А. И. Орлов

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ: НЕЧИСЛОВАЯ СТАТИСТИКА

Учебник

Москва
Ай Пи Ар Медиа
2022

УДК 004.8
ББК 32.81
О-66

Автор:
Орлов А. И. —д-р экон. наук, д-р техн. наук, канд. физ.-мат. наук,
проф. кафедры экономики и организации производства (ИБМ-2)
Московского государственного технического
университета имени Н. Э. Баумана

Орлов, Александр Иванович.
О-66 Искусственный интеллект: нечисловая статистика : учебник / А. И. Орлов. — Москва : Ай Пи Ар Медиа, 2022. — 446 с. — Текст : электронный.

ISBN 978-5-4497-1435-0

В учебнике впервые систематически рассмотрена важная составляющая искусственного интеллекта — сердцевина высоких статистических технологий, одна из четырех основных областей современной прикладной математической статистики — нечисловая статистика. Она порождена потребностями прикладных социально-экономических, технических и медико-биологических исследований. В издании раскрыты основные виды нечисловых данных, методология, процедуры и особенности их статистического анализа. Представлены статистические методы в пространствах произвольной природы, статистика нечисловых данных конкретных видов, статистика интервальных данных. Большое внимание уделяется проблемам практического применения методов и результатов нечисловой статистики.

Подготовлено в соответствии с требованиями Федерального государственного образовательного стандарта высшего образования.

Учебник рекомендуется к использованию студентам при изучении таких дисциплин, как «Прикладная статистика», «Эконометрика», «Организационно-экономическое моделирование», «Методы принятия управленческих решений», «Математические модели микроэкономики», «Математические модели макроэкономики» по направлениям подготовки 02.03.01 «Математика и компьютерные науки», 27.03.05 «Инноватика», 38.03.01 «Экономика», 38.03.02 «Менеджмент». Книга представляет интерес также для исследователей современных статистических методов в технике, экономике, управлении, медицине, социологии и иных областях, а также для разработчиков таких методов и соответствующего программного обеспечения.

Учебное электронное издание

ISBN 978-5-4497-1435-0

© Орлов А. И., 2022
© ООО Компания «Ай Пи Ар Медиа», 2022

Технический редактор, компьютерная верстка *А.В. Неверова*
Обложка *С.С. Сизиумовой*

Подписано к использованию 07.10.2022. Объем данных 9 Мб.

Издание представлено в электронно-библиотечных системах
IPR BOOKS (www.iprbookshop.ru),
Библиокомплектатор (www.bibliocomplectator.ru)

Бесплатный звонок по России: **8-800-555-22-35**
Тел.: 8 (8452) 24-77-97, 8 (8452) 24-77-96

Отдел продаж и внедрения ЭБС:
доб. 206, 213, 144, 145
E-mail: sales@iprmedia.ru

Отдел комплектования ЭБС:
доб. 224, 227, 208
E-mail: mail@iprbookshop.ru

По вопросам приобретения издания обращаться:
доб. 208, 201, 222, 224
E-mail: izdat@iprmedia.ru, author@iprmedia.ru

СОДЕРЖАНИЕ

ПРЕДИСЛОВИЕ 6

ВВЕДЕНИЕ. НЕЧИСЛОВАЯ СТАТИСТИКА — ОСНОВА ВЫСОКИХ СТАТИСТИЧЕСКИХ ТЕХНОЛОГИЙ 14

1. О развитии статистических методов 14
2. Структура нечисловой статистики 34

Литература 47

ГЛАВА 1. НЕЧИСЛОВЫЕ СТАТИСТИЧЕСКИЕ ДАННЫЕ 52

1.1. Количественные и категоризованные данные 52
1.2. Основы теории измерений 57
1.3. Виды нечисловых данных 64
1.4. Вероятностные модели порождения нечисловых данных 79
1.5. Нечеткие множества – частный случай нечисловых данных 97
1.6. Сведение нечетких множеств к случайным 109
1.7. Данные и расстояния в пространствах произвольной природы 119
1.8. Аксиоматическое введение расстояний 125

Темы докладов, рефератов, исследовательских работ 136
Контрольные вопросы и задачи 137
Литература 138

ГЛАВА 2. СТАТИСТИЧЕСКИЕ МЕТОДЫ В ПРОСТРАНСТВАХ ПРОИЗВОЛЬНОЙ ПРИРОДЫ 142

2.1. Эмпирические и теоретические средние 142
2.2. Законы больших чисел 151
2.3. Экстремальные статистические задачи 160
2.4. Одношаговые оценки 163
2.5. Непараметрические оценки плотности 173
2.6. Статистики интегрального типа 181
2.7. Методы восстановления зависимостей 201
2.8. Методы классификации 210
2.9. Методы шкалирования 230

Темы докладов, рефератов, исследовательских работ 239
Контрольные вопросы и задачи 239
Литература 240

ГЛАВА 3. СТАТИСТИКА НЕЧИСЛОВЫХ ДАННЫХ КОНКРЕТНЫХ ВИДОВ 247
3.1. Инвариантные алгоритмы и средние величины 247
3.2. Теория случайных толерантностей 254
3.3. Метод проверки гипотез по совокупности малых выборок 264
3.4. Теория люсианов 274
3.5. Метод парных сравнений 291
3.6. Статистика нечетких множеств 298
3.7. Статистика нечисловых данных в экспертных оценках 305
Темы докладов, рефератов, исследовательских работ 323
Контрольные вопросы и задачи 324
Литература 327

ГЛАВА 4. СТАТИСТИКА ИНТЕРВАЛЬНЫХ ДАННЫХ 332
4.1. Основные идеи статистики интервальных данных 332
4.2. Интервальные данные в задачах оценивания 341
4.3. Интервальные данные в задачах проверки гипотез 375
4.4. Линейный регрессионный анализ интервальных данных 379
4.5. Интервальный дискриминантный анализ 408
4.6. Интервальный кластер-анализ 410
4.7. Интервальные данные в инвестиционном менеджменте 413
4.8. Статистика интервальных данных в прикладной статистике 418
Темы докладов, рефератов, исследовательских работ 421
Контрольные вопросы и задачи 421
Литература 420

ПРИЛОЖЕНИЯ 425
Приложение 1. Теоретическая база нечисловой статистики 425
Приложение 2. Об авторе 442

ПРЕДИСЛОВИЕ

В «Национальной стратегии развития искусственного интеллекта на период до 2030 г.»[1] принято следующее определение: «...искусственный интеллект — комплекс технологических решений, позволяющий имитировать когнитивные функции человека (включая самообучение и поиск решений без заранее заданного алгоритма) и получать при выполнении конкретных задач результаты, сопоставимые, как минимум, с результатами интеллектуальной деятельности человека. Комплекс технологических решений включает в себя информационно-коммуникационную инфраструктуру, программное обеспечение (в том числе в котором используются методы машинного обучения), процессы и сервисы по обработке данных и поиску решений». В этом определении прямо не говорится про научную основу «комплекса технологических решений». По нашему мнению, в социально-экономической области в качестве такой основы можно использовать организационно-экономическое моделирование, включая высокие статистические технологии, в том числе нечисловую статистику, теорию и практику экспертных оценок, статистические методы анализа данных).

Автор занимается проблемами искусственного интеллекта около полувека (первые статьи напечатаны в 1972 г.). Настоящая книга посвящена важной составляющей искусственного интеллекта — нечисловой статистике (статистике нечисловых данных, статистике объектов нечисловой природы).

В учебнике впервые в мире систематически рассматривается одна из четырех основных областей современной прикладной статистики — нечисловая статистика. Она порождена в 70-х гг. XX в. потребностями прикладных социально-экономических, технических и медико-биологических исследований. Основой ее математического аппарата является использование расстояний между объектами нечисловой природы и решений оптимизационных задач, а не операций суммирования данных, как в других областях статистики. В учебнике рассмотрены основные виды нечисловых данных и особенности их статистического анализа. Большое внимание уделяется проблемам практического применения рассматриваемых методов и результатов.

Нечисловую статистику называют также статистикой нечисловых данных или статистикой объектов нечисловой природы. Она является сердцевиной высоких статистических технологий, т.е. современной прикладной статистики. Ее рассматривают также как одну из четырех основных обла-

[1] Указ Президента РФ от 10 октября 2019 г. № 490 «О развитии искусственного интеллекта в Российской Федерации» // СПС Гарант. URL: https://www.garant.ru/products/ipo/prime/doc/72738946/ (дата обращения: 24.09.2021).

стей статистики. Три других — это статистика чисел (случайных величин), статистика векторов (многомерный статистический анализ), статистика функций (временных рядов и случайных процессов).

Какие данные называют нечисловыми? Описание технического, социально-экономического, медицинского объекта изучения часто удается представить в виде вектора, часть координат которого измерена по количественным шкалам, а часть — по качественным, имеющим конечное число градаций. Это — наиболее распространенный тип нечисловых данных.

В общем случае под нечисловыми данными понимают элементы пространств, не являющихся линейными (векторными), в которых нет операций сложения элементов и их умножения на действительное число. Кроме результатов измерений по качественным признакам, примерами являются последовательности из 0 и 1, бинарные отношения (ранжировки, разбиения, толерантности); множества (в том числе плоские изображения и объемные тела); нечеткие (размытые, расплывчатые, *fuzzy*) числа и множества, их частный случай — интервалы; результаты парных сравнений и другие объекты, возникающие в прикладных исследованиях. Все эти виды нечисловых данных и вероятностные модели их порождения подробно рассматриваются в учебнике. Их обобщением, как и обобщением числовых данных (чисел, векторов, функций), являются элементы пространств произвольной природы.

Исторически нечисловые данные стали рассматриваться раньше, чем статистические данные в виде действительных чисел. Книга Чисел Ветхого Завета содержит обширные сведения о численностях тех или иных совокупностей. Натуральные числа можно отнести к нечисловым данным — хотя их можно складывать, но умножение на действительное число выводит за пределы натурального ряда. Теория вероятностей также начиналась с моделирования нечисловых данных, таких, как результаты бросания игральных костей и вытаскивания шаров из урн. Однако к началу XX в. основное внимание статистиков переместилось на рассмотрение числовых случайных величин, моделирующих действительнозначные результаты наблюдений.

К 70-м гг. XX в. развитие прикладных научных исследований в инженерном деле, социологии, экономике, менеджменте, психологии, медицине и других областях привело к необходимости разработки методов статистического анализа нечисловых данных. В СССР вокруг всесоюзного семинара «Экспертные оценки и нечисловая статистика» сложился неформальный научный коллектив из нескольких десятков активных исследователей.

Сначала изучались методы анализа конкретных видов нечисловых данных, устанавливались связи между ними. Затем пришло понимание статистики нечисловых данных как самостоятельной области прикладной статистики со своей внутренней структурой и разнообразными связями между подходами и результатами, относящимися к тем или иным видам нечисловых данных.

Статистика нечисловых данных была выделена нами как самостоятельная область прикладной статистики в 1979 г. За прошедшие с тех пор годы арсенал ее методов пополнился многими полезными новшествами. Но основные идеи выдержали проверку временем, что и оправдывает их изложение в настоящей книге.

О развитии нечисловой статистики. Как уже отмечалось, в СССР в 70-е гг. XX в. возник неформальный научный коллектив исследователей, изучающих методы анализа нечисловых данных различных видов. Центром являлся научный семинар «Экспертные оценки и нечисловая статистика» и одноименная комиссия в составе Научного Совета АН СССР по комплексной проблеме «Кибернетика».

Вначале разбирались подходы предшественников, в частности, аксиоматическое введение расстояний между объектами нечисловой природы и нахождение среднего по Кемени, репрезентативная теория измерений, нечеткие множества Заде, парные сравнения по Дэвиду и др. Затем были проведены многочисленные самостоятельные исследования. В частности, были установлены взаимосвязи между подходами и результатами для различных типов нечисловых данных, разработана общая теория статистического анализа нечисловых данных произвольной природы.

В итоге стало возможным говорить о новой области прикладной статистики — нечисловой статистике. Время ее окончательного формирования — первая половина 1980-х гг. — было и временем наибольшей организационной активности. Две всесоюзные конференции — в Алма-Ате (1981 г.) и в Таллинне (1984 г.) собрали по 300–500 участников.

Со второй половины 80-х гг. XX в. и до настоящего времени нечисловая статистика (статистика нечисловых данных, статистика объектов нечисловой природы) стабильно развивается. Много публикаций содержится в журналах «Заводская лаборатория. Диагностика материалов», «Научный журнал КубГАУ», «Контроллинг», «Инновации в менеджменте», «Социология: методология, методы, математические модели», периодических сборниках «Статистические методы оценивания и проверки гипотез» и «Управление большими системами». Неформальный коллектив по нечисловой стати-

стике включает в себя десятки российских исследователей, а если учитывать авторов одной-двух работ — то и сотни. За почти 30 лет выпущено несколько десятков сборников и монографий, много статей в научных журналах. Однако из-за отсутствия формальной инфраструктуры (например, Института нечисловой статистики в составе Российской академии наук) имеются лишь единичные методики и программные продукты, предназначенные для практического использования. В отличие от научных монографий практически отсутствуют учебники и учебные пособия, а также книги, содержащие введение и общий обзор нечисловой статистики.

Настоящая книга заполняет существенный пробел в литературе по нечисловой статистике. Она дает введение в предмет, позволяет познакомиться с нечисловой статистикой на современном научном уровне. Изложение доводится до переднего края ведущихся в настоящее время научных исследований. Постоянно в поле зрения находятся вопросы практического применения рассматриваемых подходов, методов, результатов. В частности, используется опыт разработки нашим коллективом автоматизированного рабочего места МАТЭК (математика в экспертизе), предназначенного для организатора экспертного опроса. В монографии отражены также работы по статистике нечисловых данных и ее применениям, за которые автору в 1992 г. была присуждена ученая степень доктора технических наук (по научному докладу об опубликованных работах, т.е. без написания диссертации классического вида).

Чтобы в сравнительно небольшой книге охватить всю статистику нечисловых данных, приходится идти на жертвы. Мы отказываемся от разбора большинства доказательств, отсылая читателей к публикациям, содержащим эти доказательства. Примерами подобного стиля изложения являются обзоры по статистике нечисловых данных, помещенные в разделе «Математические методы исследования» журнала «Заводская лаборатория» (1990, № 3; 1995, № 3, № 5; 1996, № 3; 2019, № 11).

Стиль книги. В любой математизированной области есть три уровня исследований — методологический, теоретический и практический. На методологическом уровне излагаются общие подходы и формулируются основные результаты. На теоретическом уровне, грубо говоря, доказывают теоремы. В частности, выявление необходимых и достаточных «условий регулярности» обычно осуществляется в результате цепи работ этого уровня.

Например, на методологическом уровне Центральная Предельная Теорема теории вероятностей формулируется так: «При некоторых условиях регулярности распределение центрированной и нормированной суммы незави-

симых случайных величин при росте числа слагаемых стремится к стандартному нормальному распределению». Около двухсот лет — от Муавра и Лапласа до Линдеберга и Феллера — «некоторые условия регулярности» уточнялись в работах теоретического уровня.

В настоящей книге изложение идет в основном на методологическом уровне. При спуске на теоретический уровень приводятся формулировки теорем, в основном без доказательств, но со ссылками на публикации, где они содержатся. Обоснованием для выбора такого варианта построения книги, кроме желания ограничить ее объем разумными рамками, послужило следующее представление о предпочтениях будущих читателей: большинство из них не извлечет пользы из того, что в некоторой формулировке можно заменить требование, скажем, дифференцируемости определенной функции на требование ее непрерывности. Сказанное не означает, что автор отрицает целесообразность проведения научных работ, посвященных подобным ослаблениям условий регулярности. Просто им не место в книге, предназначенной для первого знакомства с нечисловой статистикой.

На практическом уровне исследований большое внимание уделяют конкретному объекту приложений — технической, социально-экономической или медицинской системе. Для достаточно информативного описания каждого такого исследования нужна отдельная монография, которая обычно и готовится в качестве отчета по работе. Поэтому мы вынуждены ограничиться краткими замечаниями о практическом применении различных методов нечисловой статистики. Однако суммарно эти замечания составляют существенную часть как авторского замысла, так и объема книги.

Содержание книги. Во введении кратко обсуждаем историю и современное состояние статистических методов и, прежде всего, прикладной статистики, место в ней статистики нечисловых данных. Анализируется сложившаяся структура нечисловой статистики – сердцевины высоких статистических технологий.

Книга делится на главы, а главы — на разделы. В главе 1 изучаются конкретные виды нечисловых статистических данных, соответствующие вероятностные модели. Сопоставляются количественные и категоризованные данные. Разобраны основы теории измерений. Большое внимание уделено нечетким множествам как частному виду нечисловых данных. Продемонстрирована возможность сведения теории нечетких множеств к теории случайных множеств. Обсуждаются статистические данные и необходимые для их анализа расстояния в пространствах произвольной природы. Обсуждается

аксиоматический подход к введению расстояний и показателей различия в различных пространствах объектов нечисловой природы.

В главе 2 развиваются статистические методы анализа данных произвольного вида, лежащих в метрическом пространстве или в пространстве с мерой различия. Эмпирические и теоретические средние приходится определять как решения экстремальных статистических задач, и законы больших чисел оказываются частными случаями утверждений об асимптотическом поведении решений таких задач. Другие классы частных случаев подобных утверждений связаны с теорией одношаговых оценок параметров распределения вероятностей (они имеют преимущества по сравнению с оценками максимального правдоподобия) и с оптимизационными постановками основных задач прикладной статистики, в том числе задач восстановления зависимостей, классификации, шкалирования и снижения размерности. Для описания распределений нечисловых данных разработаны непараметрические оценки плотности, используемые также в регрессионном, дискриминантном и кластерном анализах. В предельной теории статистик интегрального типа найден ряд необходимых и достаточных условий.

Глава 3 посвящена статистическому анализу конкретных видов нечисловых данных. В частности, в рамках репрезентативной теории измерений получены характеризации средних величин свойством устойчивости результата сравнения средних относительно той или иной группы допустимых преобразований шкалы. Изучены случайные толерантности. Метод проверки гипотез по совокупности малых выборок применен в теории люсианов — конечных последовательностей испытаний Бернулли с, вообще говоря, различными вероятностями успеха. Люсианы находят применение в теории парных сравнений. Рассмотрены основные вопросы статистики нечетких множеств. Обсуждается использование нечисловой статистики в теории и практике экспертных оценок — области исследований, во многом стимулировавшей развитие основных идей статистического анализа нечисловых данных.

Глава 4 посвящена основным подходам и результатам статистики интервальных данных, быстро развивающейся в последние годы. Для интервальных данных решен ряд задач оценивания и проверки гипотез. Построены интервальные аналоги регрессионного, дискриминантного и кластерного анализов. Интервальные данные применены в инвестиционном менеджменте. Рассмотрена роль статистики интервальных данных в прикладной статистике.

В прил. 1 включены некоторые вопросы, относящиеся к теоретической базе нечисловой статистики. Рассмотрены классические законы больших чисел, центральные предельные теоремы, метод линеаризации и принцип инвариантности. Теоремы о наследовании сходимости сравнительно малоизвестны и могут представить особый интерес. В прил. 2 содержится информация об авторе, позволяющая читателям лучше понять происхождение идей, изложению которых посвящена настоящая книга.

Нумерация формул, определений, теорем, таблиц, рисунков — своя в каждом разделе. Литература приводится по главам в порядке первого упоминания. Списки литературы включают основные публикации по нечисловой статистике, а также те работы, на которые даются ссылки в тексте. Они не претендуют на полноту хотя бы потому, что перечень известных автору публикаций по рассматриваемой тематике по объему превысил бы настоящую книгу в несколько раз.

Для кого эта книга? Она предназначена для широкого круга читателей — студентов и преподавателей, прикладников и математиков. Для ее чтения достаточно знаний в объеме вводного курса математической статистики, включающего основные задачи описания данных, оценивания и проверки гипотез.

Эта книга — прежде всего учебник. Он предназначен для студентов различных специальностей, прежде всего технических, управленческих и экономических, слушателей институтов повышения квалификации, структур послевузовского (в том числе второго) образования, в частности, программ MBA («Мастер делового администрирования»), преподавателей вузов. Учебник будет полезен инженерам, менеджерам, экономистам, социологам, биологам, медикам, психологам, историкам, другим специалистам, самостоятельно повышающим свой научный уровень. Короче, всем научным и практическим работникам, связанным с анализом данных.

Учебник может быть использован при изучении дисциплин, полностью или частично посвященных методам анализа нечисловых результатов наблюдений (измерений, испытаний, опытов). Типовые названия таких курсов — «Прикладная статистика», «Эконометрика», «Анализ данных», «Статистический анализ», «Теория принятия решений», «Управленческие решения», «Экономико-математическое моделирование», «Прогнозирование», «Хемометрия», «Математические методы в социологии» и т.п. Учебник необходим студентам направления 15.03.01 Машиностроение (профиль подготовки «Менеджмент высоких технологий»), особенно при изучении учебной дисциплины «Организационно-экономическое моделирование».

Книга будет полезна широкому кругу специалистов, заинтересованных в применении современных статистических методов анализа нечисловых данных в любой предметной области. Она необходима разработчикам таких методов и соответствующего программного обеспечения, т.е. специалистам по прикладной статистике.

Специалистам по теории вероятностей и математической статистике эта книга также может быть интересна и полезна, поскольку в ней описан современный взгляд на прикладную математическую статистику, основные подходы и результаты в этой области, открывающие большой простор для дальнейших математических исследований.

Книга представляет интерес для исследователей — специалистов по вопросам управления, в том числе по принятию решений, методам оптимизации и математическому моделированию. Наконец, без нее не сможет обойтись ни один преподаватель прикладной или математической статистики, статистических методов для любой конкретной области применений, если он хочет, чтобы его лекционный курс был современным.

Благодарности. Автор благодарен за полезные обсуждения многочисленным коллегам по научным семинарам, по работе в Институте высоких статистических технологий и эконометрики МГТУ им. Н. Э. Баумана, в Российской ассоциации статистических методов и Российской академии статистических методов.

С текущей научной информацией по статистическим методам можно познакомиться на сайте «Высокие статистические технологии» http://orlovs.pp.ru. Достаточно большой объем информации содержит еженедельник «Эконометрика» (электронная газета кафедры «Экономика и организация производства» научно-учебного комплекса «Инженерный бизнес и менеджмент» МГТУ им. Н. Э. Баумана), выпускаемый с июля 2000 г. (о нем сказано на указанном выше сайте). Автор искренне благодарен разработчику сайта и редактору электронного еженедельника А. А. Орлову за многолетний энтузиазм.

Автор искренне благодарен сотрудникам издательства Ай Пи Ар Медиа Юлии Валентиновне Семеновой, Юлии Вадимовне Ермоловой, Анастасии Валентиновне Неверовой за большую работу по подготовке рукописи учебника к публикации.

Автор будет благодарен читателям, если они сообщат свои вопросы и замечания по адресу издательства или непосредственно автору по электронной почте: prof-orlov@mail.ru.

ВВЕДЕНИЕ. НЕЧИСЛОВАЯ СТАТИСТИКА — ОСНОВА ВЫСОКИХ СТАТИСТИЧЕСКИХ ТЕХНОЛОГИЙ

1. О РАЗВИТИИ СТАТИСТИЧЕСКИХ МЕТОДОВ

Четыре столетия статистики. Впервые термин «статистика» появился в «Гамлете» Шекспира (1602 г., акт 5, сцена 2). Смысл этого слова у Шекспира — знать, придворные. По-видимому, оно происходит от латинского слова *status*, что в оригинале означает «состояние» или «политическое состояние».

В течение следующих 400 лет термин «статистика» понимали и понимают по-разному. В работе [1] собрано более 200 определений этого термина, некоторые из них обсуждаются ниже.

Вначале под статистикой понимали описание экономического и политического состояния государства или его части. Например, к 1792 г. относится определение: «Статистика описывает состояние государства в настоящее время или в некоторый известный момент в прошлом». И в настоящее время деятельность государственных статистических служб (в нашей стране — Федеральная служба государственной статистики (Росстат)) вполне укладывается в это определение.

Однако постепенно термин «статистика» стал использоваться более широко. По Наполеону Бонапарту «Статистика — это бюджет вещей». Тем самым статистические методы были признаны полезными не только для административного управления, но и на уровне отдельного предприятия. Согласно формулировке 1833 г. «цель статистики заключается в представлении фактов в наиболее сжатой форме». Приведем еще два высказывания. Статистика состоит в наблюдении явлений, которые могут быть подсчитаны или выражены посредством чисел (1895 г.). Статистика — это численное представление фактов из любой области исследования в их взаимосвязи (1909 г.).

В XX в. статистику обычно рассматривают как самостоятельную научную дисциплину. Статистика есть совокупность методов и принципов, согласно которым проводится сбор, анализ, сравнение, представление и интерпретация числовых данных (1925 г.). В 1954 г. академик АН УССР Б. В. Гнеденко дал следующее определение: «Статистика состоит из трех разделов:

1) сбор статистических сведений, т.е. сведений, характеризующих отдельные единицы каких-либо массовых совокупностей;

2) статистическое исследование полученных данных, заключающееся в выяснении тех закономерностей, которые могут быть установлены на основе данных массового наблюдения;

3) разработка приемов статистического наблюдения и анализа статистических данных. Последний раздел, собственно, и составляет содержание математической статистики».

Термин «статистика» употребляют еще в двух смыслах. Во-первых, в обиходе под «статистикой» часто понимают набор количественных данных о каком-либо явлении или процессе. Во-вторых, в специальной литературе статистикой называют функцию от результатов наблюдений, используемую для оценивания характеристик и параметров распределений и проверки гипотез.

Чтобы подойти к термину «нечисловая статистика», кратко рассмотрим историю реальных статистических работ.

Краткая история статистических методов. Типовые примеры раннего этапа применения статистических методов описаны в Ветхом Завете (см., например, Книгу Чисел). Там, в частности, приводится число воинов в различных племенах. С математической точки зрения дело сводилось к подсчету числа попаданий значений наблюдаемых признаков в определенные градации.

