Министерство науки и высшего образования Российской Федерации

САНКТ-ПЕТЕРБУРГСКИЙ
ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО

Физико-механический институт
Высшая школа прикладной математики и вычислительной физики

А. Н. Баженов

ВВЕДЕНИЕ В АНАЛИЗ ДАННЫХ С ИНТЕРВАЛЬНОЙ НЕОПРЕДЕЛЕННОСТЬЮ

Учебное пособие

ПОЛИТЕХ-ПРЕСС
Санкт-Петербургский
политехнический университет
Петра Великого

Санкт-Петербург
2022

УДК 519.9
ББК
Б

Р е ц е н з е н т ы:
Кандидат физико-математических наук, научный сотрудник
Физико-технического института им. А. Ф. Иоффе *А. А. Красилин*
Кандидат физико-математических наук, доцент
Санкт-Петербургского политехнического университета
Петра Великого *Л. В. Павлова*

Баженов А. Н. **Введение в анализ данных с интервальной неопределенностью** : учеб. пособие / А. Н. Баженов. — СПб. : ПОЛИТЕХ-ПРЕСС, 2022. — XXX с.

Учебное пособие соответствует образовательному стандарту высшего образования Санкт-Петербургского политехнического университета Петра Великого по направлению подготовки бакалавров 01.03.02 «Прикладная математика и информатика», по дисциплине «Интервальный анализ».

Пособие посвящено введению в анализ данных с интервальной неопределенностью и демонстрации его применения в различных задачах. Важной частью пособия являются примеры, от самых простых, иллюстрирующих базовые конструкции и операции, до более сложных.

Пособие предназначено для студентов, аспирантов, научных сотрудников и инженеров, занимающихся анализом данных.

Табл. 6. Ил. 34. Библиогр.: 40 назв.

Печатается по решению
Совета по издательской деятельности Ученого совета
Санкт-Петербургского политехнического университета Петра Великого.

ISBN 978-5-7422-...
doi:10.18720/SPBPU/2/id22-...

Министерство науки и высшего образования Российской Федерации

САНКТ-ПЕТЕРБУРГСКИЙ
ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО

Физико-механический институт
Высшая школа прикладной математики и вычислительной физики

А. Н. Баженов

ВВЕДЕНИЕ В АНАЛИЗ ДАННЫХ С ИНТЕРВАЛЬНОЙ НЕОПРЕДЕЛЕННОСТЬЮ

Учебное пособие

ПОЛИТЕХ-ПРЕСС
Санкт-Петербургский
политехнический университет
Петра Великого

Санкт-Петербург
2022

УДК 519.9
ББК
Б

Р е ц е н з е н т ы:
Кандидат физико-математических наук, научный сотрудник
Физико-технического института им. А. Ф. Иоффе *А. А. Красилин*
Кандидат физико-математических наук, доцент Санкт-Петербургского
политехнического университета
Петра Великого *Л. В. Павлова*

Баженов А. Н. **Введение в анализ данных с интервальной неопределенностью** : учеб. пособие / А. Н. Баженов. — СПб. : ПОЛИТЕХ-ПРЕСС, 2022. — XXX с.

Учебное пособие соответствует образовательному стандарту высшего образования Санкт-Петербургского политехнического университета Петра Великого по направлению подготовки бакалавров 01.03.02 «Прикладная математика и информатика», по дисциплине «Интервальный анализ».

Пособие посвящено введению в анализ данных с интервальной неопределенностью и демонстрации его применения в различных задачах. Важной частью пособия являются примеры, от самых простых, иллюстрирующих базовые конструкции и операции, до более сложных.

Пособие предназначено для студентов, аспирантов, научных сотрудников и инженеров, занимающихся анализом данных.

Табл. 6. Ил. 34. Библиогр.: 40 назв.

Печатается по решению
Совета по издательской деятельности Ученого совета
Санкт-Петербургского политехнического университета Петра Великого.

ISBN 978-5-7422-...
doi:10.18720/SPBPU/2/id22-...

Оглавление

Список примеров

Введение

Учебное пособие является дополнением к книге коллектива авторов «Обработка и анализ данных с интервальной неопределенностью» [1], изложение материала в которой носит методически основательный характер. В пособии представлена согласованная система понятий и терминов, относящихся к обработке данных, имеющих интервальную неопределенность, а также дан краткий обзор основных и наиболее значимых результатов научного направления, которое можно назвать «статистикой интервальных данных», или «анализом интервальных данных». В пособии много ссылок на [1], в которой теоретические аспекты изложены обстоятельно и подробно. Задача пособия — дать обучающимся краткие сведения о теории и рассмотреть ряд примеров, иллюстрирующих интервальный подход.

Фундаментом статистики данных с интервальной неопределенностью является интервальный анализ. Основы интервального анализа представлены в [2]. В [3] дана картина применения интервальных данных и интервального анализа в более широком контексте. Наиболее полное изложение идей и методов интервального анализа дано в [4]. Изучение материала книги [4] требует более основательной математической подготовки и рекомендуется для углубленного изучения данного вопроса.

В общем виде задачи статистики данных с интервальной неопределенностью состоят в решении практических проблем в тех областях обработки данных, в которых недостаточны ранее развитые методы. В практике обработки экспериментальных данных в настоящее время широко используются статистические методы, основанные на идеях и результатах теории вероятностей. Эти методы опираются на использование ряда допущений о вероятностных свойствах погрешностей измерений, а также на наличие выборок представительной длины (как минимум в несколько десятков измерений).

Однако практики часто сталкиваются с ситуациями, когда выборки измерений коротки, а погрешности измерений не могут быть адекватно описаны с помощью инструментов теории вероятностей или же информация о вероятностных характеристиках погрешностей отсутствует.

В этих ситуациях можно применить методы интервальной статистики, основанные на идеях и результатах интервального анализа, использующих его подходы, алгоритмы и соответствующее программное обеспечение. Интервальные методы широко представлены практически для всех популярных платформ программирования. В некоторых интегрированных средах, как, например, `Mathematica`, `Octave`, поддержка базовых интервальных конструкций встроенная. Для использования наиболее популярного в настоящее время языка программирования `Python` также есть реализации основных конструкций и методов интервального анализа.

Терминология интервальной статистики наследует многое из традиционной статистики, развитый понятийный аппарат которой уже сложился. Различным аспектам анализа интервальных данных посвящены, в частности, [5]-[14], [15].

Следует отметить, что в XX в. в статистике различные математические методы продолжали развиваться и использоваться, но как будто не входя в математическую статистику. Дж. Тьюки в конце 50-х годов прошлого века предложил оформить новую научную дисциплину «анализ данных», в которой охватывались те математические методы обработки данных, которые не подпадали под математическую статистику в узком смысле этого слова [16].

Материал учебного пособия апробирован в учебных курсах для студентов Высшей школы прикладной математики и вычислительной физики Физико-механического института Санкт-Петербургского политехнического университета Петра Великого и аспирантов Физико-технического института им. А. Ф. Иоффе Российской академии наук.

Глава 1

Краткие сведения о методах статистики и обработки данных

1.1 Данные, погрешности и их обработка

На практике данные не бывают точными. В действительности нам известно приближенное значение измеряемой величины, а также некоторая информация (качественная и количественная) о погрешности этого значения. На результаты измерений могут оказывать влияние изменчивость измеряемых величин, их непостоянство во времени или пространстве. На измерения могут влиять внешние неконтролируемые факторы, так называемые «шумы». У применяемой аппаратуры имеются собственные погрешности. В процессе математической обработки данных на результат влияют неизбежные неточности расчетов (ошибки представления, округления и т. п.).

В [17] погрешности измерений и наблюдений разделяются на три класса:

1. Систематические погрешности.
2. Случайные погрешности.
3. Промахи (или выбросы).

Систематической погрешностью измерения называется составляющая погрешности измерения, которая остается постоянной или изменяется по какому-то определенному закону при повторных измерениях одной и той же величины. *Случайными погрешностями* называются неопределенные по

своей величине и природе погрешности, в появлении каждой из которых не наблюдается какой-либо явной закономерности. *Промахами* (*выбросами*) называются погрешности, приводящие к явному искажению результата измерений. Для выявления выбросов и промахов организуют специальный этап общей технологии обработки данных — *предобработку*, который предшествует применению формальных математических методов. На этом этапе промахи (выбросы) должны быть определены и удалены из обрабатываемых данных. Что касается случайных погрешностей, то в [17] указывается: «Мы считаем случайными те явления, которые определяются сложной совокупностью переменных причин, трудно поддающихся анализу; к этим явлениям индивидуальный подход невозможен, и лишь для их совокупности могут быть установлены определенные закономерности». Таким образом, термин «случайный» в этом понимании фактически означает «непредсказуемый» или же такой, в чем отсутствует закономерность.

Как учитывать случайные погрешности в данных? Прежде всего, сам факт присутствия таких погрешностей в данных можно учесть подходящей математической постановкой задачи обработки этих данных. Например, при восстановлении функциональных зависимостей (см. гл. 4) вместо задачи интерполяции данных нужно рассматривать задачу их аппроксимации (приближения), так как не имеет смысла требовать точных равенств значений функции измеренным значениям. Вообще говоря, получение результата измерения или наблюдения как решения задачи некоторого математического приближения к данным, учитывающей модель исследуемого объекта или явления, является основой *аппроксимационных методов* обработки данных.

Если о природе случайных погрешностей ничего более не известно, то на этом можно и нужно остановиться и применять далее аппроксимационные методы. Если о природе случайных погрешностей известно что-то определенное, то можно применить для обработки данных более точные методы, учитывающие дополнительную информацию.

В настоящее время существует несколько различных подходов к описанию случайности, и некоторые их них чрезвычайно развиты и популярны. Прежде всего, это теоретико-вероятностная модель погрешностей, основанная на аппарате математической теории вероятности и приводящая к *теоретико-вероятностным методам* обработки данных. Теоретико-вероятностная модель погрешностей за прошедшие два века получила большое развитие и распространение, став одним из основных инструментов обработки данных. Также следует отметить методы нечеткой статистики (см. п. 1.2.1) и *эвристические методы* обработки данных, которые применяются при анализе малоизученных явлений, когда отсутствует четкая модель и нет представления об искомых характеристиках явления или объекта.

1.2 Критика вероятностной статистики и альтернативные подходы

Развернутая критика вероятностной статистики содержится в [1]. Кратко перечислим основные пункты, по которым в в [1] проведено обсуждение.

Статистическая устойчивость. Главной интерпретацией понятия вероятности является так называемая *частотная интерпретация*, при которой вероятность понимается как предел относительной частоты рассматриваемого события в серии однородных независимых испытаний (экспериментов и т. п.).

Многие явления окружающего нас мира, в отношении которых применимо слово «случайный», не обладают свойством существования устойчивой относительной частоты, так как при росте числа наблюдений она для них не устанавливается. Для описания и анализа подобных явлений традиционная теория вероятностей непригодна.

Проблема малых выборок. Вероятностные закономерности проявляются как тенденции, которые наиболее заметны в массовых явлениях. Фактически при обработке экспериментальных данных почти всегда стоит вопрос о том, достаточен ли объем выборки (количество измерений и т. п.) для того, чтобы выводы, получаемые на основе теоретико-вероятностной модели погрешностей, имели приемлемую практическую достоверность.

Существующие промышленные стандарты и методики обработки экспериментальных данных (например, [18] -[20]) регламентируют способы работы с выборками размера лишь более 15. При этом результаты обработки выборок размера от 16 до 50 рекомендуется рассматривать как не очень надежные и сопровождать оговорками, а обработка выборок размером не более чем из 15 измерений стандартами вообще не рассматривается.

Неизвестные вероятностные характеристики распределения. Если законы теории вероятностей применимы к анализу погрешностей, то каков конкретный вид вероятностных распределений погрешностей? Каковы его числовые характеристики? Это непростые вопросы, на которые не всегда есть ответ.

Например, считается, что типичным законом вероятностного распределения погрешностей является нормальное гауссово распределение. Но насколько оно соответствует действительности? Известно высказывание А. Пуанкаре [21]: «...все верят в этот закон ... потому что экспериментаторы думают, что это математическое утверждение, а математики — что это результат экспериментов». Реальные распределения погрешностей измерений в различных ситуациях могут сильно отличаться от нормального гауссового. Для того чтобы выяснить, какое вероятностное распределение имеют анализируемые данные, часто требуется большая дополнительная работа, требующая выборок более 1000 измерений [22].

Конкретный вид функций распределения случайных величин, которые фигурируют в задачах обработки данных, может оказывать существенное влияние на способ их решения. Методология максимума правдоподобия для случая нормально распределенных погрешностей данных указывает на метод наименьших квадратов, метод наименьших модулей для распределения Лапласа или метод чебышевского сглаживания (минимаксное приближение данных) для равномерно распределенных погрешностей.

Независимость данных. Еще одна группа вопросов касается часто используемых в теории вероятностей понятий *независимости* и *корреляции* случайных величин. Имеют ли данные корреляцию между собой? Или же они независимы? Многие классические результаты вероятностной статистики требуют, как известно, независимости рассматриваемых случайных величин, представляющих результаты измерений, либо заданного уровня их корреляции. Проверка этих условий на практике почти невозможна.

Наличие погрешностей различных типов. В ходе измерений, помимо статистических погрешностей, всегда присутствуют и систематические. Последние могут имет разые источники, их весьма сложно оценить. Часто эта оценка намного сложнее получения собственно результатов. Но даже если эти оценки получены, появляется вопрос: «Каким образом можно получить совокупную ошибку?»

1.2.1 Статистика нечетких данных

При нечетком описании результатов измерений и наблюдений мы полагаем, что вместо их точных значений нам известны так называемые *функции принадлежности* нечетких чисел, получающихся в результате измерений [15]. Возникновение нечетких чисел в природных явлениях на примере спектров возбуждения и эмиссии [23] представлено на рис. 1.1.

Нечетким множеством называется множество X, образованное элементами произвольной природы, которое дополнено так называемой *функцией принадлежности* $\mu : X \to [0,1]$, значение которой $\mu(x)$ на элементе $x \in X$ показывает степень принадлежности x множеству X (рис. 1.1). У стандартной функции принадлежности множества (называемой также *индикаторной функцией* множества) значения могут быть равны только 0 или 1, поэтому допущение для функции μ непрерывного ряда значений из интервала $[0,1]$ позволяет характеризовать ситуации, когда нет уверенности в принадлежности элемента множеству, невозможно оперировать количественной мерой уверенности и строить на этой основе наши выводы и заключения.

Для построения содержательной теории нечеткого вывода и нечетких неопределенностей обычно ограничивают общность функции принадлежности μ, требуя, чтобы она была *квазивогнутой*. В одномерном случае они являются интервалами. Нечеткие множества с квазивогнутыми функциями

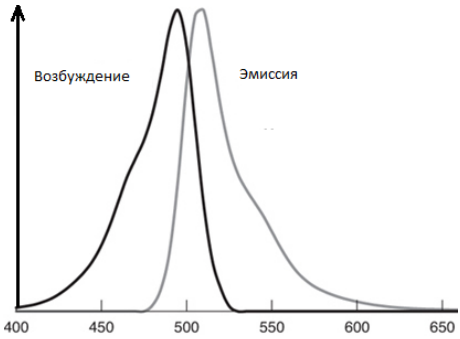

Рис. 1.1. Спектры возбуждения-эмиссии как нечеткие числа [23]

принадлежности называются *нечеткими числами*, и они могут быть эквивалентным образом заданы как семейства вложенных друг в друга интервалов, которые соответствуют различным уровням принадлежности.

Для обработки данных, имеющих нечеткую неопределенность, предложены разные подходы (см., например, [15]), в частности, большое развитие получили методы восстановления зависимостей по нечетким данным [24].

1.3 Место и особенности интервального подхода

1.3.1 Почему интервалы?

Устройство природы. Фундаментальная причина использования интервалов для описания данных состоит в том, что некоторые физические (химические, биологические и т. п.) величины принципиально не могут быть выражены точечными значениями, а лишь интервалами. Поэтому интервалы представляют собой новый удобный тип данных, которым уместно дополнить элементарные типы данных, использующиеся в метрологии. Большое количество примеров из разных областей науки и техники приведено в [3].

Пример 1.3.1 (Интервальные веса химических элементов) С 2009 г. атомные веса некоторых элементов в Периодической системе элементов Д. И. Менделеева, поддерживаемой Международным союзом теоретической и прикладной химии (ИЮПАК, IUPAC), стали выражаться интервалами [25]. Почти каждый химический элемент представлен в природе смесью своих изотопов, т. е. разновидностями атомов, сходных по своим химическим свой-

ствам (структуре электронных оболочек), но отличающихся массой ядер. Относительная доля различных изотопов существенно меняется в зависимости от места и характера взятия пробы. Например, в тканях живых организмов преобладают более легкие изотопы химических элементов, нежели в неживой природе. Отличаются друг от друга относительные доли изотопов элементов на суше и в морях и т. п.

Известны изотопы ртути с массовыми числами от 170 до 216 (количество протонов — 80, нейтронов — от 90 до 136). Природная ртуть состоит из смеси семи стабильных изотопов, гистограмма частот изотопов показана на рис. 1.2.

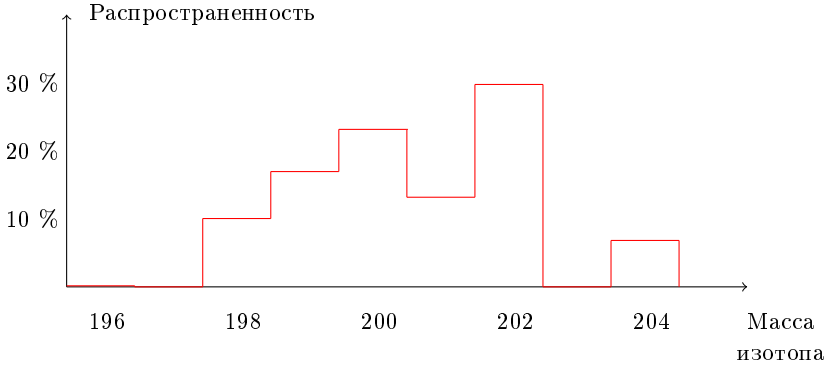

Рис. 1.2. Распространенность изотопов ртути на Земле

Математические причины. В чем преимущества и недостатки интервалов в сравнении с другими способами описания неопределенности? Имеется ряд причин, по которым интервалы нужны и важны при обработке данных.

Интервалы являются средством описания и представления типа неопределенностей, часто встречающихся в реальной жизни, ограниченных по величине неопределенностей. Интервалы проще, чем вероятностные распределения или нечеткие множества. Интервал — это «бесструктурный объект», который более сжато описывает неопределенность. Следствием этой простоты является лучшая развитость теории интервального анализа и интервальных вычислительных методов.

Дале, интервалы и интервальные арифметики оказываются уникальными во многих отношениях, в частности, по своим алгебраическим свойствам, простоте и богатству опеределения отношений между объектами и результатами операций.

Наконец, интервалы являются предельным случаем сумм независимых ограниченных величин. В большинстве практических ситуаций погрешность измерения возникает в результате накопления и наложения большого количества независимых факторов. Если некоторая величина есть сумма большого количества малых независимых слагаемых, то множество всевозможных значений этой величины близко к интервалу. Этот результат составляет содержание «предельной теоремы Крейновича» и ее обобщений [4].

Кроме того, на основе интервалов можно строить составные математические объекты, описывающие аспекты данных и вычислений, которые недоступны в рамках вещественной арифметики, твины и мультиинтервалы. Кратко рассмотрим их в п. 2.5.3.

1.3.2 Статистика интервальных данных

Интервальной неопределенностью называется состояние частичного знания о величине, которая не известна точно, но известны нижняя и верхняя границы ее возможных значений, или, иными словами, известен интервал возможных значений этой величины.

В одномерном случае интервалы являются практически наиболее важными ограниченными множествами, поэтому другие множества неопределенности используются нечасто. Но в многомерном случае множествами возможных значений величины, имеющей ограниченную неопределенность, могут быть брусы, многогранники, параллелотопы (зонотопы), эллипсоиды и прочие объекты. Мы их относим к объектам интервальной статистики и интервального анализа данных.

Отличительной чертой представляемого подхода является его применимость к выборкам любого объема, начиная с нескольких измерений (в предельном случае — одного). Как следствие, проблемы «малых выборок», характерной для вероятностной статистики, в интервальном подходе не существует. Это свойство особенно ценно, когда технические или экономические причины не позволяют проводить много экспериментов. В частности, такова ситуация с алгоритмами обработки результатов разрушающих измерений или измерений быстропротекающих процессов в реальном масштабе времени. Интервальные методы имеют *аппроксимационный* характер, т. е. осуществляют приближение (аппроксимацию) данных в нужном смысле. Следовательно, для их применения массовость не требуется.

Развиваемые идеи впервые были оформлены в пионерской работе на данную тему Л. В. Канторовича [26] и далее неоднократно использовались (или даже переоткрывались) разными авторами. В то время для обозначения аналогичных подходов в литературе по математике использовались разные термины — «минимаксный подход» и др.

Интервальный подход позволяет построить достаточно простую и эле-

гантную методику определения выбросов в данных. При анализе постоянных величин имеется ряд обобщений традиционных методов, дающих более обширную информацию.

В задаче восстановления зависимостей неопределенности входных и выходных переменных учитываются естественным образом. Оценка погрешности результатов получается автоматически в процессе вычислений, не требует дополнительного анализа и напрямую зависит от неопределенности данных задачи.

Принцип соответствия. В методологии науки *принцип соответствия* называют утверждение, что любая новая научная теория должна включать старую теорию и ее результаты как частный предельный случай. Далее будем использовать принцип соответствия как инструмент проверки адекватности используемых конструкций, понятий и методов обработки данных с интервальными неопределенностями, который позволяет отсекать заведомо «неразумные».

Глава 2

Базовые понятия и математический аппарат

2.1 Интервалы

Вещественные интервалы. Первичное понятие интервального анализа — *интервал*. Это множество, задающее целый диапазон значений интересующей нас величины, с помощью которого можно рассматривать неопределенности и неоднозначности.

Интервалы могут определяться на вещественной оси, на комплексной плоскости, а также в многомерных пространствах [4]. Далее будут рассматриваться вещественные интервалы, интервальные векторы и матрицы, так как именно они играют главную роль в измерениях и их обработке.

Определение 2.1.1 *Интервалом* $[a, b]$ *вещественной оси* $\mathbb{R}$ *называется множество всех чисел, расположенных между заданными числами* a *и* b*, 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* $[a, b]$*, левым (или нижним) и правым (или верхним) соответственно.*

Аналогичные термины, которые часто используются в математических текстах, — это *числовой промежуток* (замкнутый), *отрезок*, *сегмент* вещественной оси.

