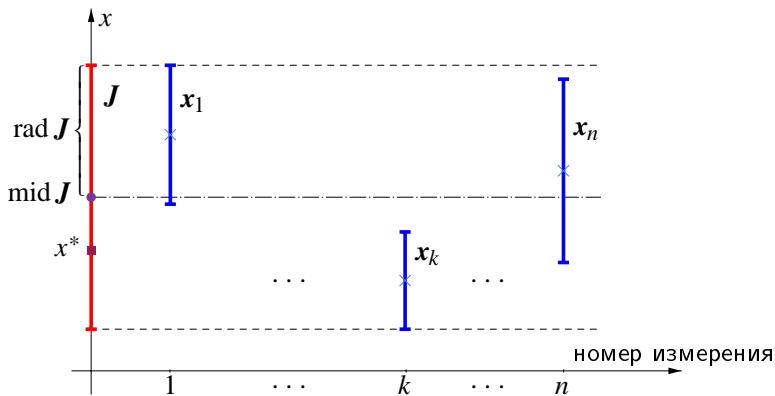


Рис.: Обработка ненакрывающей выборки интервальных измерений величины.

Уточнение априорным интервалом

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

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

Тогда границы результирующего интервала (6) могут быть уточнены пересечением

$$J = J \cap J_{\text{апр}}. \quad (8)$$

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:

$$I = \bigwedge_{1 \leq k \leq n} x_k = \left[\max_{1 \leq k \leq n} \underline{x}_k, \min_{1 \leq k \leq n} \bar{x}_k \right]. \quad (9)$$

Здесь по существу требуется использование полной интервальной арифметики Каухера, так как интервал (9) может оказаться неправильным.

Точечная оценка ненакрывающей выборки

Соответственно, точечной оценкой измеряемой величины целесообразно взять

$$x_c = \text{mid } I = \frac{1}{2} \left(\max_{1 \leq k \leq n} \underline{x}_k + \min_{1 \leq k \leq n} \bar{x}_k \right), \quad (10)$$

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

Если выборка совместна, то (10) совпадает с (4). Если же выборка несовместна, то результатом (9) является неправильный интервал I , $\text{rad } I < 0$. Соответственно, информационное множество результатов измерений по обрабатываемой выборке пусто.

Но даже когда интервал (9) неправилен, его середина (10) — 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 (9) 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 (9) теперь равен $\text{rad} + s$.

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

Можно также сказать, что в точке (10) 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].

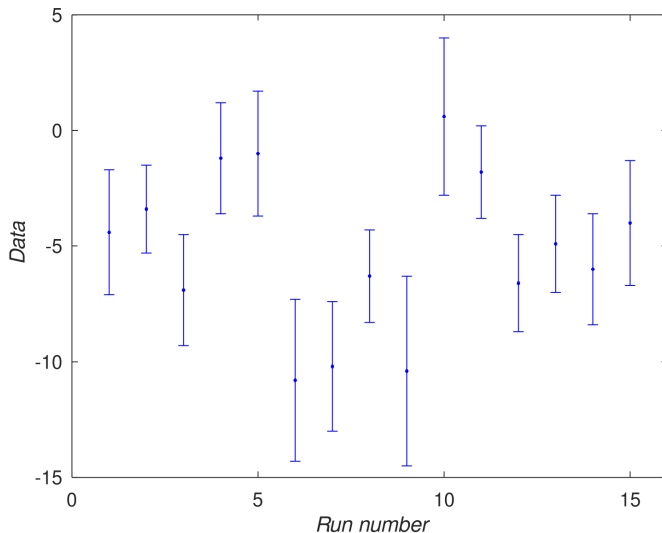


Рис.: Диаграмма рассеяния интервальных измерений [2]

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.

$$I_{Uni} = \bigcup_{1 \leq k \leq n} x_k = [-14.5, 4.0]. \quad (11)$$

Пример данных [2].

По существу измеряемая величина является константой неизвестного, но определённого знака. Оценка (5) 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 (5)

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

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 7, 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].