В дальнейшем результаты обработки статистических данных стали представлять в виде таблиц и диаграмм, как это и сейчас делает Росстат. Надо признать, что по сравнению с Ветхим Заветом есть прогресс — в Библии не было таблиц и диаграмм. Однако у Росстата нет продвижения по сравнению с работами российских статистиков конца XIX — начала XX вв. (типовой монографией тех времен можно считать книгу [2], которая в настоящее время ещё легко доступна).

Сразу после возникновения теории вероятностей (Паскаль, Ферма, XVII в.) вероятностные модели стали использоваться при обработке статистических данных. Например, изучалась частота рождения мальчиков и девочек, было установлено отличие вероятности рождения мальчика от 0,5, анализировались причины того, что в парижских приютах доля мальчиков не та, что в самом Париже, и т.д. Имеется достаточно много публикаций по истории теории вероятностей с описанием раннего этапа развития статистических методов, к лучшим из них относится очерк [3].

В 1794 г. (по другим данным — в 1795 г.) К. Гаусс разработал метод наименьших квадратов, один из наиболее популярных ныне статистических методов, и применил его при расчете орбиты астероида Церера — для борь-

бы с ошибками астрономических наблюдений [4]. В XIX в. заметный вклад в развитие практической статистики внес бельгиец А. Кетле, на основе анализа большого числа реальных данных, показавший устойчивость относительных статистических показателей, таких, как доля самоубийств среди всех смертей [5]. Интересно, что основные идеи статистического приемочного контроля и сертификации продукции обсуждались академиком Петербургской АН М. В. Остроградским (1801–1862 гг.) и применялись в российской армии ещё в середине XIX в. [3]. Статистические методы управления качеством и сертификации продукции сейчас весьма актуальны [6].

Современный этап развития статистических методов можно отсчитывать с 1900 г., когда англичанин К. Пирсон основан журнал *Biometrika*. Первая треть XX в. прошла под знаком параметрической статистики. Разрабатывались методы, основанные на анализе данных из параметрических семейств распределений, описываемых кривыми семейства Пирсона. Наиболее популярным было нормальное (гауссово) распределение. Для проверки гипотез использовались критерии Пирсона, Стьюдента, Фишера. Были предложены метод максимального правдоподобия, дисперсионный анализ, сформулированы основные идеи планирования эксперимента.

Разработанную в первой трети XX в. теорию анализа данных называем параметрической статистикой, поскольку ее основной объект изучения — это выборки из распределений, описываемых одним или небольшим числом параметров. Наиболее общим является семейство кривых Пирсона, задаваемых четырьмя параметрами. Как правило, нельзя указать каких-либо веских причин, по которым распределение результатов конкретных наблюдений должно входить в то или иное параметрическое семейство. Исключения хорошо известны: если вероятностная модель предусматривает суммирование независимых случайных величин, то сумму естественно описывать нормальным распределением; если же в модели рассматривается произведение таких величин, то итог, видимо, приближается логарифмически нормальным распределением, и т.д. Однако подобных моделей нет в подавляющем большинстве реальных ситуаций, и приближение реального распределения с помощью кривых из семейства Пирсона или его подсемейств — чисто формальная операция.

Именно из таких соображений критиковал параметрическую статистику академик АН СССР С. Н. Бернштейн в 1927 г. в своем докладе на Всероссийском съезде математиков [7]. Однако эта теория, к сожалению, до сих пор остается основой преподавания статистических методов и продолжает ис-

пользоваться основной массой прикладников, далеких от новых веяний в статистике. Почему так происходит? Чтобы попытаться ответить на этот вопрос, обратимся к наукометрии.

Наукометрия статистических исследований. В рамках движения за создание Всесоюзной статистической ассоциации (учреждена в 1990 г.) был проведен анализ статистики как области научно-практической деятельности. Он показал, в частности, что актуальными для специалистов в настоящее время являются не менее чем 100 тыс. публикаций (подробнее см. статьи [8, 9]). Реально же каждый из специалистов знаком с существенно меньшим количеством книг и статей. Так, в известном трехтомнике М. Кендалла и А. Стьюарта [10–12] — наиболее полном на русском языке издании по статистическим методам — всего около 2 тыс. литературных ссылок. При всей очевидности соображений о многократном дублировании в публикациях ценных идей приходится признать, что каждый специалист по статистическим методам владеет лишь небольшой частью накопленных в этой области знаний. Не удивительно, что приходится постоянно сталкиваться с игнорированием или повторением ранее полученных результатов, с уходом в тупиковые (с точки зрения практики) направления исследований, с беспомощностью при обращении к реальным данным, и т.д. Все это — одно из проявлений адапционного механизма торможения развития науки, о котором еще 30 лет назад писали В. В. Налимов и другие науковеды (см., например, [13]).

Традиционный предрассудок состоит в том, что каждый новый результат, полученный исследователем — это кирпич в непрерывно растущее здание науки, который непременно будет проанализирован и использован научным сообществом, а затем и при решении практических задач. Реальная ситуация — совсем иная. Основа профессиональных знаний исследователя, инженера, экономиста менеджера, социолога, историка, геолога, медика закладывается в период обучения. Затем знания пополняются в том узком направлении, в котором работает специалист. Следующий этап — их тиражирование новому поколению. В результате вузовские учебники отстоят от современного развития на десятки лет. Так, учебники по математической статистике, согласно мнению экспертов, по научному уровню в основном соответствуют 40–60-м гг. XX в. А потому середине XX в. соответствует большинство вновь публикуемых исследований и тем более — прикладных работ. Одновременно приходится признать, что результаты, не вошедшие в учебники, независимо от их ценности почти все забываются.

Активно продолжается развитие тупиковых направлений. Психологически это понятно. Приведу пример из своего опыта. По заказу Госстандарта я разработал методы оценки параметров гамма-распределения [14]. Поэтому мне близки и интересны работы по оцениванию параметров по выборкам из распределений, принадлежащих тем или иным параметрическим семействам, понятия функции максимального правдоподобия, эффективности оценок, использование неравенства Рао — Крамера и т.д. К сожалению, я знаю, что это — тупиковая ветвь теории статистики, поскольку реальные данные не подчиняются каким-либо параметрическим семействам, надо применять иные статистические методы — непараметрические. Понятно, что специалистам по параметрической статистике, потратившим многие годы на совершенствование в своей области, психологически трудно согласиться с этим утверждением. В том числе и мне. Но необходимо идти вперед.

Появление прикладной статистики. В нашей стране термин «прикладная статистика» вошел в широкое употребление в 1981 г. после издания массовым тиражом (33 940 экз.) сборника «Современные проблемы кибернетики (прикладная статистика)». В этом сборнике обосновывалась трехкомпонентная структура прикладной статистики [15]. Так, в нее входят ориентированные на прикладную деятельность статистические методы анализа данных (эту область можно назвать прикладной математической статистикой и включать также и в прикладную математику). Однако прикладную статистику нельзя целиком относить к математике. Она включает в себя две явно внематематические области. Во-первых, методологию организации статистического исследования: как планировать исследование, как собирать данные, как подготавливать данные к обработке, как представлять результаты. Во-вторых, организацию компьютерной обработки данных, в том числе разработку и использование баз данных и электронных таблиц, статистических программных продуктов, например, диалоговых систем анализа данных.

В нашей стране термин «прикладная статистика» использовался и ранее 1981 г., но лишь внутри сравнительно небольших и замкнутых групп специалистов, о некоторых из которых рассказано в статье [15].

Прикладная статистика и математическая статистика — это две разные научные дисциплины. Различие четко проявляется и при преподавании. Курс математической статистики состоит в основном из доказательств теорем, как и соответствующие учебные пособия. В курсах прикладной статистики основное — методология анализа данных и алгоритмы расчетов, а теоремы

приводятся как обоснования этих алгоритмов, доказательства же, как правило, опускаются (их можно найти в научной литературе).

Статистические методы. В области статистического анализа данных естественно выделить три вида научной и прикладной деятельности (по степени специфичности методов, сопряженной с погруженностью в конкретные проблемы):

А. Разработка и исследование методов прикладной статистики, предназначенных для анализа данных различной природы.

Б. Разработка и исследование вероятностно-статистических моделей в соответствии с конкретными потребностями науки и практики (моделей управления качеством, сбора и анализа оценок экспертов и др.).

В. Применение статистических методов и моделей для анализа конкретных данных (например, данных о росте цен с целью изучения инфляции).

Кратко рассмотрим три только что выделенных вида научной и прикладной деятельности. По мере движения от А к В сужается широта области применения статистического метода, но при этом повышается его значение для анализа конкретной ситуации. Если работам вида А соответствуют научные результаты, значимость которых оценивается по общенаучным критериям, то для работ вида В основное — успешное решение задач конкретной области. Работы вида Б занимают промежуточное положение, поскольку, с одной стороны, теоретическое изучение статистических моделей может быть достаточно сложным и математизированным (см., например, монографию [6]), с другой — результаты представляют интерес не для всей науки, а лишь для некоторого направления в ней.

Структура современной статистики. Внутренняя структура статистики как науки была выявлена и обоснована при создании в 1990 г. Всесоюзной статистической ассоциации [9]. Прикладная статистика — методическая дисциплина, являющаяся центром статистики. При применении методов прикладной статистики к конкретным областям знаний и отраслям народного хозяйства получаем научно-практические дисциплины типа «статистика в промышленности», «статистика в медицине» и др. С этой точки зрения эконометрика — это «статистические методы в экономике» [6]. Математическая статистика играет роль математического фундамента для прикладной статистики.

К настоящему времени очевидно четко выраженное размежевание этих двух научных направлений. Математическая статистика исходит из сформулированных в 1930–1950 гг. постановок математических задач, происхожде-

ние которых связано с анализом конкретных статистических данных. Начиная с 70-х гг. XX в. исследования по математической статистике посвящены обобщению и дальнейшему математическому изучению этих задач. Поток новых математических результатов (теорем) не ослабевает, но новые практические рекомендации по обработке статистических данных при этом почти не появляются. Можно сказать, что математическая статистика как научное направление замкнулась внутри себя.

Сам термин «прикладная статистика» возник как реакция на описанную выше тенденцию. Прикладная статистика нацелена на решение реальных задач. Поэтому в ней возникают новые постановки математических задач анализа статистических данных, развиваются и обосновываются новые методы. Обоснование часто проводится математическими методами, т.е. путем доказательства теорем. Большую роль играет методологическая составляющая — как именно ставить задачи, какие предположения принять с целью дальнейшего математического изучения. Велика роль современных информационных технологий, в частности, компьютерного эксперимента.

Рассматриваемое соотношение математической и прикладной статистик отнюдь не являются исключением. Как правило, математические дисциплины проходят в своем развитии ряд этапов. Вначале в какой-либо прикладной области возникает необходимость в применении математических методов и накапливаются соответствующие эмпирические приемы (для геометрии это — «измерение земли», т.е. землемерие, в Древнем Египте). Затем возникает математическая дисциплина со своей аксиоматикой (для геометрии это — время Евклида). Затем идет внутриматематическое развитие и преподавание (считается, что большинство результатов элементарной геометрии получено учителями гимназий в XIX в.). При этом на запросы исходной прикладной области перестают обращать внимание, и та порождает новые научные дисциплины (сейчас «измерением земли» занимается не геометрия, а геодезия и картография). Затем научный интерес к исходной дисциплине иссякает, но преподавание по традиции продолжается (элементарная геометрия до сих пор изучается в средней школе, хотя трудно понять, в каких практических задачах может понадобиться, например, теорема о том, что высоты треугольника пересекаются в одной точке). Следующий этап — окончательное вытеснение дисциплины из реальной жизни в историю науки (объем преподавания элементарной геометрии в настоящее время постепенно сокращается, в частности, ей все меньше уделяется внимания на вступительных экзаменах в вузах). К интеллектуальным дисциплинам, уже закончив-

шим свой жизненный путь, относится средневековая схоластика. Как справедливо отмечает проф. МГУ им. М. В. Ломоносова В. Н. Тутубалин [16], втностей и математическая статистика успешно двигаются по ее пути — вслед за элементарной геометрией.

Подведем итог. Хотя статистические данные собираются и анализируются с незапамятных времен (см., например, Книгу Чисел в Ветхом Завете), современная математическая статистика как наука была создана, по общему мнению специалистов, сравнительно недавно — в первой половине XX в. Именно тогда были разработаны основные идеи и получены результаты, излагаемые ныне в учебных курсах математической статистики. После чего специалисты по математической статистике занялись внутриматематическими проблемами, а для теоретического обслуживания проблем практического анализа статистических данных стала формироваться новая дисциплина — прикладная статистика.

В настоящее время статистическая обработка данных проводится, как правило, с помощью соответствующих программных продуктов. Разрыв между математической и прикладной статистикой проявляется, в частности, в том, что большинство методов, включенных в популярные среди исследователей статистические пакеты программ (например, в заслуженные *Statgraphics* и *SPSS* или в более новую систему *Statistica*), даже не упоминается в учебниках по математической статистике. В результате специалист по математической статистике оказывается зачастую беспомощным при обработке реальных данных, а методики статистического анализа и пакеты программ применяют (что еще хуже — и разрабатывают) лица, не имеющие необходимой теоретической подготовки. Естественно, что они допускают разнообразные ошибки, в том числе в таких ответственных документах, как государственные стандарты по статистическим методам. Анализ грубых ошибок в стандартах дан в статье [17].

Что дает прикладная статистика народному хозяйству? Так называлась статья [18], в которой приводились многочисленные примеры успешного использования методов прикладной математической статистики при решении практических задач. Перечень примеров можно продолжать практически безгранично (см., например, недавнюю сводку [19]).

Методы прикладной статистики используются в зарубежных и отечественных экономических и технических исследованиях, работах по управлению (менеджменту), в медицине, социологии, психологии, истории, геологии и других областях. Их применение дает заметный экономический эффект. Например, в США — не менее 20 млрд долл. ежегодно только в области ста-

тистического контроля качества. Недавно появилась концепция «Шесть сигм» — система управления компанией или ее подразделениями на основе интенсивного использования статистических методов [20, 41]. Внедрение «Шести сигм» дает значительный экономический эффект. Исполнительный директор *General Electric* Джек Уэлч подчеркнул в ежегодном докладе, что всего за три года «Шесть сигм» сэкономили компании более 2 млрд долл.

В 1988 г. затраты на статистический анализ данных в нашей стране оценивались в 2 млрд руб. ежегодно [21]. Согласно расчетам сравнительной стоимости валют на основе потребительских паритетов [6], эту величину можно сопоставить с 2 млрд долл. США. Следовательно, объем отечественного «рынка статистических услуг» был на порядок меньше, чем в США, что совпадает с оценками и по другим показателям, например, по числу специалистов.

Публикации по новым статистическим методам, по их применениям в технико-экономических исследованиях, в инженерном деле постоянно появляются, например, в журнале «Заводская лаборатория», в секции «Математические методы исследования». Надо назвать также журналы «Автоматика и телемеханика» (издается Институтом проблем управления Российской академии наук), «Экономика и математические методы» (издается Центральным экономико-математическим институтом РАН).

Однако необходимо констатировать, что для большинства менеджеров, экономистов и инженеров прикладная статистика и другие статистические методы — пока экзотикой. Это объясняется тем, что в вузах современным статистическим методам почти не учат. Во всяком случае, по состоянию на 2021 г. каждый квалифицированный специалист в этой области — самоучка.

Этому выводу не мешает то, что в вузовских программах обычно есть два курса, связанных со статистическими методами. Один из них — «Теория вероятностей и математическая статистика». Этот небольшой курс обычно читают специалисты с математических кафедр. Они успевают дать лишь общее представление об основных понятиях математической статистики. Кроме того, внимание математиков обычно сосредоточено на внутриматематических проблемах, их больше интересует доказательства теорем, а не применение современных статистических методов в задачах экономики и менеджмента. Другой курс — «Статистика» или «Общая теория статистики», входящий в стандартный блок экономических дисциплин. Фактически он является введением в прикладную статистику и содержит первые начала эконометрических методов (по состоянию на 1900 г.).

Прикладная статистика и другие статистические методы опираются на два названных вводных курса. Цель — вооружить специалиста современным статистическим инструментарием. Специалист — это инженер, экономист, менеджер, геолог, медик, социолог, психолог, историк, химик, физик и т.д. Во многих странах мира — Японии и США, Франции и Швейцарии, Перу и Ботсване и др. — статистическим методам обучают в средней школе. ЮНЕСКО постоянно проводят конференции по вопросам такого обучения [22]. В СССР и СЭВ, а теперь — по плохой традиции — и в России игнорируют этот предмет в средней школе и лишь слегка затрагивают его в высшей. Результат на рынке труда очевиден — снижение конкурентоспособности специалистов.

Проблемы прикладной статистики и других статистических методов постоянно обсуждаются специалистами. Широкий интерес вызвала дискуссия в журнале «Вестник статистики», в рамках которой были, в частности, опубликованы статьи [9, 18]. На появление в нашей стране прикладной статистики отреагировали и в США [23].

В нашей стране получены многие фундаментальные результаты прикладной статистики. Огромное значение имеют работы академика РАН А. Н. Колмогорова [24]. Во многих случаях именно его работы дали первоначальный толчок дальнейшему развитию ряда направлений прикладной статистики. Зачастую еще 50–70 лет назад А. Н. Колмогоров рассматривал те проблемы, которые только сейчас начинают широко обсуждаться. Как правило, его работы не устарели и сейчас. Свою жизнь посвятили прикладной статистике члены-корреспонденты АН СССР Н. В. Смирнов и Л. Н. Большев. В настоящем учебнике постоянно встречаются ссылки на лучшую публикацию XX в. по прикладной статистике — составленные ими подробно откомментированные «Таблицы» [25].

Основное продвижение в статистике конца XX в. — это создание нечисловой статистики. Ее называют также статистикой нечисловых данных или статистикой объектов нечисловой природы.

Высокие статистические технологии. Термин «высокие технологии» популярен в современной научно-технической литературе. Он используется для обозначения наиболее передовых технологий, опирающихся на последние достижения научно-технического прогресса. Есть такие технологии и среди технологий статистического анализа данных — как в любой интенсивно развивающейся научно-практической области. В учебнике [6] при обсуждении «точек роста» нашей научно-практической дисциплины в качестве

«высоких статистических технологий» выделены технологии непараметрического анализа данных; устойчивые (робастные) технологии; технологии, основанные на размножении выборок, на использовании достижений статистики нечисловых данных и статистики интервальных данных.

Обсудим пока не вполне привычный термин «высокие статистические технологии». Каждое из трех слов несет свою смысловую нагрузку.

«Высокие», как и в других областях, означает, что статистическая технология опирается на современные достижения статистической теории и практики, в частности, теории вероятностей и прикладной математической статистики. При этом «опирается на современные научные достижения» означает, во-первых, что математическая основа технологии получена сравнительно недавно в рамках соответствующей научной дисциплины, во-вторых, что алгоритмы расчетов разработаны и обоснованы в соответствии с нею (а не являются так называемыми эвристическими). Со временем, если новые подходы и результаты не заставляют пересмотреть оценку применимости и возможностей технологии, заменить ее на более современную, «высокие статистические технологии» переходят в «классические статистические технологии», такие, как метод наименьших квадратов. Итак, высокие статистические технологии — плоды недавних серьезных научных исследований. Здесь два ключевых понятия — «молодость» технологии (во всяком случае, не старше 50 лет, а лучше — не старше 10 или 30 лет) и опора на «высокую науку».

Термин «статистические» привычен, но разъяснить его нелегко. Во всяком случае, к деятельности органов официальной государственной статистики высокие статистические технологии отношения не имеют. Выше уже обсуждалась эволюция терминов «статистика» и «статистические методы».

Наконец, сравнительно редко используемый применительно к статистике термин «технологии». Статистический анализ данных, как правило, включает в себя целый ряд процедур и алгоритмов, выполняемых последовательно, параллельно или по более сложной схеме. В частности, можно выделить следующие этапы:

- планирование статистического исследования;

- организация сбора необходимых статистических данных по оптимальной или рациональной программе (планирование выборки, создание организационной структуры и подбор команды статистиков, подготовка кадров, которые будут заниматься сбором данных, а также контролеров данных и т.п.);

- непосредственный сбор данных и их фиксация на тех или иных носителях (с контролем качества сбора и отбраковкой ошибочных данных по соображениям предметной области);

- первичное описание данных (расчет различных выборочных характеристик, функций распределения, непараметрических оценок плотности, построение гистограмм, корреляционных полей, различных таблиц и диаграмм и т.д.);

- оценивание тех или иных числовых или нечисловых характеристик и параметров распределений (например, непараметрическое интервальное оценивание коэффициента вариации или восстановление зависимости между откликом и факторами, т.е. оценивание функции);

- проверка статистических гипотез (иногда их цепочек — после проверки предыдущей гипотезы принимается решение о проверке той или иной последующей гипотезы);

- более углубленное изучение, т.е. применение различных алгоритмов многомерного статистического анализа, алгоритмов диагностики и построения классификации, статистики нечисловых и интервальных данных, анализа временных рядов и др.;

- проверка устойчивости полученных оценок и выводов относительно допустимых отклонений исходных данных и предпосылок используемых вероятностно-статистических моделей, в частности, изучение свойств оценок методом размножения выборок;

- применение полученных статистических результатов в прикладных целях (например, для диагностики конкретных материалов, построения прогнозов, выбора инвестиционного проекта из предложенных вариантов, нахождения оптимальных режима осуществления технологического процесса, подведения итогов испытаний образцов технических устройств и др.);

- составление итоговых отчетов, в частности, предназначенных для тех, кто не является специалистами в статистических методах анализа данных, в том числе для руководства — «лиц, принимающих решения».

Возможны и иные структуризации статистических технологий. Важно подчеркнуть, что квалифицированное и результативное применение статистических методов — это отнюдь не проверка одной отдельно взятой статистической гипотезы или оценка параметров одного заданного распределения из фиксированного семейства. Подобного рода операции — только отдельные кирпичики, из которых складывается статистическая технология. Между тем учебники и монографии по статистике обычно рассказывают об отдель-

ных кирпичиках, но не обсуждают проблемы их организации в технологию, предназначенную для прикладного использования.

Итак, процедура статистического анализа данных — это информационный технологический процесс, другими словами, та или иная информационная технология. Статистическая информация подвергается разнообразным операциям (последовательно, параллельно или по более сложным схемам). В настоящее время об автоматизации всего процесса статистического анализа данных говорить было бы несерьезно, поскольку имеется слишком много нерешенных проблем, вызывающих дискуссии среди статистиков. «Экспертные системы» в области статистического анализа данных пока не стали рабочим инструментом статистиков. Ясно, что и не могли стать. Можно сказать и жестче — это пока научная фантастика или даже вредная утопия.

В литературе статистические технологии рассматриваются явно недостаточно. В частности, обычно все внимание сосредотачивается на том или ином элементе технологической цепочки, а переход от одного элемента к другому остается в тени. Между тем проблема «стыковки» статистических алгоритмов, как известно, требует специального рассмотрения [6], поскольку в результате использования предыдущего алгоритма зачастую нарушаются условия применимости последующего. В частности, результаты наблюдений могут перестать быть независимыми, может измениться их распределение и т.п.

Например, при проверке статистических гипотез большое значение имеют такие хорошо известные характеристики статистических критериев, как уровень значимости и мощность. Методы их расчета и использования при проверке одной гипотезы обычно хорошо известны. Если же сначала проверяется одна гипотеза, а потом с учетом результатов ее проверки — вторая, то итоговая процедура, которую также можно рассматривать как проверку некоторой (более сложной) статистической гипотезы, имеет характеристики (уровень значимости и мощность), которые, как правило, нельзя просто выразить через характеристики двух составляющих гипотез, а потому они обычно неизвестны. В результате итоговую процедуру нельзя рассматривать как научно обоснованную, она относится к эвристическим алгоритмам. Конечно, после соответствующего изучения, например, методом Монте-Карло, она может войти в число научно обоснованных процедур прикладной статистики. Этот сюжет подробнее рассмотрен в учебнике [6].

Почему живучи «низкие статистические технологии»? «Высоким статистическим технологиям» противостоят, естественно, «низкие статисти-

ческие технологии». Это те технологии, которые не соответствуют современному уровню науки и техники. Обычно они одновременно и устарели, и не адекватны сути решаемых статистических задач.

Примеры таких технологий неоднократно критически рассматривались, например, в журнале «Заводская лаборатория». Достаточно вспомнить критику использования классических процентных точек критериев Колмогорова и омега-квадрат в ситуациях, когда параметры оцениваются по выборке и эти оценки подставляются в «теоретическую» функцию распределения [39]. Приходилось констатировать широкое распространение таких порочных технологий и конкретных алгоритмов, в том числе в государственных и международных стандартах (перечень ошибочных стандартов дан в [6]), учебниках и распространенных пособиях (разбор ошибок проведен в статьях [39, 40]). Тиражирование ошибок происходит обычно в процессе обучения в вузах или путем самообразования при использовании недоброкачественной литературы.

На первый взгляд вызывает удивление устойчивость «низких статистических технологий», их постоянное возрождение во все новых статьях, монографиях, учебниках. Поэтому, как ни странно, наиболее «долгоживущими» оказываются не работы, посвященные новым научным результатам, а публикации, разоблачающие ошибки, типа статьи [39]. Прошло больше 20 лет с момента ее публикации, но она по-прежнему актуальна, поскольку ошибочное применение критериев Колмогорова и омега-квадрат по-прежнему распространено.

Целесообразно указать здесь по крайней мере три обстоятельства, которые определяют эту устойчивость ошибок.

Первое обстоятельство — прочно закрепившаяся традиция. Учебники по так называемой «Общей теории статистики», написанные экономистами (поскольку учебная дисциплина «Статистика» официально относится к экономике), если беспристрастно проанализировать их содержание, состоят в основном из введения в прикладную статистику, изложенного в стиле «низких статистических технологий», на уровне 1950-х гг. К «низкой» прикладной статистике добавлена некоторая информация о деятельности государственных органов официальной статистики. Примерно таково же положение со статистическими методами в медицине — одни и те же «низкие статистические технологии» переписываются из книги в книгу. Кратко говоря, «профессора-невежды порождают новых невежд» [9]. Так мы писали в 1990 г., но никто из указанных невежд даже не поинтересовался, какие ошибки имеются в виду. Новое поколение, обучившись ошибочным алгоритмам, их использу-

ет, а с течением времени и достижением должностей, ученых званий и степеней — пишет новые учебники со старыми ошибками.