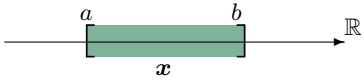

Рис. 2.1. Интервал на вещественной оси

Множество всех интервалов из $\mathbb{R}$ обозначается символом $\mathbb{IR}$. В противоположность интервалам и интервальным величинам будем называть *точечными* те величины, значениями которых являются отдельные точки вещественной оси или пространства более высокой размерности. Множество вещественных чисел $\mathbb{R}$ можно рассматривать как подмножество множества интервалов, т. е. как интервалы с совпадающими концами. Итак,

$$\mathbb{R} \subseteq \mathbb{IR}.$$

Используемая система обозначений следует неформальному международному стандарту обозначений в интервальном анализе [27]. В частности, интервалы и другие интервальные величины (векторы, матрицы и др.) всюду в тексте обозначаются полужирным шрифтом, например, $\boldsymbol{A}$, $\boldsymbol{B}$, $\ldots$, $\boldsymbol{y}$, $\boldsymbol{z}$, тогда как неинтервальные (точечные) величины никак специально не выделяются. Для интервала $\boldsymbol{a}$ посредством $\underline{\boldsymbol{a}}$ или $\inf \boldsymbol{a}$ обозначается его левый конец, тогда как $\overline{\boldsymbol{a}}$ или $\sup \boldsymbol{a}$ — это его правый конец.

В целом $\boldsymbol{a} = [\underline{\boldsymbol{a}}, \overline{\boldsymbol{a}}]$, поэтому

$$\boldsymbol{a} = \{\, x \in \mathbb{R} \mid \underline{\boldsymbol{a}} \le x \le \overline{\boldsymbol{a}} \,\}. \tag{2.1}$$

Характеристики интервала. Любой интервал полностью задается двумя числами — своими концами, но на практике широко используются также другие характеристики интервалов и представления интервалов на их основе.

Важнейшими характеристиками интервала являются его *середина* (центр)

$$\operatorname{mid} \boldsymbol{a} = \tfrac{1}{2}(\overline{\boldsymbol{a}} + \underline{\boldsymbol{a}}), \tag{2.2}$$

и его *радиус*

$$\operatorname{rad} \boldsymbol{a} = \tfrac{1}{2}(\overline{\boldsymbol{a}} - \underline{\boldsymbol{a}}). \tag{2.3}$$

В ряде случаев (например, п. 3.3) вместо радиуса рассматривается эквивалентное понятие *ширины* интервала

$$\operatorname{wid} \boldsymbol{a} = \overline{\boldsymbol{a}} - \underline{\boldsymbol{a}}. \tag{2.4}$$

В целом $\boldsymbol{a} = \operatorname{mid} \boldsymbol{a} + [-1, 1] \cdot \operatorname{rad} \boldsymbol{a}$, что равносильно представлению

$$\boldsymbol{a} = \{\, x \in \mathbb{R} \mid |x - \operatorname{mid} \boldsymbol{a}| \le \operatorname{rad} \boldsymbol{a} \,\}. \tag{2.5}$$

Таким образом, задание середины и радиуса интервала также однозначно определяет его.

Середина интервала — это точка, которая представляет его наилучшим образом, так как она наименее удалена от остальных точек этого интервала.

Радиус и ширина характеризуют разброс (рассеяние) точек интервала, т. е. абсолютную меру неопределенности или неоднозначности, выражаемой этим интервалом. Интервалы нулевой ширины (нулевого радиуса) обычно называют *вырожденными*. Они отождествляются с вещественными числами, поэтому, к примеру, $[1,1]$ — это то же самое, что и 1.

С одной стороны, важной характеристикой интервала является его *модуль* (*магнитуда*, *абсолютное значение*), определяемое как максимум модулей точек из интервала

$$|\boldsymbol{a}| \;=\; \max\{\, |a| \mid a \in \boldsymbol{a} \,\} \;=\; \max\{\, |\underline{\boldsymbol{a}}|, |\overline{\boldsymbol{a}}| \,\}.$$

Модуль интервала — это наибольшее отклонение его точек от нуля.

С другой стороны, величина, показывающая, насколько далеко отделен от нуля тот или иной интервал вне зависимости от его знака называется *мигнитудой*, которая определяется как

$$\langle \boldsymbol{a} \rangle \;=\; \min\{\, |a| \mid a \in \boldsymbol{a} \,\} \;=\; \begin{cases} \min\{\, |\underline{\boldsymbol{a}}|, |\overline{\boldsymbol{a}}| \,\}, & \text{если } 0 \notin \boldsymbol{a}, \\ 0, & \text{если } 0 \in \boldsymbol{a}. \end{cases}$$

Среди интервалов особую роль играют интервалы вида $[-a, a]$, имеющие своей серединой нуль. Их называют *уравновешенными*. Среди всех интервалов с данным абсолютным значением (модулем) именно уравновешенные интервалы имеют наибольшую ширину. И наоборот, среди интервалов фиксированной ширины уравновешенные интервалы имеют наименьшее абсолютное значение.

Интервал полностью определяется двумя своими концами и представляет собой объект, который не несет никакой дополнительной структуры. Все точки интервала равноценны (равнозначны, равновозможны), и для каждой из них интервал дает двустороннее приближение.

В частности, интервал $\boldsymbol{a}$ нельзя отождествлять с равномерным вероятностным распределением на $[\underline{\boldsymbol{a}}, \overline{\boldsymbol{a}}]$ с плотностью $1/(\mathrm{wid}\,\boldsymbol{a})$, так как в пределах $\boldsymbol{a}$ с тем же успехом может быть определено любое другое вероятностное распределение или даже какое-то распределение, меняющееся во времени, — случайный процесс.

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

$$\boldsymbol{a} \subseteq \boldsymbol{b} \;\text{ равносильно тому, что }\; \underline{\boldsymbol{a}} \geq \underline{\boldsymbol{b}} \;\text{ и }\; \overline{\boldsymbol{a}} \leq \overline{\boldsymbol{b}}. \tag{2.6}$$

Отношение включения является *частичным порядком* и превращает множество интервалов в частично упорядоченное множество.

Помимо порядка по включению на множестве интервалов большую роль играют также другие отношения, которые обобщают линейный порядок «$\leq$» на вещественной оси $\mathbb{R}$. Порядок «$\leq$» между вещественными числами может быть обобщен на интервалы многими способами. Важную роль играет следующее упорядочение.

Определение 2.1.2 *Для интервалов $\boldsymbol{a}$, $\boldsymbol{b} \in \mathbb{IR}$ условимся считать, что $\boldsymbol{a}$ не превосходит $\boldsymbol{b}$ и писать «$\boldsymbol{a} \leq \boldsymbol{b}$» тогда и только тогда, когда $\underline{\boldsymbol{a}} \leq \underline{\boldsymbol{b}}$ и $\overline{\boldsymbol{a}} \leq \overline{\boldsymbol{b}}$.*
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 $\boldsymbol{a}$ и $\boldsymbol{b}$ имеют непустое пересечение, т. е. $\boldsymbol{a} \cap \boldsymbol{b} \neq \varnothing$, то можно дать простые выражения для результатов теоретико-множественных операций пересечения и объединения через концы этих интервалов

$$\boldsymbol{a} \cap \boldsymbol{b} = \bigl[\max\{\underline{\boldsymbol{a}}, \underline{\boldsymbol{b}}\}, \min\{\overline{\boldsymbol{a}}, \overline{\boldsymbol{b}}\}\bigr], \quad \boldsymbol{a} \cup \boldsymbol{b} = \bigl[\min\{\underline{\boldsymbol{a}}, \underline{\boldsymbol{b}}\}, \max\{\overline{\boldsymbol{a}}, \overline{\boldsymbol{b}}\}\bigr].$$

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

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

$$\boldsymbol{a} \wedge \boldsymbol{b} = \bigl[\max\{\underline{\boldsymbol{a}}, \underline{\boldsymbol{b}}\}, \min\{\overline{\boldsymbol{a}}, \overline{\boldsymbol{b}}\}\bigr], \tag{2.7}$$

$$\boldsymbol{a} \vee \boldsymbol{b} = \bigl[\min\{\underline{\boldsymbol{a}}, \underline{\boldsymbol{b}}\}, \max\{\overline{\boldsymbol{a}}, \overline{\boldsymbol{b}}\}\bigr]. \tag{2.8}$$

Первая из этих операций, «$\wedge$», не всегда выполнима во множестве обычных интервалов, но это затруднение преодолевается посредством расширения множества интервалов специальными элементами — неправильными интервалами (см. п. 2.4).

2.2 Классическая интервальная арифметика

Значения физических (и иных) величин входят в математические выражения для физических законов, в различные формулы, в которых используются арифметические операции. После определения интервалов, приходим к необходимости введения операций и отношений между ними.

Наиболее естественным является определение результата интервальной операции «по представителям», как множества всевозможных результатов

этой же операции между числами из интервалов. Например, для двухместной операции «$\star$» можно считать, что

$$\boldsymbol{a} \star \boldsymbol{b} \;=\; \big\{\, a \star b \mid a \in \boldsymbol{b},\, b \in \boldsymbol{b} \,\big\}. \tag{2.9}$$

Аналогичным образом определяются интервальные аналоги для одноместных операций.

Если рассматриваются арифметические операции, т. е. $\star \in \{+, -, \cdot, /\}$, то нетрудно показать, что задаваемые правилом (2.9) множества являются интервалами, исключая случай деления на интервал $\boldsymbol{b}$, который содержит нуль [4].

Конструктивные формулы, расшифровывающие этот общий принцип для отдельных арифметических операций, выглядят следующим образом:

$$\boldsymbol{a} + \boldsymbol{b} = \big[\, \underline{\boldsymbol{a}} + \underline{\boldsymbol{b}},\, \overline{\boldsymbol{a}} + \overline{\boldsymbol{b}} \,\big]\,; \tag{2.10}$$

$$\boldsymbol{a} - \boldsymbol{b} = \big[\, \underline{\boldsymbol{a}} - \overline{\boldsymbol{b}},\, \overline{\boldsymbol{a}} - \underline{\boldsymbol{b}} \,\big]\,; \tag{2.11}$$

$$\boldsymbol{a} \cdot \boldsymbol{b} = \big[\, \min\{\underline{\boldsymbol{a}}\,\underline{\boldsymbol{b}}, \underline{\boldsymbol{a}}\,\overline{\boldsymbol{b}}, \overline{\boldsymbol{a}}\,\underline{\boldsymbol{b}}, \overline{\boldsymbol{a}}\,\overline{\boldsymbol{b}}\}\;,\; \max\{\underline{\boldsymbol{a}}\,\underline{\boldsymbol{b}}, \underline{\boldsymbol{a}}\,\overline{\boldsymbol{b}}, \overline{\boldsymbol{a}}\,\underline{\boldsymbol{b}}, \overline{\boldsymbol{a}}\,\overline{\boldsymbol{b}}\} \,\big]\,; \tag{2.12}$$

$$\boldsymbol{a}/\boldsymbol{b} = \boldsymbol{a} \cdot \big[\, 1/\overline{\boldsymbol{b}},\, 1/\underline{\boldsymbol{b}},\, \big] \qquad \text{для} \;\; \boldsymbol{b} \not\ni 0. \tag{2.13}$$

Множество всех интервалов вещественной оси с операциями сложения, вычитания, умножения и деления, которые определены посредством (2.10)–(2.13), называется *классической интервальной арифметикой*, и часто его обозначают также $\mathbb{IR}$.

Пример 2.2.2 (Расчет силы тока) Пусть максимальное напряжение в сети переменного тока находится в пределах интервала [220, 230] Вольт, а сопротивление нагревателя меняется, в пределах [20, 25] Ом. Каким будет ток в этом участке цепи?

Для расчета используем закон Ома, который дает выражение для тока в виде

$$I = \frac{U}{R},$$

где U — напряжение на участке цепи; R — ее сопротивление. Подставляя вместо этих переменных интервалы их изменения и заменяя операцию деления на интервальное деление (2.13), получим интервал значений максимального тока через нагреватель

$$\boldsymbol{I}_{\max} = \frac{[220, 240] \text{ Вольт}}{[20, 25] \text{ Ом}} \;=\; \big[\, \tfrac{220}{25},\, \tfrac{240}{20} \,\big] \text{ А} \;\approx\; [8{,}8, 12, 0] \text{ А}.$$

Это точный интервал значений тока, так как математическое выражение для него является простым и позволяет точно оценивать свою область значений с помощью классической интервальной арифметики. Для более сложных выражений оценка, полученная приведенным выше простым способом, может не быть равной области значений, но лишь содержит ее или, как часто говорят, является ее *внешней оценкой*.

Отметим важный частный случай интервального умножения, произведение числа на интервал:

$$a \cdot \boldsymbol{b} \;=\; \begin{cases} \left[a\underline{\boldsymbol{b}}, a\overline{\boldsymbol{b}}\right], & \text{если } a \geq 0, \\ \left[a\overline{\boldsymbol{b}}, a\underline{\boldsymbol{b}}\right], & \text{если } a \leq 0. \end{cases}$$

Алгебраические свойства интервальной арифметики являются необычными. Операции сложения и умножения не связаны друг с другом привычным соотношением дистрибутивности. Вместо него имеет место более слабая *субдистрибутивность*:

$$\boldsymbol{a}\,(\boldsymbol{b}+\boldsymbol{c}) \;\subseteq\; \boldsymbol{ab}+\boldsymbol{ac}.$$

Например, $[0,1]\cdot(1-1) = 0 \;\subset\; [-1,1] = [0,1]\cdot 1 + [0,1]\cdot(-1)$.

Интервальный вектор — это упорядоченный набор интервалов. Множество интервальных n-векторов, компоненты которых принадлежат $\mathbb{IR}$, обозначаем через $\mathbb{IR}^n$. Интервальные векторы называют также *брусами*, поскольку геометрическим образом векторов являются прямоугольные параллелепипеды с гранями, параллельными координатным осям в $\mathbb{R}^n$.

Интервальная матрица — это матрица с интервальными элементами, т. е. прямоугольная таблица, заполненная интервалами.

Интервальные векторы и матрицы являются специальным классом множеств в многомерных пространствах. С ними удобно работать, с их помощью оценивают другие множества, возникающие при решении математических задач. В этой связи чрезвычайно важно следующее понятие.

Определение 2.2.1 *Если S — непустое ограниченное множество в $\mathbb{R}^n$ или $\mathbb{R}^{m\times n}$, то его интервальной оболочкой $\square S$ называется наименьший по включению интервальный вектор (или матрица), содержащий S.*

2.3 Примеры интервальных расчетов

Обсуждение зависимости интервальных оценок областей значений выражений от их вида содержится в [4]. Имеет место «основная теорема» интервальной арифметики.

Теорема. *Пусть $f(x_1,\ \ldots,\ x_n)$ — рациональная функция вещественных аргументов $x_1,\ \ldots,\ x_n$ и для нее определен результат $\boldsymbol{f}_{\natural}(\boldsymbol{x}_1,\ \ldots,\ \boldsymbol{x}_n)$ подстановки вместо аргументов интервалов их изменения $\boldsymbol{x}_1,\ \boldsymbol{x}_2,\ \ldots,\ \boldsymbol{x}_n \in \mathbb{IR}$ и*

выполнения всех действий над ними по правилам интервальной арифметики. Тогда

$$\left\{ f(x_1, \ldots, x_n) \;\middle|\; x_1 \in \boldsymbol{x}_1, \ldots, x_n \in \boldsymbol{x}_n \right\} \subseteq \boldsymbol{f}_{\natural}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n), \qquad (2.14)$$

т. е. $\boldsymbol{f}_{\natural}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ *содержит множество значений функции* f *на брусе* $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. *Если выражение для* $f(x_1, \ldots, x_n)$ *содержит не более чем по одному вхождению каждой переменной в первой степени, то в* (2.14) *вместо включения выполняется точное равенство.*

Приведем пример интервальных расчетов с формулами, которые встречаются в естественнонаучных законах.

Пример 2.3.3 (Уравнение катализа.) Для описания зависимости скорости реакции, катализируемой ферментом, от концентрации субстрата, используется формула [39]:

$$v = V_{\max} \frac{S}{S + K_M}, \qquad (2.15)$$

где v — скорость реакции; $V_{\max}$ — максимальная скорость реакции; K_M — константа Михаэлиса; S — концентрация субстрата.

Для иллюстративного расчета зададимся значением $V_{\max} = 1$. Возьмем конкретную реакцию со справочным значением $K_M = 1,44 \cdot 10^{-4}$, а интервал концентрации $\boldsymbol{S}$ равным интервалу с серединой K_M и 10%-ным радиусом, т. е.

$$\boldsymbol{S} = [1,2959, 1,5841] \cdot 10^{-4}.$$

Для вычисления $\boldsymbol{v}$ выражение (2.15) представим двумя способами: в исходном виде и с делением числителя и знаменателя на S. Во втором случае переменная $\boldsymbol{S}$ входит в выражение (2.17) один раз, и согласно основной теореме интервальной арифметики результат естественного интервального оценивания совпадает с точной областью значений выражения:

$$\boldsymbol{v}_1 = V_{\max} \frac{\boldsymbol{S}}{\boldsymbol{S} + K_M} = [0,42857, 0,57895], \qquad (2.16)$$

$$\boldsymbol{v}_2 = V_{\max} \frac{1}{1 + K_M/\boldsymbol{S}} = [0,47368, 0,52381]. \qquad (2.17)$$

Средние величины интервалов (2.16) и (2.17) отличаются незначительно:

$$\operatorname{mid} \boldsymbol{v}_1 = 0,5038 \approx \operatorname{mid} \boldsymbol{v}_2 = 0,4987.$$

В то же время радиусы результатов вычислений $\boldsymbol{v}$ для выражений (2.16) и (2.17) существенно различны:

$$\operatorname{rad} \boldsymbol{v}_1 = 0,075188, \quad \operatorname{rad} \boldsymbol{v}_2 = 0,025063.$$

При этом имеет место соотношение:

$$\operatorname{rad} \boldsymbol{v}_1 > \operatorname{rad} \boldsymbol{v}_2$$

в силу неоднократного вхождения $\boldsymbol{S}$ в выражение (2.16).

2.4 Полная интервальная арифметика Каухера

Помимо классической интервальной арифметики часто возникает необходимость работать с полной интервальной арифметикой Каухера $\mathbb{KR}$. Она является алгебраическим и порядковым пополнением арифметики $\mathbb{IR}$, подобно тому, как множество целых чисел пополняет натуральный ряд.

Элементами арифметики $\mathbb{KR}$ являются пары чисел, взятые в квадратные скобки, вида $[\alpha, \beta]$, которые будем называть *интервалами*. При этом возможны ситуации, когда $\alpha \leq \beta$ или $\alpha > \beta$. Если $\alpha \leq \beta$, то $[\alpha, \beta]$ обозначает обычный интервал вещественной оси, его называют *правильным*. Если же $\alpha > \beta$, то $[\alpha, \beta]$ — *неправильный интервал*.

Таким образом,

$$\mathbb{IR} \subset \mathbb{KR}.$$

Неправильные интервалы можно рассматривать как математические абстракции (аналогичные отрицательным или мнимым числам), которым могут быть даны осмысленные физические интерпретации. В данном учебном пособии полная интервальная арифметика Каухера $\mathbb{KR}$ и неправильные интервалы, по существу, возникают при математической обработке интервальных результатов наблюдений и измерений.

Правильные и неправильные интервалы переходят друг в друга в результате отображения *дуализации* $\operatorname{dual} : \mathbb{KR} \to \mathbb{KR}$, меняющего местами концы интервала, т. е. такого, что

$$\operatorname{dual} \boldsymbol{a} := [\, \overline{\boldsymbol{a}}, \underline{\boldsymbol{a}} \,].$$

Правильной проекцией интервала $\boldsymbol{a}$ из $\mathbb{KR}$ называется интервал, обозначаемый $\operatorname{pro} \boldsymbol{a}$ и такой, что

$$\operatorname{pro} \boldsymbol{a} = \begin{cases} \boldsymbol{a}, & \text{если } \boldsymbol{a} \text{ — правильный,} \\ \operatorname{dual} \boldsymbol{a}, & \text{если } \boldsymbol{a} \text{ — неправильный.} \end{cases}$$

С помощью правильной проекции из произвольного интервала получается его «правильный образ», с которым можно обращаться как с обычным числовым интервалом в $\mathbb{R}$.

Арифметические операции между интервалами в $\mathbb{KR}$ продолжают операции в $\mathbb{IR}$. В частности, формулы (2.10), (2.11) для сложения и вычитания

также определяют сложение и вычитание в $\mathbb{KR}$. Умножение и деление между интервалами из $\mathbb{KR}$ определяется более сложно, и их описание можно найти в [4].

Наиболее важным в интервальной арифметике Каухера является обратимость арифметических операций. В частности, для любого интервала имеется противоположный ему, т. е. обратный по сложению. Для интервалов, не содержащих нуль, имеются обратные к ним по умножению. Для сложения (2.10) обратной операцией является не операция интервального вычитания (2.11), а операция «алгебраическое вычитание», которую обозначают знаком «$\ominus$»:

$$\boldsymbol{a} \ominus \boldsymbol{b} = [\underline{\boldsymbol{a}} - \underline{\boldsymbol{b}}, \overline{\boldsymbol{a}} - \overline{\boldsymbol{b}}]. \tag{2.18}$$

Иногда в математических текстах тем же символом обозначается так называемая «разность Хукухары» двух множеств (Hukuhara difference), но она имеет другие смысл и назначение. Нетрудно проверить, что для любых интервалов $\boldsymbol{a}$, $\boldsymbol{b}$ из $\mathbb{KR}$ имеют место равенства

$$\boldsymbol{a} \ominus \boldsymbol{a} = 0; \qquad (\boldsymbol{a} + \boldsymbol{b}) \ominus \boldsymbol{b} = \boldsymbol{a}; \qquad (\boldsymbol{a} \ominus \boldsymbol{b}) + \boldsymbol{b} = \boldsymbol{a}.$$

Пример 2.4.4 (Противоположный интервал) . Для интервала $[1, 2]$ противоположным по сложению является интервал $[-1, -2]$. Это неправильный интервал

$$[1, 2] + [-1, -2] = [1 - 1, 2 - 2] = [0, 0] = 0,$$

т. е. в сумме с исходным интервалом он дает нейтральный элемент 0. Отметим, что для обычного интервального вычитания

$$[1, 2] - [1, 2] = [1 - 2, 2 - 1] = [-1, 1].$$

Это иллюстрирует отмеченный ранее факт, что обычное интервальное вычитание не является операцией, обратной интервальному сложению.

Абсолютное значение интервалов из $\mathbb{KR}$ определяется как абсолютное значение их правильных проекций, т. е.

$$|\boldsymbol{a}| \;=\; \max\{\, |\underline{\boldsymbol{a}}|, |\overline{\boldsymbol{a}}| \,\}.$$

Полная интервальная арифметика Каухера $\mathbb{KR}$ пополняет классическую интервальную арифметику $\mathbb{IR}$ не только в алгебраическом смысле, но также и относительно естественного порядка по включению «$\subseteq$».