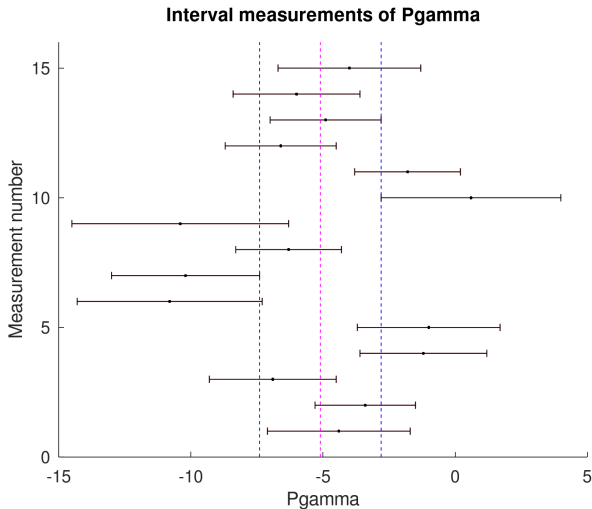


Рис.: Диаграмма рассеяния интервальных измерений величины, полоса минимума по включению (9) и точечная оценка (10).

Пример данных [2].

На Рис. 8 синими вертикальными линиями показаны границы информационного множества, полученные по формуле (9)

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

Также вычислим точечную оценку измеряемой величины по формуле (10)

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

На Рис. 8 эта величина показана вертикальной линией цветом magenda. Интервал I — неправильный. Смысл значения x_c прояснён в комментарии после формулы (10) 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.

$$\mathbf{K} = \frac{1}{n} \sum_{k=1}^n \mathbf{x}_k = [-7.77, -2.54]. \quad (12)$$

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.

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

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

$$\Delta_k = \text{dist}(\mathbf{x}_k, x_c), \quad k = 1, \dots, n. \quad (13)$$

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

$$\Delta_k = |\dot{x}_k - x_c|, \quad k = 1, \dots, n. \quad (14)$$

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 (9) — 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

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

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. 9.

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

Приём варьирования неопределённости

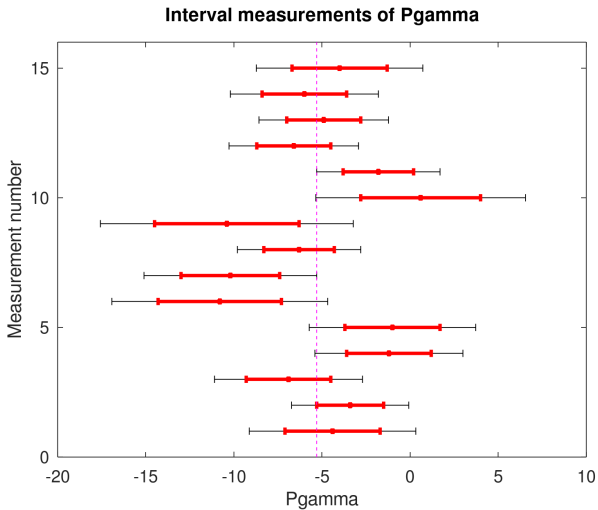


Рис.: Графическое представление интервальных данных и результаты обработки по методике [5].

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

$$I_{MM} = \bigcap_{1 \leq k \leq n} \mathbf{x}_k = x_{MM}.$$

Приём варьирования неопределённости

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 $\underline{x}_1, \overline{x}_1, \underline{x}_2, \overline{x}_2, \dots, \underline{x}_n, \overline{x}_n$

 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_k , что $\mu_k = \max_i \mu_i$;

$\text{mode } X = \bigcup_k c_k$

$\mu = \mu_k$

END IF

Рис.: Алгоритм для нахождения моды интервальной выборки.

Пример вычисления моды интервальной выборки.

Очень простой пример вычисления моды интервальной выборки.

Пример вычисления моды интервальной выборки.

Рассмотрим пример вычисления моды интервальной выборки.
Пусть имеется интервальная выборка из 4 элементов

$$X = \{ [1, 4], [5, 9], [1.5, 4.5], [6, 9] \}. \quad (15)$$

Пример вычисления моды интервальной выборки.

Диаграмма рассеяния выборки X приведена на рисунке 11

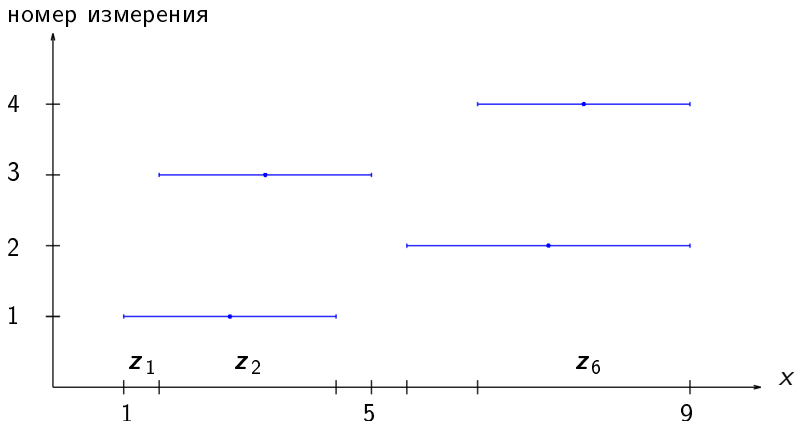


Рис.: Диаграмма рассеяния интервальной выборки (15) и элементы выборки z .