Руководство государственных органов официальной статистики РФ, воспользовавшись катаклизмами начала 1990-х гг., сделало вид, что ему неизвестно о создании в 1990 г. Всесоюзной статистической ассоциации и секции статистических методов в ее составе. Росстат по-прежнему закрыт от «высоких статистических технологий» и работает на уровне позапрошлого века. Защита стала надежнее, поскольку в соответствии с современным стилем аппаратной работы на письма и обращения можно не отвечать.

Второе обстоятельство связано с большими трудностями при оценке экономической эффективности применения статистических методов вообще и при оценке вреда от применения ошибочных методов в частности. (А без такой оценки как докажешь, что «высокие статистические технологии» лучше «низких»?) Некоторые соображения по первому из этих вопросов приведены в статье [18], содержащей оценки экономической эффективности ряда работ по применению статистических методов. При оценке вреда от применения ошибочных методов приходится учитывать, что общий успех в конкретной инженерной или научной работе вполне мог быть достигнут вопреки их применению, за счет «запаса прочности» других составляющих общей работы. Например, преимущество одного технологического приема над другим можно продемонстрировать как с помощью критерия Крамера — Уэлча проверки равенства математических ожиданий (что правильно), так и с помощью двухвыборочного критерия Стьюдента (что, вообще говоря, неверно, так как обычно не выполняются условия применимости этого критерия — нет ни нормальности распределения, ни равенства дисперсий). Кроме того, приходится выдерживать натиск невежд, защищающих свои ошибочные представления, методики и инструкции, например, государственные стандарты. Вместо исправления ошибок применяются самые разные приемы бюрократической борьбы с теми, кто разоблачает ошибки (подробнее см. [6]).

Третье существенное обстоятельство — трудности со знакомством с высокими статистическими технологиями. В течение последних 15 лет только журнал «Заводская лаборатория» предоставлял такие возможности. К сожалению, поток современных статистических книг, выпускавшихся, в частности, издательством «Финансы и статистика», практически превратился в узкий ручеек… Возможно, более существенным является влияние естественной задержки во времени между созданием «новых статистических технологий» и написанием полноценной и объемной учебной и методической лите-

ратуры. Она должна позволять знакомиться с новой методологией, новыми методами, теоремами, алгоритмами, технологиями не по кратким оригинальным статьям, а при обычном обучении.

Как ускорить внедрение «высоких статистических технологий»? Таким образом, весь арсенал используемых статистических методов можно распределить по трем потокам:

- высокие статистические технологии;
- классические статистические технологии;
- низкие статистические технологии.

Основная современная проблема статистических технологий — добиться, чтобы в конкретных статистических исследованиях использовались только технологии первых двух потоков. Под классическими статистическими технологиями понимаем технологии почтенного возраста, сохранившие свое значение для современной статистической практики. Таковы метод наименьших квадратов, статистики Колмогорова, Смирнова, омега-квадрат, непараметрические коэффициенты корреляции Спирмена и Кендалла (относить их к «ранговым» — значит делать уступку «низким статистическим технологиям») и многие другие статистические процедуры.

Каковы возможные пути решения основной современной проблемы в области статистических технологий?

Бороться с конкретными невеждами — дело почти безнадежное. Отстаивая свое положение и должности, они либо нагло игнорируют информацию о своих ошибках, как это делают авторы учебников по «Общей теории статистики», либо с помощью различных бюрократических приемов уходят и от ответственности, и от исправления ошибок по существу (как это было со стандартами по статистическим методам — см. учебник [6]). Признание и исправление ошибок встречается, увы, редко. Но встречается.

Конечно, необходима демонстрация квалифицированного применения высоких статистических технологий. В 1960–1970-х гг. этим занималась лаборатория академика А. Н. Колмогорова в МГУ им. М. В. Ломоносова. В секции «Математические методы исследования» журнала «Заводская лаборатория. Диагностика материалов», созданной в 1961 г., опубликовано более 1 000 статей в стиле «высоких статистических технологий». В настоящее время действует Институт высоких статистических технологий и эконометрики МГТУ им. Н. Э. Баумана. Есть, конечно, целый ряд других научных коллективов, работающих на уровне «высоких статистических технологий».

Но самое основное — обучение. Какие бы новые научные результаты ни были получены, если они остаются неизвестными студентам, то новое поколение исследователей и инженеров вынуждено осваивать их поодиночке, а то и переоткрывать. Т.е. практически новые научные результаты почти исчезают, едва появившись. Избыток публикаций превратился в тормоз развития. По нашим данным, к настоящему времени по статистическим технологиям опубликовано не менее миллиона статей и книг, из них не менее 100 тыс. являются актуальными для современного специалиста. Реальное число публикаций, которые способен освоить исследователь, по нашей оценке, не превышает 2–3 тыс. Во всяком случае, в наиболее «толстом» (на русском языке) трехтомнике по статистике М. Дж. Кендалла и А. Стьюарта приведено около 2 тыс. литературных ссылок. Итак, каждый исследователь знаком не более чем с 2–3 % актуальных литературных источников. Поскольку существенная часть публикаций заражена «низкими статистическими технологиями», то исследователь самоучка имеет мало шансов выйти на уровень «высоких статистических технологий». Одновременно приходится констатировать, что масса полезных результатов погребена в изданиях прошлых десятилетий и имеет мало шансов встать в ряды «высоких статистических технологий» без специально организованных усилий современных специалистов по их адаптации.

Итак, основное — обучение. Несколько огрубляя, можно сказать: что попало в учебные курсы и соответствующие учебные пособия — то сохраняется, что не попало — то пропадает. Подробнее об обучении — в конце раздела. Сейчас — об упомянутом выше Институте высоких статистических технологий и эконометрики (ИВСТЭ) МГТУ им. Н. Э. Баумана. Он был организован в 1989 г. и действует на базе факультета «Инженерный бизнес и менеджмент». Институт на хоздоговорных и госбюджетных началах занимается развитием, изучением и внедрением «высоких статистических технологий», т.е. наиболее современных технологий анализа технических, экономических, социологических, медицинских данных, ориентированных на использование в условиях современного производства и экономики. Основной интерес представляют применения «высоких статистических технологий» для анализа конкретных экономических данных, т.е. в эконометрике. Из экономических дисциплин наиболее перспективным представляется применение «высоких статистических технологий» для поддержки принятия управленческих решений, прежде всего в таком новом (для России) перспективном направлении экономической науки и практики, как контроллинг [42].

Вначале Институт действовал как Всесоюзный центр статистических методов и информатики Центрального правления Всесоюзного экономического общества. В 1990–1992 гг. было выполнено более 100 хоздоговорных работ, в том числе для НИЦентра по безопасности атомной энергетики, ВНИИ нефтепереработки, ПО «Пластик», ЦНИИ черной металлургии им. Бардина, НИИ стали, ВНИИ эластомерных материалов и изделий, НИИ прикладной химии, ЦНИИ химии и механики, НПО «Орион», ВНИИ экономических проблем развития науки и техники, ПО «Уралмаш», «АвтоВАЗ», МИИТ, Казахского политехнического института, Донецкого государственного госуниверситета и многих других.

Затем ИВСТЭ разрабатывает эконометрические методы анализа нечисловых данных, а также процедуры расчета и прогнозирования индекса инфляции и валового внутреннего продукта (для Министерства обороны РФ). Мы занимаемся методологией построения и использования математических моделей процессов налогообложения (для Министерства налогов и сборов РФ), методологией оценки рисков реализации инновационных проектов высшей школы (для Министерства промышленности, науки и технологий РФ). Институт оценивает влияние различных факторов на формирование налогооблагаемой базы ряда налогов (для Минфина РФ). Мы прорабатываем перспективы применения современных статистических и экспертных методов для анализа данных о научном потенциале (для Министерства промышленности, науки и технологий РФ). Важное направление связано с эколого-экономической тематикой — разработка методологического, программного и информационного обеспечения анализа рисков химико-технологических объектов (для Международного научно-технического центра), методов использования экспертных оценок в задачах экологического страхования (совместно с Институтом проблем рынка РАН). Институт проводит маркетинговые исследования (в частности, для Institute for Market Research GfK MR Russia, Промрадтехбанка, фирм, торгующих растворимым кофе, программным обеспечением, образовательными услугами). Интерес вызывают наши работы по прогнозированию социально-экономического развития России методом сценариев, по экономико-математическому моделированию развития малых предприятий и созданию современных систем информационной поддержки принятия решений для таких организаций.

Институт ведет фундаментальные исследования в области высоких статистических технологий и эконометрики. Информация об Институте представлена на сайте «Высокие статистические технологии» (http://orlovs.pp.ru) и

его форуме (https://orlovs.pp.ru/forum/index.php). Институтом издается компьютерный еженедельник «Эконометрика» (около одной тысячи подписчиков). Архив выпусков «Эконометрики» можно рассматривать как хрестоматию по различным разделам эконометрики, а также по высоким статистическим технологиям.

Может возникнуть естественный вопрос: зачем нужны высокие статистические технологии, разве недостаточно обычных статистических методов? Мы считаем и доказываем своими теоретическими и прикладными работами, что совершенно недостаточно. Так, многие данные в информационных системах имеют нечисловой характер, например, являются словами или принимают значения из конечных множеств. Нечисловой характер имеют и упорядочения, которые дают эксперты или менеджеры, например, выбирая главную цель, следующую по важности и т.д. Значит, нужна статистика нечисловых данных. Мы ее построили. Далее, многие величины известны не абсолютно точно, а с некоторой погрешностью — от и до. Другими словами, исходные данные — не числа, а интервалы. Нужна статистика интервальных данных. Мы ее развиваем. В монографии [42] на с. 138 хорошо сказано: «Нечеткая логика — мощный элегантный инструмент современной науки, который на Западе (и на Востоке — в Японии, Китае — А. О.) можно встретить в десятках изделий — от бытовых видеокамер до систем управления вооружениями, — у нас до самого последнего времени был практически неизвестен». Напомним, первая монография российского автора по теории нечеткости была написана нами [43]. Ни статистики нечисловых данных, ни статистики интервальных данных, ни статистики нечетких данных нет и не могло быть в классической статистике. Все это — высокие статистические технологии. Они разработаны за последние 10–30–50 лет. А обычные вузовские курсы по общей теории статистики и по математической статистике разбирают научные результаты, полученные в первой половине XX в.

Важная часть эконометрики — применение высоких статистических технологий к анализу конкретных экономических данных, что зачастую требует дополнительной теоретической работы по доработке статистических технологий применительно к конкретной ситуации. Большое значение имеют конкретные эконометрические модели, например, модели экспертных оценок или экономики качества. И конечно, такие конкретные применения, как расчет и прогнозирование индекса инфляции. Сейчас уже многим ясно, что годовой бухгалтерский баланс предприятия может быть использован для оценки его финансово-хозяйственной деятельности только с привлечением дан-

ных об инфляции. Применение эконометрики дает заметный экономический эффект. Например, в США — не менее 20 млрд долл. ежегодно только в области статистического контроля качества.

Преподавание высоких статистических технологий и их сердцевины — нечисловой статистики. Приходится с сожалением констатировать, что в России практически отсутствует подготовка специалистов по высоким статистическим технологиям. В курсах по теории вероятностей и математической статистике обычно даются лишь классические основы этих дисциплин, разработанные в первой половине XX в., а преподаватели свою научную деятельность предпочитают посвящать доказательству никому не нужных теорем, а не высоким статистическим технологиям.

В настоящее время появилась надежда на эконометрику. В России начинают развертываться эконометрические исследования и преподавание эконометрики, в том числе не только Институтом высоких статистических технологий и эконометрики. Преподавание этой дисциплины ведется в Московском государственном университете экономики, статистики и информатики (МЭСИ), на экономическом факультете МГУ им. М. В. Ломоносова и еще в нескольких экономических учебных заведениях. Среди технических вузов мы, факультет «Инженерный бизнес и менеджмент» МГТУ им. Н. Э. Баумана, имеем в настоящее время приоритет в преподавания эконометрики [6]. Мы полагаем, что экономисты, менеджеры и инженеры, прежде всего специалисты по контроллингу [42], должны быть вооружены современными средствами информационной поддержки, в том числе высокими статистическими технологиями и эконометрикой. Очевидно, преподавание должно идти впереди практического применения. Ведь как применять то, чего не знаешь?

Один раз — в 1990–1992 гг. мы уже обожглись на недооценке необходимости предварительной подготовки тех, для кого предназначены современные компьютерные средства. Наш коллектив (Всесоюзный центр статистических методов и информатики Центрального правления Всесоюзного экономического общества) разработал систему диалоговых программных систем обеспечения качества продукции. Их созданием руководили ведущие специалисты страны. Но распространение программных продуктов шло на 1–2 порядка медленнее, чем мы ожидали. Причина стала ясна не сразу. Как оказалось, работники предприятий просто не понимали возможностей разработанных систем, не знали, какие задачи можно решать с их помощью, какой экономический эффект они дадут. А не понимали и не знали потому, что в вузах никто их не учил статистическим методам управления качеством. Без

такого систематического обучения нельзя обойтись — сложные концепции «на пальцах» за 5 мин. не объяснишь.

Есть и противоположный пример — положительный. В середине 1980-х гг. в советской средней школе ввели новый предмет «Информатика». И сейчас молодое поколение превосходно владеет компьютерами, мгновенно осваивая быстро появляющиеся новинки, и этим заметно отличается от тех, кому за 40–50 лет. Если бы удалось адекватно выполнить уже принятые на государственном уровне решения и ввести в средней школе курс теории вероятностей и статистики — а такой курс есть в Японии и США, Швейцарии, Кении и Ботсване, почти во всех странах [22] — то ситуация могла бы быть резко улучшена. Надо, конечно, добиться того, чтобы такой курс был построен на высоких статистических технологиях, а не на низких. Другими словами, он должен отражать современные достижения, а не концепции пятидесятилетней или столетней давности.

2. СТРУКТУРА НЕЧИСЛОВОЙ СТАТИСТИКИ

Нечисловая статистика (статистика нечисловых данных, статистика объектов нечисловой природы) как самостоятельное научное направление была выделена в нашей стране. Термин «статистика объектов нечисловой природы» впервые появился в 1979 г. в монографии [26]. В том же году в работе [27] была сформулирована программа развития этого нового направления статистических методов.

Со второй половины 1980-х гг. существенно возрос интерес к этой тематике и у зарубежных исследователей. Это проявилось, в частности, на Первом Всемирном Конгрессе Общества математической статистики и теории вероятностей им. Бернулли, состоявшемся в сентябре 1986 г. в Ташкенте. Нечисловая статистика используется в нормативно-технической и методической документации, ее применение позволяет получить существенный технико-экономический эффект [28].

Цель настоящего раздела — дать введение в нечисловую статистику (статистику нечисловых данных, статистику объектов нечисловой природы), выделить ее структуру, указать основные идеи и результаты, подробнее рассмотренные в дальнейших главах книги.

Напомним, что объектами нечисловой природы называют элементы пространств, не являющихся линейными. Примерами являются вектора из 0 и 1, измерения в качественных шкалах, бинарные отношения (ранжировки,

разбиения, толерантности), множества, последовательности символов (тексты). Объекты нечисловой природы нельзя складывать и умножать на числа, не теряя при этом содержательного смысла. Этим они отличаются от издавна используемых в прикладной статистике (в качестве элементов выборок) чисел, векторов и функций.

Прикладную статистику по виду статистических данных принято делить на следующие направления:

- статистика случайных величин (одномерная статистика);
- многомерный статистический анализ;
- статистика временных рядов и случайных процессов;
- нечисловая статистика, или статистика нечисловых данных (ее важная часть – статистика интервальных данных).

При создании теории вероятностей и математической статистики исторически первыми были рассмотрены объекты нечисловой природы — белые и черные шары в урне. На основе соответствующих вероятностных моделей были введены биномиальное, гипергеометрическое и другие дискретные распределения. Получены теоремы Муавра — Лапласа, Пуассона и др. Современное развитие этой тематики привело, в частности, к созданию теории статистического контроля качества продукции по альтернативному признаку (годен — не годен) в работах А. Н. Колмогорова, Б. В. Гнеденко, Ю. К. Беляева, Я. П. Лумельского и многих других (см., например, классические монографии [29, 30]).

В 70-х гг. XX в. в связи с запросами практики весьма усилился интерес к статистическому анализу нечисловых данных. Московская группа, организованная Ю. Н. Тюриным, Б. Г. Литваком, А. И. Орловым, Г. А. Сатаровым, Д. С. Шмерлингом и другими специалистами вокруг созданного в 1973 г. научного семинара «Экспертные оценки и нечисловая статистика», развивала в основном вероятностную статистику нечисловых данных. Были установлены разнообразные связи между различными видами объектов нечисловой природы и изучены свойства этих объектов. Московской группой выпущены десятки сборников и обзоров, перечень которых приведен в итоговой работе [31]. Хотя в названиях многих из этих изданий стоят слова «экспертные оценки», анализ содержания сборников показывает, что подавляющая часть статей посвящена математико-статистическим вопросам, а не проблемам проведения экспертиз. Частое употребление указанных слов отражает лишь один из импульсов, стимулирующих развитие нечисловой статистики и идущих от запросов практики. При этом необходимо подчеркнуть, что получен-

ные результаты могут и должны активно использоваться в теории и практике экспертных оценок.

Новосибирская группа (Г. С. Лбов, Б. Г. Миркин и др.), как правило, не использовала вероятностные модели, т.е. вела исследования в рамках детерминированного анализа данных. В московской группе в рамках анализа данных также велись работы, в частности, Б. Г. Литваком. Исследования по статистике объектов нечисловой природы выполнялись также в Ленинграде, Ереване, Киеве, Таллинне, Тарту, Красноярске, Минске, Днепропетровске, Владивостоке, Калинине и других отечественных научных центрах.

Внутреннее деление нечисловой статистики. Внутри рассматриваемого направления прикладной статистики выделяют следующие области:

1. Статистика конкретных видов объектов нечисловой природы.

2. Статистика в пространствах общей (произвольной) природы.

3. Применение идей, подходов и результатов статистики объектов нечисловой природы в классических областях прикладной статистики.

Единство рассматриваемому направлению придает, прежде всего, вторая составляющая, позволяющая с единой точки зрения подходить к статистическим задачам описания данных, оценивания, проверки гипотез при рассмотрении выборки, элементы которой имеют ту или иную конкретную природу. Внутри первой составляющей рассматривают:

1.1) теорию измерений;

1.2) статистику бинарных отношений;

1.3) теорию люсианов (бернуллиевских векторов);

1.4) теорию парных сравнений;

1.5) статистику случайных множеств;

1.6) статистику нечетких множеств;

1.7) статистику интервальных данных;

1.8) аксиоматическое введение метрик;

1.9) многомерное шкалирование и кластер-анализ (существенную часть этой тематики относят также к многомерному статистическому анализу) и др.

Перечисленные разделы тесно связаны друг с другом, как продемонстрировано, в частности, в работах [26, 32] и дальнейших главах настоящего учебника. Вне данного перечня остались работы по хорошо развитым классическим областям — статистическому контролю, таблицам сопряженности, а также по анализу текстов и некоторые другие (см. [6, 31, 33]).

Кратко обсудим постановки задач вероятностной статистики нечисловых данных, чтобы рассмотреть как единое целое это направление прикладной статистики.

Статистика в пространствах общей природы. Пусть $x_1, x_2, \ldots, x_n$ — элементы пространства X, не являющегося линейным. Как определить среднее значение для $x_1, x_2, \ldots, x_n$? Поскольку нельзя складывать элементы X, сравнивать их по величине, то необходимы подходы, принципиально новые по сравнению с классическими. В статистике объектов нечисловой природы предложено использовать показатель различия $d: X^2 \to [0,+\infty)$ (содержательный смысл показателя различия: чем больше $d(x,y)$, тем больше различаются x и y) и определять эмпирическое среднее как решение экстремальной задачи:

$$E_n(d) = Arg\min\{\sum_{1\le i\le n} d(x_i, x), x \in X\}. \qquad (1)$$

Таким образом, среднее $E_n(d)$ — это совокупность всех тех $x \in X$, для которых функция:

$$f_n x) = \frac{1}{n}\sum_{1\le i\le n} d(x_i, x) \qquad (2)$$

достигает минимума на X.

Как известно, для классического случая $X = R^1$ при $d(x,y) = (x-y)^2$ имеем $E_n(d) = \bar{x}$. При $X = R^1$, $d(x,y)=|x-y|$ среднее $E_n(d)$ при нечетном объеме выборки совпадает с выборочной медианой. А при четном объеме — $E_n(d)$ является отрезком с концами в двух средних элементах вариационного ряда.

Для ряда конкретных объектов среднее как решение экстремальной задачи вводилось рядом авторов. В 1929 г. итальянские статистики Джини и Гальвани применили такой подход для усреднения точек на плоскости и в пространстве Американский исследователь Джон Кемени решение задачи (1) называл медианой или средним для выборки, состоящей из ранжировок (см. монографию [34]). При моделировании лесных пожаров согласно выражению (1) было введено «среднеуклоняемое множество» для описания средней выгоревшей площади (см. об этом в монографии [26]). Общее определение эмпирических средних вида (1) было впервые введено в работе [27].

Основной результат, связанный со средними вида (1) — аналог закона больших чисел. Пусть $x_1, x_2, \ldots, x_n$ — независимые одинаково распределен-

ные случайные элементы со значениями в пространстве общей природы X. Теоретическим средним, или математическим ожиданием, в статистике объектов нечисловой природы называют:

$$E_n(x_1, d) = Arg\min\{Md(x_1, x), x \in X\}. \quad (3)$$

Закон больших чисел состоит в сходимости $E_n(d)$ к $E_n(x_1, d)$ при $n \to \infty$. Поскольку и эмпирическое, и теоретическое средние — множества, то понятие сходимости требует уточнения.

Одно из возможных уточнений, впервые введенное в работе [27], таково. Для функции:

$$f(x) = Md(x_1(\omega), x), f : X \to R^1 \quad (4)$$

введем понятие «ε-пятки» (ε>0):

$$K_\varepsilon(f) = \{x \in X : f(x) < \inf\{f(y), y \in X\} + \varepsilon\}. \quad (5)$$

Очевидно, ε-пятка f — это окрестность Argmin(f) (если он достигается), заданная в терминах минимизируемой функции. Тем самым снимается вопрос о выборе метрики в пространстве X. Тогда при некоторых условиях регулярности для любого ε>0 вероятность события:

$$\{\omega : E_n(d) \subseteq K_\varepsilon(f)\} \quad (6)$$

стремится к 1 при $n \to \infty$, т.е. справедлив закон больших чисел. Подробное доказательство приведено в главе 2 ниже.

Естественное обобщение рассматриваемой задачи позволяет построить общую теорию оптимизационного подхода в статистике. Как известно, большинство задач прикладной статистики может быть представлено в качестве оптимизационных. Как себя ведут решения экстремальных задач? Частные случаи этой постановки: как ведут себя при росте объема выборки оценки максимального правдоподобия и минимального контраста (в том числе робастные в смысле Тьюки — Хьюбера)? Что можно сказать о поведении оценок нагрузок в факторном анализе и методе главных компонент при отсутствии нормальности, об оценках метода наименьших модулей в регрессии и т.д.?

Обычно легко устанавливается, что для некоторого пространства X и последовательности случайных функций $f_n(x)$ при $n \to \infty$ найдется функция $f(x)$ такая, что

$$f_n(x) \to f(x) \tag{7}$$

для любого $x \in X$ (сходимость по вероятности). Требуется вывести отсюда, что

$$Arg\min\ f_n(x) \to Arg\min\ f(x), \tag{8}$$

т.е. решения экстремальных задач также сходятся. Понятие сходимости в соотношении (8) уточняется, например, с помощью ε-пяток, как это сделано выше для закона больших чисел. Условия регулярности, при которых справедливо предельное соотношение (8), приведены в исследовании [35]. Практически для всех реальных задач эти условия выполняются.

Как оценить распределение случайного элемента в пространстве общей природы? Поскольку понятие функции распределения неприменимо, естественно использовать непараметрические оценки плотности. Что такое плотность распределения вероятностей в пространстве произвольной природы? Это функция $g: X \to [0,+\infty)$ такая, что для любого измеримого множества (т.е. случайного события) $A \subseteq X$ справедливо соотношение:

$$P(x_1(\omega) \in A) = \int_A g(x)\mu(dx)\ , \tag{9}$$

где μ — некоторая мера в X. Ряд непараметрических оценок плотности был предложен в работе [27]. Например, ядерной оценкой плотности называется оценка:

$$g_n(x) = \frac{1}{\nu(h_n, x)} \sum_{1 \le i \le n} H\left(\frac{d(x_i, x)}{h_n}\right), \tag{10}$$

где d — показатель различия; H — ядерная функция; h_n — последовательность положительных чисел; $\nu(h_n, x)$ — нормирующий множитель. Удалось установить, что, что статистики типа (10) обладают такими же свойствами, по крайней мере при фиксированном x, что и их классические аналоги при $X = R^1$. В частности, такой же скоростью сходимости. Некоторые изменения

необходимы при рассмотрении дискретных X, каковыми являются многие пространства конкретных объектов нечисловой природы (см. главу 2). С помощью непараметрических оценок плотности можно развивать регрессионный анализ, дискриминантный анализ и другие направления в пространствах общей природы.

Для проверки гипотез согласия, однородности, независимости в пространствах общей природы могут быть использованы статистики интегрального типа:

$$\int f_n(x,\omega)dF_n(x,\omega), \qquad (11)$$

где $f_n(x,\omega)$ — последовательность случайных функций на X; $F_n(x,\omega)$ — последовательность случайных распределений (или зарядов). Обычно $f_n(x,\omega)$ при $n\to\infty$ сходится по распределению к некоторой случайной функции $f(x,\omega)$, а $F_n(x,\omega)$ — к распределению $F(x)$. Тогда распределение статистики интегрального типа (11) сходится к распределению случайного элемента:

$$\int f(x,\omega)dF(x). \qquad (12)$$

Условия, при которых это справедливо, даны в главе 2 на основе работы [36]. Пример применения — вывод предельного распределения статистики типа омега-квадрат для проверки симметрии распределения.