Определение 2.4.1 *Будем говорить, что для интервалов $\boldsymbol{a}$, $\boldsymbol{b} \in \mathbb{KR}$ выполняется включение $\boldsymbol{a} \subseteq \boldsymbol{b}$, если*

$$\underline{\boldsymbol{a}} \geq \underline{\boldsymbol{b}} \quad \text{и} \quad \overline{\boldsymbol{a}} \leq \overline{\boldsymbol{b}},$$

т. е. справедливы те же соотношения (2.6) *между концами интервалов, что и в случае классической интервальной арифметики.*

Относительно введенного таким образом отношения включения в $\mathbb{KR}$ для любых двух интервалов существует минимальный и максимальный по включению, т. е. результаты операций $\boldsymbol{a} \wedge \boldsymbol{b}$ и $\boldsymbol{a} \vee \boldsymbol{b}$ всегда определены.

Пример 2.4.5 (Минимум и максимум по включению в полной интервальной арифметике) :

$$[1,2] \wedge [3,4] = [3,2], \qquad [1,2] \vee [3,4] = [1,4].$$

Расстояние на множестве интервалов. Расстояние между интервалами $\boldsymbol{a}$ и $\boldsymbol{b}$ из $\mathbb{IR}$ или $\mathbb{KR}$ определяется как

$$\operatorname{dist}(\boldsymbol{a}, \boldsymbol{b}) \;=\; \max\bigl\{|\underline{\boldsymbol{a}} - \underline{\boldsymbol{b}}|, |\overline{\boldsymbol{a}} - \overline{\boldsymbol{b}}|\bigr\}. \tag{2.19}$$

Расстояние (2.19) обладает всеми свойствами абстрактного расстояния (метрики). Легко убедиться, что

$$\operatorname{dist}(\boldsymbol{a}, \boldsymbol{b}) \;=\; |\boldsymbol{a} \ominus \boldsymbol{b}|.$$

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

Справедливо также следующее равносильное представление расстояния (2.19) между интервалами:

$$\operatorname{dist}(\boldsymbol{a}, \boldsymbol{b}) \;=\; |\operatorname{mid} \boldsymbol{a} - \operatorname{mid} \boldsymbol{b}| + |\operatorname{rad} \boldsymbol{a} - \operatorname{rad} \boldsymbol{b}|.$$

Пример 2.4.6 (Расстояния между интервалами) . Рассмотрим интервал $[3,7]$ и точку 4 внутри него. Расстояние от этой точки, отождествляемой с вырожденным интервалом $[4,4]$, до данного интервала равно

$$\operatorname{dist}(4, [3,7]) = \max\bigl\{|4-3|, |4-7|\bigr\} = 3.$$

Рассмотрим дуальный интервал к интервалу $[3,7]$. Это интервал $\operatorname{dual}[3,7] = [7,3]$. Расстояние его до исходного интервала равно $\operatorname{dist}([3,7],[7,3]) = 4$.

Расстояние важно для определения отклонения интервалов друг от друга и, как следствие, для определения погрешности интервальных измерений. Полная интервальная арифметика реализована С. И. Жилиным на языке `Octave` [28].

2.5 Оценки и погрешности измерений

2.5.1 Оценки точечные и интервальные

Следуя [1], опишем два вида оценок в традиционной и интервальной статистиках. Оценки величин могут быть *точечными* или *интервальными*.

Точечные оценки, т. е. оценки в виде точек — чисел, векторов или матриц, — соответствуют, как правило, тому типу данных, который используется в модели рассматриваемого объекта или явления, и могут непосредственно использоваться при его дальнейшем исследовании, прогнозировании его поведения и т. п.

Интервальные оценки дают области возможных значений точечных оценок и нужны для характеризации их возможного разброса и изменчивости (*вариабельность*, см. пп. 3.5 и 3.5). Так как в традиционной вероятностной статистике оценки параметров являются случайными величинами, а носители их вероятностных распределений могут быть неограниченными, то при определении интервальных оценок обычно задают некоторый *уровень значимости* или *доверительной вероятности*, с помощью которых выполняют усечение вероятностного распределения. Тем самым всегда обеспечивается ограниченность интервальных оценок и их практичность.

В интервальном анализе данных оценки величин также могут быть *точечными* либо *интервальными* или даже иметь форму других множеств. Точечная оценка несет тот же смысл, что и в традиционной статистике, а интервальная оценка тоже дает область возможных значений точечных оценок, характеризуя их возможный разброс и вариабельность. Многомерные интервальные оценки удобнее всего брать в форме брусов.

Но есть и существенные отличия от вероятностной статистики. Во-первых, задание уровня значимости не требуется, так как множества значений оценки, как правило, являются ограниченными. Во-вторых, интервальные оценки могут иметь различных смысл — быть внутренними, внешними или какими-нибудь другими, сообразно чему их смысл различен. В-третьих, в пределах внутренней интервальной оценки все значения равноценны и тоже могут являться точечными оценками рассматриваемой величины. Напротив, в традиционной вероятностной статистике точечные значения внутри интервальной оценки не вполне равноценны друг другу.

2.5.2 Измерения и их результаты

Основным понятием теории обработки наблюдений является понятие *«измерения»* (*«наблюдения»*). Слово «измерение» имеет много значений. Оно может обозначать как процесс измерения или наблюдения, так и его результат. Из контекста обычно бывает ясно, какое значение слова имеется в виду [1].

Определение 2.5.1 *Измерением (замером, наблюдением) будем называть измеренное значение величины.*

По способу получения результата измерения все процессы измерения разделяются в [17] на *прямые*, *косвенные* и *совокупные*.

При прямых измерениях объект исследования приводят в непосредственное взаимодействие со средством измерений, которое выдает результат. При косвенных измерениях значение измеряемой величины находят на основании известной зависимости между измеряемой величиной и искомой величинами.

При совокупных измерениях значения искомых величин определяются из системы (совокупности) уравнений.

Приведенная классификация весьма условна. Следует отметить, что результат измерения является *итогом* какого-то физического эксперимента, в котором получаются первичные измерения, и *последующего применения некоторого способа математической обработки* первичных измерений.

На практике измерение (замер, наблюдение) может представлять собой вещественное число или интервал или же составленные из них многомерные объекты (вектор, матрицу, интервальный вектор, интервальную матрицу и т. п.). Вещественный тип данных для измерений является традиционным. Каким образом в результате измерений могут быть получены интервалы? Приведем ряд примеров.

Погрешности квантования. Это инструментальная погрешность, возникающая при преобразовании величины, принимающей непрерывный ряд значений, в цифровую форму, которая может принимать дискретный набор допустимых уровней. Значение преобразуемого аналогового сигнала заменяется ближайшим разрешенным уровнем цифрового сигнала, что дает погрешность квантования. Ее часто называют также *погрешностью оцифровки*.

Погрешности квантования присущи всем аналого-цифровым преобразователям. Если мы используем интервальный тип данных, интервальные результаты измерений, то результат представления непрерывного сигнала t, не равный точно какому-либо допустимому уровню t_0, t_1, $\ldots$, t_p, может быть записан как интервал $[t_i, t_{i+1}] \ni t$, и такое представление точное (см. п. 3.4).

Неопределенность измерения нуля. Согласно [31] *погрешностью нуля* называется погрешность средства измерений в контрольной точке, когда заданное значение измеряемой величины равно нулю. Следовательно, неопределенность измерений нуля — это неопределенность измерений, когда заданное значение измеряемой величины равно нулю (см. п. 3.4).

Агрегирование результатов многократных наблюдений. Во многих практических ситуациях измерение интересующей величины выполняется для надежности многократно. Тем не менее повторные измерения одних и тех же явлений не показывает в пределах точности измерений совпадений результатов. В данном случае 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 $x_1, x_2, \ \ldots, \ x_n$, то интервальным результатом следует взять

$$\boldsymbol{x} = \big[\min_{1 \le i \le n} x_i, \max_{1 \le i \le n} x_i \big].$$

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

Используя введенные выше (см. п. 2.2) операции взятия интервальной оболочки множества и максимума по включению (2.8), этот результат можно записать следующим равносильным образом:

$$\boldsymbol{x} = \square\{x_1, x_2, \ \ldots, \ x_n\}$$

или

$$\boldsymbol{x} = \bigvee_{1 \le i \le n} x_i.$$

Эти представления хороши тем, что могут быть обобщены на более сложные случаи (см. п. 3.4).

Погрешности измерений и наблюдений. 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 распространенных способов получения интервальных результатов в первичных измерениях — это обинтерваливание точечных значений, когда к точечному *базовому значению* $\mathring{x}$, которое считывается по показаниям измерительного прибора, прибавляется *интервал погрешности* $\boldsymbol{\epsilon}$:

$$\boldsymbol{x} = \mathring{x} + \boldsymbol{\epsilon}. \tag{2.20}$$

Интервал погрешности, вообще говоря, может быть произвольным, но если он уравновешен, т. е.

$$\boldsymbol{\epsilon} = [-\epsilon, \epsilon] \quad \text{для некоторого } \epsilon > 0,$$

то иногда для прямых измерений это можно трактовать, как отсутствие систематических погрешностей.

Пример 2.5.7 (Интервал показаний измерителя силы тока) . Предположим, что в процессе измерения силы тока мы смотрим на шкалу амперметра и считываем измеренное значение — 5,4 А. Класс точности используемого прибора — 2, и это, по определению, максимально допустимое значение

основной приведенной погрешности, выраженной в процентах. Следовательно, истинное значение измеряемого тока должно лежать в интервале

$$[5,4 - 0,02 \cdot 5, 4, 5,4 + 0,02 \cdot 5,4] \text{ А } = [5,292, 5,508] \text{ А}.$$

Интервал измерения строится с целью оценить истинное значение измеряемой величины, и получаемые при этом приближения могут быть качественно различными. Они могут включать (накрывать) истинное значение, но они также могут его и не содержать.

Выборка в вероятностной статистике — это часть генеральной совокупности элементов, которая охватывается экспериментом (наблюдением, опросом). В данной работе будем называть *выборкой* совокупность результатов измерений. Абстрактное понятие *генеральной совокупности*, которая представляет собой совокупность всех мыслимых (но реально не существующих) наблюдений интересующей нас величины при заданных условиях эксперимента, в анализе интервальных данных не используется.

Каждое измерение из выборки в случае неточности описывается своим интервалом неопределенности. Погрешности и неопределенности многомерных величин могут описываться интервальными векторами, которые обычно называют брусами.

Существуют также другие подходы к классификации интервальных данных, что может использоваться при выборе способа их обработки.

Классификация по ширине интервалов. Прежде всего, имеет смысл различать интервальные данные по ширине интервалов, т. е. по величине имеющейся у них неопределенности.

Если интервальные данные являются узкими, почти совпадая с точечными величинами, то для них интервальная специфика выражена слабо или вовсе не выражена. В некоторых ситуациях для их обработки можно даже применять те подходы и алгоритмы, которые используются для неинтервальных (точечных) данных. При этом «небольшая интервальность» узких интервалов позволяет выполнять с ними упрощенные, но достаточно точные приемы обработки, основанные на асимптотических разложениях, пренебрежении членами высокого порядка и т. п.

При увеличении ширины интервальных данных их уже нельзя рассматривать как «приближенно точечные», они становятся «существенно интервальными», но несут некоторые черты, присущие точечным данным. Отсутствие пересечений интервальных измерений выборки является признаком того, что интервальные данные все еще «не слишком широки» и не слишком сильно отличаются от точечных данных.

Наконец, при дальнейшем увеличении ширины интервальных измерений в выборке они начинают пересекаться друг с другом, и это служит признаком следующего качественного состояния, — когда интервальные данные являются «широкими интервальными», т. е. интервальная неопределенность

велика. Этот случай является специфически интервальным, к которому точечные методы обработки данных уже принципиально неприменимы.

Классификация по способу измерения. В [15] вводится классификация интервальных данных по способу их получения — с помощью одного или нескольких измерительных устройств.

С учетом различных способов измерений и различной точности измерительных инструментов нужно по-разному применять прием варьирования неопределенности в выборке (см. п. 3.5).

2.5.3 Более сложные типы интервалов

Для описания данных разработана теория и более сложных типов интервалов. Популярное описание двух таких типов, твинов и мультиинтервалов, дано в [3].

Твины — интервалы интервалов. На практике концы интервалов, представляющие результаты измерений, могут быть известны неточно, поэтому возникает необходимость работы с интервалами, имеющими интервальные концы. В интервальном анализе такие объекты называются *твинами* (по-англ. twin, сокращение фразы twice interval, «двойной интервал»). Развернутое изложении теории твинов дано в диссертации [32].

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

$$\boldsymbol{X} = [\boldsymbol{a}, \boldsymbol{b}] = \big[\, [\underline{\boldsymbol{a}}, \overline{\boldsymbol{a}}], [\underline{\boldsymbol{b}}, \overline{\boldsymbol{b}}] \,\big]. \tag{2.21}$$

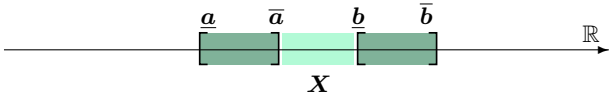

Рис. 2.2. Твины на вещественной оси

На рис. 2.2 твин $\boldsymbol{X}$ представлен в графической форме. Концы твина, т. е. интервалы $\boldsymbol{a}$ и $\boldsymbol{b}$, представлены более темной заливкой, чем остальная часть твина.

Твин является множеством всех интервалов, больших или равных $[\underline{a}, \overline{a}]$ и меньших или равных $[\underline{b}, \overline{b}]$, 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.

Пример 2.5.8 (Измерение температуры термометром сопротивления в виде твина) . 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

$$\Delta t = \pm 0,35 \text{ ° C}.$$

Пусть измеряемая температура находится в диапазоне [19, 5, 20, 5] ° C, которую представим как интервал $\boldsymbol{t}$:

$$\boldsymbol{t} = [19,5, 20,5] \text{ ° C}. \tag{2.22}$$

Представим границы $\underline{\boldsymbol{t}}$, $\overline{\boldsymbol{t}}$ интервала $\boldsymbol{t}$ как интервалы. С учетом систематической погрешности твин температур $\boldsymbol{T}$, даваемый датчиком, составит

$$\boldsymbol{T} = \big[\,[19,15, 19,85],\ [20,15, 20,85]\,\big] \text{ ° C}. \tag{2.23}$$

Графическое представление твина $\boldsymbol{T}$ (2.23) дано на рис. 2.3. Форма записи

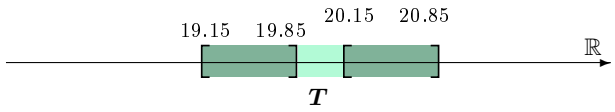

Рис. 2.3. Температура как твин.

температуры в виде твина $\boldsymbol{T}$ (2.23) четко и полно представляет информацию об измеряемых данных. В случае если концы интервала в выражении (2.22) могут меняться независимо, возможны различные ситуации. Например, 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.

Согласно определению, приведенному в [4], *мультиинтервал* — это объединение конечного числа несвязных интервалов числовой оси (рис. 2.4).

Рис. 2.4. Мультиинтервал в $\mathbb{R}$.

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

на множестве интервалов. Мультиинтервалы можно получать при решении уравнений и систем уравнений.

Пример 2.5.9 (Пример мультиинтервалов и их преобразований) Приведем модельный пример, в котором появляются неодносвязные интервалы. Рассмотрим задачу нахождения корня уравнения второй степени с различными значениями параметра a.

$$a \cdot x^2 = [0,5, 1,5] \tag{2.24}$$

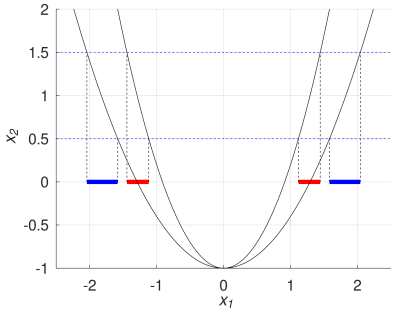

Рис. 2.5. Решение уравнения (2.24)
с разными значениями параметра a.

Возьмем для определенности значения $a = 0,6$ и $a = 1,2$. Получим решения уравнения (2.24):

$$a = 0,6: \quad \boldsymbol{X}_1 = [\ [-2,04, -1,58],\ [1,58, 2,04]\ ],$$
$$a = 1,2: \quad \boldsymbol{X}_2 = [\ [-1,44, -1,12],\ [1,12, 1,44]\ ].$$

Мультиинтервалы $\boldsymbol{X}_1, \boldsymbol{X}_2$ показаны на рис. 2.5 соответственно парами отрезков синего и красного цвета. При изменении коэффициента при старшей степени полинома компоненты мультиинтервалов — решений уравнения (2.24) меняют и размер, и положение на вещественной оси.

2.6 Описание измерений

2.6.1 Накрывающие и ненакрывающие измерения

Результат измерения интересующей нас величины может получиться либо равным, либо не равным ее истинному значению. В случае измере-

ния непрерывных физических величин, принадлежащих вещественному типу данных, равенство является исключительным событием, неустойчивым к сколь угодно малым возмущениям или погрешностям в вычислительных алгоритмах.

Принципиально другая ситуация возникает, если результат измерения может быть интервалом. Невырожденный интервал по своей сути является представительным множеством на вещественной оси (имеющим ненулевую меру), и оно, как правило, устойчиво к малым возмущениям и погрешностям вычислений. Для обработки интервальных данных фундаментальный характер имеет следующее определение ([1], [40]):

Определение 2.6.1 *Накрывающее измерение (накрывающий замер) — это интервальный результат измерения, который гарантированно содержит истинное значение измеряемой величины. Измерение, для которого нельзя утверждать, что оно содержат истинное значение измеряемой величины, будем называть ненакрывающим (рис.* 2.6 *и* 2.7*).*

Отметим, что с точки зрения формальной логики понятия накрывающего и ненакрывающего измерений являются противоположными, но при этом не противоречащими.

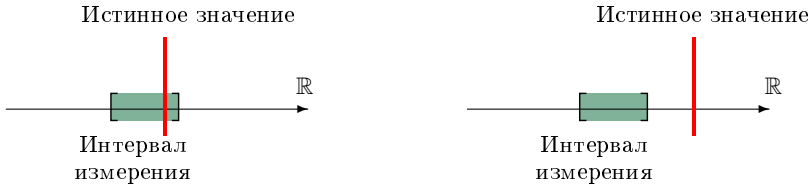

Рис. 2.6. Накрывающее (слева) и ненакрывающее (справа) измерения точечного истинного значения величины

Накрывающее измерение является гарантированной двусторонней вилкой значений измеряемой величины, тогда как для ненакрывающего измерения подобное утверждать нельзя. При перенесении свойства накрытия истинного значения на выборки простейший путь — объявить накрывающей выборкой совокупность накрывающих измерений, тогда как выборки, в которых присутствует хотя бы одно ненакрывающее измерение, станут ненакрывающими. Погрешности и выбросы (промахи) неотъемлемо присутствуют в данных, и проверка свойства «накрытия истинного значения» является нетривиальной.

Далее мы будем называть *накрывающей выборкой* совокупность измерений, в которой доминирующая часть (большинство и т. п.) измерений (наблюдений) являются накрывающими.

Рис. 2.7. Накрывающее (слева) и ненакрывающее (справа) измерения интервального истинного значения величины

Напротив, выборка называется *ненакрывающей*, если преобладающая часть входящих в нее измерений ненакрывающие.

Данное определение нестрогое и использует расплывчатые понятия «большинство», «доминирующая часть» и т. п., которые должны уточняться каждый раз в процессе применения.

Важность введенных понятий обусловлена тем, что накрывающее (охватывающее) измерение дает не только приближение к интересующему истинному значению физической величины, но и двустороннюю оценку этого значения, т. е. его гарантированные оценки снизу и сверху. Это обстоятельство позволяет привлечь для обработки накрывающих измерений более сильные средства, качественно другой математический аппарат (в частности, некоторые специфичные методы интервального анализа) и получить в результате уточненные оценки для истинного значения также в виде двусторонней оценки. Для ненакрывающих измерений и выборок это не всегда достижимо.

Тот факт, что интервальный результат измерения не является накрывающим, может быть вызван различными причинами: измерение может оказаться выбросом, погрешность измерения недооценена, неадекватность выбранной модели объекта (непостоянство во времени, выбор неверной функциональной зависимости).

Проверка того, является ли данное измерение или выборка накрывающими, находится вне рамок математической теории интервальных измерений, и решается в каждом случае конкретно. Для традиционных точечных измерений аналога введенных понятий не существует, так как все точечные измерения, как правило, ненакрывающие.

Для достижения свойства накрытия нередко прибегают к специальным приемам в процессе предобработки данных (см. пп. 3.5 и 4.5.1).

2.6.2 Информационное множество

Данные измерений, которые были описаны ранее, можно называть *первичными*, так как чаще всего они подвергаются дальнейшей обработке. Та-

ким образом, для определения окончательного результата измерения необходимо дополнить наши конструкции моделью обработки данных (способом обработки данных). Это математическая модель, формализующая требования к результату обработки измерения и оформленная в виде системы уравнений, задачи оптимизации и т. п., которая определяет то, что должно считаться результатом обработки измерений.

Информационное множество для интервальных данных — это множество значений параметров, удовлетворяющих математической системе отношений, полученной в результате агрегирования информации о математической модели объекта, первичных данных измерений и модели их обработки. Информационное множество зависит от выбранной модели обработки данных, и потому даже для одних и тех же данных может быть определено неединственным образом в зависимости от того, как эти данные обрабатываются и интерпретируются.

Пример 2.6.10 (Различные походы к задаче восстановления зависимости.) Предположим, что мы решаем задачу восстановления зависимости некоторого заданного вида по данным измерений. Эта зависимость может восстанавливаться, например, методом наименьших квадратов (МНК), методом наименьших модулей (МНМ) или с помощью чебышевского (минимаксного) сглаживания. Перечисленные методы представляют собой разные модели обработки данных.

Для одних и тех же данных измерений, т. е. первичных данных, итоговый результат измерения будет разным в зависимости от того, какая именно методика их обработки применяется — МНК, МНМ или минимаксное приближение. Следовательно, мы получим три различных информационных множества, которые в обычном случае неинтервальных данных, скорее всего, будут одноточечными множествами. (См. п. 4.6.19).

Неформально говоря, *информационное множество — это множество параметров задачи, которые совместны с данными измерений в рамках выбранной модели их обработки.* В гл. 3 и 4 приведены конкретные определения информационных множеств, возникающих в задаче оценивания постоянной величины и в задаче восстановления линейной зависимости.

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.

2.7 Выбросы и промахи

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

Другое популярное определение выбросов (промахов) состоит в том, что это результаты измерений, которые для данных условий резко отличаются от остальных результатов общей выборки. Выбросы нарушают некоторую однородность (согласованность, непротиворечивость), характерную для большинства наблюдений выборки по отношению к заданной математической модели (см. п. 4.5).