Пример вычисления моды интервальной выборки.

В соответствии с алгоритмом ??, проверим совместность **X**.
Пересечение элементов выборки пусто

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

Таким образом, необходимо выполнить шаги алгоритма после ключевого слова ELSE.

Сформируем массив интервалов **z** из концов интервалов **X**

$$z = \{ [1, 1.5], [1.5, 4], [4, 4.5], [4.5, 5], [5, 6], [6, 9] \}. \quad (16)$$

Мощность N массива **z** равна 6, и она меньше $2n - 1 = 7$ из-за совпадения правых концов интервалов $[5, 9]$ и $[6, 9]$.

Пример вычисления моды интервальной выборки.

Для каждого интервала z_i подсчитываем число μ_i интервалов из выборки \mathbf{X} , включающих z_i , получаем массив μ_i в виде

$$\{1, 2, 1, 0, 1, 2, 1\}. \quad (17)$$

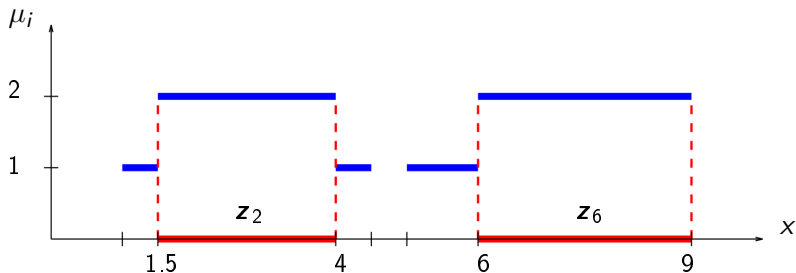
Максимальные μ_i , равные 2, достигаются для индексного множества

$$K = \{2, 6\},$$

так что частота моды равна $\mu = 2$. Как итог, мода является мультиинтервалом

$$\text{mode } \mathbf{X} = \bigcup_{k \in K} z_k = [1.5, 4] \cup [6, 9]. \quad (18)$$

Пример вычисления моды интервальной выборки.



Тот факт, что выборка не является унимодальной, может служить признаком сложной внутренней структуры описываемого ею явления. Получается, что из всего диапазона охватываемых выборкой значений выделяются тогда два или более изолированных друг от друга участка, одинаково доминирующих над остальными значениями по частоте.

Если это доминирование велико, то исследуемая величина может, к примеру, не быть постоянной, а является «смесью» нескольких близких постоянных величин.

Мода интервальной выборки

Так как сама выборка, очевидно, является своей подвыборкой, то понятие моды совпадает с пересечением всех интервалов выборки в случае её совместности.

Если же выборка несовместна, то мода может быть мультиинтервалом. Это совершенно аналогично ситуации с обычными неинтервальными данными, где мод у выборки или у распределения может быть несколько.

Пример вычисления моды интервальной выборки.

Реальный пример вычисления моды интервальной выборки.

Диаграмма рассеяния интервальных измерений.

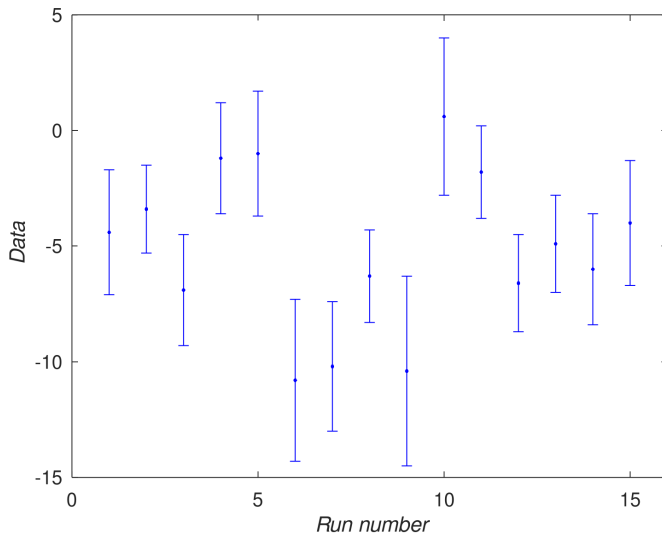


Рис.: Диаграмма рассеяния интервальных измерений [2]

Массив подинтервалов.