Перейдем к статистике конкретных видов объектов нечисловой природы.

Теория измерений. Цель теории измерений — борьба с субъективизмом исследователя при приписывании численных значений реальным объектам. Так, расстояния можно измерять в верстах, аршинах, саженях, метрах, микронах, милях, парсеках и других единицах измерения. Выбор единиц измерения зависит от исследователя, т.е. субъективен. Статистические выводы могут быть адекватны реальности только тогда, когда они не зависят от того, какую именно единицу измерения предпочтет исследователь, т.е. когда они инвариантны относительно допустимого преобразования шкалы.

Теория измерений известна в нашей стране уже более 40 лет. С начала 1970-х гг. активно работают отечественные исследователи. В настоящее время изложение основ теории измерений включают в справочные издания, помещают в научно-популярные журналы и книги для детей. Однако она еще

не стала общеизвестной среди специалистов. Поэтому опишем одну из задач теории измерений (ср. раздел 3.1 ниже).

Как известно, шкала задается группой допустимых преобразований (прямой в себя). Номинальная шкала (шкала наименований) задается группой всех взаимно-однозначных преобразований, шкала порядка — группой всех строго возрастающих преобразований. Это — шкалы качественных признаков. Группа линейных возрастающих преобразований $\varphi(x) = ax + b, a > 0,$ задает шкалу интервалов. Группа $\varphi(x) = ax, a > 0,$ определяет шкалу отношений. Наконец, группа, состоящая из одного тождественного преобразования, описывает абсолютную шкалу. Это — шкалы количественных признаков. Используют и некоторые другие шкалы.

Практическую пользу теории измерений обычно демонстрируют на примере задачи сравнения средних значений для двух совокупностей $x_1, x_2, \ldots, x_n$ и $y_1, y_2, \ldots, y_n$. Пусть среднее вычисляется с помощью функции $f : R^n \to R^1$. Если

$$f(x_1, x_2, \ldots, x_n) < f(y_1, y_2, \ldots, y_n), \tag{13}$$

то необходимо, чтобы

$$f(\varphi(x_1), \varphi(x_2), \ldots, \varphi(x_n)) < f(\varphi(y_1), \varphi(y_2), \ldots, \varphi(y_n)) \tag{14}$$

для любого допустимого преобразования φ из задающей шкалу группы Ф. (В противном случае результат сравнения будет зависеть от того, какое из эквивалентных представлений шкалы выбрал исследователь.)

Требование равносильности неравенств (13) и (14) вместе с некоторыми условиями регулярности приводит к тому, что в порядковой шкале в качестве средних можно использовать только члены вариационного ряда, в частности, медиану, но нельзя использовать среднее геометрическое, среднее арифметическое и т.д. В количественных шкалах это требование выделяет из всех обобщенных средних по А. Н. Колмогорову в шкале интервалов — только среднее арифметическое, а в шкале отношений — только степенные средние и среднее геометрическое. Кроме средних, аналогичные задачи рассмотрены в статистике нечисловых данных для расстояний, мер связи случайных признаков и других процедур анализа данных [26].

Приведенные результаты о средних величинах применялись, например, при проектировании системы датчиков в АСУ ТП доменных печей. Велико

прикладное значение теории измерений в задачах стандартизации и управления качеством, в частности, в квалиметрии. Так, В. В. Подиновский показал, что любое изменение коэффициентов весомости единичных показателей качества продукции приводит к изменению упорядочения изделий по средневзвешенному показателю, а Н. В. Хованов развил одну из возможных теорий шкал измерения качества. Теория измерений полезна и в других прикладных областях.

Статистика бинарных отношений. Оценивание центра распределения случайного бинарного отношения проводят обычно с помощью медианы Кемени. Состоятельность вытекает из закона больших чисел [26]. Разработаны различные вычислительные процедуры нахождения медианы Кемени.

Методы проверки гипотез развиты отдельно для каждой разновидности бинарных отношений. В области статистики ранжировок, или ранговой корреляции, классической является книга Кендалла [37]. Современные достижения отражены в работах Ю. Н. Тюрина и Д. С. Шмерлинга. Статистика случайных разбиений развита А. В. Маамяги. Статистика случайных толерантностей (рефлексивных симметричных отношений) впервые изложена в работе [26]. Многие ее задачи являются частными случаями задач теории люсианов.

Теория люсианов (бернуллиевских векторов). Люсиан (бернуллиевский вектор) — это последовательность испытаний Бернулли, вообще говоря, различными вероятностями успеха. Реализация люсиана (бернуллиевского вектора) — это последовательность из 0 и 1. Люсианы (бернуллиевские вектора) рассматривались при статистическом анализе случайных множеств с независимыми элементами, а также результатов независимых парных сравнений. Последовательность результатов контроля качества единиц продукции по альтернативному признаку — также реализация люсиана (бернуллиевского вектора). Случайная толерантность может быть записана в виде люсиана. Поскольку один и тот же математический объект необходим в различных прикладных областях, естественно для его наименования применять специально введенный термин «бернуллиевский вектор». Используется также более краткий термин «люсиан».

В рассматриваемой теории разработаны методы проверки согласованности (одинаковой распределенности), однородности двух выборок, независимости люсианов. Методы проверки указанных гипотез нацелены на ситуацию, когда число бернуллиевских векторов фиксировано, а их длина растет. При этом число неизвестных параметров возрастает пропорционально объе-

му данных, т.е. теория построена в асимптотике растущего числа параметров. Ранее подобная асимптотика под названием асимптотики А. Н. Колмогорова использовалась в дискриминантном анализе, но там применялись совсем другие методы для решения иных задач прикладной статистики.

Непараметрическая теория парных сравнений (в предположении независимости результатов отдельных сравнений) — часть теории люсианов. В параметрической теории выражают вероятности того или иного исхода через значения гипотетических или реальных параметров сравниваемых объектов. Известны модели Терстоуна, Бредли — Терри — Льюса и др. В нашей стране построен ряд новых моделей парных сравнений. В частности, имеются модели парных сравнений с тремя исходами (больше, меньше, неразличимо), модели зависимых сравнений, сравнений нескольких объектов (сближающие рассматриваемую область с теорией случайных ранжировок) и т.д.

Статистика случайных и нечетких множеств. Давнюю историю имеет статистика случайных геометрических объектов (отрезков, треугольников, кругов и т.д.). Современная теория случайных множеств сложилась при изучении пористых сред и объектов сложной природы в таких областях, как металлография, петрография, биология. Различные направления внутри этой теории рассмотрены в работе [26, гл. 4]. Укажем основные из них.

Случайные множества, лежащие в евклидовом пространстве, можно складывать: сумма множеств A и B — это объединение всех векторов $x + y$, где $x \in A, y \in B$. Н. Н. Ляшенко получил аналоги законов больших чисел, центральной предельной теоремы, ряда методов прикладной статистики, систематически используя подобные суммы.

Для нечисловой статистики интереснее подмножества пространств, не являющихся линейными. В работе [26] рассмотрены некоторые задачи теории конечных случайных множеств. Позже ряд интересных результатов получил С. А. Ковязин, в частности, он доказал нашу гипотезу о справедливости закона больших чисел при использовании расстояния между множествами:

$$d(a,b) = \mu(A \Delta B), \tag{15}$$

где μ — некоторая мера; Δ — знак симметрической разности. Расстояние (15) выведено из некоторой системы аксиом в монографии [26]. Прикладники также делают попытки развивать и применять методы статистики случайных множеств.

С теорией случайных множеств тесно связана теория нечетких множеств, начало которой положено статьей Л. А. Заде 1965 г. Это направление прикладной математики получило бурное развитие — к настоящему времени число публикаций измеряется десятками тысяч, имеются международные журналы, постоянно проводятся конференции, практические приложения дали ощутимый технико-экономический эффект. При изложении теории нечетких множеств обычно не подчеркивается связь с вероятностными моделями. Между тем еще в первой половине 1970-х гг. было установлено [26], что теория нечеткости в определенном смысле сводится к теории случайных множеств, хотя эта связь, возможно, имеет в основном теоретическое значение.

С точки зрения нечисловой статистики нечеткие множества — лишь один из видов объектов нечисловой природы. Поэтому к ним применима общая теория, развитая для пространств произвольной природы. Имеются работы, в которых совместно используются соображения вероятности и нечеткости.

Многомерное шкалирование и аксиоматическое введение метрик. Многомерное шкалирование имеет целью представление объектов точками в пространстве небольшой размерности (1–3) с максимально возможным сохранением расстояний между точками.

Из сказанного выше ясно, какое большое место занимают в статистике объектов нечисловой природы метрики (расстояния). Как их выбрать? Предлагают выводить вид метрик из некоторых систем аксиом. Аксиоматически получена метрика в пространстве ранжировок, которая оказалась линейно связанной с коэффициентом ранговой корреляции Кендалла [34]. Метрика (15) в пространстве множеств получена в работе [26] также исходя из некоторой системы аксиом. Г. В. Раушенбахом [38] дана сводка по аксиоматическому подходу к введению метрик в пространствах нечисловой природы. К настоящему времени практически для каждой используемой в прикладных работах метрики удалось подобрать систему аксиом, из которой чисто математическими средствами можно вывести именно эту метрику.

Применения статистики объектов нечисловой природы. Идеи, подходы, результаты статистики объектов нечисловой природы оказались полезными и в классических областях прикладной статистики. Статистика в пространствах общей природы позволила с единых позиций рассмотреть всю прикладную статистику, в частности, показать, что регрессионный, дисперсионный и дискриминантный анализы являются частными случаями общей схемы регрессионного анализа в пространстве произвольной природы. По-

скольку структура модели — объект нечисловой природы, то ее оценивание, в частности, оценивание степени полинома в регрессии, также относится к статистике нечисловых данных. Если учесть, что результаты измерения всегда имеют погрешность, т.е. являются не числами, а интервалами или нечеткими множествами, то приходим к необходимости разработки статистики интервальных данных. Ее развитие заставило пересмотреть некоторые выводы теоретической статистики. Например, в статистике интервальных данных отсутствует состоятельность оценок, нецелесообразно увеличивать объем выборок сверх некоторого предела (см. главу 4).

Технико-экономическая эффективность от применения методов статистики нечисловых данных достаточно высока. К сожалению, из-за изменения экономической ситуации, в частности, из-за инфляции трудно сопоставлять конкретные экономические результаты в разные моменты времени. Кроме того, методы нечисловой статистики составляют часть методов прикладной статистики. А те, в свою очередь — часть методов, входящих в систему информационной поддержки принятия решений на предприятии. Какую часть приращения прибыли предприятия надо отнести на эту систему? Можно проанализировать, как работает система управления фирмой в настоящее время. Но можно только оценивать, скорее всего, с помощью экспертных оценок, каковы были бы результаты финансово-хозяйственной деятельности предприятия, если бы система управления фирмой была бы иной, например, содержала бы методы нечисловой статистики.

Нечисловая статистика как часть прикладной статистики продолжает бурно развиваться. В частности, постоянно увеличивается количество ее практически полезных применений при анализе конкретных технических, экономических, медицинских данных — в научных, инженерно-технических, социологических, исторических, маркетинговых исследованиях, в контроллинге, при управлении качеством и предприятием в целом, в макроэкономике, при проведении научных медицинских работ и др.

Нечисловая статистика и концепция устойчивости. Основой для развития нечисловой статистики послужили результаты, полученные в монографии [26]. Судя по названию, она посвящена проблемам устойчивости в математических моделях социально-экономических явлений и процессов. Устойчивость выводов и принимаемых решений рассматривается относительно допустимых отклонений исходных данных и предпосылок модели. Как связаны проблемы устойчивости с нечисловой статистикой?

Во-первых, результаты объективного измерения нечисловых объектов обычно более устойчивы, чем числовых величин. Например, заключение о качестве изделия (годно — дефектно) более устойчиво, чем результат измерения его числового параметра (например, массы). Из-за погрешности повторного измерения масса изделия будет описываться несколько иным числом, а вывод о дефектности при повторной проверке сохранится.

Во-вторых, человеку свойственно использовать в своем мышлении нечисловые величины, прежде всего слова, а не появившиеся исторические недавно числовые системы. Именно поэтому для описания лингвистических переменных стали использовать нечеткие множества. Нечисловые оценки и выводы — первичны, их числовая оболочка — вторична. Поэтому нечисловая сердцевина устойчивее числовой периферии мышления и принятия решений. Другими словами, результаты субъективного измерения нечисловых объектов также более устойчивы, чем результаты субъективного измерения числовых величин.

В-третьих, многие постановки, приведенные выше, приобретают естественный вид в рамках концепции устойчивости. Например, требование устойчивости результата сравнения средних приводит к характеризации средних величин шкалами измерений, в которых их можно использовать. Любая предельная теорема — это утверждение об устойчивости того или иного математического объекта относительно изменения объема выборки или другого параметра, по которому происходит переход к пределу. Много подобных примеров приведено в монографии [26].

Таким образом, нечисловая статистика — это не только наиболее современная область статистических методов, но и центральная часть этой научно-практической дисциплины, наиболее важная как с теоретической, так и с прикладной точки зрения. Нечисловая статистика – сердцевина высоких статистических технологий [44].

В настоящее время нечисловая статистика (статистика нечисловых данных, статистика объектов нечисловой природы) — весьма развитая область искусственного интеллекта. К ней относятся посвящено большинство новых публикаций по прикладной статистике [45]. Развитию нечисловой статистики посвящена, в частности, часть I монографии [46], статьи [47, 48].

Развитию нечисловой статистики (статистики нечисловых данных, статистики объектов нечисловой природы) как области прикладной статистики посвящены работы [47–49]. В аналитическом обзоре [45] продемонстрировано, что статьи по нечисловой статистике составляют основную часть новых научных публикации по прикладной статистике. Понятие «высокие стати-

стические технологии» раскрыто в статьях [44, 50] и монографии [46]. Нечисловая статистика является частью системной нечеткой интервальной математики [51]. Мы рассматриваем системную нечеткую интервальную математику как основу математики XXI в. [52].

В литературе встречаются различные представления об искусственном интеллекте. Некоторые формулировки вызвали наше скептическо отношение к этому течению научной и практической мысли [53, 54]. Раскрытое в предисловии к этой книге современное представление об искусственном интеллекте позволяет констатировать, что наши работы, опубликованные в 1970–2021 гг., могут рассматриваться как вклад в разработку научного обеспечения искусственного интеллекта [55, 56]. Речь идет, прежде всего, о развитии и применении информационно-коммуникационных технологий в экономике и управлении [57]. Велико значение информационно-коммуникационных технологий для математических методов исследований [58]. Отметим роль метода статистических испытаний (метода Монте-Карло) как одной из информационно-коммуникационных технологий, наиболее полезных для нечисловой статистики [59, 60]. Место нечисловой статистики в эконометрике раскрыто в учебнике [61].

Литература

1. *Никитина, Е. П.* Коллекция определений термина «статистика» / Е. П. Никитина, В. Д. Фрейдлина, А. В. Ярхо. — Москва : Изд-во МГУ, 1972. — 46 с.

2. *Ленин, В. И.* Развитие капитализма в России. Процесс образования внутреннего рынка для крупной промышленности / В. И. Ленин. — Москва : Политиздат, 1986. — 610 с.

3. *Гнеденко, Б. В.* Очерк по истории теории вероятностей / Б. В. Гнеденко. — Москва : УРСС, 2001. — 88 с.

4. *Клейн, Ф.* Лекции о развитии математики в XIX столетии. Часть I / Ф. Клейн. — Москва : Ленинград : Объединенное научно-техническое изд-во НКТП СССР, 1937. — 432 с.

5. *Плошко, Б. Г.* История статистики : учебное пособие / Б. Г. Плошко, И. И. Елисеева. — Москва : Финансы и статистика, 1990. — 295 с.

6. *Орлов, А. И.* Эконометрика : учебник для вузов / А. И. Орлов. —3-е изд., испр. и доп. — Москва : Экзамен, 2004. — 576 с.

7. *Бернштейн, С. Н.* Современное состояние теории вероятностей и ее приложений / С. Н. Бернштейн // Труды Всероссийского съезда математиков

в Москве 27 апреля — 4 мая 1927 г. : сборник. — Москва : Ленинград : ГИЗ, 1928. — С. 50–63.

8. *Орлов, А. И.* О современных проблемах внедрения прикладной статистики и других статистических методов / А. И. Орлов // Заводская лаборатория. — 1992. — Т. 58. — № 1. — С. 67–74.

9. *Орлов, А. И.* О перестройке статистической науки и её применений / А. И. Орлов // Вестник статистики. — 1990. — № 1. — С. 65–71.

10. *Кендалл, М.* Теория распределений / М. Кендалл, А. Стьюарт. — Москва : Наука, 1966. — 566 с.

11. *Кендалл, М.* Статистические выводы и связи / М. Кендалл, А. Стьюарт. — Москва : Наука, 1973. — 899 с.

12. *Кендалл, М.* Многомерный статистический анализ и временные ряды / М. Кендалл, А. Стьюарт. — Москва : Наука, 1976. — 736 с.

13. *Налимов, В. В.* Наукометрия. Изучение развития науки как информационного процесса / В. В. Налимов, З. М. Мульченко. — Москва : Наука, 1969. — 192 с.

14. ГОСТ 11.011-83. Прикладная статистика. Правила определения оценок и доверительных границ для параметров гамма-распределения = Applied statistics. Regulations for determinations of estimates and confidence limits for parameters of gamma distribution : государственный стандарт Союза ССР : издание официальное : утвержден Постановлением Государственного комитета СССР по стандартам от 27 июня 1983 г. № 2684 : введен впервые : дата введения 1 января 1985 г. — Москва : Изд-во стандартов, 1984. — 53 с. (В настоящее время отменен как нормативный документ, но может использоваться как научная публикация.)

15. *Орлов, А. И.* О развитии прикладной статистики / А. И. Орлов // Современные проблемы кибернетики (прикладная статистика). — Москва : Знание, 1981. — С. 3–14.

16. *Тутубалин, В. Н.* Границы применимости (вероятностно-статистические методы и их возможности) / В. Н. Тутубалин. — Москва : Знание, 1977. — 64 с.

17. *Орлов, А. И.* Сертификация и статистические методы / А. И. Орлов // Заводская лаборатория. — 1997. — Т. 63. — № 3. — С. 55–62.

18. *Орлов, А. И.* Что дает прикладная статистика народному хозяйству? / А. И. Орлов // Вестник статистики. — 1986. — № 8. — С. 52–56.

19. *Орлов, А. И.* Применение эконометрических методов при решении задач контроллинга / А. И. Орлов, Л. А. Орлова // Контроллинг. — 2003. — № 4. — С. 50–54.

20. *Панде, П.* Что такое «Шесть сигм»? Революционный метод управления качеством / П. Панде, Л. Холп. — Москва : Альпина Бизнес Букс, 2004. — 158 с.

21. *Комаров, Д. М.* Роль методологических исследований в разработке методоориентированных экспертных систем (на примере оптимизационных и статистических методов) / Д. М. Комаров, А. И. Орлов // Вопросы применения экспертных систем. — Минск : Центросистем, 1988. — С. 151–160.

22. The teaching of statistics. Vol. 7. Studies in mathematical education. — Paris : UNESCO, 1991. — 258 p.

23. *Котц, С.* Пространство Хаусдорфа и прикладная статистика: точка зрения ученых СССР / С. Котц, К. Смит // The American Statistician. — 1988. Vol. 42. — № 4. — P. 241–244.

24. Кудлаев, Э. М. Вероятностно-статистические методы исследования в работах А. Н. Колмогорова / Э. М. Кудлаев, А. И. Орлов // Заводская лаборатория. — 2003. — Т. 69. — № 5. — С. 55–61.

25. *Большев, Л. Н.* Таблицы математической статистики / Л. Н. Большев, Н. В. Смирнов. — 3-е изд. — Москва : Наука, 1983.

26. *Орлов, А. И.* Устойчивость в социально-экономических моделях / А. И. Орлов. — Москва : Наука, 1979. — 296 с.

27. *Орлов, А. И.* Статистика объектов нечисловой природы и экспертные оценки / А. И. Орлов // Экспертные оценки. Вопросы кибернетики. — Вып. 58. — Москва : Научный Совет АН СССР по комплексной проблеме «Кибернетика», 1979. — С. 17–33.

28. *Кривцов, В. С.* Современные статистические методы в стандартизации и управлении качеством продукции / В. С. Кривцов, А. И. Орлов, В. Н. Фомин // Стандарты и качество. — 1988. — № 3. — С. 32–36.

29. *Беляев, Ю. К.* Вероятностные методы выборочного контроля / Ю. К. Беляев. — Москва : Наука, 1975. — 408 с.

30. *Лумельский, Я. П.* Статистические оценки результатов контроля качества / Я. П. Лумельский. — Москва : Изд-во стандартов, 1979. — 200 с.

31. *Орлов, А. И.* Статистика объектов нечисловой природы (обзор) / А. И. Орлов // Заводская лаборатория. — 1990. — Т. 56. — № 3. — С. 76–83.

32. Вероятность и математическая статистика : энциклопедия / главный редактор Ю. В. Прохоров. — Москва : Большая Российская энциклопедия, 1999. — 910 с.

33. *Толстова, Ю. Н.* Анализ социологических данных / Ю. Н. Толстова. — Москва : Научный мир, 2000. — 352 с.

34. *Кемени, Дж.* Кибернетическое моделирование. Некоторые приложения / Дж. Кемени, Дж. Снелл. — Москва : Советское радио, 1972. — 192 с.

35. *Орлов, А. И.* Асимптотика решений экстремальных статистических задач / А. И. Орлов // Анализ нечисловых данных в системных исследованиях : сборник трудов. — Вып. 10. — Москва : Всесоюзный научно-исследовательский институт системных исследований, 1982. — С. 4–12.

36. *Орлов, А. И.* Асимптотическое поведение статистик интегрального типа / А. И. Орлов // Вероятностные процессы и их приложения : межвузовский сборник. — Москва : МИЭМ, 1989. — С. 118–123.

37. *Кендэл, М.* Ранговые корреляции / М. Кендэл. — Москва : Статистика, 1975. — 216 с.

38. *Раушенбах, Г. В.* Меры близости и сходства / Г. В. Раушенбах // Анализ нечисловой информации в социологических исследованиях. — Москва : Наука, 1985. — С. 169–203.

39. *Орлов, А. И.* Распространенная ошибка при использовании критериев Колмогорова и омега-квадрат / А. И. Орлов // Заводская лаборатория. — 1985. — Т. 51. — № 1. — С. 60–62.

40. *Орлов, А. И.* Какие гипотезы можно проверять с помощью двухвыборочного критерия Вилкоксона? / А. И. Орлов // Заводская лаборатория. — 1999. — Т. 65. — № 1. — С. 51–55.

41. *Орлов, А. И.* «Шесть сигм» — новая система внедрения математических методов исследования / А. И. Орлов / Заводская лаборатория. — 2006. — Т. 72. — № 5. — С. 50–53.

42. Контроллинг в бизнесе. Методологические и практические основы построения контроллинга в организациях / А. М. Карминский, Н. И. Оленев, А. Г. Примак, С. Г.Фалько. — Москва : Финансы и статистика, 1998. — 256 с.

43. *Орлов, А. И.* Задачи оптимизации и нечеткие переменные / А. И. Орлов. — Москва : Знание, 1980. — 64 с.

44. *Орлов, А. И.* Высокие статистические технологии / А. И. Орлов // Заводская лаборатория. Диагностика материалов. — 2003. — Т. 69. — № 11. С. 55–60.

45. *Орлов, А. И.* Развитие математических методов исследования (2006–2015 гг.) / А. И. Орлов // Заводская лаборатория. Диагностика материалов. — 2017. — Т. 83. — № 1. — Ч. 1. — С. 78–86.

46. *Лойко, В. И.* Высокие статистические технологии и системно-когнитивное моделирование в экологии : монография / В. И. Лойко, Е. В. Луценко, А. И. Орлов. — Краснодар : КубГАУ, 2019. — 258 с.

47. *Орлов, А. И.* Статистика нечисловых данных за сорок лет (обзор) / А. И. Орлов // Заводская лаборатория. Диагностика материалов. — 2019. — Т. 85. — № 11. — С. 69–84.

48. *Орлов, А. И.* Статистика нечисловых данных — центральная часть современной прикладной статистики / А. И. Орлов // Научный журнал КубГАУ. — 2020. — № 156. — С. 111–142.

49. *Орлов, А. И.* О развитии статистики объектов нечисловой природы / А. И. Орлов // Научный журнал КубГАУ. — 2013. — № 93. — С. 41–50.

50. *Орлов, А. И.* О высоких статистических технологиях / А. И. Орлов // Научный журнал КубГАУ. — 2015. — № 105. — С. 14–38.

51. *Орлов, А. И.* Системная нечеткая интервальная математика : монография / А. И. Орлов, Е. В. Луценко. — Краснодар : КубГАУ, 2014. — 600 с.

52. *Орлов, А. И.* Системная нечеткая интервальная математика — основа математики XXI в. / А. И. Орлов // Научный журнал КубГАУ. — 2021. — № 165. — С. 111–130.

53. *Орлов, А. И.* Миф XX в.: искусственный интеллект / А. И. Орлов // Подводная лодка. — 2003. — № 11. — С. 102–103.

54. *Орлов, А. И.* Искусственный интеллект или мощный калькулятор? / А. И. Орлов // Магия ПК. — 2003. — № 3 (59). — С. 42–45.

55. *Орлов, А. И.* Организационно-экономическое моделирование и искусственный интеллект в организации производства в эпоху цифровой экономики / А. И. Орлов // Инновации в менеджменте. — 2021. — № 2 (28). — С. 36–45.