Оба приведенных определения неформальны, так как, по-видимому, одно формальное определение для данного важнейного понятия дать нельзя.

Как правило, выбросы стремятся удалить из выборки на этапе ее предварительной обработки (предобработки), т.е. перед применением формальных математических методов, так как присутствие выбросов существенно искажает оценки истинных значений параметров. Выявление выбросов является нетривиальной и, как правило, труднoформализуемой процедурой, которая опирается на опыт и т. п. Для вероятностной статистики выявление выбросов входит необходимой составной частью в обработку данных, а некоторые процедуры даже рекомендованы в стандартах [20].

Что считать выбросом в случае интервальных результатов измерений? 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 п. 1.3.2. При стремлении ширины интервальных измерений к нулю они переходят в точечные измерения, которые, как правило, всегда ненакрывающие.

Если априори известно, что измерение, производимое данным инструментом с помощью некоторой определенной методики должно быть накрывающим, то получение ненакрывающего результата является признаком выброса (промаха).

Более подробное выбросы и промахи, а также методики их выявления будут подробнее в гл. 3 и 4.

Глава 3

Измерение постоянной величины

3.1 Выборка измерений и интервалы их неопределенности

Постоянная величина — это величина, которая в рассматриваемом процессе сохраняет свое значение неизменным. Например, рост человека не меняется заметно в процессе его измерения, и поэтому может считаться постоянной величиной, хотя на протяжении жизни человека рост, конечно же, непостоянен.

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, с помощью какой-либо из процедур, описанных в предыдущих главах. 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 называть *длиной выборки* (или объемом выборки). По интервальным результатам измерений или наблюдений требуется найти оценку для интересующей нас величины.

Табличные данные В качестве источника данных для ряда примеров будем использовать [33]. В Табл. 3.1 воспроизведена часть данных из [33].

Для наглядного представления выборки часто чертят образующие ее интервалы в виде графика, изображенного на рис. 3.1, который по статистиче-

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 3.1. Данные табл. 1 для величины $\delta \times 10^5$ [33]

ской традиции называют *диаграммой рассеяния* (см. также рис. 3.2 и 3.7). Можно повернуть картинку и представлять интервалы данных горизонтально (см. рис. 3.4).

Значения rad $\boldsymbol{x}_k$, $k = 1, 2, \ldots, n$, показывают величины интервальной неопределенности отдельных измерений выборки. Величину неопределенности всей выборки характеризует вектор радиусов

$$\text{rad}\, \boldsymbol{x} = (\text{rad}\, \boldsymbol{x}_1, \text{rad}\, \boldsymbol{x}_2, \ \ldots, \ \text{rad}\, \boldsymbol{x}_n)$$

. Но часто такая детальность не требуется в представлении неопределенности выборки, а нужна какая-либо одна величина, которая агрегированным образом представляет данную неопределенность. В этом случае можно взять какую-либо норму вектора rad $\boldsymbol{x}$.

По аналогии с традиционной метрологией будем называть измерения выборки *равноширинными*, если неопределенность всех этих измерений одинакова, т. е. rad $\boldsymbol{x}_k = r = \text{const}$, $k = 1, \ \ldots, \ n$. Напротив, *неравноширинными* (разноширинными) называем измерения, в которых величина неопределенности rad $\boldsymbol{x}_k$ может меняться в зависимости от измерения выборки, $k = 1, \ \ldots, \ n$. Фактически эти термины означают «имеющие равную неопределенность» и «имеющие неодинаковые неопределенности».

Информационным множеством в случае оценивания единичной постоянной величины по выборке интервальных данных будет также интервал, который будем называть *информационным интервалом*. Неформально говоря, это интервал, содержащий значения оцениваемой величины, которые совместны с измерениями выборки (согласуются с данными этих измерений).

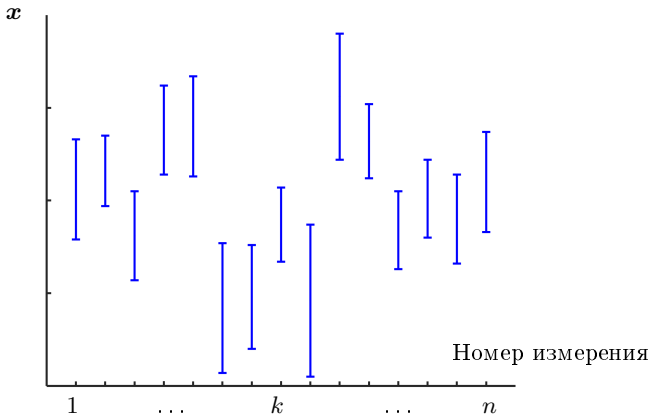

Рис. 3.1. Диаграмма рассеяния интервальных измерений постоянной величины

Но 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.

3.2 Обработка накрывающих выборок

Если истинное значение величины содержится во всех интервалах измерений выборки $\{ \boldsymbol{x}_k \}_{k=1}^{n}$, то оно должно принадлежать также пересечению этих интервалов. Следовательно, уточненным интервалом принадлежности истинного значения может быть объединение

$$\boldsymbol{I} = \bigcap_{1 \le k \le n} \boldsymbol{x}_k. \tag{3.2}$$

Это и будет информационным множеством $\boldsymbol{I}$ оценки измеряемой физической величины (см. рис. 3.2) — *информационный интервал*. Явные выражения для его левой (нижней) и правой (верхней) границ выражены следующими формулами:

$$\underline{\boldsymbol{I}} = \max_{1 \le k \le n} \underline{\boldsymbol{x}}_k; \qquad\qquad \overline{\boldsymbol{I}} = \min_{1 \le k \le n} \overline{\boldsymbol{x}}_k. \tag{3.3}$$

В силу сделанного допущения о том, что выборка накрывает истинное значение величины, имеем $\underline{\boldsymbol{I}} \le \overline{\boldsymbol{I}}$. При этом заслуживает внимания предельный случай совместной выборки, когда $\underline{\boldsymbol{I}} = \overline{\boldsymbol{I}} = x^*$. Тогда выборка совместна,

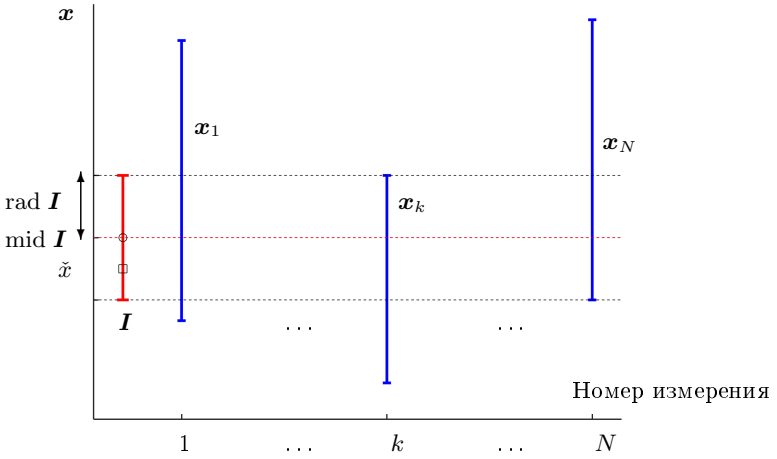

Рис. 3.2. Обработка накрывающей выборки интервальных измерений величины

но на пределе совместности, и информационный интервал $\boldsymbol{I}$ вырождается в точку.

Если известен некоторый априорный интервал возможных значений оцениваемой постоянной величины $\boldsymbol{I}_{\text{апр}} = [\underline{\boldsymbol{I}}_{\text{апр}}, \overline{\boldsymbol{I}}_{\text{апр}}]$, который должен гарантированно содержать ее, то границы результирующего интервала (3.2) могут быть уточнены пересечением

$$\boldsymbol{I} = \boldsymbol{I} \cap \boldsymbol{I}_{\text{апр}}. \tag{3.4}$$

Отметим, что априорный интервал $\boldsymbol{I}_{\text{апр}}$ может задавать одностороннее ограничение, если он имеет вид $[\underline{\boldsymbol{I}}_{\text{апр}}, +\infty]$ или $[-\infty, \overline{\boldsymbol{I}}_{\text{апр}}]$.

На практике часто необходимо работать не с интервалами интересующей нас величины — (3.2) или (3.4), а с некоторой точечной оценкой $\check{x}$. Все точки информационного интервала равноценны друг другу, поэтому точечную оценку $\check{x}$ можно выбирать произвольно (см. рис. 3.2). 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},

$$x_{\text{c}} \;=\; \operatorname{mid} \boldsymbol{I} \;=\; \tfrac{1}{2}\left(\underline{\boldsymbol{I}} + \overline{\boldsymbol{I}}\right). \tag{3.5}$$

Середина интервала обладает определенной оптимальностью, являясь точкой, которая наименее удалена от других точек этого интервала.

Пример 3.2.11 (Обработка накрывающей выборки.) Выберем из данных табл. 3.1 накрывающую подвыборку. Это замеры с номерами

$$\{1, 2, 3, 8, 12, 13, 14, 15\}. \tag{3.6}$$

Диаграмма рассеяния выборки (3.6) приводится на рис. 3.3. Также на рисунке приведены оценки границы информационного множества (3.3). Численно оценки границ 'этого информационного множества в единицах 10^{-5} составляют

$$\underline{\boldsymbol{I}} = \max_{1 \le k \le n} \underline{\boldsymbol{x}}_k = -5,3; \qquad \overline{\boldsymbol{I}} = \min_{1 \le k \le n} \overline{\boldsymbol{x}}_k = -4,5.$$

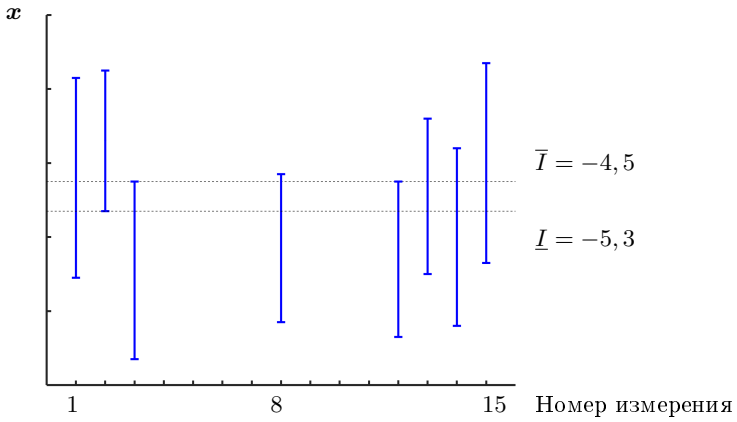

Рис. 3.3. Пример обработки накрывающей выборки интервальных измерений

Центральная оценка (3.5) в единицах 10^{-5} равна

$$x_{\mathrm{c}} = \operatorname{mid} \boldsymbol{I} = \tfrac{1}{2}\left(\underline{\boldsymbol{I}} + \overline{\boldsymbol{I}}\right) = -4,9.$$

В дальнейшем сравним полученные оценки с другими способами оценивания по полной ненакрывающей выборке табл. 3.1.

3.3 Оценки интервальных выборок

В п. 3.3 будут введены различные оценки интервальных выборок, рассмотрены конкретные примеры и взаимные отношения различных мер.

3.3.1 Мода интервальной выборки

В традиционной статистике важной характеристикой выборки является ее *мода* — значение из выборки, которое встречается наиболее часто. Для непрерывного вероятностного распределения мода — точка с наибольшей плотностью вероятности.

Имеет смысл распространить понятие моды на обработку интервальных данных, где оно будет обозначать интервал тех значений, которые встречаются в интервалах обрабатываемых данных наиболее часто. Фактически это означает, что точки из моды интервальной выборки накрываются наибольшим числом интервалов этой выборки. Ясно, что по самому своему определению понятие моды имеет содержательный смысл лишь для накрывающих выборок. Следуя [34], введем следующее определение

Определение 3.3.1 *Модой интервальной выборки назовем интервал пересечения ее наибольшей совместной подвыборки.*

Псевдокод алгоритма для нахождения моды выборки интервальных измерений приведен в табл. **??**.

Пример 3.3.12 (Пример вычисления моды интервальной выборки.) Рассмотрим пример вычисления моды интервальной выборки. Пусть имеется интервальная выборка из четырех элементов

$$\boldsymbol{X} = \{\ [1,4], [5,9], [1,5,4,5], [6,9]\ \}. \tag{3.7}$$

Диаграмма рассеяния выборки $\boldsymbol{X}$ приведена на рис. 3.4.

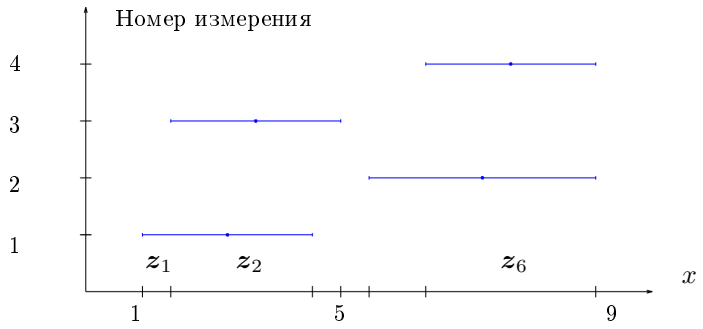

Рис. 3.4. Диаграмма рассеяния интервальной выборки (3.7) и элементы выборки $\boldsymbol{z}$

Алгоритм нахождения моды
интервальной выборки

Вход

Интервальная выборка $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^{n}$ длины n.

Выход

Мода $\operatorname{mode}\boldsymbol{X}$ выборки $\boldsymbol{X}$ и ее частота μ.

Алгоритм

$\boldsymbol{I} \leftarrow \bigcap_{i=1}^{n} \boldsymbol{x}_i$;

IF $\boldsymbol{I} \neq \varnothing$ THEN

$\operatorname{mode}\boldsymbol{X} \leftarrow \boldsymbol{I}$;

$\mu \leftarrow n$

ELSE

объединяем все концы $\underline{\boldsymbol{x}}_1$, $\overline{\boldsymbol{x}}_1$, $\underline{\boldsymbol{x}}_2$, $\overline{\boldsymbol{x}}_2$, $\ldots$, $\underline{\boldsymbol{x}}_n$, $\overline{\boldsymbol{x}}_n$ интервалов рассматриваемой выборки $\boldsymbol{X}$ в один массив $Y = \{y_1, y_2, \ldots, y_N\}$, где $N \leq 2n$;

упорядочиваем элементы Y по возрастанию значений;

порождаем интервалы $\boldsymbol{z}_i = [y_i, y_{i+1}]$, $i = 1, 2, \ldots, N-1$;

для каждого $\boldsymbol{z}_i$ подсчитываем число μ_i интервалов из выборки $\boldsymbol{X}$, включающих интервал $\boldsymbol{z}_i$;

вычисляем $\mu \leftarrow \max\limits_{1 \leq i \leq N-1} \mu_i$;

выбираем номера k интервалов $\boldsymbol{z}_k$, для которых μ_k равно максимальному, т. е. $\mu_k = \mu$, и формируем из таких k множество $K = \{k\} \subseteq \{1, 2, \ldots, N-1\}$;

$\operatorname{mode}\boldsymbol{X} \leftarrow \bigcup_{k \in K} \boldsymbol{z}_k$

END IF.

В соответствии с алгоритмом **??**, проверим совместность $\boldsymbol{X}$. Пересечение элементов выборки пусто

$$\boldsymbol{I} = \bigcap_{i=1}^{n} \boldsymbol{x}_i = \varnothing.$$

Таким образом, необходимо выполнить шаги алгоритма после ключевого слова ELSE. Сформируем массив интервалов $\boldsymbol{z}$ из концов интервалов $\boldsymbol{X}$:

$$\boldsymbol{z} = \{\ [1,0,1,5], [1,5,4,0], [4,0,4,5], [4,5,5,0], [5,0,6,0], [6,0,9,0], [9,0,9,0]\ \}. \tag{3.8}$$

Мощность N массива $\boldsymbol{z}$ равна 7.

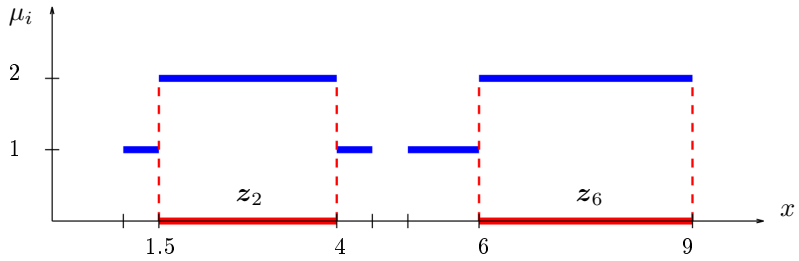

Рис. 3.5. Значения частот μ_i, интервальная мода $\operatorname{mode}\boldsymbol{X}$ выборки (3.7) и элементы выборки $\boldsymbol{z_k} : k \in K$

Для каждого интервала $\boldsymbol{z}_i$ подсчитываем число μ_i интервалов из выборки $\boldsymbol{X}$, включающих $\boldsymbol{z}_i$, получаем массив μ_i в виде

$$\{1,2,1,0,1,2,2\}. \tag{3.9}$$

Максимальные μ_i, равные 2, достигаются для индексного множества

$$K = \{2,6,7\},$$

поэтому частота моды равна $\mu = 2$. Как итог, мода является мультиинтервалом (см. п. 2.5.3)

$$\operatorname{mode}\boldsymbol{X} = \bigcup_{k\in K} \boldsymbol{z}_k = [1,5,4,0] \cup [6,0,9,0]. \tag{3.10}$$

На Рис. 3.5 значения частот μ_i (3.9) показаны синим цветом, а интервальная мода $\operatorname{mode}\boldsymbol{X}$ (3.10) — красным цветом.

Выборки унимодальные и мультимодальные. Тот факт, что выборка не является унимодальной, может служить признаком сложной внутренней структуры описываемого ею явления. Исследуемая величина может,

к примеру, не быть постоянной, а является композицией нескольких близких постоянных величин. Примером может быть природное распределение изотопов ртути (п. 1.3.1, рис. 1.2).

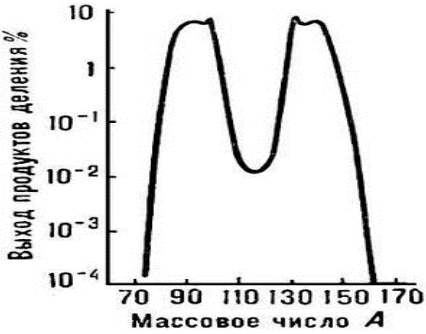

Рис. 3.6. Бимодальное распределение масс осколков в делении ядра урана [35]

Так как выборка, очевидно, является своей подвыборкой, то понятие моды совпадает с пересечением всех интервалов выборки в случае ее совместности. Если же выборка несовместна, то мода может быть мультиинтервалом. Это аналогично ситуации с обычными неинтервальными данными, где мод у выборки или у распределения может быть несколько.

При обработке неинтервальных (точечных) данных распределения вида, показанного на рис. 3.6, обоснованно считаются уже не унимодальными, так как имеют более одного явно выраженного пика, хотя и разной высоты. В таком случае требуется дополнительная обработка не только основной моды, но и более слабых.

3.3.2 Медиана интервальной выборки

В изложении следуем материалу [36]. Для вариационного ряда $\{x_i\}_{i=1}^n, x_i \in \mathbb{R}$ существует несколько определений медианы. Приведем две из них:

Медиана это такое значение (члена вариационного ряда), для которого

M1 половина членов ряда (с учетом их частот) лежит слева, а половина - справа от него

M2 минимальна сумма расстояний от него до других членов ряда с учетом их частот.

В качестве интервальной медианы для выборки $\{\boldsymbol{x}_i\}_{i=1}^n, \boldsymbol{x}_i \in \mathbb{IR}$ в [36] предлагается использовать следующие определения как аналог определениям M1 и M2:

Интервальная медиана — это интервал $\boldsymbol{r}_m$ со средней (геометрически) накопленной частотой, т.е. сумма накопленных частот слева равна сумме накопленных частот справа:

$$\sum_{i=1}^{m-1} f_i = \sum_{i=m+1}^{n} f_i, \tag{3.11}$$

где f_i — частота интервала $\boldsymbol{r}_i$ — количество интервалов из заданного вариационного ряда, в которых содержится $\boldsymbol{r}_i$. Если оказалось так, что

$$\sum_{i=1}^{m} f_i = \sum_{i=m+1}^{n} f_i,$$

то интервальная медиана вычисляется по формуле

$$\operatorname{med}(\boldsymbol{X}) = \frac{\boldsymbol{r}_m + \boldsymbol{r}_{m+1}}{2}. \tag{3.12}$$

Интервальная медиана — это интервал $\boldsymbol{r}_m$ такой, что выполнено:

$$\sum_{i=1, i\neq m}^{n} \rho(\boldsymbol{r}_m, \boldsymbol{x}_i) = \min_{\{\boldsymbol{r}_j\}} \sum_{i=1, i\neq j}^{n} \rho(\boldsymbol{r}_j, \boldsymbol{x}_i), \tag{3.13}$$

где ρ хаусдорфово расстояние $\rho(\boldsymbol{a}, \boldsymbol{b})$ между интервалами $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{IR}$.

При этом интервальная медиана, вычисленная по формулам (3.11), (3.12), может отличаться от интервальной медианы, вычисленной по формуле (3.13).

Конструктивное построение интервальной медианы интервального вариационного ряда во многом сходно с построением интервальной моды, рассмотренной в п. 3.3.1. В табл. 3.2 приведен алгоритм для нахождения медианы интервальной выборки.

Стоит отметить, что в разбиение $\boldsymbol{r}_i$ интервалов из ряда $\{\boldsymbol{x}_i\}_{i=1}^n$ могут войти интервалы, которых нет в исходных данных. В таком случае интервальной медианой может оказаться интервал, не входящий в исходную выборку. Если в качестве медианы нужен интервал, который будет пересекаться с какими-либо интервалами из исходных данных, то можно провести регуляризацию данных. Тогда в качестве интервальной медианы можно будет взять интервальную медиану, построенную для регуляризованных данных.

Пример 3.3.13 (Пример медианы интервальной выборки.) Пусть имеется интервальная выборка

$$\boldsymbol{X} = \{\ [5, 10], [3, 9]\, [1, 4]\ \}. \tag{3.14}$$

Таблица 3.2. Алгоритм для нахождения медианы интервальной выборки

Вход

Интервальная выборка $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^{n}$ длины n.

Выход

Медиана $\mathrm{med}(\boldsymbol{X})$ выборки $\boldsymbol{X}$.