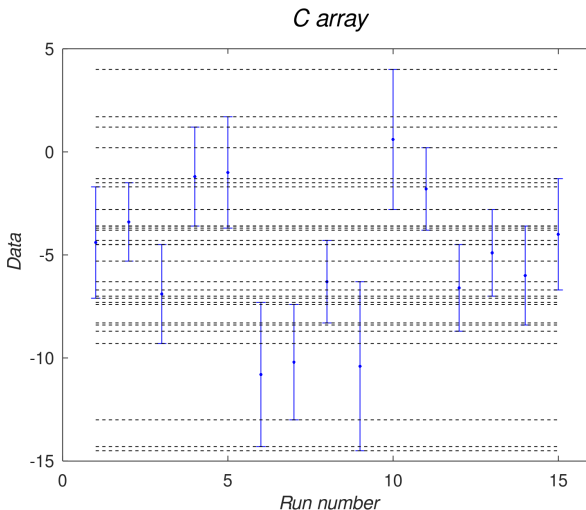


Рис.: Массив *c*.

Массив подинтервалов.

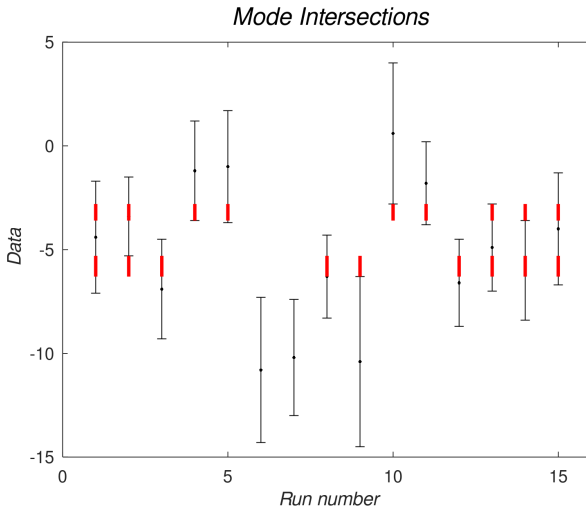


Рис.: Массив \mathbf{c} .

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

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

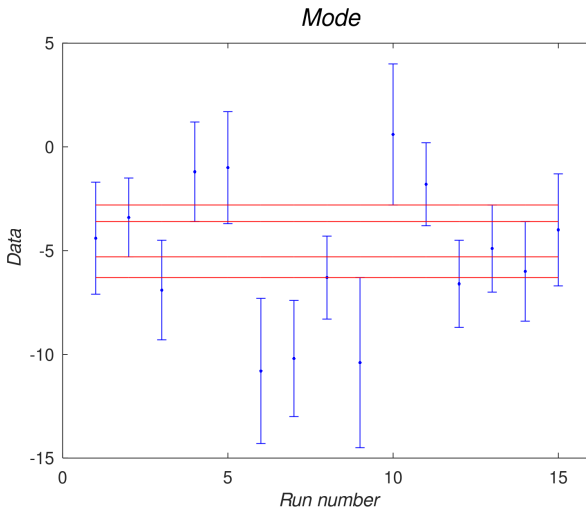


Рис.: Интервальная мода выборки данных таблицы 1.

Диаграмма рассеяния интервальных измерений.

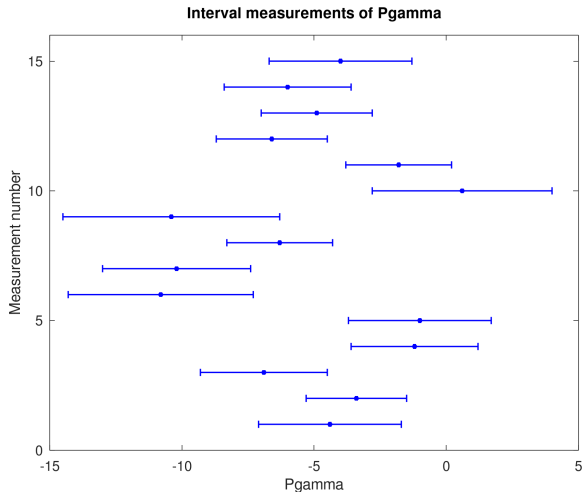


Рис.: Диаграмма рассеяния интервальных измерений [2].