56. *Орлов, А. И.* Организационно-экономическое моделирование и искусственный интеллект в цифровой экономике (на примере управления качеством) / А. И. Орлов // Научный журнал КубГАУ. — 2021. — № 169. — С. 216–242.

57. *Орлов, А. И.* Прогноз развития информационно-коммуникационных технологий / А. И. Орлов // Научный журнал КубГАУ. — 2016. — № 116. — С. 435–461.

58. *Орлов, А. И.* Значение информационно-коммуникационных технологий для математических методов исследования / А. И. Орлов // Заводская лаборатория. Диагностика материалов. — 2017. — Т. 83. — № 7. — С. 5–6.

59. *Орлов, А. И.* Предельные теоремы и метод Монте-Карло / А. И. Орлов // Заводская лаборатория. Диагностика материалов. — 2016. — Т. 82. — № 7. — С. 67–72.

60. *Орлов, А. И.* Метод статистических испытаний в прикладной статистике / А. И. Орлов // Заводская лаборатория. Диагностика материалов. — 2019. — Т. 85. — № 5. — С. 67–79.

61. *Агаларов, З. С.* Эконометрика : учебник / З. С. Агаларов, А. И. Орлов. — Москва : Дашков и К, 2021. — 380 с.

ГЛАВА 1. НЕЧИСЛОВЫЕ СТАТИСТИЧЕСКИЕ ДАННЫЕ

1.1. КОЛИЧЕСТВЕННЫЕ И КАТЕГОРИЗОВАННЫЕ ДАННЫЕ

Статистические методы — это методы анализа данных, причем обычно достаточно большого количества данных. Статистические данные могут иметь различную природу. Исторически самыми ранними были два вида данных — сведения о числе объектов, удовлетворяющих тем или иным условиям, и числовые результаты измерений.

Первый из этих видов данных до сих пор главенствует в статистических сборниках Росстата. Такого рода данные часто называют *категоризованными*, поскольку о каждом из рассматриваемых объектов известно, в какую из нескольких заранее заданных категорий он попадает. Примером является информация Росстата о населении страны, с разделением по возрастным категориям и полу. Часто при составлении таблиц жертвуют информацией, заменяя точное значение измеряемой величины на указание интервала группировки, в которую это значение попадает. Например, вместо точного возраста человека используют лишь один из указанных в таблице возрастных интервалов.

Второй наиболее распространенный вид данных — количественные данные, рассматриваемые как действительные числа. Таковы результаты измерений, наблюдений, испытаний, опытов, анализов. Количественные данные обычно описываются набором чисел (выборкой), а не таблицей.

Нельзя утверждать, что категоризованные данные соответствуют первому этапу исследования, а числовые — следующему, на котором используются более совершенные методы измерения. Дело в том, что человеку свойственно давать качественные ответы на возникающие в его практической деятельности вопросы. Примером является таблица[2], посвященная анализу сильных и слабых сторон конкретной Компании (табл. 1). Она составлена одним из руководителей этой Компании и предназначена для использования при управлении Компанией.

[2] Данные взяты из выпускной работы А. А. Пивня «Анализ и перспективы развития маркетинга ЗАО "Компания Новгородский завод ГАРО"» (Академия народного хозяйства при правительстве Российской Федерации, 2003).

Таблица 1

Оценка сильных и слабых сторон Компании

Показатели, описывающие различные стороны работы Компании	Оценка показателя (по отношению к предприятиям отрасли)				Важность (вес) показателя			
	Очень высокая	Высокая	Средняя	Низкая	Очень низкая	Высокая	Средняя	Низкая
1	2	3	4	5	6	7	8	9
1. Финансы								
1.1. Оценка структуры активов			X			X		
1.2. Инвестиционная привлекательность			X			X		
1.3. Доход на активы				X		X		
1.4. Норма прибыли					X	X		
1.5. Доход на вложенный капитал				X			X	
2. Производство								
2.1. Использования оборудования			X				X	
2.2. Производственные мощности			X					X
2.3. Численность			X				X	
2.4. Система контроля качества		X				X		
2.5. Возможность расширения производства			X			X		
2.6. Износ оборудования				X		X		
3. Организация и управление								
3.1. Численность ИТР и управленческого персонала			X			X		
3.2. Скорость реакции управления на изменения во внешней среде			X			X		

Показатели, описывающие различные стороны работы Компании	Оценка показателя (по отношению к предприятиям отрасли)				Важность (вес) показателя			
	Очень высокая	Высокая	Средняя	Низкая	Очень низкая	Высокая	Средняя	Низкая
1	2	3	4	5	6	7	8	9
3.3. Четкость разделения полномочий и функций				X			X	
3.4. Качество используемой в управлении информации			X			X		
3.5. Гибкость оргструктуры управления		X				X		
4. Маркетинг								
4.1. Доля рынка		X				X		
4.2. Репутация Компании		X				X		
4.3. Престиж торговой марки			X				X	
4.4. Стимулирование сбыта		X				X		
4.5. Численность сбытового персонала				X				X
4.6. Уровень цен			X			X		
4.7. Уровень сервиса		X				X		
4.8. Число клиентов		X					X	
4.9. Качество поступающей информации			X				X	
5. Кадровый состав								
5.1. Уровень квалификации производственного персонала		X				X		
5.2. Расходы по подготовке и переподготовке персонала		X				X		

Показатели, описывающие различные стороны работы Компании	Оценка показателя (по отношению к предприятиям отрасли)				Важность (вес) показателя			
	Очень высокая	Высокая	Средняя	Низкая	Очень низкая	Высокая	Средняя	Низкая
1	2	3	4	5	6	7	8	9
5.3. Уровень подготовки сбытового персонала в технической области				X			X	
6. Технология								
6.1. Применяемые стандарты		X						X
6.2. Новые продукты			X				X	
6.3. Расходы на НИОКР		X					X	

Ясно, что вполне можно превратить в числа значения признаков, названия которых приведены в столбце «Показатели Компании», однако этот переход будет зависеть от исследователя, носить неизбежный налет субъективизма. Отметим, что важность (вес) показателей также оценивается качественно, а не количественно.

Иногда нецелесообразно однозначно относить данные к категоризованным или количественным. Например, в Ветхом Завете, в Четвертой книге Моисеевой «Числа» указывается количество воинов в различных коленах. С одной стороны, это типичные категоризованные данные, градациями служат названия колен. С другой стороны, эти данные можно рассматривать как количественные, как выборку, их вполне естественно складывать, вычислять среднее арифметическое и т.п.

Описанная ситуация типична. Существует весьма много различных видов статистических данных. Это связано, в частности, со способами их получения. Например, если испытания некоторых технических устройств продолжаются до определенного момента, то получаем так называемые *цензурированные* данные, состоящие из набора чисел — продолжительности работы ряда устройств до отказа, и информации о том, что остальные устройства продолжали работать в момент окончания испытания. Такого рода данные часто используются при оценке и контроле надежности технических устройств.

Описание вида данных и, при необходимости, механизма их порождения — начало любого статистического исследования.

В простейшем случае статистические данные — это значения некоторого признака, свойственного изучаемым объектам. Значения могут быть количественными или представлять собой указание на категорию, к которой можно отнести объект. Во втором случае говорят о качественном признаке. Используют и более сложные признаки, перечень которых будет расширяться по мере развертывания изложения в учебнике.

При измерении по нескольким количественным или качественным признакам в качестве статистических данных об объекте получаем вектор. Его можно рассматривать как новый вид данных. В таком случае выборка состоит из набора векторов. Есть часть координат — числа, а часть — качественные (категоризованные) данные, то говорим о векторе разнотипных данных.

Одним элементом выборки, т.е. одним измерением, может быть и функция в целом. Например, электрокардиограмма больного или амплитуда биений вала двигателя. Или временной ряд, описывающий динамику показателей определенной фирмы. Тогда выборка состоит из набора функций.

Элементами выборки могут быть и бинарные отношения. Например, при опросах экспертов часто используют упорядочения (ранжировки) объектов экспертизы — образцов продукции, инвестиционных проектов, вариантов управленческих решений. В зависимости от регламента экспертного исследования элементами выборки могут быть различные виды бинарных отношений (упорядочения, разбиения, толерантности), множества, нечеткие множества и т.д.

Итак, математическая природа элементов выборки в различных задачах прикладной статистики может быть самой разной. Однако можно выделить два класса статистических данных — числовые и нечисловые. Соответственно прикладная статистика разбивается на две части — числовую статистику и нечисловую статистику (ее называют также статистикой нечисловых данных или статистикой объектов нечисловой природы).

Числовые статистические данные — это числа, вектора, функции. Их можно складывать, умножать на коэффициенты. Поэтому в числовой статистике большое значение имеют разнообразные суммы. Математический аппарат анализа сумм случайных элементов выборки — это (классические) законы больших чисел и центральные предельные теоремы (см. прил. 1).

Нечисловые статистические данные — это категоризованные данные, вектора разнотипных признаков, бинарные отношения, множества, нечеткие

множества и др. Их нельзя складывать и умножать на коэффициенты. Поэтому не имеет смысла говорить о суммах нечисловых статистических данных. Они являются элементами нечисловых математических пространств (множеств). Математический аппарат анализа нечисловых статистических данных основан на использовании расстояний между элементами (а также мер близости, показателей различия) в таких пространствах. С помощью расстояний определяются эмпирические и теоретические средние, доказываются законы больших чисел, строятся непараметрические оценки плотности распределения вероятностей, решаются задачи диагностики и кластерного анализа и т.д.

Сведем информацию об основных областях прикладной статистики в табл. 2. Отметим, что модели порождения цензурированных данных входят в состав каждой из рассматриваемых областей.

Таблица 2

Области прикладной статистики

№ п/п	Вид статистических данных	Область прикладной статистики
1	Числа	Статистика (случайных) величин
2	Конечномерные вектора	Многомерный статистический анализ
3	Функции	Статистика случайных процессов и временных рядов
4	Объекты нечисловой природы	Нечисловая статистика

1.2. ОСНОВЫ ТЕОРИИ ИЗМЕРЕНИЙ

Почему необходима теория измерений? Теория измерений (в дальнейшем сокращенно ТИ) является одной из составных частей прикладной статистики. Она входит в состав *статистики объектов нечисловой природы (нечисловой статистики)*.

Использование чисел в жизни и хозяйственной деятельности людей отнюдь не всегда предполагает, что эти числа можно складывать и умножать, производить иные арифметические действия. Что бы вы сказали о человеке, который занимается умножением телефонных номеров? И отнюдь не всегда 2 + 2 = 4. Если вы вечером поместите в клетку двух животных, а потом еще двух, то отнюдь не всегда можно утром найти в этой клетке четырех животных. Их может быть и много больше — если вечером вы загнали в клетку ов-

цематок или беременных кошек. Их может быть и меньше — если к двум волкам вы поместили двух ягнят. Числа используются гораздо шире, чем арифметика.

Так, например, мнения экспертов часто выражены в *порядковой шкале* (подробнее о шкалах говорится ниже), т.е. эксперт может сказать (и обосновать), что один показатель качества продукции более важен, чем другой, первый технологический объект более опасен, чем второй, и т.д. Но он не в состоянии сказать, *во сколько раз* или *на сколько* более важен, соответственно, более опасен. Экспертов часто просят дать ранжировку (упорядочение) объектов экспертизы, т.е. расположить их в порядке возрастания (или убывания) интенсивности интересующей организаторов экспертизы характеристики. Ранг — это номер (объекта экспертизы) в упорядоченном ряду значений характеристики у различных объектов. Такой ряд в статистике называется вариационным. Формально ранги выражаются числами 1, 2, 3, ..., но с этими числами нельзя делать привычные арифметические операции. Например, хотя в арифметике 1 + 2 = 3, но нельзя утверждать, что для объекта, стоящем на третьем месте в упорядочении, интенсивность изучаемой характеристики равна сумме интенсивностей объектов с рангами 1 и 2. Так, один из видов экспертного оценивания — оценки учащихся. Вряд ли кто-либо будет утверждать, что знания отличника равны сумме знаний двоечника и троечника (хотя 5 = 2 + 3), хорошист соответствует двум двоечникам (2 + 2 = 4), а между отличником и троечником такая же разница, как между хорошистом и двоечником (5 – 3 = 4 – 2). Поэтому очевидно, что для анализа подобного рода качественных данных необходима не всем известная с начальной школы арифметика, а другая теория, дающая базу для разработки, изучения и применения конкретных методов расчета. Это и есть теория измерений (ТИ).

При чтении литературы надо иметь в виду, что в настоящее время термин «теория измерений» применяется для обозначения целого ряда научных дисциплин. А именно, классической метрологии (науки об измерениях физических величин), рассматриваемой здесь (репрезентативной) ТИ, некоторых других научных направлений, например, алгоритмической теории измерений. Обычно из контекста понятно, о какой конкретно теории идет речь.

Краткая история теории измерений. Сначала ТИ развивалась как теория психофизических измерений. В послевоенных публикациях[3] американский психолог С. С. Стивенс основное внимание уделял шкалам измерения (в основном связям между объективной величиной физического воздействия и его

[3] Вышедших сразу после Второй мировой войны.

субъективным восприятием для различных видов воздействий). Во второй половине XX в. сфера применения ТИ стремительно расширяется. Посмотрим, как это происходило. Один из томов выпущенной в США в 1950-х гг. «Энциклопедии психологических наук» назывался «Психологические измерения». Значит, составители этого тома расширили сферу применения репрезентативной ТИ с психофизики на психологию в целом. А в основной статье в этом сборнике под названием, обратите внимание, «Основы теории измерений», изложение шло на абстрактно-математическом уровне, без привязки к какой-либо конкретной области применения. В этой статье [1] упор был сделан на «гомоморфизмах эмпирических систем с отношениями в числовые» (в эти математические термины здесь вдаваться нет необходимости), и математическая сложность изложения заметно возросла по сравнению с работами С. С. Стивенса.

Уже в одной из первых отечественных статей по репрезентативной ТИ (конец 1960-х гг.) утверждалось, что баллы, присваиваемые экспертами при оценке объектов экспертизы, как правило, измерены в порядковой шкале. Дальнейшие работы, появившиеся в начале 1970-х гг., привели к существенному расширению области использования репрезентативной ТИ. Ее применяли к педагогической квалиметрии (измерению качества знаний учащихся), в системных исследованиях, в различных задачах теории экспертных оценок, для агрегирования показателей качества продукции, в социологических исследованиях и др.

Итоги этого этапа были подведены в монографии [2]. В качестве одной из двух основных проблем репрезентативной ТИ наряду с *установлением типа шкалы* измерения конкретных данных был выдвинут поиск алгоритмов анализа данных, результат работы которых не меняется при любом допустимом преобразовании шкалы (т.е. является *инвариантным* относительно этого преобразования).

Метрологи вначале резко возражали против использования термина «измерение» для качественных признаков. Однако постепенно возражения сошли на нет, и к концу XX в. все научные школы стали рассматривать ТИ как общенаучную теорию.

Шесть типов шкал. В соответствии с ТИ при математическом моделировании реального явления или процесса следует, прежде всего, установить *типы шкал*, в которых измерены те или иные переменные. Тип шкалы задает *группу допустимых преобразований шкалы*. Допустимые преобразования не меняют соотношений между объектами измерения. Например, при измерении

длины переход от аршин к метрам не меняет соотношений между длинами рассматриваемых объектов — если первый объект длиннее второго, то это будет установлено и при измерении в аршинах, и при измерении в метрах. Обратите внимание, что при этом численное значение длины в аршинах отличается от численного значения длины в метрах — не меняется лишь результат сравнения длин двух объектов.

Укажем основные виды шкал измерения и соответствующие группы допустимых преобразований.

В *шкале наименований* (другое название этой шкалы — *номинальная*; это — переписанное русскими буквами английское название шкалы) **допустимыми** являются все взаимно-однозначные преобразования. В этой шкале числа используются лишь как метки. Примерно так же, как при сдаче белья в прачечную, т.е. лишь для различения объектов. В шкале наименований измерены, например, номера телефонов, автомашин, паспортов, студенческих билетов. Номера страховых свидетельств государственного пенсионного страхования, медицинского страхования, ИНН (индивидуальный номер налогоплательщика), штрих-коды товаров измерены в шкале наименований. Пол людей тоже измерен в шкале наименований, результат измерения принимает два значения — мужской, женский. Раса, национальность, цвет глаз, волос — номинальные признаки. Номера букв в алфавите — тоже измерения в шкале наименований. Никому в здравом уме не придет в голову складывать или умножать номера телефонов, такие операции не имеют смысла. Сравнивать буквы и говорить, например, что буква П лучше буквы С, также никто не будет. Единственное, для чего годятся результаты измерений в шкале наименований — для различения объектов. Во многих случаях только это от них и требуется. Например, шкафчики для одежды в раздевалках для взрослых различают по номерам, т.е. числам, а в детских садах используют рисунки, поскольку дети еще не знают чисел.

В *порядковой шкале* числа используются не только для различения объектов, но и для установления порядка между объектами. Простейшим примером являются оценки знаний учащихся. Символично, что в средней школе применяются оценки 2, 3, 4, 5, а в высшей школе ровно тот же смысл выражается словесно — неудовлетворительно, удовлетворительно, хорошо, отлично. Этим подчеркивается «нечисловой» характер оценок знаний учащихся. В порядковой шкале **допустимыми** являются все строго возрастающие преобразования.

Установление типа шкалы, т.е. задания группы допустимых преобразований шкалы измерения — дело специалистов соответствующей прикладной области. Так, оценки привлекательности профессий мы в монографии [2], выступая в качестве социологов, считали измеренными в порядковой шкале. Однако отдельные социологи не соглашались с нами, полагая, что выпускники школ пользуются шкалой с более узкой группой допустимых преобразований, например, шкалой интервалов. Очевидно, эта проблема относится не к математике, а к наукам о человеке. Для ее решения может быть поставлен достаточно трудоемкий эксперимент. Пока же он не поставлен, целесообразно принимать порядковую шкалу, так как это гарантирует от возможных ошибок.

Оценки экспертов, как уже отмечалось, часто следует считать измеренными в порядковой шкале. Типичным примером являются задачи ранжирования и классификации промышленных объектов, подлежащих экологическому страхованию.

Почему мнения экспертов естественно выражать именно в порядковой шкале? **Как показали многочисленные опыты, человек более правильно (и с меньшими затруднениями) отвечает на вопросы качественного, например, сравнительного, характера, чем количественного.** Так, ему легче сказать, какая из двух гирь тяжелее, чем указать их примерный вес в граммах.

В различных областях человеческой деятельности применяется много других видов порядковых шкал. Так, например, в минералогии используется шкала Мооса, по которому минералы классифицируются согласно критерию твердости. А именно: тальк имеет балл 1, гипс — 2, кальций — 3, флюорит — 4, апатит — 5, ортоклаз — 6, кварц — 7, топаз — 8, корунд — 9, алмаз — 10. Минерал с большим номером является более твердым, чем минерал с меньшим номером, при нажатии царапает его.

Порядковыми шкалами в географии являются — бофортова шкала ветров («штиль», «слабый ветер», «умеренный ветер» и т.д.), шкала силы землетрясений. Очевидно, нельзя утверждать, что землетрясение в 2 балла (лампа качнулась под потолком — такое бывает и в Москве) ровно в 5 раз слабее, чем землетрясение в 10 баллов (полное разрушение всех построек на поверхности земли).

В медицине порядковыми шкалами являются — шкала стадий гипертонической болезни (по Мясникову), шкала степеней сердечной недостаточности (по Стражеско — Василенко — Лангу), шкала степени выраженности ко-

ронарной недостаточности (по Фогельсону), и т.д. Все эти шкалы построены по одной схеме: заболевание не обнаружено; первая стадия заболевания; вторая стадия; третья стадия… Иногда выделяют стадии 1а, 1б и др. Каждая стадия имеет свойственную только ей медицинскую характеристику. При описании групп инвалидности числа используются в противоположном порядке: самая тяжелая — первая группа инвалидности, затем — вторая, самая легкая — третья.

Номера домов также измерены в порядковой шкале — они показывают, в каком порядке стоят дома вдоль улицы. Номера томов в собрании сочинений писателя или номера дел в архиве предприятия обычно связаны с хронологическим порядком их создания.

При оценке качества продукции и услуг, в так называемой квалиметрии (буквальный перевод: измерение качества) популярны порядковые шкалы. А именно, единица продукции оценивается как годная или не годная. При более тщательном анализе используется шкала с тремя градациями: есть значительные дефекты — присутствуют только незначительные дефекты — нет дефектов. Иногда применяют четыре градации: имеются критические дефекты (делающие невозможным использование) — есть значительные дефекты — присутствуют только незначительные дефекты — нет дефектов. Аналогичный смысл имеет сортность продукции — высший сорт, первый сорт, второй сорт и т. д.

При оценке экологических воздействий первая, наиболее обобщенная оценка — обычно порядковая, например: природная среда стабильна — природная среда угнетена (деградирует). Аналогично в эколого-медицинской шкале: нет выраженного воздействия на здоровье людей — отмечается отрицательное воздействие на здоровье.

Порядковая шкала используется и во многих иных областях. Отметим различные методы экспертных оценок (см. посвященный им раздел в главе 3).

Все шкалы измерения делят на две группы: шкалы качественных признаков и шкалы количественных признаков.

Порядковая шкала и шкала наименований — основные шкалы качественных признаков. Поэтому во многих конкретных областях науки и практики результаты качественного анализа можно рассматривать как измерения по этим шкалам.

Шкалы количественных признаков — это шкалы интервалов, отношений, разностей, абсолютная. По шкале *интервалов* измеряют величину потенциальной энергии или координату точки на прямой. В этих случаях на

шкале нельзя отметить ни естественное начало отсчета, ни естественную единицу измерения. Исследователь должен сам задать точку (начало) отсчета и сам выбрать единицу измерения. Допустимыми преобразованиями в шкале интервалов являются линейные возрастающие преобразования, т.е. линейные функции. Температурные шкалы Цельсия и Фаренгейта связаны именно такой зависимостью: $^{\circ}C = 5/9\ (^{\circ}F - 32)$, где $^{\circ}C$ — температура (в градусах) по шкале Цельсия, а $^{\circ}F$ — температура по шкале Фаренгейта.

Из количественных шкал наиболее распространенными в науке и практике являются шкалы *отношений*. В них есть естественное начало отсчета — нуль, т.е. отсутствие величины, но нет естественной единицы измерения. По шкале отношений измерены большинство физических единиц: масса тела, длина, заряд, а также цены (и различные стоимостные характеристики) в экономике. Допустимыми преобразованиями в шкале отношений являются подобные преобразования (изменяющие только масштаб). Другими словами, линейные возрастающие преобразования без свободного члена. Примером является пересчет цен из одной валюты в другую по фиксированному курсу.

Предположим, мы сравниваем экономическую эффективность двух инвестиционных проектов, используя цены в рублях. Пусть первый проект оказался лучше второго. Теперь перейдем на валюту самой экономически мощной державы мира — юани[4], используя фиксированный курс пересчета. Очевидно, первый проект должен опять оказаться более выгодным, чем второй. Это очевидно из общих соображений. Однако алгоритмы расчета не обеспечивают автоматически выполнения этого очевидного условия. Надо проверять, что оно выполнено. Результаты подобной проверки для алгоритмов расчета средних величин описаны ниже (раздел 3.1).

В шкале *разностей* есть естественная единица измерения, но нет естественного начала отсчета. Допустимые преобразования — сдвиги, т.е. линейные функции с единичным коэффициентом линейного члена, свободный же член произволен. Время измеряется по шкале *разностей*, если год (или сутки — от полудня до полудня) принимаем естественной единицей измерения, и по шкале интервалов в общем случае. На современном уровне знаний естественного начала отсчета указать нельзя. Дату сотворения мира различные авторы рассчитывают по-разному, равно как и момент рождества Христова. Так, согласно статистической хронологии [4], разработанной группой известного историка, академика РАН А. Т. Фоменко, Господь Иисус Хри-

[4] При использовании курсов валют, основанных на их реальной покупательной способности, по валовому внутреннему продукту (ВВП) Китай лидирует в мире, его ВВП больше, чем у любой другой страны, в частности, больше, чем у США [3].

стос родился примерно в 1054 г. по принятому ныне летоисчислению[5] в Стамбуле (он же — Царьград, Византия, Троя, Иерусалим, Рим).

Только для *абсолютной* шкалы результаты измерений — числа в обычном смысле слова. Примером является число людей в комнате. Для абсолютной шкалы допустимым является только тождественное преобразование.

В процессе развития соответствующей области знания тип шкалы может меняться. Так, сначала температура измерялась по *порядковой* шкале (холоднее — теплее). Затем — по *интервальной* (шкалы Цельсия, Фаренгейта, Реомюра). Наконец, после открытия абсолютного нуля температуру можно считать измеренной по шкале *отношений* (шкала Кельвина). Надо отметить, что среди специалистов иногда имеются разногласия по поводу того, по каким шкалам следует считать измеренными те или иные реальные величины. Другими словами, процесс измерения включает в себя и определение типа шкалы (вместе с обоснованием выбора определенного типа шкалы). Кроме перечисленных шести основных типов шкал, иногда используют и иные шкалы.

Обсуждение шкал измерения будет продолжено далее в более широком контексте — как одного из понятий нечисловой статистики (статистики нечисловых данных).

1.3. ВИДЫ НЕЧИСЛОВЫХ ДАННЫХ

Статистика нечисловых данных — это направление в прикладной математической статистике, в котором в качестве исходных статистических данных (результатов наблюдений) рассматриваются объекты нечисловой природы. Так принято называть объекты, которые нецелесообразно описывать числами, в частности элементы различных нелинейных пространств. Примерами являются бинарные отношения (ранжировки, разбиения, толерантности и др.), результаты парных и множественных сравнений, множества, нечеткие множества, измерение в шкалах, отличных от абсолютных. Этот перечень примеров не претендует на законченность. Он складывался постепенно, по мере того, как развивались теоретические исследования в области нечисловой статистики (статистики нечисловых данных) и расширялся опыт применений этого направления прикладной математической статистики.

[5] Позже те же авторы указали несколько иную дату — 1152 г. (см. http://chronologia.org).