Алгоритм

объединяем все концы $\underline{\boldsymbol{x}}_1$, $\overline{\boldsymbol{x}}_1$, $\underline{\boldsymbol{x}}_2$, $\overline{\boldsymbol{x}}_2$, $\ldots$, $\underline{\boldsymbol{x}}_n$, $\overline{\boldsymbol{x}}_n$
интервалов рассматриваемой выборки $\boldsymbol{X}$ в один
массив $Y = \{y_1, y_2, \ldots, y_N\}$, где $N \leq 2n$;
упорядочиваем элементы Y по возрастанию значений;
порождаем интервалы $\boldsymbol{z}_i = [y_i, y_{i+1}]$, $i = 1, 2, \ldots, N-1$;
для каждого $\boldsymbol{z}_i$ подсчитываем число μ_i интервалов
из выборки $\boldsymbol{X}$, включающих интервал $\boldsymbol{z}_i$;
для каждого $\boldsymbol{z}_i$ подсчитываем сумму частот слева и справа ;
выбираем интервал $\boldsymbol{z}_k$, для которого выполнено условие (3.11)
или
для каждого $\boldsymbol{z}_i$ подсчитываем
сумму расстояний до элементов выборки ;
выбираем интервал $\boldsymbol{z}_k$, для которого выполнено условие (3.13) ;

$\mathrm{med}(\boldsymbol{X}) \leftarrow \boldsymbol{z}_k$.

Найдем ее медиану двумя способами, следуя алгоритму табл. 3.2.

Используя первый способ, построим массив $\boldsymbol{z}_i$ из концов интервалов выборки $\boldsymbol{X}$ (3.14):

$$\boldsymbol{z} = \{\ [1,3],[3,4],[4,5],[5,9],[9,10]\ \} \tag{3.15}$$

и на его основе — массив частот

$$\{\mu\} = \{1,2,1,2,1\}.$$

Согласно (3.11) имеем

$$m = 3, \quad \mathrm{med}(\boldsymbol{X}) = \boldsymbol{z}_3 = [4,5].$$

По второму способу, по формуле (3.13) ищем $\boldsymbol{z}_i$, наименее удаленный от интервалов исходной выборки. Массив расстояний

$$\{d\} = \{14,13,12,8,18\}.$$

Согласно (3.13) имеем

$$m = 4; \quad \mathrm{med}(\boldsymbol{X}) = \boldsymbol{z}_4 = [5,9].$$

В данном примере медианы, вычисленные на основе частот и с учетом хаусдорфовых расстояний, оказались различны.

3.3.3 Мера совместности интервальной выборки

Для описания выборок, помимо оценок их размеров, желательно иметь дополнительную информацию о мере сходства элементов выборки. В п. 2.5.2 был рассмотрен вопрос о классификации выборок в зависимости от соотношения ширин интервалов в выборке по отношению к их полной вариабельности. При определении накрывающих выборок в п. 2.6.1 отмечалось, что понятие невозможно определить строго, поскольку жесткие требования к «накрытию» приводят к исключению из рассмотрения подавляющего большинства практических ситуаций.

В различных областях анализа данных в науках о Земле, биологии, информатике используют множество мер сходства множеств [37]. Мера сходства бинарная: $S(A,B) \to [0,1]$ — это вещественная функция между объектами A, B. Формально принадлежность к мерам сходства определяется системой аксиом:

- ограниченность $0 \le S(A,B) \le 1$;
- симметрия $S(A,B) = S(B,A) \le 1$;
- рефлексивность $S(A,B) = 1 \iff A = B$;
- транзитивность $A \subseteq B \subseteq C \Longrightarrow S(A,B) \ge S(A,C)$.

Эти свойства также называют t-нормой. Существуют и иные системы аксиом сходства. В компьютерных приложениях (обработка изображений, машинное обучение) меру сходства множеств обозначают как *IoU* (*Intersection over Union*). В математике часто используют наименование *индекс Жаккара*, по имени математика, предложившего подобную меру.

По мере развития интервального анализа были введены различные определения и конструкции оценки меры совместности интервальных объектов. Вместе с тем в практике обработки данных часто необходимо оперировать относительными величинами. В частности, это нужно в связи с необходимостью сопоставления допусков и размеров деталей, погрешности измерителей и значений измеряемых величин и т.п. [38].

Введем базовую конструкцию совместности для двух интервалов. Для иллюстрации идеи рассмотрим следующую числовую характеристику степени совместности двух интервалов $\boldsymbol{x}, \boldsymbol{y}$:

$$\mathrm{JK}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\mathrm{wid}\ (\boldsymbol{x} \wedge \boldsymbol{y})}{\mathrm{wid}\ (\boldsymbol{x} \vee \boldsymbol{y})}. \tag{3.16}$$

В выражении (3.16) используется ширина интервала (см. п. 2.1), а вместо операций пересечения и объединения множеств — операции взятия минимума ($\wedge$) (2.7) и максимума ($\vee$) (2.8) по включению двух величин в полной интервальной арифметике Каухера. В наименовании $\mathrm{JK}(\boldsymbol{x}, \boldsymbol{y})$ буква J указывает на фамилию Jaccard, а K — на арифметику Каухера. В общем случае минимум по включению в выражении (3.16) может быть неправильным интервалом.

Рассмотренная мера обобщает обычное понятие меры совместности на различные типы взаимной совместности интервалов. В случае $\boldsymbol{x} \cap \boldsymbol{y} = \varnothing$, $\boldsymbol{x} \wedge \boldsymbol{y}$ — неправильный интервал, числитель (3.16) имеет отрицательное значение. В предельном случае вещественных значений $x \neq y$ имеем

$$\mathrm{JK}(x, y) = -1.$$

В целом получаем

$$-1 \leq \mathrm{JK}(\boldsymbol{x}, \boldsymbol{y}) \leq 1. \tag{3.17}$$

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

Мера совместности, введенная для двух интервалов в форме (3.16), допускает естественное обобщение в случае интервальной выборки. Пусть имеется интервальная выборка $\boldsymbol{X} = \{\boldsymbol{x}_i\},\ i = 1, 2,\ \ldots,\ n$. Определим меру $\mathrm{JK}(\boldsymbol{X})$ как

$$\mathrm{JK}(\boldsymbol{X}) = \frac{\mathrm{wid}\ (\bigwedge_i \boldsymbol{x}_i)}{\mathrm{wid}\ (\bigvee_i \boldsymbol{x}_i)}. \tag{3.18}$$

Важно, что выражение (3.18) переходит в случае интервальной выборки из двух элементов в выражение (3.16).

Пример 3.3.14 (Пример вычисления меры совместности для накрывающей выборки.) Пусть имеется интервальная выборка из четырех элементов (3.7), рассмотренная при вычислении интервальной моды в п. 3.3.1

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

Диаграмма рассеяния выборки $\boldsymbol{X}$ приведена на рис. 3.4. Выберем из нее накрывающую подвыборку

$$\boldsymbol{X}_{\text{c}} = \{[5,9],[6,9]\}.$$

Для выборки $\boldsymbol{X}_{\text{c}}$ имеем согласно (3.18)

$$\text{JK}(\boldsymbol{X}_{\text{c}}) = \frac{9-6}{9-5} = 0,75.$$

Значение $\text{JK}(\boldsymbol{X}_{\text{c}})$ демонстрирует высокую меру сходства элементов выборки $\boldsymbol{X}$.

3.4 Обработка ненакрывающих выборок

Если выборка — ненакрывающая, т. е. некоторые из ее измерений не содержат истинного значения измеряемой величины, то приведенные в п. 3.2 рассуждения и приемы частично теряют свой смысл. Уточнение пересечением здесь уже неуместно, и информационное множество для истинного значения величины имеет смысл взять в виде объединения всех интервалов выборки, т. е. как

$$\bigcup_{1 \leq k \leq n} \boldsymbol{x}_k. \tag{3.19}$$

Это множество может не быть единым интервалом на вещественной оси (подобное часто случается, к примеру, если выборка несовместна). Следует воспользоваться вместо объединения обобщающей его операцией «$\vee$» (см. (2.8)), т. е. взятием максимума по включению, и вместо (3.19) использовать информационный интервал в виде

$$\boldsymbol{J} = \bigvee_{1 \ leq k \leq n} \boldsymbol{x}_k = \Big[\min_{1 \leq k \leq n} \underline{\boldsymbol{x}}_k, \max_{1 \leq k \leq n} \overline{\boldsymbol{x}}_k \Big]. \tag{3.20}$$

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

$$x_{\text{c}} = \operatorname{mid} \boldsymbol{J} = \tfrac{1}{2}\Big(\min_{1 \leq k \leq n} \underline{\boldsymbol{x}}_k + \max_{1 \leq k \leq n} \overline{\boldsymbol{x}}_k \Big). \tag{3.21}$$

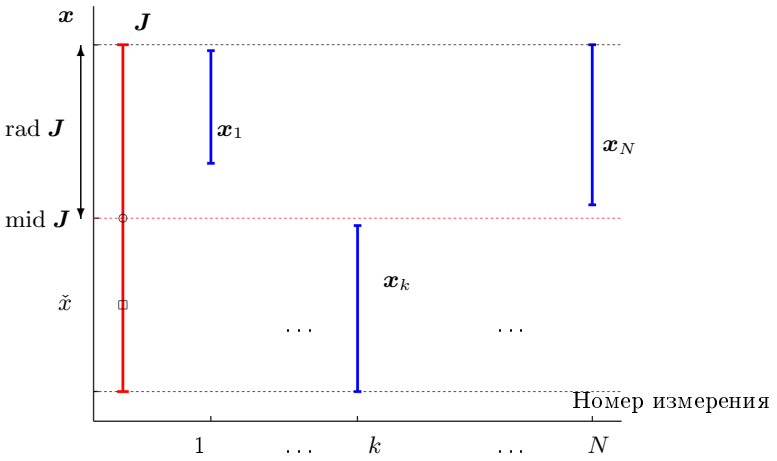

Рис. 3.7. Обработка ненакрывающей выборки интервальных измерений величины.

Как и ранее, нам может быть известен некоторый априорный интервал возможных значений оцениваемой постоянной величины $\boldsymbol{J}_{\text{апр}} = [\underline{\boldsymbol{J}}_{\text{апр}}, \overline{\boldsymbol{J}}_{\text{апр}}]$, который должен гарантированно содержать ее. Его могут задавать внешние физические (химические, биологические, экономические и т. п.) условия или ограничения. Тогда границы результирующего интервала (3.20) могут быть уточнены пересечением

$$\boldsymbol{J} = \boldsymbol{J} \cap \boldsymbol{J}_{\text{апр}}. \tag{3.22}$$

Пример 3.4.15 (Неопределенность измерения нуля цифрового измерителя напряжения.) Рассмотрим неопределенность измерения нуля цифрового измерителя напряжения. Это явление может быть причиной возникновения интервальной неопределенности результатов измерений и упоминается в п. 2.5.2.

Пусть в качестве измерителя используется микросхема аналоговой памяти DRS4 для записи коротких сигналов [10]. Перед проведением основных измерений необходимо вычислить неопределенность измерения нуля. Для этого на вход измерителя подают нулевое значение напряжения и получают выборку замеров.

При дальнейшей обработке данных полученной таким образом выборки возможны различные варианты, соотносящиеся с шириной интервалов по отношению к разбросу средних значений в выборке, как это рассматривалось в п. 2.5.2. Дело в том, что измерение выборки электрического сигнала может

производиться с различной точностью, причем точность измерения может варьироваться в весьма широких пределах.

В цифровых измерителях напряжения для грубых измерений типичными являются измерители с восемью двоичными разрядами, что соответствует амплитудному разрешению, равному $1/2^8 \cdot 100\,\% \simeq 0,4\,\%$. Для более точных измерений разрядность измерителя зависит от частоты проводимых измерений и варьируется от 10 ($\simeq 0,1\,\%$) до 24 ($\simeq 10^{-5}\,\%$) двоичных разрядов.

В п. 2.5.2 введено понятие модели погрешности измерений. В конкретном случае можно в качестве модели измерения (2.20) принять выражение

$$\boldsymbol{x} = \mathring{x} + \boldsymbol{\epsilon}, \tag{3.23}$$

$\mathring{x}$ — значение, выданное измерителем, а интервал погрешности принять в виде уравновешенного интервала

$$\boldsymbol{\epsilon} = [-\epsilon, \epsilon] \quad \epsilon = \frac{1}{2^{NOB}}, \tag{3.24}$$

где NOB (number of bits) — разрядность измерителя. При этом предполагается, что систематические погрешности отсутствуют или неизвестны.

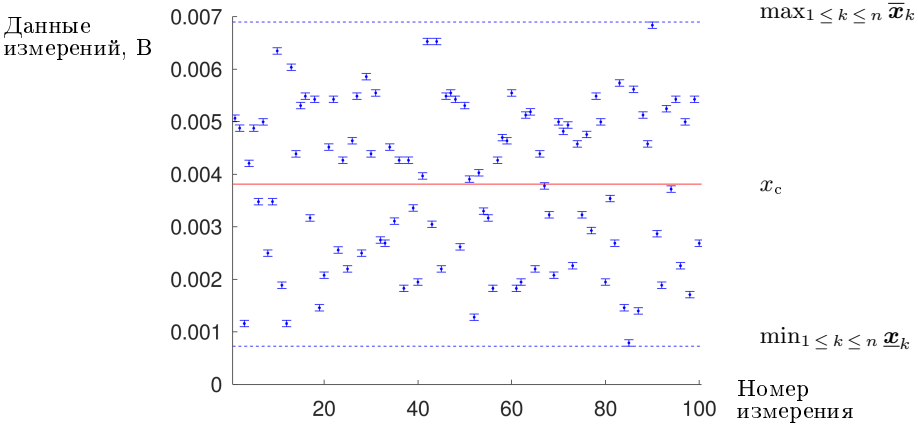

Рис. 3.8. Диаграмма рассеяния интервальных измерений неопределенности нуля. Разрядность измерителя $NOB = 14$

Пусть паспортная разрядность цифрового измерителя равна 14 двоичным разрядам. На рис. 3.8 представлены данные для 100 измерений неопределенности нуля $\{\mathring{x}_k\}_{k=1}^{100}$. По характеру данных, представленных на рис. 3.8

видно, что для конкретной точности измерителя, ширины интервалов отдельных измерений по модели (3.24) малы в сравнении с полным диапазоном значений в выборке.

В п. 2.5.2, говорится о том, что в такой ситуации следует обратить внимание на характер пересечений пар результатов отдельных замеров $\boldsymbol{x}_i \cap \boldsymbol{x}_j$. Из рис. 3.8 видно, что число непустых пересечений относительно невелико. В этом случае можно применять подходы и алгоритмы, которые используются для неинтервальных (точечных) данных. Информативно построение гистограммы множества $\{\mathring{x}_k\}_{k=1}^{100}$.

Гистограмма для выборки $\{\mathring{x}_k\}_{k=1}^{100}$ представлена на рис. 3.9 и демонстрирует несимметричное распределение величины неопределенности нуля и непохожа на популярные теоретико-вероятностные распределения. В такой ситуации для того, чтобы не привносить в обработку данных необоснованных модельных представлений, имеет смысл ограничиться наиболее общими оценками, рассмотренными в начале п. 3.4.

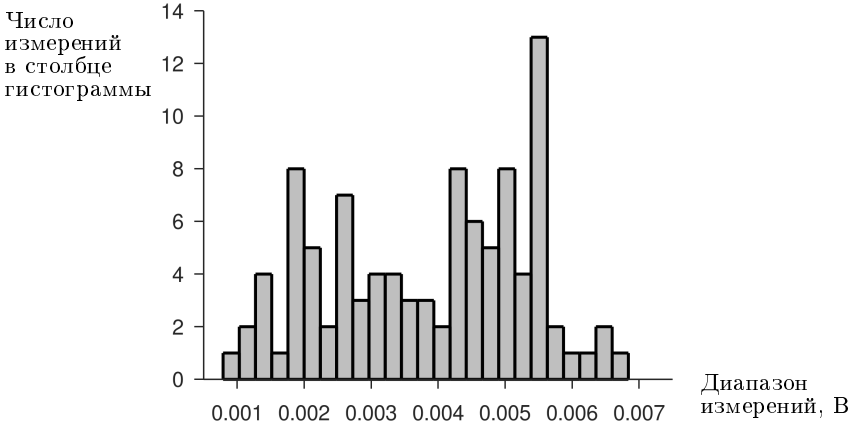

Рис. 3.9. Гистограмма данных $\{\mathring{x}_k\}_{k=1}^{100}$ интервальных измерений неопределенности нуля. Разрядность измерителя $NOB = 14$

Согласно выражению (3.20), имеем оценку информационного множества неопределенности нуля

$$\boldsymbol{J} = \bigvee_{1 \le k \le n} \boldsymbol{x}_k = \Big[\min_{1 \le k \le n} \underline{\boldsymbol{x}}_k, \max_{1 \le k \le n} \overline{\boldsymbol{x}}_k \Big] = \big[0{,}73 \cdot 10^{-3},\, 6{,}90 \cdot 10^{-3}\big].$$

Точечная оценка неопределенности нуля (3.21) равна

$$x_{\mathrm{c}} = \operatorname{mid} \boldsymbol{J} = \tfrac{1}{2}\Big(\min_{1 \le k \le n} \underline{\boldsymbol{x}}_k + \max_{1 \le k \le n} \overline{\boldsymbol{x}}_k \Big) = 3{,}82 \cdot 10^{-3}.$$

В целом вид диаграммы рассеяния на рис. 3.8 и гистограммы распределения значений $\{\mathring{x}_k\}_{k=1}^{100}$, на рис. 3.9 свидетельствует о переоценке точности представления результатов измерения или недоучете систематических погрешностей. Иначе говоря, модель ошибки (3.24), включающая только погрешность квантования цифрового измерителя, не описывает корректно данные, и суммарная ошибка в каждом измерении больше. В таком случае можно применить к выборке процедуру варьирования неопределенности, описанную в п. 3.5, и добиться совместности данных (получить накрывающую выборку).

Другой возможный сценарий обработки данных ненакрывающей выборки может состоять в том, что вместо пересечения интервальных измерений (как в п. 3.2) используем обобщающую ее операцию «$\wedge$», т. е. взятие минимума всех интервальных результатов измерений относительно упорядочения по включению, которое задается как

$$\boldsymbol{I} \;=\; \bigwedge_{1\le k\le n} \boldsymbol{x}_k \;=\; \Big[\max_{1\le k\le n} \underline{\boldsymbol{x}}_k,\; \min_{1\le k\le n} \overline{\boldsymbol{x}}_k\Big]. \tag{3.25}$$

В данном случае требуется использование полной интервальной арифметики Каухера, так как интервал (3.25) может оказаться неправильным. Следовательно, в качестве точечной оценки измеряемой величины целесообразно взять

$$x_c \;=\; \operatorname{mid} \boldsymbol{I} \;=\; \tfrac{1}{2}\Big(\max_{1\le k\le n} \underline{\boldsymbol{x}}_k + \min_{1\le k\le n} \overline{\boldsymbol{x}}_k\Big), \tag{3.26}$$

т. е. середину интервала, который получается как минимум по включению всех интервалов выборки (см. (2.7)). Если выборка совместна, то (3.26) совпадает с (3.5). Если же выборка несовместна, то результатом (3.25) является неправильный интервал $\boldsymbol{I}$, $\operatorname{rad} \boldsymbol{I} < 0$. Следовательно, информационное множество результатов измерений по обрабатываемой выборке пусто.

Но даже когда интервал (3.25) неправилен, его середина (3.26) — это точка, обладающая определенными условиями оптимальности. Она первой появляется в непустом пересечении интервалов выборки, если равномерно уширять их, увеличивая неопределенность измерений (см. п. 3.5). Если увеличить радиусы всех интервалов выборки на s и взять s таким, чтобы $s \ge |\operatorname{rad} \boldsymbol{I}|$, то получившийся интервал станет правильным, и точка x_c будет лежать в нем. Можно также сказать, что в точке (3.26) минимизируется равномерное уширение интервалов данных рассматриваемой выборки, необходимое для достижения ее совместности.

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

$$\boldsymbol{K} = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{x}_k.$$

Середина $\boldsymbol{K}$ может служить точечной оценкой измеряемой величины.

Все три рассмотренных приема обработки ненакрывающей выборки при стремлении ширины интервальных данных к нулю переходят в методы оценивания постоянной величины по точечным данным. То есть эти методы удовлетворяют принципу соответствия, рассмотренному в п.1.3.2.

Пример 3.4.16 (Пример вычисления меры совместности для ненакрывающей выборки) Пусть имеется интервальная выборка из четырех элементов (3.7)

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

Диаграмма рассеяния выборки $\boldsymbol{X}$ приведена на рис. 3.4. Для выборки $\boldsymbol{X}$ (3.7) имеем согласно (3.18)

$$\mathrm{JK}(\boldsymbol{X}) = \frac{4-6}{9-1} = -0{,}25.$$

Отрицательность JK говорит о несовместности выборки $\boldsymbol{X}$, а абсолютная величина — о степени несовместности ее элементов.

3.5 Вариабельность оценки и варьирование неопределенности.

Рассмотрим характеристики разброса оценок постоянной величины, полученных для интервальной выборки. Ее наиболее естественной мерой, если информационный интервал непуст, является *радиус* ϱ, т. е.

$$\varrho \;=\; \mathrm{rad}\,\boldsymbol{I} \;=\; \tfrac{1}{2}\left(\overline{\boldsymbol{I}} - \underline{\boldsymbol{I}}\right).$$

Фактически это максимальное отклонение границ информационного интервала от центральной оценки.

При анализе данных необходимо 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

$$\Delta_k = \mathrm{dist}\,(\boldsymbol{x}_k, x_{\mathrm{c}}), \qquad k = 1,\ \ldots,\ n. \tag{3.27}$$

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

$$\Delta_k = |\mathring{x}_k - x_{\mathrm{c}}|, \qquad k = 1,\ \ldots,\ n. \tag{3.28}$$

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

Пример 3.5.17 (Пример вычисления вариабельности оценки.) Рассмотрим данные табл. 3.1 из п. 3.2. Точечная оценка равна

$$x_c = \operatorname{mid} \boldsymbol{X} = -5{,}15. \tag{3.29}$$

Информационный интервал множества пуст, поэтому непосредственно вычислить вариабельность невозможно. Однако можно произвести вычисление максимального отклонения границ информационного интервала от центральной оценки.

Возьмем максимум по включению элементов множества интервальных данных из табл. 3.1

$$\boldsymbol{X}_U = \bigvee_i \boldsymbol{x}_i = [-14{,}4, 4{,}1],$$

и вычислим радиус $\boldsymbol{X}_U$:

$$\varrho \ = \tfrac{1}{2} \left(\overline{\boldsymbol{X}}_U - \underline{\boldsymbol{X}}_U \right) = 9{,}25.$$

В данном случае это максимальное отклонение границ интервала максимума по включению от центральной оценки.