График частоты пересечений подинтервалов с исходными интервалами.

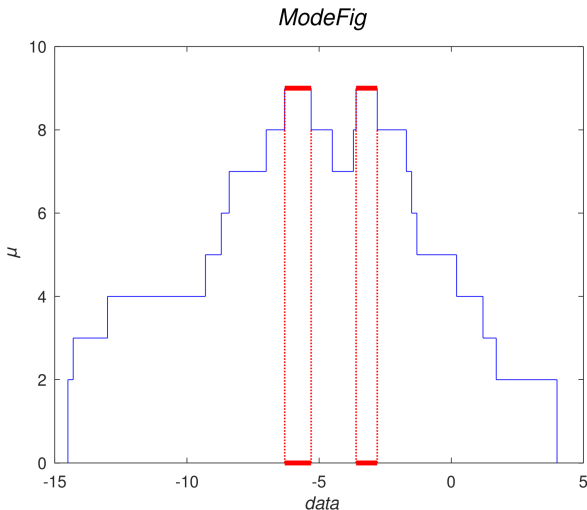


Рис. : График частоты пересечений подинтервалов с исходными интервалами.

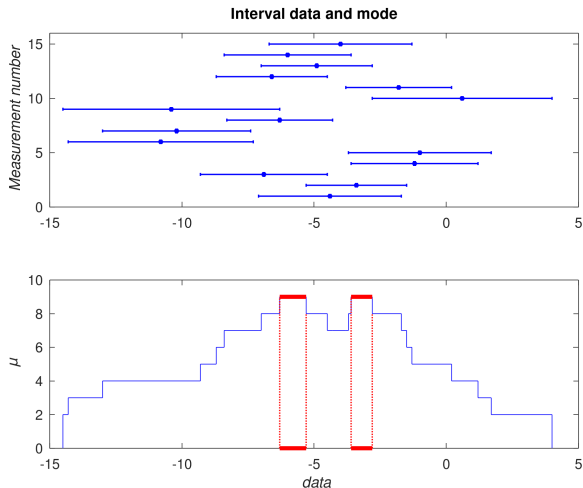


Рис.: Данные и мода.

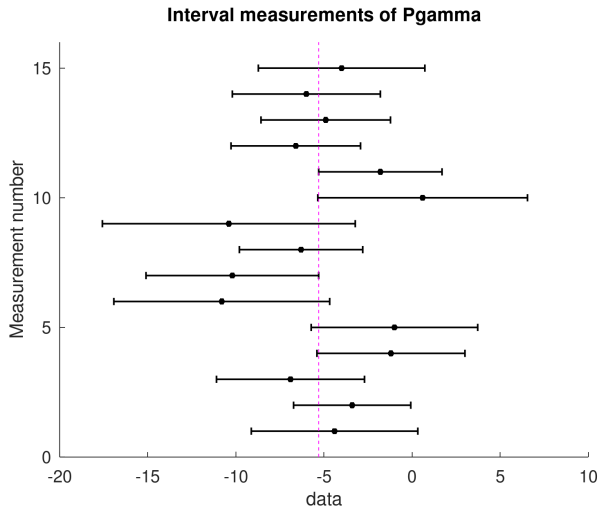


Рис.: Регуляризованные данные.

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.

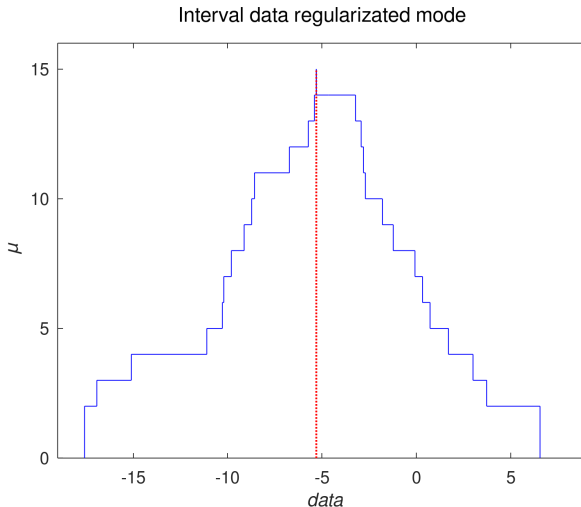


Рис.: Мода выборки с регуляризованными данными.

Регуляризованные данные и мода.