Объекты нечисловой природы широко используются в теоретических и прикладных исследованиях по экономике, менеджменту и другим проблемам управления, в частности управления качеством продукции, в технических науках, социологии, психологии, медицине и т.д., а также практически во всех отраслях народного хозяйства.

Начнем с первоначального знакомства с основными видами объектов нечисловой природы.

Результаты измерений в шкалах, отличных от абсолютной. Рассмотрим подробнее, чем выше, конкретное исследование в области маркетинга образовательных услуг, послужившее поводом к развитию одного из направлений отечественных исследований по теории измерений. При изучении привлекательности различных профессий для выпускников новосибирских школ был составлен список из 30 профессий. Опрашиваемых просили оценить каждую из этих профессий одним из баллов 1, 2, ..., 10 по правилу: чем больше нравится, тем выше балл. Для получения социологических выводов необходимо было дать единую оценку привлекательности определенной профессии для совокупности выпускников школ. В качестве такой оценки в работе [5] использовалось среднее арифметическое баллов, выставленных профессии опрошенными школьниками. В частности, физика получила средний балл 7,69, а математика — 7,50. Поскольку 7,69 больше, чем 7,50, был сделан вывод, что физика более предпочтительна для школьников, чем математика.

Однако этот вывод противоречит данным работы [6], согласно которым ленинградские школьники средних классов больше любят математику, чем физику. Как объяснить это противоречие? Есть много подходов к выяснению причин различия выводов новосибирских и ленинградских исследователей. Здесь обсудим одно из возможных объяснений этого противоречия, основанное на идеях нечисловой статистики. Оно сводится к указанию на неадекватность (с точки зрения теории измерений) методики обработки статистических данных о предпочтениях выпускников школ, примененной в работе [5].

Дело в том, что баллы 1, 2, ..., 10 введены конкретными исследователями, т.е. субъективно. Если одна профессия оценена в 10 баллов, а вторая — в 2, то из этого нельзя заключить, что первая ровно в 5 раз привлекательней второй. Другой коллектив социологов мог бы принять иную систему баллов, например 1, 4, 9, 16, ..., 100. Естественно предположить, что упорядочивание профессий по привлекательности, присущее школьникам, не зависит от того, какой системой баллов им предложит пользоваться социолог. Раз так, то рас-

пределение профессий по градациям десятибалльной системы не изменится, если перейти к другой системе баллов с помощью любого допустимого преобразования в порядковой шкале, т.е. с помощью строго возрастающей функции $g: R^1 \to R^1$. Если $Y_1, Y_2, ..., Y_n$ — ответы n выпускников школ, касающиеся математики, а $Z_1, Z_2, ..., Z_n$ — физики, то после перехода к новой системе баллов ответы относительно математики будут иметь вид $g(Y_1)$, $g(Y_2)$, ..., $g(Y_n)$, а относительно физики — $g(Z_1)$, $g(Z_2)$, ..., $g(Z_n)$.

Пусть единая оценка привлекательности профессии вычисляется с помощью функции $f(X_1, X_2, ..., X_n)$. Какие требования естественно наложить на функцию $f: R^n \to R^1$, чтобы полученные с ее помощью выводы не зависели от того, какой именно системой баллов пользовался социолог (в рассматриваемом исследовании он выступал как специалист по маркетингу образовательных услуг)?

Замечание. Обсуждение можно вести в терминах экспертных оценок. Тогда вместо сравнения математики и физики n экспертов (а не выпускников школ) оценивают по конкурентоспособности на мировом рынке, например, две марки стали. Однако в настоящее время маркетинговые и социологические исследования более привычны, чем экспертные.

Единая оценка вычислялась для того, чтобы сравнивать профессии по привлекательности. Пусть $f(X_1, X_2, ..., X_n)$ — среднее по Коши[6]. Пусть среднее по первой совокупности меньше среднего по второй совокупности:

$$f(Y_1, Y_2, ..., Y_n) < f(Z_1, Z_2, ..., Z_n).$$

Тогда согласно теории измерений необходимо потребовать, чтобы для любого допустимого преобразования g из группы допустимых преобразований в порядковой шкале было справедливо также неравенство

$$f(g(Y_1), g(Y_2), ..., g(Y_n)) < f(g(Z_1), g(Z_2), ..., g(Z_n)),$$

т.е. среднее преобразованных значений из первой совокупности также было меньше среднего преобразованных значений для второй совокупности. Более того, два рассматриваемых неравенства должны быть равносильны. Причем сформулированное условие должно быть верно для любых двух совокупностей $Y_1, Y_2, ..., Y_n$ и $Z_1, Z_2, ..., Z_n$ и, напомним, любого допустимого преобразо-

[6] Среднее по Коши — любая функция $f(X_1, X_2, ...,X_n)$ такая, что $\min(X_1, X_2, ...,X_n) \le f(X_1, X_2, ...,X_n) \le \max(X_1, X_2, ..., X_n)$ при всех $X_1, X_2, ..., X_n$.

вания. Средние величины, удовлетворяющие сформулированному условию, называют допустимыми (в порядковой шкале). Согласно теории измерений только такими средними можно пользоваться при анализе мнений выпускников школ или экспертов, обработке иных данных, измеренных в порядковой шкале.

Какие единые оценки привлекательности профессий $f(X_1, X_2, ..., X_n)$ устойчивы относительно сравнения? Ответ на этот вопрос дается ниже в разделе 3.1. В частности, оказалось, что средним арифметическим, как в работе [5] новосибирских социологов (специалистов по маркетингу образовательных услуг), пользоваться нельзя. А порядковыми статистиками, т.е. членами вариационного ряда (и только ими) — можно.

Методы анализа конкретных статистических данных, измеренных в шкалах, отличных от абсолютной, являются предметом изучения в нечисловой статистике. Как описано выше, основные шкалы измерения делятся на качественные (шкалы наименований и порядка) и количественные (шкалы интервалов, отношений, разностей, абсолютная). Методы анализа статистических данных в количественных шкалах сравнительно мало отличаются от таковых в абсолютной шкале. Добавляется только требование инвариантности относительно преобразований сдвига и/или масштаба. Методы анализа качественных данных — принципиально иные.

Напомним, что исходным понятием теории измерений является совокупность $\Phi = \{\varphi\}$ допустимых преобразований шкалы (обычно Ф — группа), $\varphi : R^1 \to R^1$. Алгоритм обработки данных W, т.е. функция $W : R^n \to A$ (здесь A-множество возможных результатов работы алгоритма) называется *адекватным* в шкале с совокупностью допустимых преобразований Ф, если:

$$W(x_1, x_2, ..., x_n) = W(\varphi(x_1), \varphi(x_2), ..., \varphi(x_n)) \qquad (1)$$

для всех $x_i \in R^1$, $i = 1,2,...,n$, и всех $\varphi \in \Phi$. Таким образом, теорию измерений рассматриваем как теорию инвариантов относительно различных совокупностей допустимых преобразований Ф. Интерес вызывают две задачи:

а) дана группа допустимых преобразований Ф (т.е. задана шкала измерения); какие алгоритмы анализа данных W из определенного класса являются адекватными (т.е. удовлетворяют тождеству (1))?

б) дан алгоритм анализа данных W; для каких шкал (т.е. групп допустимых преобразований Ф) он является адекватным?

В разделе 3.1 первая задача рассматривается для алгоритмов расчета средних величин. Информацию о других результатах решения задач указанных типов можно найти в работах [2, 7, 8].

Бинарные отношения. Пусть $W : R^n \to A$ — адекватный алгоритм в шкале наименований. Можно показать, что этот алгоритм задается некоторой функцией от матрицы $B = \|b_{ij}\| = B(x_1, x_2, ..., x_n)$ порядка $n \times n$, где:

$$b_{ij} = \begin{cases} 1, & x_i = x_j, i, j = 1,2,...,n, \\ 0, & x_i \neq x_j, i, j = 1,2,...,n. \end{cases}$$

Если $W : R^n \to A$ — адекватный алгоритм в шкале порядка, то этот алгоритм задается некоторой функцией от матрицы $C = \|c_{ij}\| = C(x_1, x_2, ..., x_n)$ порядка $n \times n$, где:

$$c_{ij} = \begin{cases} 1, & x_i \leq x_j, i, j = 1,2,...,n, \\ 0, & x_i > x_j, i, j = 1,2,...,n. \end{cases}$$

Матрицы B и C можно проинтерпретировать в терминах бинарных отношений. Пусть некоторая характеристика измеряется у n объектов q_1, q_2, ..., q_n, причем x_i — результат ее измерения у объекта q_i. Тогда матрицы B и C задают бинарные отношения на множестве объектов $Q = \{q_1, q_2, ..., q_n\}$. Поскольку бинарное отношение можно рассматривать как подмножество декартова квадрата $Q \times Q$, то любой матрице $D = ||d_{ij}||$ порядка $n \times n$ из 0 и 1 соответствует бинарное отношение $R(D)$, определяемое следующим образом: $(q_i, q_j) \in R(D)$ тогда и только тогда, когда $d_{ij} = 1$.

Бинарное отношение $R(B)$ — отношение эквивалентности, т.е. симметричное рефлексивное транзитивное отношение. Оно задает разбиение Q на классы эквивалентности. Два объекта q_i и q_j входят в один класс эквивалентности тогда и только тогда, когда $x_i = x_j$, $b_{ij} = 1$.

Выше показано, как разбиения возникают в результате измерений в шкале наименований. Разбиения могут появляться и непосредственно. Так, при оценке качества промышленной продукции эксперты дают разбиение показателей качества на группы. Для изучения психологического состояния людей их просят разбить предъявленные рисунки на группы сходных между

собой. Аналогичная методика применяется и в иных экспериментальных психологических исследованиях, необходимых для оптимизации управления персоналом.

Во многих эконометрических задачах разбиения получаются «на выходе» (например, в кластерном анализе) или же используются на промежуточных этапах анализа данных (например, сначала проводят классификацию с целью выделения однородных групп, а затем в каждой группе строят регрессионную зависимость).

Бинарное отношение $R(C)$ задает разбиение Q на классы эквивалентности, между которыми введено отношение строгого порядка. Два объекта q_i и q_j входят в один класс тогда и только тогда, когда $c_{ij} = 1$ и $c_{ji} = 1$, т.е. $x_i = x_j$. Класс эквивалентности Q_1 предшествует классу эквивалентности Q_2 тогда и только тогда, когда для любых $q_i \in Q_1$, $q_j \in Q_2$ имеем $c_{ij} = 1$, $c_{ji} = 0$, т.е. $x_i < x_j$. Такое бинарное отношение в статистике часто называют ранжировкой со связями; связанными считаются объекты, входящие в один класс эквивалентности. В литературе встречаются и другие названия: кластеризованная ранжировка, линейный квазипорядок, упорядочение, квазисерия, ранжирование. Если каждый из классов эквивалентности состоит только из одного элемента, то имеем обычную ранжировку (другими словами, строгий линейный порядок).

Как известно, ранжировки возникают в результате измерений в порядковой шкале. Так, при описанном выше опросе ответ выпускника школы — это ранжировка (со связями) профессий по привлекательности. Ранжировки часто возникают и непосредственно, без промежуточного этапа — приписывания объектам квазичисловых оценок — баллов. Многочисленные примеры тому даны английским статистиком М. Кендэлом [9]. При оценке качества промышленной продукции широко применяемые нормативные и методические документы предусматривают использование ранжировок.

Для прикладных областей, кроме ранжировок и разбиений, представляют интерес толерантности, т.е. рефлексивные симметричные отношения. Толерантность — математическая модель для выражения представлений о сходстве (похожести, близости). Разбиения — частный вид толерантностей. Толерантность, обладающая свойством транзитивности — это разбиение. Однако в общем случае толерантность не обязана быть транзитивной. Толерантности появляются во многих постановках теории экспертных оценок, например, как результат парных сравнений (см. ниже).

Напомним, что любое бинарное отношение на конечном множестве может быть описано матрицей из 0 и 1.

Дихотомические (бинарные) данные. Это данные, которые могут принимать одно из двух значений (0 или 1), т.е. результаты измерений значений альтернативного признака. Как уже было показано, измерения в шкале наименований и порядковой шкале приводят к бинарным отношениям, а те могут быть выражены как результаты измерений по нескольким альтернативным признакам, соответствующим элементам матриц, описывающих отношения. Дихотомические данные возникают в прикладных исследованиях и многими иными путями.

В настоящее время в большинстве технических регламентов, стандартов, технических условий, договоров на поставку конкретной продукции предусмотрен контроль по альтернативному признаку. Это означает, что единица продукции относится к одной из двух категорий — «годных» или «дефектных», т.е. соответствующих или не соответствующих требованиям стандарта. Отечественными специалистами проведены обширные теоретические исследования проблем статистического приемочного контроля по альтернативному признаку. Основополагающими в этой области являются работы академика А. Н. Колмогорова. Подход советской вероятностно-статистической школы к проблемам контроля качества продукции отражен в монографиях [10, 11] (см. также главу 13 учебника [3]).

Дихотомические данные — давний объект математической статистики. Особенно большое применение они имеют в экономических и социологических исследованиях, в которых большинство переменных, интересующих специалистов, измеряется по качественным шкалам. При этом дихотомические данные зачастую являются более адекватными, чем результаты измерений по методикам, использующим большее число градаций. В частности, психологические тесты типа MMPI (расшифровывается как Миннесотское многофакторное личностное исследование) используют только дихотомические данные. На них опираются и популярные в технико-экономическом анализе методы парных сравнений [12].

Элементарным актом в методе парных сравнений является предъявление эксперту для сравнения двух объектов (сравнение может проводиться также прибором). В одних постановках эксперт должен выбрать из двух объектов лучший по качеству, в других — ответить, похожи объекты или нет. В обоих случаях ответ эксперта можно выразить одной из двух цифр (меток) — 0 или 1. В первой постановке: 0, если лучшим объявлен первый объект;

1 — если второй. Во второй постановке: 0, если объекты похожи, схожи, близки; 1 — в противном случае.

Подводя итоги, можно сказать, что рассмотренные выше виды данные могут быть представлены в виде векторов из 0 и 1 (при обосновании этого утверждения используется тот очевидный факт, что матрицы могут быть записаны в виде векторов). Более того, поскольку все мыслимые результаты наблюдений имеют лишь несколько значащих цифр, то, используя двоичную систему счисления, любые виды анализируемых статистическими методами данных можно записать в виде векторов конечной длины (размерности) из 0 и 1. Представляется, однако, что эта возможность в большинстве случаев имеет лишь академический интерес. Но во всяком случае можно констатировать, что анализ дихотомических данных необходим во многих прикладных постановках.

Множества. Совокупность X^n векторов $X = (x_1, x_2, ..., x_n)$ из 0 и 1 размерности n находится во взаимно-однозначном соответствии с совокупностью всех 2^n подмножеств множества $N = \{1, 2, ..., n\}$. При этом вектору $X = (x_1, x_2, ..., x_n)$ соответствует подмножество $N(X) \subseteq N$, состоящее из тех и только из тех i, для которых $x_i = 1$. Это объясняет, почему изложение вероятностных и статистических результатов, относящихся к анализу данных, являющихся объектами нечисловой природы перечисленных выше видов, можно вести на языке конечных случайных множеств, как это было сделано в монографии [2].

Множества как исходные данные появляются и в иных постановках. Из геологических задач исходил Ж. Матерон, из электротехнических — Н. Н. Лященко и др. Случайные множества применялись для описания процесса случайного распространения, например распространения информации, слухов, эпидемии или пожара, а также в математической экономике. В монографии [2] рассмотрены приложения случайных множеств в теории экспертных оценок и в теории управления запасами и ресурсами (логистике).

Отметим, что с точки зрения математики реальные объекты можно моделировать случайными множествами как из конечного числа элементов, так и из бесконечного, однако при компьютерных расчетах неизбежна дискретизация, т.е. переход к первой из названных возможностей.

Объекты нечисловой природы как статистические данные. В теории и практике статистических методов наиболее распространенный объект изучения и применения — выборка $x_1, x_2, ..., x_n$, т.е. совокупность результатов n наблюдений (измерений, испытаний. анализов, опытов). В различных обла-

стях статистики результат наблюдения — это или число, или конечномерный вектор, или функция. Соответственно проводится, как уже отмечалось, деление прикладной статистики: одномерная статистика, многомерный статистический анализ, статистика временных рядов и случайных процессов... В нечисловой статистике (статистике нечисловых данных) в качестве результатов наблюдений рассматриваются объекты нечисловой природы. В частности, математические объекты перечисленных выше видов — измерения в шкалах, отличных от абсолютной, бинарные отношения, вектора из 0 и 1, множества. А также нечеткие множества, о которых речь пойдет ниже. Выборка может состоять из n ранжировок, или n толерантностей, или n множеств, или n нечетких множеств и т.д.

Отметим необходимость развития методов статистической обработки «разнотипных данных», обусловленную большой ролью в прикладных исследованиях «признаков смешанной природы». Речь идет о том, что результат наблюдения состояния объекта зачастую представляет собой вектор, у которого часть координат измерена по шкале наименований, часть — по порядковой шкале, часть — по шкале интервалов и т.д. Классические статистические методы ориентированы обычно либо на абсолютную шкалу, либо на шкалу наименований (анализ таблиц сопряженности), а потому зачастую непригодны для обработки разнотипных данных. Есть и более сложные модели разнотипных данных, например, когда некоторые координаты вектора наблюдений описываются нечеткими множествами.

Для обозначения подобных неклассических результатов наблюдений в 1979 г. в монографии [2] предложен собирательный термин — объекты нечисловой природы. Термин «нечисловой» означает, что структура пространства, в котором лежат результаты наблюдений, не является структурой действительных чисел, векторов или функций, она вообще не является структурой линейного (векторного) пространства. В памяти компьютеров и при расчетах объекты числовой природы, разумеется, изображаются с помощью чисел, но эти числа нельзя складывать и умножать.

С целью «стандартизации математических орудий» (выражение группы французских математиков, действовавшей в середине XX в. под псевдонимом Н. Бурбаки) целесообразно разрабатывать методы статистического анализа данных, пригодные одновременно для всех перечисленных выше видов результатов наблюдений. Кроме того, в процессе развития теоретических и прикладных исследований выявляется необходимость использования новых видов объектов нечисловой природы, отличных от рассмотренных выше,

например, в связи с развитием статистических методов обработки текстовой информации. Поэтому целесообразно ввести еще один вид объектов нечисловой природы — объекты произвольной природы, т.е. элементы множеств, на которые не наложено никаких условий (кроме «условий регулярности», необходимых для справедливости доказываемых теорем). Другими словами, в этом случае предполагается, что результаты наблюдений (элементы выборки) лежат в произвольном пространстве X. Для получения теорем необходимо потребовать, чтобы X удовлетворяло некоторым внутриматематическим условиям, например, было так называемым топологическим пространством. Как известно, ряд результатов классической математической статистики получен именно в такой постановке. Так, при изучении оценок максимального правдоподобия элементы выборки могут лежать в пространстве произвольной природы. Это не влияет на рассуждения, поскольку в них рассматривается лишь зависимость плотности вероятности от параметра. Методы классификации, использующие лишь расстояние между классифицируемыми объектами, могут применяться к совокупностям объектов произвольной природы, лишь бы в пространстве, где они лежат, была задана метрика. Цель нечисловой статистики (в некоторых литературных источниках используются термины «статистика нечисловых данных» и «статистика объектов нечисловой природы») состоит в том, чтобы систематически рассматривать методы статистической обработки данных как произвольной природы, так и относящихся к указанным выше конкретным видам объектов нечисловой природы, т.е. методы описания данных, оценивания и проверки гипотез. Взгляд с общей точки зрения позволяет получить новые результаты и в других областях прикладной статистики.

Объекты нечисловой природы при формировании статистической или математической модели реального явления. Использование объектов нечисловой природы часто порождено желанием обрабатывать более объективную, более освобожденную от погрешностей информацию. Как показали многочисленные опыты, человек более правильно (и с меньшими затруднениями) отвечает на вопросы качественного например, сравнительного, характера, чем количественного. Так, ему легче сказать, какая из двух гирь тяжелее, чем указать их примерный вес в граммах. Другими словами, использование объектов нечисловой природы — средство повысить устойчивость эконометрических и экономико-математических моделей реальных явлений. Сначала конкретные области статистики объектов нечисловой природы (а именно, прикладная теория измерений, нечеткие и случайные множества)

были рассмотрены в монографии [2] при анализе частных постановок проблемы устойчивости математических моделей социально-экономических явлений и процессов к допустимым отклонениям исходных данных и предпосылок модели. А затем была понята необходимость проведения работ по развитию статистики объектов нечисловой природы как самостоятельного научного направления.

Обсуждение начнем со шкал измерения. Науку о единстве мер и точности измерений называют метрологией. Таким образом, репрезентативная теория измерений — часть метрологии. Методы обработки данных должны быть адекватны относительно допустимых преобразований шкал измерения в смысле репрезентативной теории измерений. Однако установление типа шкалы, т.е. задание группы преобразований Ф — дело специалиста соответствующей прикладной области. Так, оценки привлекательности профессий мы считали измеренными в порядковой шкале [2]. Однако отдельные социологи не соглашались с этим, считая, что выпускники школ пользуются шкалой с более узкой группой допустимых преобразований, например, интервальной шкалой. Очевидно, эта проблема относится не к математике, а к наукам о человеке. Для ее решения может быть поставлен достаточно трудоемкий эксперимент. Пока же он не поставлен, целесообразно принимать порядковую шкалу, так как это гарантирует от возможных ошибок.

Как уже отмечалось, номинальные и порядковые шкалы широко распространены не только в социально-экономических исследованиях. Они применяются в медицине, минералогии, географии и т.д. Напомним, что по шкале интервалов измеряют величину потенциальной энергии или координату точки на прямой, на которой не отмечены ни начало, ни единица измерения; по шкале отношений — большинство физических единиц: массу тела, длину, заряд, а также цены в экономике. Время измеряется по шкале разностей, если год принимаем естественной единицей измерения, и по шкале интервалов в общем случае. В процессе развития соответствующей области знания тип шкалы может меняться. Так, сначала температура измерялась по порядковой шкале (холоднее — теплее), затем — по интервальной (шкалы Цельсия, Фаренгейта, Реомюра) и, наконец, после открытия абсолютного нуля температур — по шкале отношений (шкала Кельвина). Следует отметить, что среди специалистов иногда имеются разногласия по поводу того, по каким шкалам следует считать измеренными те или иные реальные величины.

Отметим, что термин «репрезентативная» использовался, чтобы отличить рассматриваемый подход к теории измерений от классической метроло-

гии, а также от работ А. Н. Колмогорова и А. Лебега, связанных с измерением геометрических величин, от «алгоритмической теории измерения», и от других научных направлений.

Необходимость использования в математических моделях реальных явлений таких объектов нечисловой природы, как бинарные отношения, множества, нечеткие множества, кратко была показана выше. Здесь же обратим внимание, что анализируемые в классической статистике результаты наблюдений также «не совсем числа». А именно, любая величина X измеряется всегда с некоторой погрешностью ΔX и результатом наблюдения является:

$$Y = X + \Delta X.$$

Как уже отмечалось, погрешностями измерений занимается метрология. Отметим справедливость следующих фактов:

а) для большинства реальных измерений невозможно полностью исключить систематическую ошибку, т.е. $M(\Delta X) \neq 0$;

б) распределение ΔX в подавляющем большинстве случаев не является нормальным [3];

в) измеряемую величину X и погрешность ее измерения ΔX обычно нельзя считать независимыми случайными величинами;

г) распределение погрешностей оценивается по результатам специально проведенных измерений, следовательно, полностью известным считать его нельзя; зачастую исследователь располагает лишь границами для систематической погрешности и оценками таких характеристик случайной погрешности, как дисперсия или размах.

Приведенные факты показывают ограниченность области применимости распространенной модели погрешностей, в которой X и ΔX рассматриваются как независимые случайные величины, причем ΔX имеет нормальное распределение с нулевым математическим ожиданием.

Строго говоря, результаты наблюдения всегда имеют дискретное распределение, поскольку описываются числами, у которых немного значащих цифр (обычно от 1 до 5). Возникает дилемма: либо признать, что непрерывные распределения — внутриматематическая фикция, и прекратить ими пользоваться, либо считать, что непрерывные распределения имеют «реальные» величины X, которые наблюдаются с принципиально неустранимой погрешностью ΔX. Первый выход в настоящее время нецелесообразен, так как

он требует отказа от большей части разработанного математического аппарата. Из второго следует необходимость изучения влияния неустранимых погрешностей на статистические выводы.

Погрешности ΔX можно учитывать либо с помощью вероятностной модели (ΔX — случайная величина, имеющая функцию распределения, вообще говоря, зависящую от X), либо с помощью нечетких множеств. Во втором случае приходим к теории нечетких чисел и к ее частному случаю — статистике интервальных данных.

Другой источник появления погрешности ΔX связан с принятой в конструкторской и технологической документации системой допусков на контролируемые параметры изделий и деталей, с использованием шаблонов при проверке контроля качества продукции [13]. В этих случаях характеристики ΔX определяются не свойствами средств измерения, а применяемой технологией проектирования и производства. В терминах прикладной статистики сказанному соответствует группировка данных, при которой мы знаем, какому из заданных интервалов принадлежит наблюдение, но не знаем точного значения результата наблюдения. Применение группировки может дать экономический эффект, поскольку зачастую легче (в среднем) установить, к какому интервалу относится результат наблюдения, чем точно измерить его.

Объекты нечисловой природы как результат статистической обработки данных. Объекты нечисловой природы появляются не только на «входе» статистической процедуры, но и в процессе обработки данных, и на «выходе» в качестве итога статистического анализа.

Рассмотрим простейшую прикладную постановку задачи регрессии (см. также [3]). Исходные данные имеют вид $(x_i, y_i) \in R^2$, $i = 1,2,...,n$. Цель состоит в том, чтобы с достаточной точностью описать y как многочлен (полином) от x, т.е. модель имеет вид:

$$y_i = \sum_{k=0}^{m} a_k x_i^k + \varepsilon_i, \quad i = 1,2,...,n, \tag{2}$$

где m — неизвестная степень полинома; $a_0, a_1, a_2, ..., a_m$ — неизвестные коэффициенты многочлена; ε_i, $i = 1,2,...,n$ — погрешности, которые для простоты примем независимыми и имеющими одно и то же нормальное распределение.