Вычисления по формуле (3.28) дают вектор

$$\begin{aligned} \Delta_k = \{&3{,}45, 3{,}65, 4{,}14, 6{,}35, 6{,}85, 9{,}14, 7{,}84, 3{,}14, \\ &9{,}34, 9{,}15, 5{,}35, 3{,}54, 2{,}35, 3{,}24, 3{,}85\}. \end{aligned}$$

Приведем пример вычисления различных норм вектора рассеяния Δ_k

$$|\Delta| := \begin{cases} |\Delta|_1 = 81{,}45, \\ |\Delta|_2 = 22{,}98, \\ |\Delta|_\infty = 9{,}35. \end{cases}$$

Ранее показано, что величина реальной неопределенности измерения, т. е. радиуса интервала измерения, определяется непросто и иногда неоднозначно. Однако, он сильно влияет на свойства как отдельного измерения, так и выборки интервальных измерений.

Изложенное ранее приводит к мысли о том, что при обработке интервальных данных величиной неопределенности можно управлять, варьируя ее с целью исследования интервальных измерений, их выборок и построения оценок с нужными свойствами. В этом состоит суть приема варьирования неопределенности [45].

Если выборка интервальных измерений несовместна, то, увеличивая одновременно величину неопределенности всех измерений, можно добиться того, чтобы выборка стала совместной, т. е. чтобы пересечение интервалов стало непустым, а интервал минимума по включению (3.25) — правильным. Кроме того, точка (или точки), которая первой появляется в непустом пересечении интервалов при расширении интервальных измерений и тем самым требует наименьшего увеличения неопределенности измерений для достижения совместности выборки, является наименее несовместной. Ее разумно исользовать в качестве оценки величины (или оценки параметров зависимости).

В конкретной ситуации данных табл. 3.1, измерения выборки являются существенно неравноширинными. Одновременное изменение величины неопределенности для всех измерений на одно и то же значение может оказаться неразумным. Пусть задан некоторый положительный весовой вектор $w = (w_1, w_2, \ \ldots, \ w_n)$, $w_k > 0$, размерность которого равна длине исследуемой выборки, причем изменение величины неопределенности k-го измерения $\mathrm{rad}\, \boldsymbol{x}_k$ должно быть пропорциональным w_k, т. е. для любых k и l справедливо

$$\frac{\text{Изменение } \mathrm{rad}\, \boldsymbol{x}_k}{\text{Изменение } \mathrm{rad}\, \boldsymbol{x}_l} = \frac{w_k}{w_l}.$$

Пример 3.5.18 (Пример варьирования неопределенности.) Применительно к данным табл. 3.1, использование методики приведено на рис. 3.10. Красным цветом представлены исходные данные табл. 3.1, а черным цветом — расширенные интервалы данных при выбранном коэффициенте расширения.

Вычисления проведены с использованием кода `Octave` С. И. Жилина [29]. При этом решается задача линейного программирования, в ходе которой вычисляются два параметра: оптимальное положение центра неопределенности и коэффициент расширения радиусов замеров.

$$x_{MM} = \texttt{oskorbin\_center} = -5{,}30; \quad \texttt{k} = 1{,}75.$$

В даном случаев индексе x_{MM} обозначение MM соответствует Minimal Module — функции оптимизации задачи линейного программирования.

Информационное множество представляет точку

$$\boldsymbol{I}_{MM} = \bigcap_{1 \le k \le n} \boldsymbol{x}_k \ = \ x_{MM}.$$

Содержательным результатом вычислений является уточнение положения наиболее вероятной точечной оценки физической величины [33] и вычисление дополнительной погрешности для каждого элемента выборки, необходимой для достижения совместности данных.

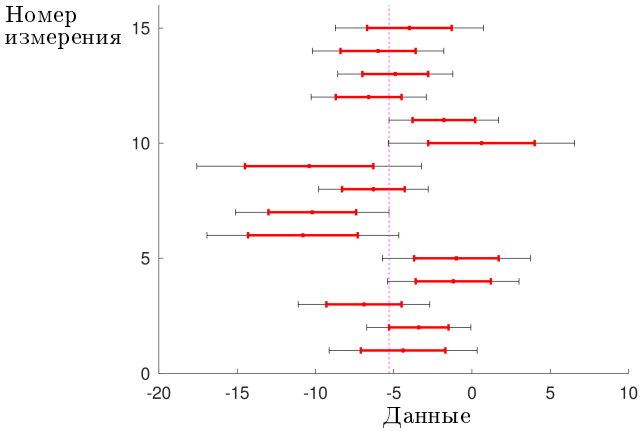

Рис. 3.10. Графическое представление интервальных данных и результаты обработки по методике [45]

Глава 4

Задача восстановления зависимостей

В гл. 4 даны определения новых терминов и понятий, которые возникают в связи с восстановлением функциональных зависимостей по данным их измерений, имеющих интервальную неопределенность. Представлены основные идеи и типичные приемы восстановления зависимостей по интервальным данным и возникающие при этом проблемы. Подробно исследуется линейная зависимость, но большинство построений и рассуждений соотвествует общему нелинейному случаю.

В пп. 4.1 — 4.4 кратко рассмотрена схема восстановления зависимостей по интервальным данным из [1]. Далее приводятся примеры решения конкретных задач.

4.1 Постановка задачи

Пусть величина y является функцией некоторого заданного вида от независимых переменных $x_1, x_2, \ \ldots, \ x_m$, т. е.

$$y = f(x, \beta), \tag{4.1}$$

где $x = (x_1, \ \ldots, \ x_m)$ — вектор независимых переменных; $\beta = (\beta_1, \ \ldots, \ \beta_l)$ — вектор параметров функции. Имея набор значений переменных x и y, нужно найти $\beta_1, \ \ldots, \ \beta_l$, которые соответствуют конкретной функции f из параметрического семейства (4.1). Эту задачу называют *задачей восстановления зависимости*, и она будет основным предметом рассмотрения в гл. 4.

Широко используются также другие названия — «задача идентификации параметров», «задача подгонки данных», «задача подгонки кривой», «задача сглаживания данных» (соответствующие англоязычные термины — identification problem, data fitting problem, curve fitting problem) и т. п. В вероятностной статистике рассматриваемую задачу называют «задачей построения регрессии» или «задачей регрессионного анализа», а соответствующая математическая дисциплина называется регрессионным анализом. Еще одно название задачи — «задача построения эмпирических формул». Исходя из контекста или предметной области, где рассматривается поставленная задача, для переменных в рассматриваемой функциональной зависимости используют также различные термины. Независимые переменные часто называют *экзогенными*, *предикторными* или *входными* переменными, а зависимая переменная называется также *эндогенной*, *критериальной* или *выходной* переменной.

Важнейший частный случай рассматриваемой задачи — определение параметров линейной функции вида

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m, \tag{4.2}$$

в которой x_1, x_2, $\ldots$, x_m — независимые переменные; y — зависимая переменная; β_0, β_1, $\ldots$, β_m — некоторые коэффициенты. Эти неизвестные коэффициенты должны быть определены из ряда измерений значений x_1, x_2, $\ldots$, x_m и y.

Результаты измерений неточны, и предполагается, что они имеют *ограниченную неопределенность* (см. п. 1.3.2), когда известны лишь некоторые интервалы, дающие двусторонние границы измеренных значений. Таким образом, результатом i-го измерения являются такие интервалы $\boldsymbol{x}_1^{(i)}$, $\boldsymbol{x}_2^{(i)}$, $\ldots$, $\boldsymbol{x}_m^{(i)}$, $\boldsymbol{y}^{(i)}$, относительно которых предполагается, что истинное значение x_1 лежит в пределах $\boldsymbol{x}_1^{(i)}$, истинное значение x_2 лежит в $\boldsymbol{x}_2^{(i)}$ и т. д., вплоть до y, истинное значение которого находится в интервале $\boldsymbol{y}^{(i)}$. В целом имеется n измерений, поэтому индекс i может принимать значения из множества натуральных чисел $\{1, 2, \ldots, n\}$.

Далее для удобства построений и выкладок обозначим номер измерения i не верхним, а нижним индексом, который поставим первым при обозначении входов. Таким образом, полный набор данных для восстановления зависимости будет иметь вид

$$\begin{array}{ccccc} \boldsymbol{x}_{11}, & \boldsymbol{x}_{12}, & \ldots, & \boldsymbol{x}_{1m}, & \boldsymbol{y}_1, \\ \boldsymbol{x}_{21}, & \boldsymbol{x}_{22}, & \ldots, & \boldsymbol{x}_{2m}, & \boldsymbol{y}_2, \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{x}_{n1}, & \boldsymbol{x}_{n2}, & \ldots, & \boldsymbol{x}_{nm}, & \boldsymbol{y}_n. \end{array} \tag{4.3}$$

Необходимо найти или как-то оценить коэффициенты β_j, $j = 0, 1, \ldots, m$,

для которых линейная функция (4.2) наилучшим образом приближала бы интервальные данные измерений (4.3).

Для обозначения $n \times m$-матрицы, составленной из данных (4.3) для независимых переменных, часто используют термины *матрица плана эксперимента* или *матрица плана*, которые появились в теории планирования эксперимента. Интервалы $\boldsymbol{x}_{i1}$, $\boldsymbol{x}_{i2}$, $\ldots$, $\boldsymbol{x}_{im}$, $\boldsymbol{y}_i$ будем называть, как и раньше, *интервалами неопределенности i-го измерения*. Но кроме них также потребуется обращаться ко всему множеству, ограничиваемому в многомерном пространстве $\mathbb{R}^{m+1}$ этими интервалами по отдельным координатным осям.

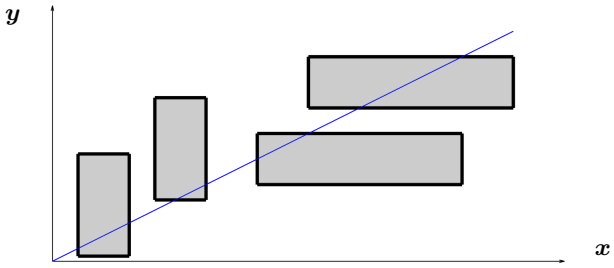

Рис. 4.1. Иллюстрация задачи восстановления линейной зависимости по данным с интервальной неопределенностью

Определение 4.1.1 *Брусом неопределенности i-го измерения функциональной зависимости будем называть интервальный вектор-брус* $(\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{im}, \boldsymbol{y}_i) \subset \mathbb{R}^{m+1}$, $i = 1, 2, \ldots, n$, *образованный i-й строкой таблицы данных* (4.3).

Таким образом, каждый брус неопределенности измерения зависимости является прямым декартовым произведением интервалов неопределенности независимых переменных и зависимой переменной. На рис. 4.1 на плоскости $0xy$ наглядно показаны брусы неопределенности измерений и график искомой линейной функции. Далее рассматриваем данные (4.3) как уже существующие и не обсуждаем их получение, выбор или оптимизацию.

4.2 Накрывающие и ненакрывающие измерения и выборки

Как и в п. 2.6.1, принимаем следующее определение

Определение 4.2.1 *Брус неопределенности измерения функциональной зависимости называется накрывающим, если он гарантированно содержит*

истинные значения измеряемых величин входных и выходных переменных зависимости.

Условимся называть брус неопределенности измерения *ненакрывающим*, если нельзя утверждать, что он наверняка содержит истинное значение. Иными словами, ненакрывающий брус может включать истинное значение, а может и не включать его.

Накрывающей выборкой будем называть совокупность измерений, т. е. выборку, в которой *доминирующая* часть измерений являются накрывающей. Ненакрывающей называем выборку, большинство составляющих которую измерений могут не содержать истинных значений измеряемой зависимости.

О ненакрывающей выборке можно сказать то же самое, что и по поводу ненакрывающего бруса. Удобно называть этим термином выборку, для большинства измерений который не гарантировано свойство накрытия истинного значения. Итак, *ненакрывающая выборка* — это выборка, для измерений которой нельзя утверждать, что они наверняка являются накрывающими.

Для визуализации интервальных данных аналогично традиционному точечному случаю используют *диаграммы рассеяния*. В традиционном понимании диаграмма рассеяния используется в статистике и анализе данных для визуализации значений двух переменных в виде облака точек на декартовой плоскости и позволяет оценить наличие или отсутствие корреляции и других взаимосвязей между двумя переменными. На диаграмме рассеяния для интервальных данных каждое интервальное наблюдение отображается в виде бруса (бруса неопределенности). При отсутствии неопределенности по одной из переменных, брусы наблюдений могут превращаться в одномерные вертикальные или горизонтальные отрезки («ворота» для случая накрывающих измерений). Примерами являются диаграмм рассеяния, изображенные на рис. 4.1 и 4.4.

4.3 Информационное множество задачи

Существует большое количество стандартных подходов к решению задачи восстановления зависимостей для обычных точечных данных. В практической обработке данных широко используются метод наименьших квадратов, метод наименьших модулей, чебышевское (минимаксное) сглаживание. Все эти методы основаны на нахождении минимума какой-либо количественной меры отклонения конструируемой функции от приближаемых данных, которую часто называют *функционалом качества*. Ищут набор параметров, который доставляет минимум этой мере отклонения.

Для интервальных данных реализация описанного общего принципа становится затруднительной, поскольку не вполне ясно, как именно выбирать отклонение функции от приближаемых интервальных данных. Это особенно

характерно для накрывающих измерений и накрывающих выборок, которые представляют собой множества возможных значений измеряемой величины.

Для анализа ситуации, который приведет к фундаментальному понятию информационного множества задачи восстановления зависимости, необходимо начать рассмотрение задачи с самого начала. Пусть имеется набор экспериментальных данных

$$\begin{array}{ccccc} x_{11}, & x_{12}, & \ldots, & x_{1m}, & y_1, \\ x_{21}, & x_{22}, & \ldots, & x_{2m}, & y_2, \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{n1}, & x_{n2}, & \ldots, & x_{nm}, & y_n \end{array} \tag{4.4}$$

и формула для функциональной зависимости, зависящая от параметров (4.1).

Мы подставляем данные в формулу для зависимости (4.2) и получаем для каждого измерения одно уравнение вида

$$f(x_i, \beta) = y_i,$$

где $x_i = (x_{i1}, x_{i2}, \ldots, x_{im})$. В целом в результате этой процедуры возникает система уравнений, решив которую относительно β, найдем параметры зависимости.

В традиционном случае обработки точечных данных полученная система уравнений является, как правило, несовместной и решений в обычном смысле не имеет. При подстановке любого набора параметров β в уравнения (4.1), получаем ненулевое расхождение левой и правой частей, которое в традиционном регрессионнном анализе называется *остатком*

$$f(x_i, \beta) - y_i = \varepsilon_i.$$

Поэтому вместо обычных решений системы рассматривают решения в обобщенном смысле — *псевдорешения*, т. е. векторы, на которых достигается минимальное отклонение левой и правой частей системы уравнений:

$$\hat{\beta} = \arg\min_{\beta} \|\varepsilon_i\|.$$

Фактически задача нахождения псевдорешения — это и есть задача минимизации функционала качества в какой-то конкретной норме, в которой он является величиной вектора остатков измерений.

Таким образом, имеется два общих подхода к нахождению параметров зависимости по эмпирическим данным. На основе вида искомой функциональной зависимости и обрабатываемых данных составляется

- система уравнений и находится ее решение;

- задача минимизации отклонения функции от эмпирических данных и находится ее решение.

В случае точечных данных преимущественное значение имеет второй способ, так как система уравнений почти всегда не имеет обычных решений. Но для интервальных данных ситуация меняется на противоположную.

Что следует считать решением задачи восстановления зависимости по интервальным данным (4.3)? Очевидно, что функцию вида (4.1) или (4.2) нужно считать точным решением задачи восстановления искомой зависимости, если ее график проходит через все брусы неопределенности данных. В случае точечных данных эта идеальная ситуация почти никогда не реализуется и неустойчива к малым возмущениям в данных. Но для данных с существенной интервальной неопределенностью прохождение графика функции через брусы данных (4.3) может реализовываться, и оно устойчиво к возмущениям в данных. Кроме того, брусы неопределенности данных (4.3), в отличие от бесконечно малых и бесструктурных точек, получают структуру, и потому нужно различать, как именно проходит график функции через эти брусы.

В интервальном случае, подставляя данные в равенство (4.1) для искомой функциональной зависимости, получим интервальную систему уравнений. Ее решением будет вектор оценки параметров восстанавливаемой зависимости (4.1). При этом разрешимость системы интервальных уравнений не является исключительным событием, тогда как определение задачи минимизации отклонения функциональной зависимости от данных сталкивается с трудностями.

В соответствии с терминологией, введенной в п. 2.6.2, *информационным множеством* задачи восстановления зависимости нужно называть множество значений параметров зависимости, совместных с данными в каком-то определенном смысле. Информационное множество задачи восстановления функциональной зависимости по интервальным данным — это множество решений интервальной системы уравнений, неравенств и т. п. условий, вытекающих из постановки задачи восстановления зависимостей, т. е. вида функциональной зависимости и обрабатываемых данных.

Почему понятие информационного множества столь важно при обработке интервальных измерений? Дело в том, что именно информационное множество учитывает специальный характер накрывающих интервальных измерений, когда они являются не просто большими «раздувшимися точками», а еще включают и возможные точные значения измеряемых величин.

В оптимизационном подходе учет «накрытия/ненакрытия» отодвигается на второй план, а потому его нужно сочетать с проверкой существования непустого информационного множества задачи.

4.4 Прогнозный коридор и коридор совместных зависимостей

Определение параметров функциональной зависимости производится, как правило, для того, чтобы затем найденную формулу использовать для предсказания значений зависимости в других точках из области определения или вне нее. Такое предсказание будет осуществляться с некоторой погрешностью, вызванной неопределенностями данных, неоднозначностью процедуры восстановления и т. п.

Пусть дана задача восстановления функциональной зависимости вида $y = f(x, \beta)$, где областью определения независимой переменной x является множество X, а значения зависимой переменной y принимаются во множестве Y. Будем называть *прогнозным коридором* для задачи восстановления зависимостей по интервальным данным многозначное отображение $\Pi : X \to Y$, которое каждой точке области определения X восстанавливаемой зависимости сопоставляет множество возможных значений отображений, которые в рамках рассматриваемой модели могут принимать функциональные зависимости, восстановленные по данным задачи.

Если информационное множество задачи восстановления зависимостей непусто, то обычно оно задает целое семейство зависимостей, совместных с данными задачи, которое имеет смысл рассматривать вместе — как единое целое. Как следствие, возникает необходимость рассматривать вместе единым целым, множество всех функций, совместных с интервальными данными задачи восстановления зависимости. Будем будем называть его *коридором совместных зависимостей* (см. рис. 4.2).

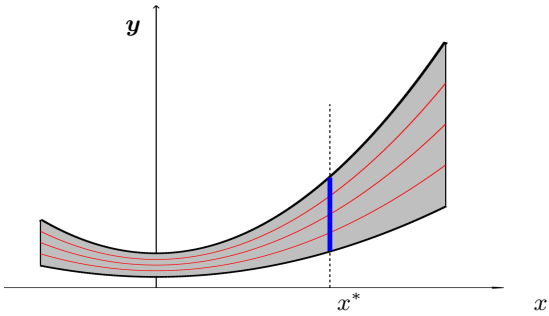

Рис. 4.2. Коридор совместных зависимостей и его сечение для какого-то значения аргумента x^*

В специализированнлй литературе использовались также другие термины — «трубка» совместных зависимостей (имеет происхождение в теории

управления), «полоса» или «слой неопределенности», «коридор неопределенности» и т. п. Строгое определение коридора совместных зависимостей может быть дано на основе математического понятия многозначного отображения. Для произвольных множеств X и Y *многозначным отображением* F из X в Y называется соответствие (правило), сопоставляющее каждой точке $x \in X$ непустое подмножество $F(x) \subset Y$, называемое *значением*, или *образом* x.

Определение 4.4.1 *Пусть в задаче восстановления зависимостей информационное множество Ω параметров зависимостей $y = f(x, \beta)$, совместных с данными, является непустым. Коридором совместных зависимостей рассматриваемой задачи называется многозначное отображение Υ, сопоставляющее каждому значению аргумента x множество*

$$\Upsilon(x) \;=\; \bigcup_{\beta \in \Omega} f(x, \beta).$$

Значение $\Upsilon(\tilde{x})$ коридора совместных зависимостей при каком-то определенном аргументе $\tilde{x}$ (сечение коридора) — это множество $\cup_{\beta \in \Omega} f(\tilde{x}, \beta)$, образованное всевозможными значениями, которые принимают на этом аргументе функциональные зависимости, совместные с интервальными данными измерений.

Это множество описывает неопределенность прогноза на аргументе $\tilde{x}$. Его нужно уметь вычислять или каким-либо образом оценивать. В частности, необходимо знать внешние оценки интервала

$$\Big[\min_{\beta \,\in\, \Omega} f(\tilde{x}, \beta), \max_{\beta \,\in\, \Omega} f(\tilde{x}, \beta) \Big].$$

В ряде задач необходимо также знать внутреннюю оценку коридора совместных зависимостей. На рис. 4.2 изображен коридор совместных зависимостей в задаче восстановления нелинейной зависимости, но для рассматриваемого линейного случая границы коридора совместных зависимостей являются кусочно-линейными (см. рис. 4.9). С примерами использования коридора совместных зависимостей можно ознакомиться в [13].

Понятие прогнозного коридора шире понятия коридора совместных зависимостей. Если информационное множество задачи пусто, то и о коридоре совместных зависимостей не имеет смысл говорить, но, как правило, оценку параметров при этом все равно необходимо получить, и какая-то функциональная зависимость будет построена. У этого решения задачи некоторая неопределенность все равно присутствует, а потому имеет смысл и прогнозный коридор.

4.5 Выбросы и их выявление

Общие идеи выявления выбросов Понятие «выброс» в статистике и анализе данных, как правило, определяется неформально, поскольку критерии для признания измерения выбросом лежат вне формальной математической постановки задачи анализа данных. Существует много подходов, в которых общим является указание на нарушение измерением-выбросом согласованности, ожидаемой для большинства наблюдений выборки по отношению к конкретной математической модели.

Формальным индикатором согласованности данных, модели и априорной информации является непустота информационного множества, соответствующего задаче. Пустота информационного множества свидетельствует о наличии тех или иных противоречий между данными и моделью. Поиск причин появления противоречий, а также выбор путей их преодоления — процесс неформальный.

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

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

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

Среди внешних измерений особым образом выделяют *выбросы* (промахи). Построение модели по выборке, пополненной таким наблюдением, приводит не только к уменьшению информационного множества, но и к его пустоте ($\Omega(\mathcal{S}_n \cup s) = \varnothing$), т.е. к разрушению модели.