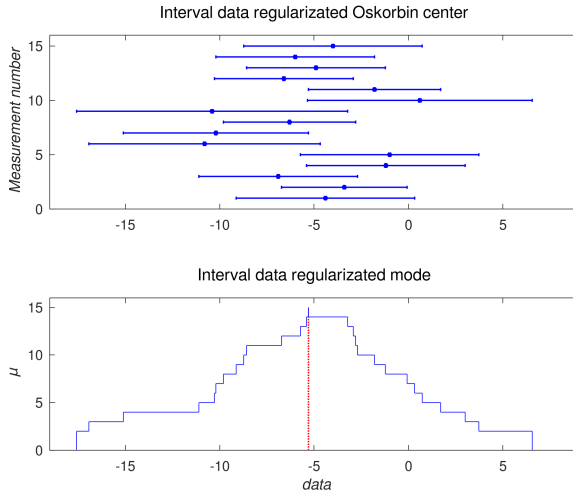
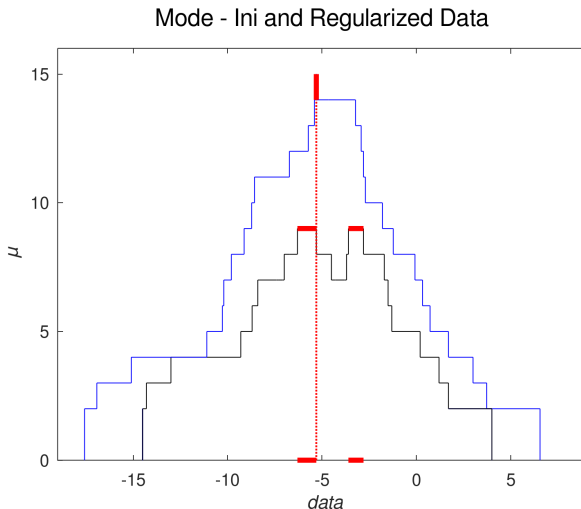
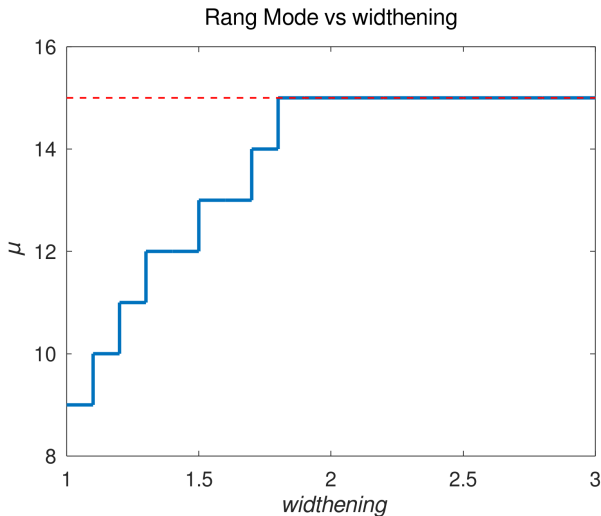


Рис.: Регуляризованные данные и мода.

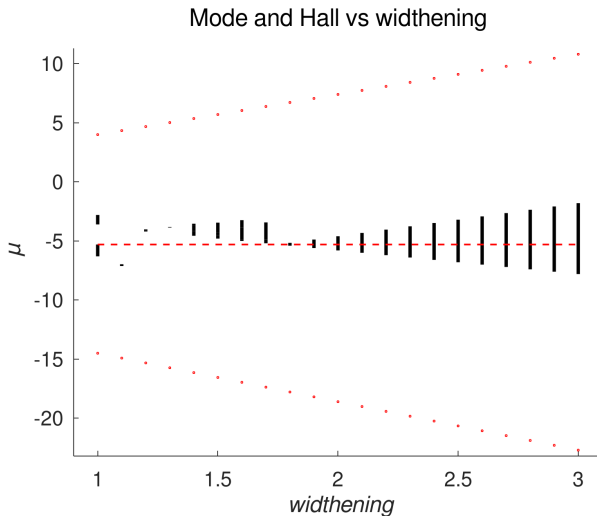
Моды выборки с исходными и регуляризованными данными.



Ранг моды выборки с регуляризованными данными.



Мода и оболочка выборки с регуляризованными данными.



Мера совместности интервалов и выборок.

В различных областях анализа данных в науках о Земле, биологии, информатике используют множество мер сходства множеств [?]. Иначе их называют коэффициентами сходства. Большинство коэффициентов нормированы и находятся в диапазоне от 0 (сходство отсутствует) до 1 (полное сходство). Наиболее часто используются бинарные меры сходства.