Здесь наглядно проявляется одна из причин живучести статистических моделей на основе нормального распределения. Такие модели, хотя и, как правило, неадекватны реальной ситуации [3], с математической точки зрения

позволяет проникнуть глубже в суть изучаемого явления. Поэтому они пригодны для первоначального анализа ситуации, как и в рассматриваемом случае. Дальнейшие научные исследования должны быть направлены на снятие нереалистического предположения нормальности и переход к непараметрическим моделям погрешности.

Распространенная процедура восстановления зависимости с помощью многочлена такова: сначала пытаются применить модель (2) для линейной функции (m = 1), при неудаче (неадекватности модели) переходят к многочлену второго порядка (m = 2), если снова неудача, то берут модель (2) с m = 3 и т.д. (адекватность модели проверяют по F-критерию Фишера).

Обсудим свойства этой процедуры в терминах прикладной статистики. Если степень полинома задана ($m = m_0$), то его коэффициенты оценивают методом наименьших квадратов, свойства этих оценок хорошо известны (см., например, учебник [3] или монографию [14, гл. 26]). Однако в описанной выше реальной постановке m тоже является неизвестным параметром и подлежит оценке. Таким образом, требуется оценить объект (m, a_0, a_1, a_2, ..., a_m), множество значений которого можно описать как $R^1 \bigcup R^2 \bigcup R^3 \bigcup ...$ Это — объект нечисловой природы, обычные методы оценивания для него неприменимы, так как m — дискретный параметр. В рассматриваемой постановке разработанные к настоящему времени методы оценивания степени полинома носят в основном эвристический характер (см., например, гл. 12 монографии [15]). Свойства описанной выше распространенной процедуры рассмотрены в [3]. Показано, что обычно используемыми методами степень полинома m оценивается несостоятельно, и найдено предельное распределение оценок этого параметра, оказавшееся геометрическим. Отметим, что для степени многочлена давно предложены состоятельные оценки [16].

В более общем случае линейной регрессии данные имеют вид (y_i, X_i), $i = 1,2,...,n$, где $X_i = (x_{i1}, x_{i2}, ..., x_{iN}) \in R^N$ — вектор предикторов (факторов, объясняющих переменных), а модель такова:

$$y_i = \sum_{j \in K} a_j x_{ij} + \varepsilon_i, \quad i = 1,2,...,n \qquad (3)$$

(здесь K — некоторое подмножество множества $\{1,2,\ldots,n\}$; ε_i — те же, что и в модели (2); a_j — неизвестные коэффициенты при предикторах с номерами из K). Модель (2) сводится к модели (3), если:

$$x_{i1} = 1, \; x_{i2} = x_i, \; x_{i3} = x_i^2, \; x_{i4} = x_i^3, ..., x_{ij} = x_i^{j-1}.$$

В модели (2) есть естественный порядок ввода предикторов в рассмотрение — в соответствии с возрастанием степени, а в модели (3) естественного порядка нет, поэтому здесь стоит произвольное подмножество множества предикторов. Есть только частичный порядок — чем мощность подмножества меньше, тем лучше. Модель (3) особенно актуальна в технических исследованиях (см. многочисленные примеры в журнале «Заводская лаборатория»). Она применяется в задачах управления качеством продукции и других технико-экономических исследованиях, в медицине, экономике, маркетинге и социологии, когда из большого числа факторов, предположительно влияющих на изучаемую переменную, надо отобрать по возможности наименьшее число значимых факторов и с их помощью сконструировать прогнозирующую формулу (3).

Задача оценивания модели (3) разбивается на две последовательные задачи: оценивание множества K — подмножества множества всех предикторов, а затем — неизвестных параметров a_j. Методы решения второй задачи хорошо известны и подробно изучены (обычно используют метод наименьших квадратов). Гораздо хуже обстоит дело с оцениванием объекта нечисловой природы K. Как уже отмечалось, существующие методы — в основном эвристические, они зачастую не являются даже состоятельными. Даже само понятие состоятельности в данном случае требует специального определения. Пусть K_0 — истинное подмножество предикторов, т.е. подмножество, для которого справедлива модель (3), а подмножество предикторов K_n — его оценка. Оценка K_n называется состоятельной, если:

$$\lim_{n\to\infty} Card(K_n \Delta K_0) = 0 ,$$

где Δ — символ симметрической разности множеств; $Card(K)$ означает число элементов множества K, а предел понимается в смысле сходимости по вероятности.

Задача оценивания в моделях регрессии, таким образом, разбивается на две — оценивание структуры модели и оценивание параметров при заданной структуре. В модели (2) структура описывается неотрицательным целым числом m, в модели (3) — множеством K. Структура — объект нечисловой природы. Задача ее оценивания сложна, в то время как задача оценивания численных параметров при заданной структуре хорошо изучена, разработаны эффективные (в смысле прикладной математической статистики) методы.

Такова же ситуация и в других методах многомерного статистического анализа — в факторном анализе (включая метод главных компонент) и в многомерном шкалировании, в иных оптимизационных постановках проблем прикладного многомерного статистического анализа.

Перейдем к объектам нечисловой природы на «выходе» статистической процедуры. Примеры многочисленны. Разбиения — итог работы многих алгоритмов классификации, в частности, алгоритмов кластер-анализа. Ранжировки — результат упорядочения профессий по привлекательности или автоматизированной обработки мнений экспертов — членов комиссии по подведению итогов конкурса научных работ. (В последнем случае используются ранжировки со связями; так, в одну группу, наиболее многочисленную, попадают работы, не получившие наград.) Из всех объектов нечисловой природы, видимо, наиболее часты на «выходе» дихотомические данные — принять или не принять гипотезу, в частности, принять или забраковать партию продукции. Результатом статистической обработки данных может быть множество, например зона наибольшего поражения при аварии, или последовательность множеств, например, «среднемерное» описание распространения пожара (см. главу 4 в монографии [2]). Нечетким множеством Э. Борель [17] еще в начале XX в. предлагал описывать представление людей о числе зерен, образующем «кучу». С помощью нечетких множеств формализуются значения лингвистических переменных, выступающих как итоговая оценка качества систем автоматизированного проектирования, сельскохозяйственных машин, бытовых газовых плит, надежности программного обеспечения или систем управления. Можно констатировать, что все виды объектов нечисловой природы могут появляться « на выходе» статистического исследования.

1.4. ВЕРОЯТНОСТНЫЕ МОДЕЛИ ПОРОЖДЕНИЯ НЕЧИСЛОВЫХ ДАННЫХ

Рассмотрим основные вероятностные модели порождения нечисловых данных. А именно, дихотомических данных, результатов парных сравнений, бинарных отношений, рангов, объектов общей природы. Обсудим различные варианты вероятностных моделей и их практическое использование (см. также обзор [18]).

Дихотомические данные. Рассмотрим базовую вероятностную модель дихотомических данных — *бернуллиевский вектор* (в терминологии энциклопедии [19] — *люсиан*), т.е. конечную последовательность $X = (X_1, X_2, ..., X_k)$ не-

зависимых испытаний Бернулли X_i, для которых $P(X_i = 1) = p_i$ и $P(X_i = 0) = 1 - p_i$, $i = 1,2,...,k,$ причем вероятности p_i могут быть различны.

Бернуллиевские вектора часто применяются при практическом использовании эконометрических методов. Так, они использованы в монографии [2] для описания равномерно распределенных случайных толерантностей. Как известно, толерантность на множестве из m элементов можно задать симметричной матрицей $\| \delta_{ij} \|$ из 0 и 1, на главной диагонали которой стоят 1. Тогда случайная толерантность описывается распределением $m(m - 1) / 2$ дихотомических случайных величин δ_{ij}, $1 \le i < j \le m,$ а для равномерно распределенной (на множестве всех толерантностей) толерантности эти случайные величины, как можно доказать, оказываются независимыми и принимают значения 0 и 1 с равными вероятностями 1/2. Записав элементы δ_{ij} задающей такую толерантность матрицы в строку, получим бернуллиевский вектор с $k = m(m - 1) / 2$ и $p_i = 1/2,$ $i = 1,2,...,k.$

В связи с оцениванием по статистическим данным функции принадлежности нечеткой толерантности в 1970-е гг. была построена теория случайных толерантностей с такими независимыми $\delta_{ij},$ что вероятности $P(\delta_{ij} = 1) = p_{ij}$ произвольны [2]. Случайные множества с независимыми элементами использовались как общий язык для описания парных сравнений и случайных толерантностей. В некоторых публикациях термин «люсиан» применялся как сокращение для выражения «случайные множества с независимыми элементами».

Был выявлен ряд областей, в которых полезен математический аппарат решения различных статистических задач, связанных с бернуллиевскими векторами. Перечислим эти области, включая ранее названные:

- анализ случайных толерантностей;
- случайные множества с независимыми элементами;
- обработка результатов независимых парных сравнений;
- статистические методы анализа точности и стабильности технологического процесса,
- анализ и синтез планов статистического приемочного контроля (по альтернативным, т.е. дихотомическим, признакам);
- обработка маркетинговых и социологических анкет (с закрытыми вопросами типа «да» — «нет»);

- обработка социально-психологических и медицинских данных, в частности, ответов на психологические тесты типа MMPI (используемых в задачах управления персоналом);

- анализ топографических карт (применяемых для анализа и прогноза зон поражения при технологических авариях, распространении коррозии, распространении экологически вредных загрязнений и в других ситуациях) и т.д.

Теорию бернуллиевских векторов можно выразить в терминах любой из этих теоретических и прикладных областей. Однако терминология одной из этих областей «режет слух» и приводит к недоразумениям в другой из них. Поэтому целесообразно использовать термин «бернуллиевский вектор» в указанном выше значении, не связанном ни с какой из перечисленных областей приложения этой теории (в ряде публикаций в том же значении использовался термин «люсиан»).

Распределение бернуллиевского вектора X полностью описывается векторным параметром $P = (p_1, p_2, ..., p_k)$, т.е. нечетким подмножеством множества {1, 2, ..., k}. Действительно, для любого детерминированного вектора $x = (x_1, x_2, ..., x_k)$ из 0 и 1 имеем:

$$P(X = x) = \prod_{1 \le j \le k} h(x_j, p_j),$$

где $h(x, p) = p$ при $x = 1$ и $h(x,p) = 1 - p$ при $x = 0$.

Теперь можно уточнить способы использования люсианов в прикладной статистике. Бернуллиевскими векторами можно моделировать:

- результаты статистического контроля (0 — годное изделие, 1 — дефектное);

- результаты маркетинговых и социологических опросов (0 — опрашиваемый выбрал первую из двух подсказок, 1 — вторую);

- распределение посторонних включений в материале (0 — нет включения в определенном объеме материала, 1 — есть);

- результаты испытаний и анализов (0 — нет нарушений требований нормативно-технической документации, 1 — есть такие нарушения);

- процессы распространения, например, пожаров (0 — нет загорания, 1 — есть; подробнее см. [2, с. 215–223]);

- состояние технологического процесса (0 — процесс находится в границах допуска, 1 — вышел из них);

- ответы экспертов (опрашиваемых) о сходстве объектов (проектов, образцов) и т.д.

Парные сравнения. Общую модель парных сравнений опишем согласно монографии Г. Дэвида [12, с. 9]. Предположим, что t объектов $A_1, A_2, ..., A_t$ сравниваются попарно каждым из n экспертов. Всего возможных пар для сравнения имеется $s = t(t-1)/2$. Эксперт с номером γ делает r_γ повторных сравнений для каждой из s возможностей. Пусть $X(i, j, \gamma, \delta)$, $i, j = 1, 2, ..., t$, $i \neq j$, $\gamma = 1, 2, ..., n$; $\delta = 1, 2, ..., r_\gamma$, — случайная величина, принимающая значение 1 или 0 в зависимости от того, предпочитает ли эксперт с номером γ объект A_i или объект A_j в δ-м сравнении двух объектов. Предполагается, что все сравнения проводятся независимо друг от друга, так что случайные величины $X(i, j, \gamma, \delta)$ независимы в совокупности, если не считать того, что $X(i, j, \gamma, \delta) + X(j, i, \gamma, \delta) = 1$. Положим:

$$P(X(i, j, \gamma, \delta) = 1) = \pi(i, j, \gamma, \delta).$$

Ясно, что описанная модель парных сравнений представляет собой частный случай бернуллиевского вектора. В этой модели число наблюдений равно числу неизвестных параметров, поэтому для получения статистических выводов необходимо наложить априорные условия на вероятности $\pi(i, j, \gamma, \delta)$, например [12, с. 9]:

- $\pi(i, j, \gamma, \delta) = \pi(i, j, \gamma)$ (нет эффекта от повторений);
- $\pi(i, j, \gamma, \delta) = \pi(i, j)$ (нет эффекта от повторений и от экспертов).

Теорию независимых парных сравнений целесообразно разделить на две части — непараметрическую, в которой статистические задачи ставятся непосредственно в терминах $\pi(i, j, \gamma, \delta)$, и параметрическую, в которой вероятности $\pi(i, j, \gamma, \delta)$ выражаются через меньшее число иных параметров. Ряд результатов непараметрической теории парных сравнений непосредственно вытекает из теории бернуллиевских векторов.

В параметрической теории парных сравнений наиболее популярна так называемая линейная модель [12, с. 11], в которой предполагается , что каждому объекту A_i можно сопоставить некоторую «ценность» V_i так, что вероятность предпочтения $\pi(i, j)$ (т.е. предполагается дополнительно, что эффект от повторений и от экспертов отсутствует) выражается следующим образом:

$$\pi(i, j) = H(V_i - V_j), \qquad (1)$$

где $H(x)$ — функция распределения, симметричная относительно 0, т.е.

$$H(-x) = 1 - H(x) \qquad (2)$$

при всех x.

Широко применяются модели Терстоуна — Мостеллера и Брэдли — Терри, в которых $H(x)$ — функции нормального и логистического распределений соответственно. Поскольку функция $\Phi(x)$ стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1 и функция:

$$\Psi(x) = e^x (1 + e^x)^{-1}$$

стандартного логистического распределения удовлетворяют (см., например, [20]) соотношению:

$$\sup_{x \in R^1} |\Phi(x) - \Psi(1{,}7x)| < 0{,}01,$$

то для обоснованного выбора по статистическим данным между моделями Терстоуна — Мостеллера и Брэдли — Терри необходимо не менее тысячи наблюдений [3, п. 4.1].

Соотношение (1) вытекает из следующей модели поведения эксперта: он измеряет «ценность» V_i и V_j объектов A_i и A_j, но с ошибками ε_i и ε_j соответственно, а затем сравнивает свои оценки ценности объектов $y_i = V_i + \varepsilon_i$ и $y_j = V_j + \varepsilon_j$. Если $y_i > y_j$, то он предпочитает A_i, в противном случае — A_j. Тогда

$$\pi(i, j) = P(\varepsilon_j - \varepsilon_i < V_i - V_j) = H(V_i - V_j). \qquad (3)$$

Обычно предполагают, что субъективные ошибки эксперта ε_i и ε_j независимы и имеют одно и то же непрерывное распределение. Тогда функция распределения $H(x)$ из соотношения (3) непрерывна и удовлетворяет функциональному уравнению (2).

Существует много разновидностей моделей парных сравнений, постоянно предполагаются новые. В качестве примера опишем модель парных сравнений, основанную не на процедуре упорядочения, а на определении сходства объектов. Пусть каждому объекту A_i соответствует точка a_i в

r-мерном евклидовом пространстве R^r. Эксперт «измеряет» a_i и a_j с ошибками ε_i и ε_j соответственно и в случае, если евклидово расстояние между $a_i + \varepsilon_i$ и $a_j + \varepsilon_j$ меньше 1, заявляет о сходстве объектов A_i и A_j, в противном случае — об их различии. Предполагается, что ошибки ε_i и ε_j независимы и имеют одно и то же распределение, например, круговое нормальное распределение с нулевым математическим ожиданием и дисперсией координат σ^2. Целью статистической обработки является определение по результатам парных сравнений оценок параметров a_1, a_2, ..., a_t и σ^2, а также проверка согласия опытных данных с моделью.

Рассмотренные модели парных сравнений могут быть обобщены в различных направлениях. Так, можно ввести понятие «ничья» — ситуации, когда эксперт оценивает объекты одинаково. Модели с учетом «ничьих» предполагают, что эксперт может отказаться от выбора одного из объектов и заявить об их эквивалентности, т. е. число возможных ответов увеличивается с 2 до 3. В моделях множественных сравнений эксперту представляется не два объекта, а три или большее число.

Модели, учитывающие «ничьи», строятся обычно с помощью используемых в психофизике «порогов чувствительности»: если $|y_i - y_j| \le d$ (где d — порог чувствительности), то объекты A_i и A_j эксперт объявляет неразличимыми. Приведем пример модели с «ничьими», основанной на другом принципе. Пусть каждому объекту A_i соответствует точка a_i в r-мерном линейном пространстве. Как и прежде, эксперт «измеряет» объектные точки a_i и a_j с ошибками ε_i и ε_j соответственно, т.е. принимает решение на основе $y_i = a_i + \varepsilon_i$ и $y_j = a_j + \varepsilon_j$. Если все координаты y_i больше соответствующих координат y_j, то A_i предпочитается A_j. Соответственно, если каждая координата y_i меньше координаты y_j с тем же номером, то эксперт считает наилучшим объект A_j. Во всех остальных случаях эксперт объявляет о ничейной ситуации. Эта модель при $r = 1$ переходит в описанную выше линейную модель. Она связана с принципом Парето в теории группового выбора и предусматривает выбор оптимального по Парето объекта, если он существует (роль согласуемых критериев играют процедуры сравнения значений отдельных координат), и отказ от выбора, если такого объекта нет.

Можно строить модели, учитывающие порядок предъявления объектов при сравнении, зависимость результата сравнения от результатов предшествующих сравнений. Опишем одну из подобных моделей.

Пусть эксперт сравнивает три объекта — A, B, C, причем сначала сравниваются A и B, потом — B и C и, наконец, A и C. Для определенности пусть $A > B$ будет означать, что A более предпочтителен, чем B. Пусть при предъявлении двух объектов:

$$P(A > B) = \pi_{AB}, P(B > C) = \pi_{BC}, P(A > C) = \pi_{AC}.$$

Теперь пусть пара B, C предъявляется после пары A, B. Естественно предположить, что высокая оценка B в первом сравнении повышает вероятность предпочтения B и во втором, и, наоборот, отрицательное мнение о B в первом сравнении сохраняется и при проведении второго сравнения. Это предположение проще всего учесть в модели следующим образом:

$$P(B > C \mid B > A) = \pi_{BC} + \delta, \quad P(B > C \mid A > B) = \pi_{BC} - \delta,$$

где δ — некоторое положительное число, показывающее степень влияния первого сравнения на второе. По аналогичным причинам вероятности исхода третьего сравнения в зависимости от результатов первых двух можно описать так:

$$P(A > C \mid A > B, B > C) = \pi_{AC} + 2\delta, \quad P(A > C \mid A > B, B < C) = \pi_{AC},$$
$$P(A > C \mid A < B, B > C) = \pi_{AC}, \quad P(A > C \mid A < B, B < C) = \pi_{AC} - 2\delta.$$

Статистическая задача состоит в определении параметров π_{AB}, π_{BC}, π_{AC} и δ по результатам сравнений, проведенных n экспертами, и в проверке адекватности модели.

Ясно, что можно рассматривать и другие модели, в частности, учитывающие тягу экспертов к транзитивности ответов. Очевидно, что проблемы построения моделей парных сравнений относятся не к нечисловой статистике, а к тем прикладным областям, для решения задач которых развиваются методы парных сравнений, например, к организации машиностроительного производства, экономике предприятия, стратегическому менеджменту, производственной психологии, изучению поведения потребителей, экспертным оценкам и т.д.

Метод парных сравнений был введен в 1860 г. Г. Т. Фехнером для решения задач психофизики. Расскажем об этом несколько подробнее. Как из-

вестно, основателем психофизики по праву считается Густав Теодор Фехнер (1801–1887 гг.), а год выхода в свет его фундаментальной работы «Элементы психофизики» (1860 г.) — датой рождения новой науки. В этой работе широко применялся предложенный Г. Т. Фехнером метод парных сравнений (обсуждение событий тех лет с современных позиций дано в монографии [12, с. 14–16]).

С точки зрения математической статистики, приведенные выше модели не представляют большого теоретического интереса: оценки параметров находятся обычно методом максимального правдоподобия или асимптотически эквивалентным ему методом одношаговых оценок (см. ниже главу 2), а проверка согласия проводится по критерию отношения правдоподобия или асимптотически эквивалентными ему критериями типа хи-квадрат [12]. При этом вычислительные процедуры обычно достаточно сложны и плохо исследованы.

Отметим некоторые сложности при обосновании возможности использовании линейных моделей типа (1) — (3). Вероятностно-статистическая теория достаточно проста, когда предполагается, что каждому отдельному сравнению двух объектов соответствуют свои собственные ошибки экспертов, причем все ошибки независимы в совокупности. Однако это предположение отнюдь не очевидно с содержательной точки зрения. В качестве примера рассмотрим три объекта A, B и C, которые сравнивают попарно: A и B, B и C, A и C. В соответствии со сказанным, в рассмотрение вводят 6 ошибок одного и того же эксперта: ε_A и ε_B в первом сравнении, $\varepsilon_B^{'}$ и ε_C — во втором, $\varepsilon_A^{'}$ и $\varepsilon_C^{'}$ — в третьем, причем все эти 6 случайных величин независимы в совокупности. Между тем естественно думать, что мнения эксперта об одном и том же объекте связаны между собой. Т. е. ε_A и $\varepsilon_A^{'}$ зависимы, равно как ε_B и $\varepsilon_B^{'}$, а также ε_C и $\varepsilon_C^{'}$. Более того, если принять, что точка зрения эксперта полностью определена для него самого, то следует положить $\varepsilon_A=\varepsilon_A^{'}$ и соответственно $\varepsilon_B=\varepsilon_B^{'}$ и $\varepsilon_C=\varepsilon_C^{'}$. При этом, напомним, случайные величины ε_A, ε_B и др. интерпретируется как отклонения мнений отдельных экспертов от истины. Видимо, ошибку эксперта целесообразно считать состоящей из двух слагаемых, а именно: отклонения от истины, вызванного внутренними особенностями эксперта (систематическая погрешность) и колебания мнения эксперта в связи с очередным парным сравнением (случайная погрешность). Игнорирование систематической погрешности облегчает развитие математи-

ко-статистической теории, а ее учет приводит к необходимости изучения зависимых парных сравнений.

При обработке результатов парных сравнений первый этап — проверка согласованности. Понятие согласованности уточняется различными способами, но все они имеют один и тот же смысл проверки однородности обрабатываемого материала, т.е. того, что целесообразно агрегировать мнения отдельных экспертов, объединить данные и совместно их обрабатывать. При отсутствии однородности данные разбиваются на группы (классы, кластеры, таксоны) с целью обеспечения однородности внутри отдельных групп. Естественно, согласованность целесообразно проверять, вводя возможно меньше гипотез о структуре данных. Следовательно, целесообразно пользоваться для этого непараметрической теорией парных сравнений, основанной на теории бернуллиевских векторов.

Хорошо известно, что модели парных сравнений с успехом применяются в экспертных и экспериментальных процедурах упорядочивания и выбора. В частности, для анализа голосований, турниров, выбора наилучшего объекта (проекта, образца, кандидатуры); в планировании и анализе сравнительных экспериментов и испытаний; в органолептической экспертизе (в частности, дегустации); при изучении поведения потребителей; визуальной колоритмии (принятии решений на основе цвета), определении индивидуальных рейтингов и вообще изучении предпочтений при выборе и т. д. (подробнее см. [2, 3, 12]).

Бинарные отношения. Теорию ранговой корреляции можно рассматривать как теорию статистического анализа случайных ранжировок, равномерно распределенных на множестве всех ранжировок. Так, при обработке данных классического психофизического эксперимента по упорядочению кубиков соответственно их весу, подробно описанного в работе [21], оказалась адекватной следующая так называемую *T*-модель ранжирования.

Пусть имеется *t* объектов $A_1, A_2, ..., A_t$, причем каждому объекту A_i соответствует число a_i, описывающее его положение на шкале изучаемого признака. Испытуемый упорядочивает объекты так, как если бы оценивал соответствующие им значения с ошибками, т.е. находил $y_i = a_i + \varepsilon_i$, $i = 1, 2, \ldots n$, где ε_i — ошибка при рассмотрении *i*-го объекта, а затем располагал бы объекты в том порядке, в каком располагаются $y_1, y_2, ..., y_t$. В этом случае вероятность появления упорядочения $A_{i1}, A_{i2}, ...A_{it}$ есть $P(y_{i1} < y_{i2} < ... < y_{it})$, а ранги $R_1, R_2, ...R_t$ объектов являются рангами случайных величин $y_1, y_2, ..., y_t$, получен-

ными при их упорядочении в порядке возрастания. Кроме того, для простоты расчетов в модели предполагается, что ошибки испытуемого $\varepsilon_1, \varepsilon_2, ..., \varepsilon_t$ независимы и имеют нормальное распределение с математическим ожиданием 0 и дисперсией σ^2.