Анализ взаимоотношений информационных множеств $\Omega(\mathcal{S}_n)$ и $\Omega(\mathcal{S}_n \cup s)$ или $\Omega(\mathcal{S}_n)$ и $\Omega(s)$ можно заменить выяснением отношений интервала неопределенности $\boldsymbol{y}$ анализируемого измерения $s = (x, \boldsymbol{y})$ и интервального прогнозного значения рассматриваемой модели в той же точке $\Upsilon(x; \mathcal{S}_n)$. На рис. 4.3 анализируемые измерения показаны линиями, а соответствующие им интервалы прогнозов — широкими линиями. Их ширина не имеет содержательного

смысла, а лишь упрощает восприятие наложенных друг на друга интервалов.

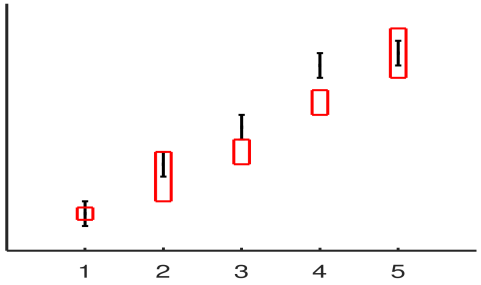

Рис. 4.3. Интервальные наблюдения с различными статусами: *внутреннее* ($n = 1$); *граничные* ($n = 2$); *внешние* ($n = 3$); *строго внешнее* ($n = 5$); *выбросы* ($n = 4$)

Внутреннее интервальное измерение $s = (x, \boldsymbol{y})$ полностью содержит в себе прогнозный интервал, оцененный с помощью модели $\Upsilon(x; \mathcal{S}_n)$, или, иными словами, пересечение двух этих интервалов совпадает с прогнозным: $\boldsymbol{y} \cap \Upsilon(x; \mathcal{S}_n) = \Upsilon(x; \mathcal{S}_n)$. Будучи перестроенной по выборке, пополненной подобным измерением, модель не претерпит изменений, поскольку соответствующее ей информационное множество окажется внутри ограничения, порожденного добавленным внутренним измерением, 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:

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

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 анализе данных вводят специальные величины *размаха* (от англ. — high leverage — плечо) и *относительного остатка* (от англ. relative residual — относительное остаточное отклонение, относительное смещение,). Размах и остаток позволяют установить статус наблюдения посредством проверки выполнения некоторых простых неравенств [1]. Следует отметить, что характеризация наблюдений в терминах размахов и остатков не зависит от размерности входной переменной x.

4.5.1 Варьирование величины неопределенности измерений

Один из приемов выявления выбросов в задаче построения зависимости по интервальным наблюдениям основан на интерпретации выбросов как наблюдений с недооцененной величиной неопределенности [12, 43]. Закономерным шагом в этом случае становится поиск некоторой минимальной коррекции величин неопределенности 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 $\mathcal{S}_n$ 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 посредством решения задачи условной оптимизации

$$\text{найти} \quad \min_{w,\,\beta} \Sigma_{i=1}^{n} w_i \tag{4.5}$$

при ограничениях

$$\begin{cases} \mathring{y}_i - w_i\,\epsilon_i \leq f(x_i, \beta) \leq \mathring{y}_i + w_i\,\epsilon_i, \\ w_i \geq 1, \end{cases} \qquad i = 1, \ \ldots, \ n. \tag{4.6}$$

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

Следует отметить значительную гибкость языка неравенств. Он дает возможность переформулировать и расширять систему ограничений (4.6) для учета специфики данных и задачи при поиске допустимой коррекции данных, приводящей к разрешению исходных противоречий. Например, если имеются основания считать, что величина неопределенности некоторой группы наблюдений одинакова и при коррекции должна увеличиваться синхронно, то система ограничений (4.6) может быть пополнена равенствами вида

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

где $i_1, \ \ldots, \ i_K$ — номера наблюдений группы. В случае, когда в надежности каких-либо наблюдений исследователь уверен полностью, при решении задачи (4.5), (4.6) соответствующие им величины w_i можно положить равными единице, т. е. запретить варьировать их неопределенность.

Задача поиска коэффициентов масштабирования величины неопределенности (4.5), (4.6) сформулирована для распространенного случая уравновешенных интервалов погрешности и подразумевает синхронную подвижность верхней и нижней границ интервалов неопределенности 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 $\epsilon^- \neq \epsilon^+$), то границы интервальных измерений можно варьировать независимо, масштабируя величины неопределенности ϵ_i^- и ϵ_i^+ с помощью отдельных коэффициентов w_i^- и w_i^+:

$$\text{найти} \quad \min_{w^-,\, w^+,\, \beta} \ \Sigma_{i=1}^n (w_i^- + w_i^+) \tag{4.7}$$

при ограничениях

$$\begin{cases} \mathring{y}_i - w_i^-\, \epsilon_i^- \le f(x_i, \beta) \le \mathring{y}_i + w_i^+\, \epsilon_i^+, \\ w_i^- \ge 1, \qquad\qquad i = 1, \ldots, n. \\ w_i^+ \ge 1, \end{cases} \tag{4.8}$$

Для линейной по параметрам β зависимости $y = f(x, \beta)$ задача (4.5), (4.6) представляет собою задачу линейного программирования, для решения которой широко доступны программы в составе библиотек на различных языках программирования, в виде стандартных процедур систем компьютерной математики, а также в виде интерактивных подсистем электронных таблиц.

4.6 Случай точных измерений входных переменных

Важнейшим и часто встречающимся частным случаем рассмотренной задачи является ситуация, когда независимые (экзогенные, предикторные, входные) переменные $x_1,\ x_2,\ \ldots,\ x_m$ измеряются точно, и вместо телесных брусов неопределенности измерений (как на рис. 4.1) имеем отрезки прямых $(x_{i1}, x_{i2},\ \ldots,\ x_{im}, \boldsymbol{y}_i)$, $i = 1, 2, \ldots, n$, параллельные оси зависимой (эндогенной, критериальной, выходной) переменной (рис. 4.4). Впервые такая постановка задачи была рассмотрена в работе Л. В. Канторовича [26].

Популярность задачи восстановления зависимости в такой постановке определяется несколькими факторами. В широком классе случаев входные переменные определены точно (номер измерения, задание входной переменной в целочисленной арифметике) или с очень малой погрешностью. Погрешность может быть неизвестна, и непонятно, как ее оценивать. Измерения могут быть настолько грубыми, что погрешности во входных данных заведомо пренебрежимы.

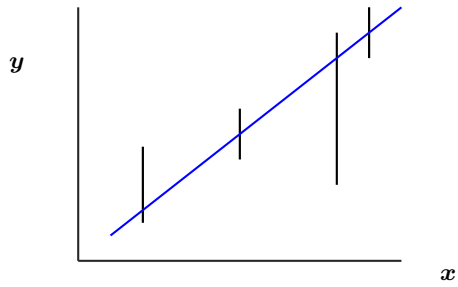

Рис. 4.4. Частный случай задачи восстановления линейной зависимости по неточным данным, когда входные переменные измеряются точно

Отсутствие неопределенности значений независимых переменных приводит к кардинальному упрощению математической модели. Брусы неопределенности измерений зависимости, введенные ранее, схлопываясь по независимым переменным, превращаются в *отрезки неопределенности*. Для решения и полного исследования этого частного случая, начиная с работы [26], предложено большое количество эффективных вычислительных методов.

Линейная зависимость (4.2) *совместна* (согласуется) с интервальными данными измерений, если ее график проходит через все отрезки неопределенности, задаваемые интервалами измерений выходной переменной y, как это изображено на рис. 4.4). Подобное понимание совместности (согласования) является прямым обобщением того понимания совместности, которое традиционно для неинтервального случая и используется, к примеру, в постановке задачи интерполяции.

Подставляя в зависимость (4.2) данные для входных переменных x_1, x_2, $\ldots$, x_m в i-е измерение и требуя включения полученного значения в интервалы $\boldsymbol{y}_i$, получим

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_m x_{im} \in \boldsymbol{y}_i, \qquad i = 1, 2, \ldots, n. \tag{4.9}$$

С одной стороны, это интервальная система линейных алгебраических уравнений

$$\begin{cases} \beta_0 + x_{11}\beta_1 + x_{12}\beta_2 + \ldots + x_{1m}\beta_m = \boldsymbol{y}_1, \\ \beta_0 + x_{21}\beta_1 + x_{22}\beta_2 + \ldots + x_{2m}\beta_m = \boldsymbol{y}_2, \\ \vdots \quad\quad \vdots \quad\quad \vdots \quad\quad \ddots \quad\quad \vdots \quad\quad \vdots \\ \beta_0 + x_{n1}\beta_1 + x_{n2}\beta_2 + \ldots + x_{nm}\beta_m = \boldsymbol{y}_n, \end{cases}$$

у которой интервальность присутствует только в правой части. С другой стороны, система (4.9) равносильна системе (4.10)

$$\begin{cases} \underline{\boldsymbol{y}}_1 \le \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_m x_{1m} \le \overline{\boldsymbol{y}}_1, \\ \underline{\boldsymbol{y}}_2 \le \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_m x_{2m} \le \overline{\boldsymbol{y}}_2, \\ \quad\vdots \quad\vdots \qquad\qquad\qquad \ddots \qquad\qquad\quad \vdots \quad\vdots \\ \underline{\boldsymbol{y}}_n \le \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_m x_{nm} \le \overline{\boldsymbol{y}}_n. \end{cases} \tag{4.10}$$

Это система двусторонних линейных неравенств относительно неизвестных параметров β_0, β_1, β_2, $\ldots$, β_m, решив которую, мы можем найти искомую линейную зависимость. Множество решений системы неравенств (4.10) является информационным множеством параметров восстанавливаемой зависимости для рассматриваемого случая.

Для i-го двустороннего неравенства из системы (4.10) множество решений — это полоса в пространстве $\mathbb{R}^{m+1}$ параметров $(\beta_0, \beta_1, \ldots, \beta_m)$, ограниченная с двух сторон гиперплоскостями с уравнениями

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_m x_{im} = \underline{\boldsymbol{y}}_i,$$

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_m x_{im} = \overline{\boldsymbol{y}}_i.$$

Множество решений системы неравенств (4.10) является пересечением n штук таких полос, отвечающих отдельным измерениям. Можно рассматривать эти полосы как информационные множества отдельных измерений. На рис. 4.5 изображено формирование множества решений системы неравенств (4.10) для случая двух параметров (то есть $m = 2$) и $n = 3$.

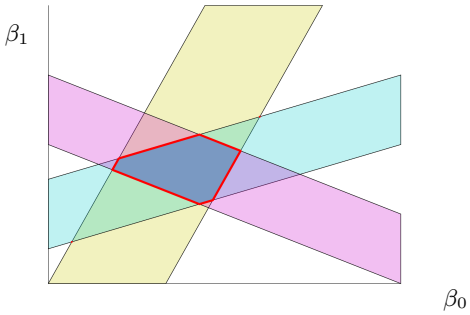

Рис. 4.5. Образование информационного множества параметров линейной зависимости (ограничено красной линией) для случая точных входных переменных

В целом множество решений системы линейных алгебраических неравенств (4.10) является выпуклым многогранным множеством в пространстве $\mathbb{R}^{m+1}$. Распознавание его пустоты или непустоты, а также нахождение какой-либо точки из него являются задачами, сложность которых ограничена полиномом от их размера. Существуют эффективные и хорошо разработанные вычислительные методы для решения этих вопросов и для нахождения оценок множества решений.

4.6.1 Пример решения задачи для случая точных измерений входных переменных

Рассмотрим конкретный пример решения задачи для случая точных измерений входных переменных, для которого используется аппарат линейного программирования для достижения совместности информационного множества [11, 12]. Технологическая схема вычислений представлена в виде блокнота на ресурсе С. И. Жилина [29].

В целом при восстановлении зависимости по интервальным измерениям нужно решить следующие задачи:
— построение модели данных согласно п. 2.5.2;
— построение функциональной модели;
— определение параметров модели;
— построение информационного множества и коридора совместности;
— построение прогноза внутри и за пределами экспериментальных данных;
— нахождение граничных точек множества совместности.

Пример 4.6.19 (Пример восстановления зависимости.) При измерении параметров шагового двигателя была получена зависимость положения вала от номера шага [?] — рис. 4.6.

Номер	1	2	3	4	5	6	7	8	9	10
Данные	388	737	951	1354	1756	1970	2399	2801	3204	3606

Таблица 4.1. Подвыборка данных рис. 4.6

Для облегчения восприятия, выберем 10 значений замеров из числа данных, представленных на рис. 4.6. Конкретно выбрано 10 первых нечетных значений для статических положений вала двигателя и вычтена аддитивная константа, отвечающая числу поворотов вала. Получившиеся результаты представлены далее в табл. 4.1.

В данном случае имеем дело с типичной ситуацией при работе с приборами, выдающими цифровые значения измерений. Данные энкодера доступны

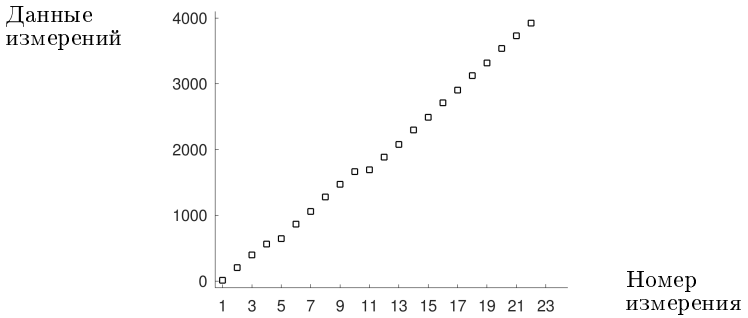

Рис. 4.6. Зависимость положения вала двигателя от номера шага

в виде целых значений и паспортная неопределенность измерений равна 0,5° [46] при дискретности измерений 12 бинарных разрядов на 360°. Таким образом, имеем тип погрешности данных согласно (3.24)

$$\boldsymbol{y} = \mathring{y} + \boldsymbol{\epsilon},$$

$\mathring{y}$ — значение, выданное измерителем, а интервал погрешности примем в виде

$$\boldsymbol{\epsilon} = [-\epsilon, \epsilon]; \quad \epsilon = \lceil 0{,}5 \cdot \tfrac{2^{12}}{360} \rceil = 6.$$

В данном случае $\lceil \cdot \rceil$ означает округление в большую сторону.

Например, для первой строчки в табл. 4.1 имеем $\boldsymbol{y}_1 = [382, 394]$. Реально погрешность, как увидим, существенно выше и включает много факторов, о части которых недостаточно сведений, и можно судить только об их совокупности по результату измерений.

Точечная оценка параметров регрессии. Сначала проведем точечную оценку параметров регрессии. Пусть модель задается в классе линейных функций

$$y = \beta_1 + \beta_2 x, \tag{4.11}$$

где x — номер измерения в выборке; y — угол поворота вала двигателя.

Для согласования с данными поставим задачу оптимизации и решим методами линейного программирования [1]. В соотвествии с подходом к варьированию величины неопределенности п. 4.5.1 поставим задачу (4.6) в виде

$$\begin{cases} \operatorname{mid} \boldsymbol{y}_i - w_i \cdot \operatorname{rad} \boldsymbol{y}_i \le X\beta \le \operatorname{mid} \boldsymbol{y}_i + w_i \cdot \operatorname{rad} \boldsymbol{y}_i, \quad i = 1, m, \\ \sum\limits_{i=1}^{m} w_i \longrightarrow \min \\ w_i \ge 0, \quad i = 1, 2, \ \ldots, \ m, \\ w, \beta = ? \end{cases}$$

Здесь X — матрица $m \times 2$, в первом столбце которой элементы, равные 1, во втором — значения x_i. В качестве значений середины и радиуса возьмем $\operatorname{mid} \boldsymbol{y}_i = y_i$ и $\operatorname{rad} \boldsymbol{y}_i = 1$.

Решив поставленную задачу с помощью программных средств на языке `Octave`, доступных на ресурсе [29], получим уравнение регрессионной в виде

$$y = -11{,}7 + 352{,}3 \cdot x. \tag{4.12}$$

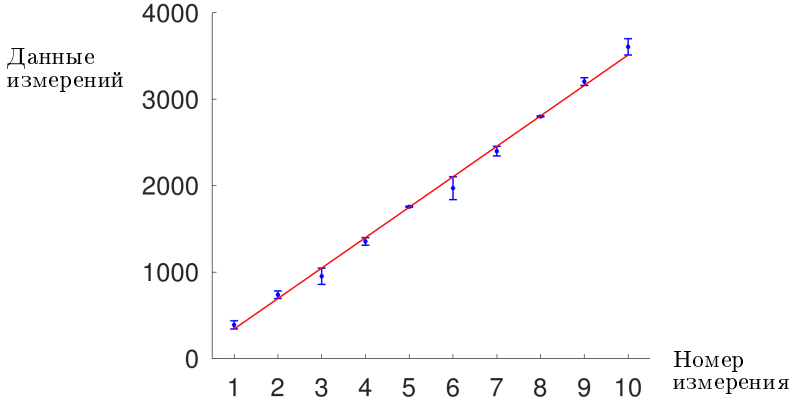

Рис. 4.7. Регрессия с оценкой по норме L_1 (данные табл. 4.1)

Вектор весов w радиусов отдельных замеров приведен в (4.13). Рис. 4.7 и высокая неоднородность значений w свидетельствуют о разной по величине степени отклонении данных от регрессионной прямой на разных участках оси абсцисс

$$w = \{4{,}8, 4{,}8, 17{,}7, 8{,}7, 0{,}17, 22{,}3, 9{,}0, 0{,}17, 8{,}8, 17{,}7\}. \tag{4.13}$$

Наибольшее отклонение от регрессионной прямой и максимальные веса, необходимые для достижения совместности, имеет измерение в середине рассматриваемого участка.

Интервальная оценка параметров регрессии. Приступим к интервальной оценке параметров регрессии. Ясно, что при достаточно высокой погрешности данных выборка станет *накрывающей* или по крайней мере совместной согласно п. 2.6.1.

Для этого необходимо приписать данным какие-то дополнительные погрешности, помимо погрешностей квантования. Значения компонент вектора w несут индивидуальную информацию о каждом измерении. Такая информация обладает высокой степенью избыточности, и желательно ее заменить на более экономное представление. Имеет смысл в качестве первой оценки

реалистичной погрешности данных взять близкую к максимальному значению εw в (4.13). Итак, примем для всех измерений значение

$$\operatorname{rad} \boldsymbol{y}_i := \varepsilon = \max_i \varepsilon_i w_i \simeq 150.$$

Информационное множество параметров $\boldsymbol{I}$. Определим интервальные параметры регрессии по методике [29]. На рис. 4.8 изображено информационное множество п. 4.3 параметров модели (4.11) — сдвигов и наклонов регрессионной прямой. Оно ограничено многоугольником и выделено заливкой. Также на рис. 4.8 приведены различные точечные оценки. Они определены в результате вычисления максимальной диагонали, центра тяжести, методом наименьших квадратов, точечной регрессией. Для заданного значения погрешности данных все точечные оценки содержатся в информационном множестве.

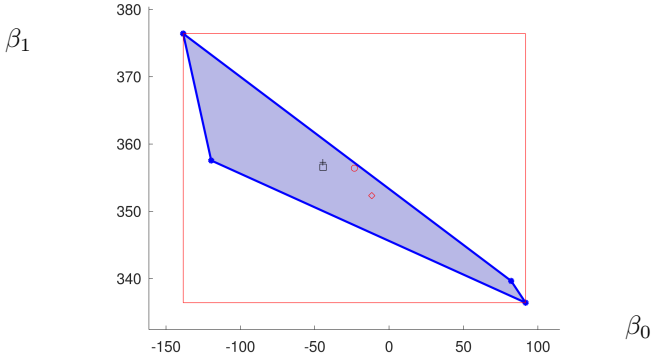

Рис. 4.8. Информационное множество $\boldsymbol{I}$, погрешность $\varepsilon = 150$

Коридор совместности Υ. На рис. 4.9 изображены диаграмма рассеяния данных и коридор совместности п. 4.4 для полученных параметров модели регрессии для заданной модели погрешности данных.

Также дана прямая регрессии по параметрам, соответствующим центру тяжести множества, показанного на рис. 4.8. Для значения независимой переменной, равной 6, эта прямая касается границ коридора совместности. То есть в этом месте имеется «излом» множества Υ.

Прогноз значений выходной переменной. Важнейшим назначением регрессионной модели является предсказание значений выходной переменной для заданных значений входной.

С помощью информационного множества $\boldsymbol{I}$ для построенной модели

$$\boldsymbol{y}(x) = [-138{,}4, 91{,}7] + [336{,}4, 376{,}4] \cdot x \tag{4.14}$$

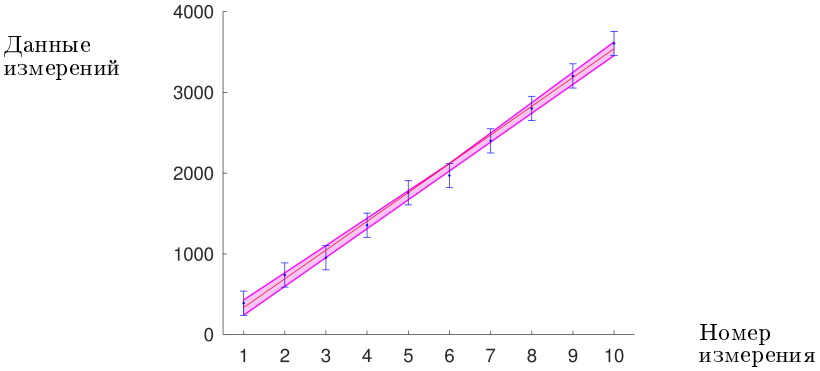

Рис. 4.9. Диаграмма рассеяния и коридор совместности Υ, $\varepsilon = 150$.

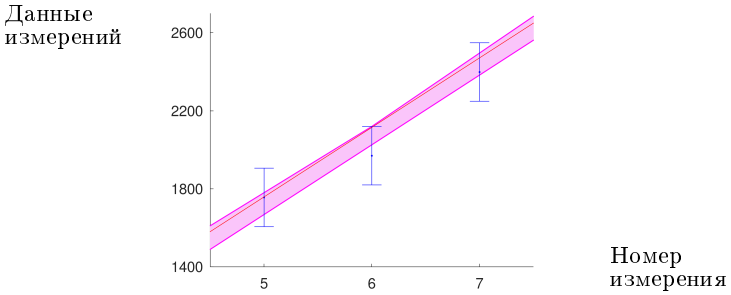

Рис. 4.10. «Излом» множества Υ.

можно получить прогнозные значения выходной переменной в точках эксперимента. В (4.14) для величин параметров регресссии $(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1)$ взяты величины интервальной облочки $\square\boldsymbol{I}$ см. рис. 4.8.

Ценность модели заключается в возможности ее примененения для предсказания выходной переменной в точках, где измерения не производились. Для иллюстрации приведем прогнозы в одной точке внутри диапазона $x = 5$ и двух точках за его границами $x = -1$, $x = 15$. Графически результат расчета представлен на рис. 4.11.