Мера сходства бинарная .

Мера сходства бинарная $S(A, B) \rightarrow [0, 1]$ — это вещественная функция между объектами A, B . Формально принадлежность к мерам сходства определяется системой аксиом:

- ограниченность $0 \leq S(A, B) \leq 1$;
- симметрия $S(A, B) = S(B, A) \leq 1$;
- рефлексивность $S(A, B) = 1 \iff A = B$;
- транзитивность $A \subseteq B \subseteq C \implies S(A, B) \geq S(A, C)$.

Эти свойства также называют t -норма. Существуют и иные системы аксиом сходства.

В компьютерных приложениях (обработка изображений, машинное обучение) меру сходства множеств часто обозначают как IoU — (*Intersection over Union*).

В математике часто используют именование *индекс Жаккара*, по имени математика, предложившего подобную меру.

Мера сходства бинарная.

Введём базовую конструкцию совместности для двух интервалов.

Для иллюстрации идеи, рассмотрим следующую числовую характеристику степени совместности двух интервалов \mathbf{x}, \mathbf{y}

$$\frac{\text{wid}(\mathbf{x} \wedge \mathbf{y})}{\text{wid}(\mathbf{x} \vee \mathbf{y})}. \quad (19)$$

В выражении (19) используется ширина интервала, а вместо операций пересечения и объединения множеств — операции взятия минимума (\wedge) и максимума (\vee) по включению двух величин в полной интервальной арифметике (Каухера).

В общем случае, минимум по включению в выражении (19) может быть неправильным интервалом. При этом его ширина не определена и нужно использовать либо его правильную проекцию, либо задать нужную конструкцию в явном виде.

В связи с этим, выпишем относительную меру покрытия как

$$JK(\underline{x}, \underline{y}) = \frac{\min\{\overline{x}, \overline{y}\} - \max\{\underline{x}, \underline{y}\}}{\max\{\overline{x}, \overline{y}\} - \min\{\underline{x}, \underline{y}\}}. \quad (20)$$

Мера сходства бинарная.

В записи формулы (20), вместо ширин интервалов используются явные выражения взятия минимума и максимума по включению, обеспечивающие универсальный характер конструкции, независимо от того, является ли результат операции взятия минимума по включению (\wedge) правильным или неправильным интервалом.

В именовании $JK(x, y)$ буква J отвечает фамилии Jaccard, а K указывает на арифметику Каухера.

Мера сходства бинарная.

Рассмотренная мера обобщает обычное понятие меры совместности на различные типы взаимной совместности интервалов.

В случае $\mathbf{x} \cap \mathbf{y} = \emptyset$, $\mathbf{x} \wedge \mathbf{y}$ — неправильный интервал, числитель (20) имеет отрицательное значение.

В предельном случае вещественных значений $x \neq y$, имеем

$$JK(x, y) = -1.$$

В целом имеем,

$$-1 \leq JK(\mathbf{x}, \mathbf{y}) \leq 1. \quad (21)$$

Таким образом, величина R непрерывно описывает ситуации от полной несовместности вещественных значений $x \neq y$ до полного перекрытия интервалов $\mathbf{x} = \mathbf{y}$.

Мера сходства интервальных выборок.

Мера совместности, введенная для двух интервалов в форме (20), допускает естественное обобщение на случай интервальной выборки.

Пусть имеется интервальная выборка $\mathbf{X} = \{\mathbf{x}_i\}$, $i = 1, 2, \dots, n$.
Определим меру JK

$$JK(\mathbf{X}) = \frac{\min_i \bar{\mathbf{x}}_i - \max_i \underline{\mathbf{x}}_i}{\max_i \bar{\mathbf{x}}_i - \min_i \underline{\mathbf{x}}_i}. \quad (22)$$

Важно, что выражение (22) переходит в случае интервальной выборки из 2 элементов в выражение (20). Таким образом, принцип соответствия выполнен.

Пример меры сходства интервальных выборок.

Пусть имеется интервальная выборка из 4 элементов (15), рассмотренная при вычислении интервальной моды

$$X = \{ [1, 4], [5, 9], [1.5, 4.5], [6, 9] \}.$$

Пример меры сходства интервальных выборок.