Численные результаты расчетов представлены в табл. 4.2. Как видно, чем более удалена точка прогноза от области данных, тем больше предсказываемая погрешность.

Уточнение модели погрешности данных. Итак, при значении погрешности данных, равной $\varepsilon = 150$, получены согласованные оценки пара-

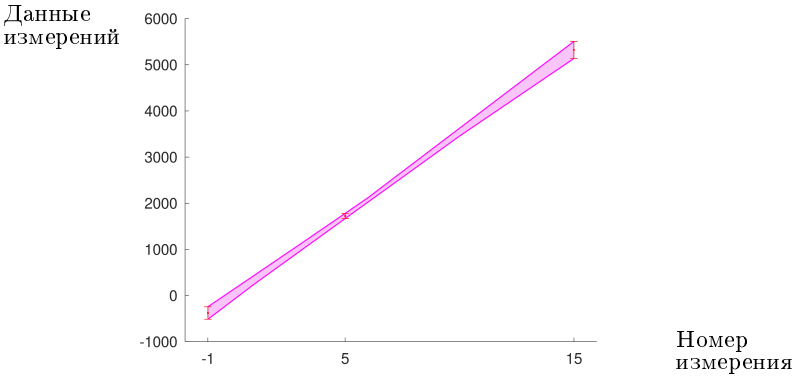

Рис. 4.11. Прогноз значений внутри и вне интервала имеющихся данных, погрешность данных $\varepsilon = 150$

i	x_i	mid $\boldsymbol{y}$	rad $\boldsymbol{y}_i$	$\underline{\boldsymbol{y}}_i$	$\overline{\boldsymbol{y}}_i$
1	-1	-380	148	-515	-245
2	5	1724	56	1689	1780
3	15	5323	185	5318	5508

Таблица 4.2. Прогноз измерений по модели (4.14).

метров линейной модели данных (4.14). Напомним, что величина ε выбрана с запасом для обеспечения заведомого согласования данных и линейной модели. Посмотрим, что произойдет при попытке уменьшить эту неопределенность. Пусть $\varepsilon = 100$.

Определим интервальные параметры регрессии. На рис. 4.12 приведено новое информационное множество сдвигов и наклонов регрессионной прямой.

Множество параметров линейной модели на рис. 4.12 существенно меньше аналогичного множества рис. 4.8. Конкретные значения ширин параметров β приведены в табл. 4.3. Согласование модели и данных в таких условиях становится проблематично. В частности, оценка точечных параметров модели методом наименьших квадратов (черный квадратик на рис. 4.12) находится за пределами $\boldsymbol{I}$.

Уменьшение информационного множества приводит к сужению коридора совместности параметров модели. На рис. 4.13 приведены диаграмма рассеяния данных и коридор совместности параметров модели регрессии Υ для

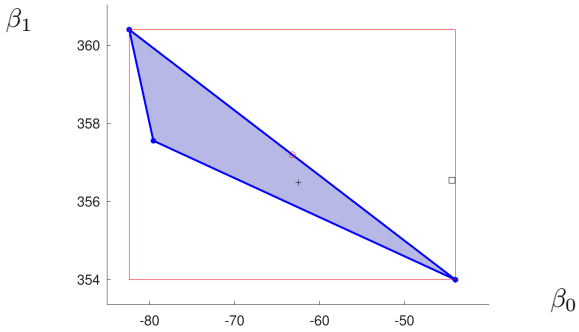

Рис. 4.12. Информационное множество, погрешность данных $\varepsilon = 110$

ε	wid $\boldsymbol{\beta}_1$	wid $\boldsymbol{\beta}_2$
100	$\simeq 29$	$\simeq 4$
150	$\simeq 250$	$\simeq 38$

Таблица 4.3. Размеры множества параметров линейной модели данных

заданной погрешности данных. Коридор совместности Υ представляет собой узкую полосу, проходящую через крайние значения нескольких брусов. Коридор совместности касается множества вершин брусов

$$\Omega_B = \{\underline{\boldsymbol{y}}_1, \overline{\boldsymbol{y}}_6, \underline{\boldsymbol{y}}_{10}\}. \tag{4.15}$$

Как было отмечено ранее, в середине графика для измерения 6 имеется «излом». Дальнейшее уменьшение ε приводит к пустоте множества параметров. При $\varepsilon = 100$ выборка становится *ненакрывающей*.

В п. 4.5 дана классификация данных выборки по отношению к формированию информационного множества и введено понятие *граничных* измерений информационного множества. Подмножество всех граничных наблюдений в S_n играет особую роль, поскольку оно является *минимальной подвыборкой, полностью определяющей модель*. В рассмотренном примере это множество Ω_B (4.15). Удаление неграничных наблюдений из выборки не изменяет модель.

В приведенном примере продемонстрирована технология обработки выборки с точными значениями входных переменных и *неизвестной заранее*

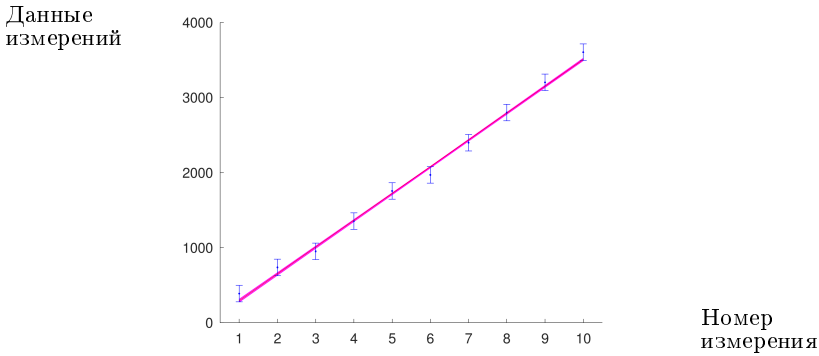

Рис. 4.13. Диаграмма рассеяния и коридор совместности Υ, погрешность данных $\varepsilon = 110$

погрешностью данных. Выбором модели погрешностей выборка была сделана *накрывающей*. Инструментом служил аппарат линейного программирования. Далее было показано, что при занижении погрешности данных происходит уменьшение информационного множества вплоть до его пустоты.

Помимо техники линейного программирования, можно проводить вычисления посредством нахождения максимума функционала специального вида, что носит название «метод максимума согласования» [41]. Программно метод поддерживается свободно распространяемыми программами, доступными на сайте [42]. Для данных табл. 4.1 нахождение неизвестных β_0, β_1 с использованием программы `tolsolvty` [42] для прямой (4.11) дает уравнение регресиии

$$y = -72{,}4 + 357{,}6 \cdot x. \tag{4.16}$$

Прямая для (4.16) близка к прямой на рис. 4.7 по (4.12).

4.7 Общий случай задачи восстановления зависимостей

Рассмотрим случай, когда неопределенность присутствует как в измерениях значений зависимой переменной, так и в измерениях значений аргументов (рис. 4.1). Это может быть вызвано различными причинами, отчасти рассмотренными в [3] . Например, существенно неточное измерение входных переменных происходит в ситуациях, когда они должны устанавливаться в течение значительного времени. Тогда их уместно выразить интервалами, а не точечными значениями.

Отметим, что этот класс задач сложнее, чем рассмотренный выше случай точных измерений входных переменных п. 4.6. Эта сложность относится как к постановке задачи, так и методам решения. Развернутое, хотя и не исчерпывающее всех аспектов проблемы, обсуждение проведено в [1].

Если выборка измерений независимых переменных и зависимой переменной — накрывающая, то

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_m x_{im} \in \boldsymbol{y}_i, \qquad i = 1, 2, \ldots, n,$$

где все x_{i1} могут принимать значения из соответствующих интервалов $\boldsymbol{x}_{i1}$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$. Как следствие, получаем интервальную систему линейных алгебраических уравнений (ИСЛАУ), сходную с (4.10).

Это формальная запись, означающая совокупность обычных (точечных) систем линейных алгебраических уравнений того же размера и с теми же неизвестными переменными, у которых коэффициенты и правые части лежат в предписанных им интервалах (см. [4]). Восстановление параметров линейной зависимости можно рассматривать как решение выписанной интервальной системы уравнений.

В случае присутствия погрешностей как в измерениях аргумента, так и в измерениях зависимости, множество параметров зависимостей, совместных (согласующихся) с данными, характеризуются новыми свойствами. Множества решений отдельных интервальных уравнений уже не являются полосами в пространстве $\mathbb{R}^n$, вроде тех, что изображены на рис. 4.5. Их конкретный вид зависит от того, какой смысл вкладывается в понятие совместности (согласования) параметров и данных, т. е. от того, какое множество решений ИСЛАУ взято в качестве информационного множества.

Понятие совместности (согласования) параметров и данных должно быть расширено и переосмыслено. В обычном неинтервальном случае результаты измерений — это точки, и прохождение через них графика функциональной зависимости описывается двумя значениями — «да» или «нет». Для брусов неопределенности имеются различные варианты прохождения через них графика зависимости. Брус неопределенности измерений $(\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{im}, \boldsymbol{y}_i)$ является прямым декартовым произведением интервалов по различным осям координат, и эти оси имеют разный смысл: интервалы $\boldsymbol{x}_{i1}$, $\boldsymbol{x}_{i2}$, $\ldots$, $\boldsymbol{x}_{im}$ соответствуют входным переменным, а интервал $\boldsymbol{y}_i$ — выходной переменной. При этом становится важным, как именно проходит график восстанавливаемой зависимости через брусы неопределенности измерений (рис. 4.14).

Функциональную зависимость называют *слабо совместной* с интервальными данными, если ее график проходит через каждый брус неопределенности измерений хотя бы для одного значения аргумента. График зависимости пересекает брусы неопределенности, но как именно — неважно (рис. 4.14, средний брус), достаточно не менее одной точки пересечения.

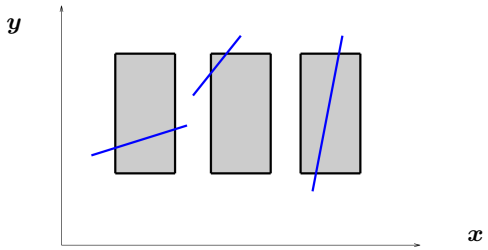

Рис. 4.14. Различные способы пересечения линии с брусом неопределенности измерения зависимости

Функциональную зависимость назовем *сильно совместной* с интервальными данными, если ее график проходит через каждый брус неопределенности измерений для любого значения аргумента из интервалов неопределенности входных переменных. График зависимости целиком содержится в коридорах, задаваемых интервалами выходной переменной при всех значениях входных переменных из соответствующих им интервалов (рис. 4.14, левый брус).

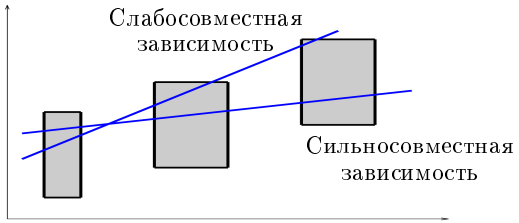

Рис. 4.15. Линейные зависимости с разными типами согласования с данными.

На рис. 4.14 правый брус соответствует ситуации, когда график зависимости лежит в коридоре, задаваемом интервалом входной переменной $\boldsymbol{x}$, при любых значениях выходной переменной y из соответствующего ей интервала $\boldsymbol{y}$.

Сильная совместность при интервальной неопределенности данных означает, что выходная величина остается в пределах измеренного для нее интервала вне зависимости от конкретных значений входных переменных внутри их интервала. В работах С. П. Шарого, например, в [5], показано, что требование сильной совместности параметров и данных позволяет обрабатывать

различные сложные случаи восстановление зависимостей по широким и существенно «перекрывающимся» интервальным данным.

В настоящее время предложено несколько методов восстановления линейных зависимостей: метод центра неопределенности [43, 11], метод максимума совместности (максимума согласования) [41, 6, 9], метод парциальных информационных множеств [13, 14] и др.

Восстановление зависимостей по ненакрывающим выборкам. В случае восстановления зависимостей по ненакрывающим выборкам не может быть универсальных подходов. В [1] предлагается в виде критерия, по которому можно определять пригодность восстанавливаемой зависимости, использовать *расстояние до брусов* данных. Там же обсуждаются возникающие сложные, иногда парадоксальные ситуации при обработке ненакрывающих выборок. В целом в данной области еще немного результатов, и это передний фронт исследований в области анализа данных с интервальной неопределенностью [47].

Восстановление нелинейных зависимостей. Восстановление нелинейной зависимости в принципиальном плане не отличается от линейного случая. Пример применения интервального подхода с использованием полиномов второго порядка содержится в работе [48].

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Обработка и анализ данных с интервальной неопределенностью / **А. Н. Баженов [и др.]**. — Ижевск : РХД, 2022 - 270 с.
2. **Баженов А. Н.** Интервальный анализ. Основы теории и учебные примеры : учеб. пособие / А.Н. Баженов. — СПб., 2020. URL: https://elib.spbstu.ru/dl/2/s20-76.pdf/info
3. **Баженов А. Н.** Естественнонаучные и технические применения интервального анализа. : учеб. пособие / А.Н. Баженов. — СПб., 2021. URL: https://elib.spbstu.ru/dl/5/tr/2021/tr21-169.pdf/info
4. **Шарый С. П.** Конечномерный интервальный анализ. / С. П. Шарый [Электронный ресурс] ;– ФИЦ ИВТ. Новосибирск, 2022. URL: http://www.nsc.ru/interval/Library/InteBooks/SSharyBook.pdf
5. **Шарый С. П.** Сильная согласованность в задаче восстановления зависимостей при интервальной неопределенности данных / С. П. Шарый. // Вычислительные технологии. – 2017. – Т. 2, № 2. – С. 150–172.
6. **Шарый С. П.** Метод максимума согласования для восстановления зависимостей по данным с интервальной неопределенностью / С. П. Шарый. // Известия академии наук. Теория и системы управления. – 2017. – № 6. – С. 3–19.
7. **С. П. Шарый.** Выявление выбросов в методе максимума согласования при анализе интервальных данных / С. П. Шарый // МАК-2018 Сборник трудов Всероссийской конференции по математике с международным участием. – Барнаул : Изд-во АлтГУ, 2018. – С. 215–218. URL: http://elibrary.asu.ru/handle/asu/6303
8. **С. П. Шарый.** О мере вариабельности оценки параметров в статистике интервальных данных / С. П. Шарый // Вычислительные технологии. – 2019. – Т. 24, № 5. – С. 90–108.
9. **С. П. Шарый.** Задача восстановления зависимостей по данным с интервальной неопределенностью / С. П. Шарый // Заводская лаборатория. Диагностика материалов. – 2020. – Т. 86, № 1. – С. 62–74.
10. DRS4 microchip. URL: https://www.psi.ch/en/drs/documentation

11. **Zhilin S. I.** On fitting empirical data under interval error / S. I. Zhilin // Reliable Computing. – 2005. – Vol. 11. – P. 433–442.
12. **Zhilin S. I.** Simple method for outlier detection in fitting experimental data under interval error. / S. I. Zhilin // Chemometrics and Intelligent Laboratory Systems. – 2007. – Vol. 88, No. 1. – P. 60-68
13. **Кумков С. И.** Обработка экспериментальных данных ионной проводимости расплавленного электролита методами интервального анализа. / С. И. Кумков // Расплавы. – 2010. – №3. – С. 79–89.
14. **Kumkov, S.I.** Interval approach to identification of catalytic process parameters./ S. I. Kumkov, Yu. V. Mikushina // Reliable Computing. – 2013. – Vol. 19. – P. 197–214.
15. Computing Statistics under Interval and Fuzzy Uncertainty. **H. T. Nguyen [et al.]** // Applications to Computer Science and Engineering. – Springer. Berlin; Heidelberg, 2012.
16. **Tukey J. W.** The Future of Data Analysis. / J. W. Tukey // Annals of Mathematical Statistics – 1962. – Vol. 33, Issue 1. – P. 1–67
17. **Маликов М. Ф.** Основы метрологии. Ч1. Учение об измерении. / М. Ф. Маликов ; – Комитет по делам мер и измерительных приборов при Совете Министров СССР. М., 1949. – 479 с.
18. ГОСТ 34100.3—2017/ISO/IEC Guide 98-3:2008. Неопределенность измерения. Ч3. Руководство по выражению неопределенности измерения (ISO/IEC Guide 98-3:2008, ЮТ). Межгосударственный стандарт. – М. Стандартинформ, 2017.
19. ГОСТ 34100.3.2–2017/ISO/IEC Guide 98-3:2008. неопределенность измерения. Ч3. Руководство по выражению неопределенности измерения. Дополнение 2. Обобщение на случай произвольного числа выходных величин (ISO/IEC Guide 98-3:2008, ЮТ). Идентичен Guide 98-3/Suppl 2:2011. Межгосударственный стандарт. Издание официальное. – М. Стандартинформ, 2017.
20. ГОСТ Р 8.736–2011 Государственная система обеспечения единства измерений (ГСИ). Измерения прямые многократные. Методы обработки результатов измерений. Основные положения. – М. Стандартинформ, 2019.
21. **Пуанкаре А.** Теория вероятностей. / А. Пуанкаре – Ижевск : РХД, 1999.
22. **Орлов А. И.** Распределения реальных статистических данных не являются нормальными // Научный журнал КубГАУ. – 2016. – № 117 (03). – С. 71-90.
23. **А.Н. Баженов** Тензорные разложения и их применение в флуориметрии: учеб. пособие. / А.Н. Баженов, Т. О. Яворук. – СПб, 2021. URL: https://elib.spbstu.ru/dl/5/tr/2021/tr21-170.pdf/info

24. **R.Boukezzoula.** Gradual interval arithmetic and fuzzy interval arithmetic / R. Boukezzoula. // Granular Computing (2021) 6:451–471. URL: https://doi.org/10.1007/s41066-019-00208-z

25. Atomic weights of the elements 2013 (IUPAC Technical Report) / J. Meija [et al] // Pure and Applied Chemistry. – 2016. – Vol. 88, Issue 3. – P. 265–291.

26. **Канторович Л. В.** О некоторых новых подходах к вычислительным методам и обработке наблюдений / Л. В. Канторович // Сибирский Математический Журнал. – 1962. – Т. 3, № 5. – С. 701–709.

27. Standardized notation in interval analysis / Kearfott, R.B. [et al] // Вычислительные Технологии. – 2010. – Т. 15, № 1. – С. 7–13.

28. **Жилин С. И.** Реализация интервальной арифметики Каухера в системе компьютерной математики Octave / С. И. Жилин [Электронный ресурс]— URL: https://github.com/szhilin/kinterval

29. **Жилин С. И.** Примеры и программы анализа интервальных данных в системе компьютерной математики Octave / С. И. Жилин [Электронный ресурс] — URL: https://github.com/szhilin/octave-interval-examples

30. Уравнение Михаэлиса-Ментен.
— URL: https://en.wikipedia.org/wiki/Michaelis-Menten_kinetics

31. РМГ 29–2013 ГСИ. Рекомендации по межгосударственной стандартизации. Государственная система обеспечения единства измерений. Метрология. Основные термины и определения. – М.: Стандартинформ, 2014.

32. **Нестеров В. М.** Твинные арифметики и их применение в методах и алгоритмах двустороннего интервального оценивания : дис. ... д.ф.-м.н. - СПб : СПИИРАН, 1999, 234 с.

33. Circular polarization of γ-quanta in the $np \to d\gamma$ reactions with polarized neutrons / A. N. Bazhenov [et al] // Physics Letters B,3 September 1992, Vol 289, 1–2, Pages 17-21.

34. **Hu C.** On statistics, probability, and entropy of interval-valued datasets / C. Hu, Z. H. Hu // Lesot MJ. et al. Communications in Computer and Information Science, – Cham: Springer, 2020. vol. 1239.

35. Спектр масс осколков ядра при захвате нейтронов. URL: https://ru.wikipedia.org/Деление_ядра

36. **Пролубников А В.** Попытка построения интервальной медианы. / А. В. Пролубников // Презентация на Всероссийском вебинаре по интервальному анализу: 2022

37. Коэффициент Жаккара
URL: https://en.wikipedia.org/wiki/Jaccard_index

38. Novel Similarity Measure for Interval-Valued Data Based on Overlapping Ratio / S. Kabir [et al] // IEEE International Conference on Fuzzy Systems — 2017.

39. Уравнение Михаэлиса-Ментен.
URL: https://en.wikipedia.org/wiki/Michaelis-Menten_kinetics

40. **Shary S.** Enclosing vs. Non-enclosing Measurements in Interval Data Processing / S. Shary // January 2022.
URL https://www.researchgate.net/publication/357857123

41. **Шарый С. П.** Распознавание разрешимости интервальных уравнений и его приложения к анализу данных / С. П. Шарый, И. А. Шарая // Вычислительные технологии. – 2013. – Т. 18. – № 3. – С. 80–109.

42. **Шарый С. П.** Программы анализа интервальных данных / С. П. Шарый [Электронный ресурс] //
URL http://www.nsc.ru/interval/shary/index.html

43. **Жилин С.И.** Нестатистические методы и модели построения и анализа зависимостей / С. И. Жилин : – lис. .... к.ф.-м.н. Барнаул, 2004.
URL http://www.nsc.ru/interval/Library/ApplDiss/Zhilin.pdf

44. **Вощинин А. П.** Оптимизация в условиях неопределенности / А. П. Вощинин, Г. Р. Сотиров – М: Изд-во МЭИ ; София : Техника, 1989. – 224 с.

45. **Оскорбин Н. М.** Построение и анализ эмпирических зависимостей методом центра неопределенности / Н. М. Оскорбин, А. В. Максимов, С. И. Жилин // Барнаул : Изв. Алтайского гос. университета. – 1998. – № 1. – С. 35–38.

46. **Ермаков Н. В.** Стенд для испытаний шаговых двигателей / Н. В. Ермаков [и др] ПТЭ, № 1. 2023.

47. **Звягин М. А.** Об одном подходе к восстановлению зависимостей по ненакрывающим интервальным данным / М. А. Звягин , С. П. Шарый // Вычислительные Технологии. –2022.

48. **Коваленко Н. С.** Визуальная одометрия в условиях интервальной неопределенности. / Н. С. Коваленко СПбПУ . — СПб, 2021.
URL https://elib.spbstu.ru/dl/3/2021/vr/vr21-4531.pdf/info

Предметный указатель

Министерство науки и высшего образования Российской Федерации

САНКТ-ПЕТЕРБУРГСКИЙ
ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО

Физико-механический институт
Высшая школа прикладной математики и вычислительной физики

А. Н. Баженов

ВВЕДЕНИЕ В АНАЛИЗ ДАННЫХ С ИНТЕРВАЛЬНОЙ НЕОПРЕДЕЛЕННОСТЬЮ

Учебное пособие

ПОЛИТЕХ-ПРЕСС
Санкт-Петербургский
политехнический университет
Петра Великого

Санкт-Петербург

2022