Диаграмма рассеяния выборки X

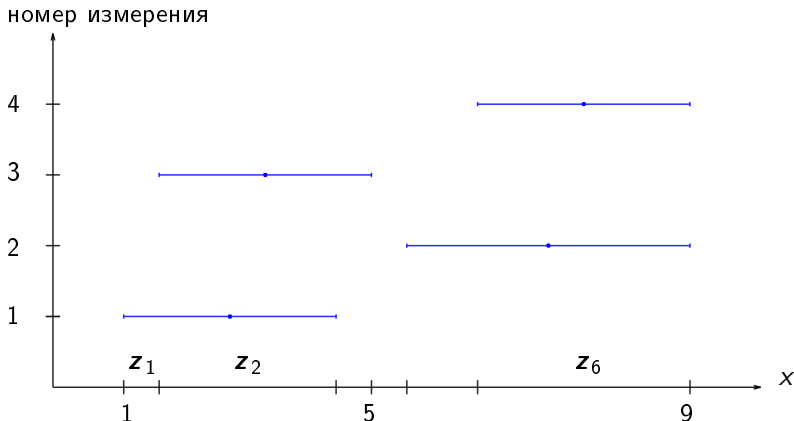


Рис.: Диаграмма рассеяния интервальной выборки (15) и элементы выборки z .

Пример меры сходства интервальных выборок.

Выберем из нее накрывающую подвыборку

$$\mathbf{X}_c = \{ [5, 9], [6, 9] \}.$$

Для выборки \mathbf{X}_c имеем согласно (22)

$$JK(\mathbf{X}_c) = \frac{9 - 6}{9 - 5} = 0.75.$$

Значение $JK(\mathbf{X}_c)$ демонстрирует высокую меру сходства элементов выборки \mathbf{X}_c .

Характеризация области значения переменной.

Измерения температуры двумя датчиками.

На верхнем рисунке — данные за весь период проведения эксперимента, около 2-х суток. Серым залит временной диапазон среднего графика.

На среднем рисунке — данные в окрестности стационарного режима, серая заливка — временной диапазон нижнего графика.

На нижнем рисунке — данные за 2 часа проведения испытаний.

Совместность показаний обоих датчиков, т.е. $A \cap B \neq \emptyset$, имеет место около 1 часа.

Характеризация области значения переменной.

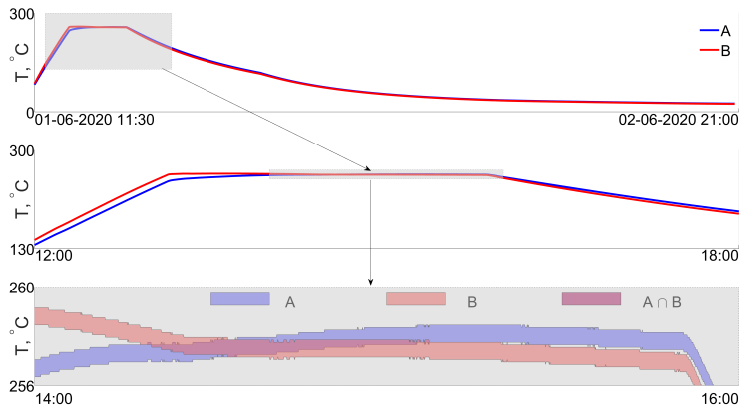
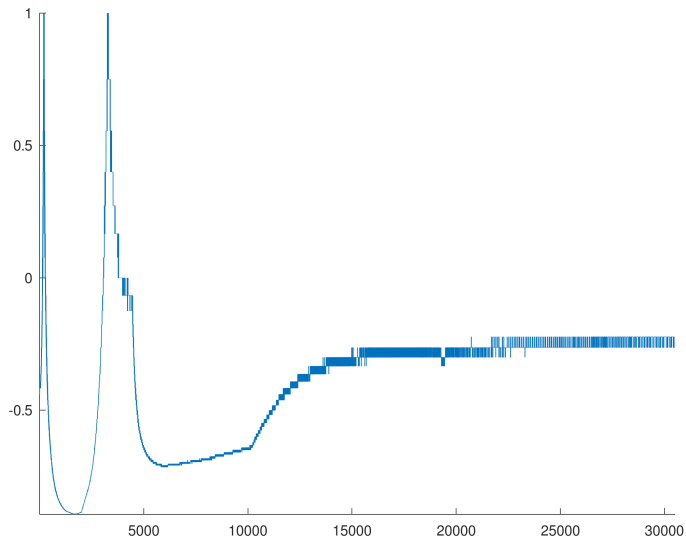







График значений функционала JK.



-  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.