Как уже отмечалось, бинарное отношение на множестве из t элементов полностью описывается матрицей из 0 и 1 порядка $t \times t$. Поэтому задать распределение случайного бинарного отношения — это то же самое, что задать распределение вероятностей на множестве всех матриц описанного вида, состоящем из $2^{(t^2)}$ элементов. Пространства ранжировок, разбиений, толерантностей зачастую удобно считать подпространствами пространства всех бинарных отношений, тогда распределения вероятностей на них — частные случаи описанного выше распределения, выделенные тем, что вероятности принадлежности случайного бинарного отношения соответствующим подпространствам равны 1. Распределение произвольного бинарного отношения описывается $2^{(t^2)} - 1$ параметрами, распределение случайной ранжировки (без связей) — $(t! - 1)$ параметрами, а описанная выше T-модель ранжирования задается $(t + 1)$ параметром. При $t = 4$ эти числа равны соответственно 65 535, 23 и 5. Первое из этих чисел показывает реальную невозможность использования произвольных бинарных отношений в предназначенных для практического применения вероятностно-статистических моделях, поскольку по имеющимся данным невозможно оценить столь большое число параметров. Приходится ограничиваться теми или иными семействами бинарных отношений — ранжировками, разбиениями, толерантностями и др. Модель произвольной случайной ранжировки при $t = 5$ описывается 119 параметрами, при $t = 6$ — уже 719 параметрами, при $t = 7$ число параметром достигает 5 049, что уже явно за пределами возможности оценивания. В то же время T-модель ранжирования при $t = 7$ описывается всего 8-ю параметрами, а потому может быть кандидатом для практического использования.

Что естественно предположить относительно распределения случайного элемента со значениями в том или ином пространстве бинарных отношений? Зачастую целесообразно считать, что распределение имеет некий центр, попадание в который наиболее вероятно, а по мере удаления от центра вероятности убывают. Это соответствует естественной модели измерения с ошибкой. В классическом одномерном случае результат подобного измерения обычно описывается унимодальной симметричной плотностью, монотонно возрастающей слева от модального значения, в котором плотность максимальна, и монотонно убывающей справа от него. Чтобы ввести понятие

монотонного распределения в пространстве бинарных отношений, будем исходить из метрики в этом пространстве. Воспользовавшись тем, что бинарные отношения C и D однозначно описываются матрицами $\| c_{ij} \|$ и $\| d_{ij} \|$ порядка $t \times t$, рассмотрим расстояние (в несколько другой терминологии — метрику) в пространстве бинарных отношений:

$$d(C,D) = \sum_{1 \le i,j \le t} | c_{ij} - d_{ij} |. \qquad (4)$$

Метрика (4) в различных пространствах бинарных отношений — ранжировок, разбиений, толерантностей — может быть введена с помощью соответствующих систем аксиом (см. ниже). В настоящее время метрику (4) обычно называют расстоянием Кемени в честь американского исследователя Джона Кемени, впервые получившего эту метрику исходя из предложенной им системы аксиом для расстояния между упорядочениями (ранжировками).

В статистике нечисловых данных используются и иные метрики, отличающиеся от расстояния Кемени. Более того, для введения понятия монотонного распределения, о котором сейчас идет речь, нет необходимости требовать выполнения неравенства треугольника, а достаточно, чтобы $d(C, D)$ можно было рассматривать как показатель различия. Под показателем различия понимаем такую функцию $d(C, D)$ двух бинарных отношений C и D, что $d(C, D) = 0$ при $C = D$ и увеличение $d(C, D)$ интерпретируется как возрастание различия между C и D.

Определение 1. Распределение бинарного отношения X называется монотонным с центром в C_0 относительно расстояния (показателя различия) d, если из $d(C, C_0) < d(D, C_0)$ следует, что $P(X = C) > P(X = D)$.

Это определение впервые введено в монографии [2, с. 196]. Оно может использоваться в любых пространствах бинарных отношений и, более того, в любых пространствах из конечного числа элементов, лишь бы в них была введена функция $d(C, D)$ — показатель различия элементов C и D этого пространства. Монотонное распределение унимодально, мода находится в C_0.

Определение 2. Распределение бинарного отношения X называется симметричным относительно расстояния d с центром в C_0, если существует такая функция $f : R_+^1 \to [0,1]$, что

$$P(X = C) = f(d(C,C_0)). \qquad (5)$$

Если распределение X монотонно и таково, что из $d(C, C_0) = d(D, C_0)$ следует $P(X = C) = P(X = D)$, то оно симметрично. Если функция f в формуле (5) монотонно строго убывает, то соответствующее распределение монотонно в смысле определения 1.

Поскольку толерантность на множестве из t элементов задается $0{,}5t(t-1)$ элементами матрицы из 0 и 1 порядка $t \times t$, лежащими выше главной диагонали, то всего толерантностей имеется $2^{0,5t(t-1)}$, а потому распределение на множестве толерантностей задается в общем случае $2^{0,5t(t-1)} - 1$ параметрами. Естественно выделить семейство распределений, соответствующее независимым элементам матрицы. Оно задается бернуллиевским вектором (люсианом) с $0{,}5t(t-1)$ параметрами (выше бернуллиевские вектора рассмотрены подробнее). Математическая техника, необходимая для изучения толерантностей с независимыми элементами, существенно проще, чем в случае ранжировок и разбиений. Здесь легко отказаться от условия равномерности распределения. Этому условию соответствует $p_{ij} \equiv 1/2$, в то время как статистические методы анализа люсианов, развитые в статистике нечисловых данных (см., например, работы [2, 22, 23]) не налагают никаких существенных ограничений на p_{ij}.

Как уже отмечалось, при обработке мнений экспертов сначала проверяют согласованность. В частности, если мнения экспертов описываются монотонными распределениями, то для согласованности необходимо совпадение центров этих распределений. К сожалению, рассмотренные выше классические методы проверки согласованности для ранжировок, основанные на коэффициентах ранговой корреляции и конкордации, позволяют лишь отвергнуть гипотезу о равнораспределенности. Но не установить, можно ли считать, что центры соответствующих экспертам распределений совпадают или же, например, существует две группы экспертов, каждая со своим центром. Теория случайных толерантностей лишена этого недостатка. Отсюда вытекают следующие практические рекомендации.

Пусть цель обработки экспертных данных состоит в получении ранжировки, отражающей групповое мнение. Однако согласно рекомендуемой процедуре экспертного опроса пусть эксперты не упорядочивают объекты, а проводят парные сравнения, сравнивая каждый из рассматриваемых объектов со всеми остальными, причем ровно один раз. Тогда ответ эксперта — толерантность, но, вообще говоря, не ранжировка, поскольку в ответах эксперта может нарушаться транзитивность.

Возможны два пути обработки данных. Первый — превратить ответ эксперта в ранжировку (тем или иным способом «спроектировав» его на пространство ранжировок), а затем проверять согласованность ранжировок с помощью известных критериев. При этом от толерантности перейти к ранжировке можно, например, так. Будем выбирать ближайшую (в смысле применяемого расстояния) матрицу к матрице ответов эксперта из всех, соответствующих ранжировкам без связей.

Второй путь — проверить согласованность случайных толерантностей, а групповое мнение искать с помощью медианы Кемени (подробнее см. ниже) непосредственно по исходным данным, т.е. по толерантностям. Групповое мнение при этом может быть найдено в пространстве ранжировок. Второй путь мы считаем более предпочтительным, поскольку при этом обеспечивается более адекватная проверка согласованности и исключается процедура укладывания мнения эксперта в «прокрустово ложе» ранжировки (эта процедура может приводить как к потере информации, так и к принципиально неверным выводам, вызванным искажениями мнений экспертов).

Области применения статистики бинарных отношений многообразны: ранговая корреляция — оценка величины связи между переменными, измеренными в порядковой шкале; анализ экспертных или экспериментальных упорядочений; анализ разбиений технико-экономических показателей на группы сходных между собой; обработка данных о сходстве (взаимозаменяемости); статистический анализ классификаций; математические вопросы теории менеджмента и др.

Случайные множества. Будем рассматривать случайные подмножества некоторого множества Q. Если Q состоит из конечного числа элементов, то считаем, что случайное подмножество S — это случайный элемент со значениями в 2^Q — множестве всех подмножеств множества Q, состоящем из $2^{card(Q)}$ элементов. Чтобы удовлетворить математиков, считаем, что все подмножества Q измеримы (другими словами, σ-алгебра измеримых множеств совпадает с совокупностью всех подмножеств рассматриваемого конечного множества). Тогда распределение случайного подмножества $S = S(\omega)$ множества Q — это:

$$P_S(A) = P(S = A) = \mathrm{P}(\{\omega : S(\omega) = A\}), A \subseteq Q. \qquad (6)$$

В формуле (6) предполагается, что $S : \Omega \to 2^Q$, где (Ω, F, P) — вероятностное пространство, на котором определен случайный элемент $S(\omega)$. (Здесь

Ω — пространство элементарных событий, F — σ-алгебра случайных событий, P — вероятностная мера на F.) Через распределение $P_S(A)$ выражаются вероятности различных событий, связанных с S. Так, чтобы найти вероятность накрытия фиксированного элемента q случайным множеством S, достаточно вычислить:

$$P(q \in S) = \mathrm{P}(\{\omega : q \in S(\omega)\}) = \sum_{A: q \in A, A \subseteq 2^Q} P(S = A),$$

где суммирование идет по всем подмножествам A множества Q, содержащим q. Пусть $Q = \{q_1, q_2, ..., q_k\}$. Рассмотрим случайные величины, определяемые по случайному множеству S следующим образом:

$$\chi_i(\omega) = \begin{cases} 1, q_i \in S(\omega), \\ 0, q_i \notin S(\omega). \end{cases}$$

Определение 3. Случайное множество S называется случайным множеством с независимыми элементами, если случайные величины $\chi_i(\omega), i = 1,2,...,k,$ независимы (в совокупности).

Последовательность случайных величин $\chi_1, \chi_2, ..., \chi_k$ — бернуллиевский вектор с $X_i = \chi_i$ и $p_i = P(q_i \in S(\omega)), i = 1,2,...,k.$ Из сказанного выше следует, что распределение случайного множества с независимыми элементами задается формулой:

$$P(S = A) = \prod_{q_i \in A} p_i \prod_{q_i \in Q \setminus A} (1 - p_i),$$

т.е. такие распределения образуют $k = card(Q)$ — мерное параметрическое семейство, входящее в $(2^{card(Q)} - 1)$ — одномерное семейство всех распределений случайных подмножеств множества Q.

При исследовании случайных подмножеств произвольного множества Q будем рассматривать их как случайные величины со значениями в некотором пространстве подмножеств множества Q, например, в пространстве замкнутых подмножеств 2^Q множества Q.

Представляющими интерес лишь для математиков способами введения измеримой структуры в 2^Q интересоваться не будем. Отсутствие специального интереса к проблеме измеримости связано с тем, что при вероятностно-

статистическом моделировании и обработке на компьютере все случайные подмножества рассматриваются как конечные (т.е. подмножества конечного множества).

Случайные множества находят разнообразные применения в многообразных проблемах эконометрики и математической экономики. В том числе в задачах управлении запасами и ресурсами (см. об этом главу 5 в монографии [2]), в задачах менеджмента и, в частности, маркетинга, в экспертных оценках, например, при анализе мнений голосующих или опрашиваемых, каждый из которых отмечает несколько пунктов из списка и т.д. Кроме того, случайные множества применяются в гранулометрии, при изучении пористых сред и объектов сложной природы в таких областях, как металлография, петрография, биология, в частности, математическая морфология. Они используются при изучении структуры веществ и материалов, в исследовании процессов распространения, в том числе просачивания, распространения пожаров, экологических загрязнений, при районировании, в изучении областей поражения, например, поражения металла коррозией и сердечной мышцы при инфаркте миокарда и т.д., и т.п. Можно вспомнить о компьютерной томографии, о наглядном представлении сложной информации на экране компьютера, об изучении распространения рекламной информации, о картах Кохонена (популярный метод представления информации при применении нейросетей) и т.д.

Ранговые методы. Ранее установлено, что любой адекватный алгоритм в порядковой шкале является функцией от некоторой матрицы C. Пусть никакие два из результатов наблюдений $x_1, x_2, ..., x_n$ не совпадают, а $r_1, r_2, ..., r_n$ — их ранги. Тогда элементы матрицы C и ранги результатов наблюдений связаны взаимно однозначным соответствием:

$$r_i = 1 + \sum_{1 \le j \le n} (1 - c_{ij}),$$

а c_{ij} через ранги выражаются так: $c_{ij} = 1$, если $r_i < r_j$, и $c_{ij} = 0$ в противном случае.

Сказанное означает, что при обработке данных, измеренных в порядковой шкале, могут применяться только ранговые статистические методы. Отметим, что часто используемое в непараметрической статистике преобразование $Y = F(X)$ (здесь $F(x)$ — непрерывная функция распределения случайной величины X, причем F предполагается произвольной) фактически означает переход к порядковой шкале, поскольку статистические вы-

воды при этом инвариантны относительно допустимых преобразований в порядковой шкале.

Разумеется, ранговые статистические методы могут применяться не только при обработке данных, измеренных в порядковой шкале. Так, для проверки независимости двух количественных признаков в случае, когда нет уверенности в нормальности соответствующего двумерного распределения, целесообразно пользоваться коэффициентами ранговой корреляции Кендалла или Спирмена.

В настоящее время с помощью непараметрических и прежде всего ранговых методов можно решать тот же набор задач прикладной статистики, что и с помощью параметрических методов, в частности, основанных на предположении нормальности. Однако параметрические методы вошли в массовое сознание исследователей и инженеров и мешают широкому внедрению более обоснованной и прогрессивной ранговой статистики. Так, при проверке однородности двух выборок вместо критерия Стьюдента целесообразно использовать ранговые методы [3], но пока это делается редко.

Объекты общей природы. Вероятностная модель объекта нечисловой природы в общем случае — случайный элемент со значениями в пространстве произвольного вида, а модель выборки таких объектов — совокупность независимых одинаково распределенных случайных элементов. Именно такая модель была использована для обработки наблюдений, каждое из которых — нечеткое множество [24].

Из-за имеющего разнобоя в терминологии приведем математические определения из справочника по теории вероятностей академика РАН Ю. В. Прохорова и профессора Ю. А. Розанова [25].

Пусть (X,B) — некоторое измеримое пространство; (F,B) — измеримая функция $\xi=\xi(\omega)$ на пространстве элементарных событий (Ω,F,P) (где P — вероятностная мера на σ-алгебре F-измеримых подмножеств Ω, называемых событиями) со значениями в (X,B) называется случайной величиной (чаще этот математический объект называют случайным элементом, оставляя термин «случайная величина» за частным случаем, когда X — числовая прямая — А. О.) в фазовом пространстве (X,B). Распределением вероятностей этой случайной величины ξ называется функция $P_\xi=P_\xi(B)$ на σ-алгебре B фазового пространства, определенная как:

$$P_\xi=P\{\xi\in B\}\quad (B\in\mathrm{B}) \tag{7}$$

(распределение вероятностей P_ξ представляет собой вероятностную меру в фазовом пространстве (X,B)) [25, с. 132].

Пусть $\xi_1,\xi_2,...,\xi_n$ — случайные величины на пространстве случайных событий (Ω,F,P) в соответствующих фазовых пространствах $(\mathrm{X}_k,\mathrm{B}_k)$. Совместным распределением вероятностей этих величин называется функция $P_{\xi_1,\xi_2,...,\xi_n}=P_{\xi_1,\xi_2,...,\xi_n}(B_1,B_2,...,B_n)$, определенная на множествах $B_1\in\mathrm{B}_1$, $B_2\in\mathrm{B}_2$, ..., $B_n\in\mathrm{B}_n$ как

$$P_{\xi_1,\xi_2,...,\xi_n}(B_1,B_2,...,B_n)=P_{\xi_1,\xi_2,...,\xi_n}(\xi_1\in B_1,\xi_2\in B_2,...,\xi_n\in B_n). \tag{8}$$

Распределение вероятностей $P_{\xi_1,\xi_2,...,\xi_n}$ как функция на полукольце множеств вида $B_1\times B_2\times...\times B_n$, $B_1\in\mathrm{B}_1,B_2\in\mathrm{B}_2,...,B_n\in\mathrm{B}_n$, в произведении пространств $\mathrm{X}_1,\mathrm{X}_2,...,\mathrm{X}_n$ представляет собой функцию распределения. Случайные величины $\xi_1,\xi_2,...,\xi_n$ называются независимыми, если при любых B_1, B_2, ..., B_n (см. [25, с. 133]):

$$P_{\xi_1,\xi_2,...,\xi_n}(B_1,B_2,...,B_n)=P_{\xi_1}(B_1)P_{\xi_2}(B_2)...P_{\xi_n}(B_n). \tag{9}$$

Предположим, что совместное распределение вероятностей $P_{\xi,\eta}(A,B)$ случайных величин ξ и η абсолютно непрерывно относительно некоторой меры Q на произведении пространств $\mathrm{X}\times\mathrm{Y}$, являющейся произведением мер Q_X и Q_Y, т.е.

$$P_{\xi,\eta}(A,B)=\int\limits_{A\times B}p(x,y)Q(dx,dy) \tag{10}$$

для любых $A\in\mathrm{A}$ и $B\in\mathrm{B}$, где $p(x,y)$ — соответствующая плотность распределения вероятностей [25, с. 145].

В формуле (10) предполагается, что $\xi=\xi(\omega)$ и $\eta=\eta(\omega)$ — случайные величины на одном и том же пространстве элементарных событий Ω со значениями в фазовых пространствах (X,A) и (Y,B). Существование плотности $p(x,y)$ вытекает из абсолютной непрерывности $P_{\xi,\eta}(A,B)$ относительно Q в соответствии с теоремой Радона — Никодима.

Условное распределение вероятностей $P_\xi(A \mid \eta), A \in \mathrm{A}$, может быть выбрано одинаковым для всех $\omega \in \Omega$, при которых случайная величина $\eta = \eta(\omega)$ сохраняет одно и то же значение: $\eta(\omega) = y$. При почти каждом $y \in \mathrm{Y}$ (относительно распределения P_η в фазовом пространстве (Y, B)) условное распределение вероятностей $P_\xi(A \mid y) = P_{\omega,\xi}(A)$, где $\omega \in \{\eta = y\}$ и $A \in \mathrm{A}$, будет абсолютно непрерывно относительно меры Q_X:

$$Q_\mathrm{X}(A) = \int\limits_{A \times \mathrm{X}} Q(dx, dy).$$

Причем соответствующая плотность условного распределения вероятностей будет иметь вид (см. [25, с. 145–146]):

$$p_\xi(x \mid y) = \frac{P_\xi(dx \mid y)}{Q_\mathrm{X}(dx)} = \frac{p(x,y)}{\int\limits_{\mathrm{X}} p(x,y) Q_\mathrm{X}(dx)}. \qquad (11)$$

При построении вероятностных моделей реальных явлений важны вероятностные пространства из конечного числа элементарных событий. Для них перечисленные выше общие понятия становятся более прозрачными, в частности, снимаются вопросы измеримости (все подмножества конечного множества обычно считаются измеримыми). Вместо плотностей и условных плотностей рассматриваются вероятности и условные вероятности. Отметим, что вероятности можно рассматривать как плотности относительно меры, приписывающей каждому элементу пространства элементарных событий вес 1, т.е. считающей меры:

$$Q(A) = Card(A)$$

(мера каждого множества равна числу его элементов). В целом ясно, что определения основных понятий теории вероятностей в общей ситуации практически не отличаются от таковых в элементарных курсах, во всяком случае с идейной точки зрения.

За последние тридцать лет в прикладной статистике сформировалась новая область — нечисловая статистика, или статистика нечисловых данных, она же — статистика объектов нечисловой природы. К настоящему времени она развита не менее, чем ранее выделенные статистика случайных

величин, многомерный статистический анализ, статистика временных рядов и случайных процессов. Краткая сводка основных постановок и результатов прикладной статистики в пространствах нечисловой природы дана в настоящей книге.

Теория, построенная для результатов наблюдений, лежащих в пространствах общей природы, является центральным стержнем в нечисловой статистике. В ее рамках удалось разработать и изучить методы оценивания параметров и характеристик, проверки гипотез (в частности, с помощью статистик интегрального типа), параметрической и непараметрической регрессии (восстановления зависимостей), непараметрического оценивания плотности, дискриминантного и кластерного анализов и т.д.

Вероятностно-статистические методы, развитые для результатов наблюдений, принадлежащих пространствам произвольного вида, позволяют единообразно проводить анализ данных из любого конкретного пространства. Так, в монографии [2] они применены к конечным случайным множествам, в работе [24] — к нечетким множествам. С их помощью установлено поведение обобщенного мнения экспертной комиссии (медианы Кемени) при увеличении числа экспертов, когда ответы экспертов лежат в том или ином пространстве бинарных отношений. Методы классификации могут быть основаны на непараметрических оценках плотности распределения вероятностей в пространстве общей природы. Такие методы были применены для медицинской диагностики в пространстве разнотипных данных, когда часть координат вектора измерена по количественным шкалам, а часть — по качественным и т.д.

1.5. НЕЧЕТКИЕ МНОЖЕСТВА — ЧАСТНЫЙ СЛУЧАЙ НЕЧИСЛОВЫХ ДАННЫХ

Уже много раз упоминались нечеткие множества как практически важный вид объектов нечисловой природы. Что же это такое? Познакомимся с основами теории нечетких множеств.

Нечеткие множества. Пусть A — некоторое множество. Подмножество B множества A характеризуется своей характеристической функцией:

$$\mu_B(x) = \begin{cases} 1, & x \in B, \\ 0, & x \notin B. \end{cases} \qquad (1)$$

Что такое нечеткое множество? Обычно говорят, что нечеткое подмножество C множества A характеризуется своей функцией принадлежности $\mu_C : A \to [0;1]$. Значение функции принадлежности в точке x показывает степень принадлежности этой точки нечеткому множеству. Нечеткое множество описывает неопределенность, соответствующую точке x — она одновременно и входит, и не входит в нечеткое множество C. За вхождение — $\mu_C(x)$ шансов, за второе, т.е. за то, что точка не входит в множество, $(1 - \mu_C(x))$ шансов.

Если функция принадлежности $\mu_C(x)$ имеет вид (1) при некотором B, то C есть обычное (четкое) подмножество A. Таким образом, теория нечетких множество является более общей или хотя бы не менее общей математической дисциплиной, чем обычная теория множеств, поскольку обычные множества – частный случай нечетких. Соответственно можно ожидать, что теория нечеткости как целое обобщает классическую математику. Однако позже мы увидим, что теория нечеткости в определенном смысле сводится к теории случайных множеств и тем самым является частью классической математики. Другими словами, по степени общности обычная математика и нечеткая математика эквивалентны. Однако для практического применения, например, в теории принятия решений описание и анализ неопределенностей с помощью теории нечетких множеств весьма плодотворны.

Обычное подмножество можно было бы отождествить с его характеристической функцией. Этого математики не делают, поскольку для задания функции (в ныне принятом подходе) необходимо сначала задать множество. Нечеткое же подмножество с формальной точки зрения можно отождествить с его функцией принадлежности. Однако термин «нечеткое подмножество» предпочтительнее при построении математических моделей реальных явлений.

Теория нечеткости является обобщением интервальной математики. Действительно, функция принадлежности:

$$\mu_B(x) = \begin{cases} 1, & x \in [a;b], \\ 0, & x \notin [a;b] \end{cases}$$

задает интервальную неопределенность — про рассматриваемую величину известно лишь, что она лежит в заданном интервале $[a, b]$. Тем самым описание неопределенностей с помощью нечетких множеств является более общим, чем с помощью интервалов.

Начало современной теории нечеткости положено работой 1965 г. американского ученого азербайджанского происхождения Л. А. Заде. К настоящему времени по этой теории опубликованы тысячи книг и статей, издается несколько международных журналов, выполнено достаточно много как теоретических, так и прикладных работ. Первая книга российского автора по теории нечеткости вышла в 1980 г. [24].

Л. А. Заде рассматривал теорию нечетких множеств как аппарат анализа и моделирования гуманистических систем, т.е. систем, в которых участвует человек. Его подход опирается на предпосылку о том, что элементами мышления человека являются не числа, а элементы некоторых нечетких множеств или классов объектов, для которых переход от «принадлежности» к «непринадлежности» не скачкообразен, а непрерывен. В настоящее время методы теории нечеткости используются почти во всех прикладных областях, в том числе при управлении предприятиями, качеством продукции и технологическими процессами, при описании предпочтений потребителей и оптимизации процессов варки стали.

Л. А. Заде использовал термин «fuzzy set» (нечеткое множество). На русский язык термин «fuzzy» переводили как нечеткий, размытый, расплывчатый, и даже как пушистый и туманный.

Аппарат теории нечеткости громоздок. В качестве примера дадим определения теоретико-множественных операций над нечеткими множествами. Пусть C и D — два нечетких подмножества A с функциями принадлежности $\mu_C(x)$ и $\mu_D(x)$соответственно. Пересечением $C \cap D$, произведением CD, объединением $C \cup D$, отрицанием $\overline{C}$, суммой $C + D$ называются нечеткие подмножества A с функциями принадлежности:

$$\mu_{C\cap D}(x) = \min(\mu_C(x), \mu_D(x)), \quad \mu_{CD}(x) = \mu_C(x)\mu_D(x), \quad \mu_{\overline{C}}(x) = 1 - \mu_C(x),$$

$$\mu_{C\cup D}(x) = \max(\mu_C(x), \mu_D(x)), \quad \mu_{C+D}(x) = \mu_C(x) + \mu_D(x) - \mu_C(x)\mu_D(x), \quad x \in A,$$

соответственно.

Как уже отмечалось, теория нечетких множеств в определенном смысле сводится к теории вероятностей, а именно, к теории случайных множеств. Соответствующий цикл теорем приведен в следующем разделе. Однако при решении прикладных задач вероятностно-статистические методы и методы теории нечеткости обычно рассматриваются как различные.

Для знакомства со спецификой нечетких множеств изучим некоторые их свойства.

В дальнейшем считаем, что все рассматриваемые нечеткие множества являются подмножествами одного и того же множества Y.

Законы де Моргана для нечетких множеств. Как известно, законами же Моргана называются следующие тождества алгебры множеств:

$$\overline{A \cup B} = \overline{A} \cap \overline{B}, \quad \overline{A \cap B} = \overline{A} \cup \overline{B}. \tag{2}$$

Теорема 1. Для нечетких множеств справедливы тождества:

$$\overline{A \cup B} = \overline{A} \cap \overline{B}, \quad \overline{A \cap B} = \overline{A} \cup \overline{B}, \tag{3}$$

$$\overline{A + B} = \overline{A}\,\overline{B}, \quad \overline{AB} = \overline{A} + \overline{B}. \tag{4}$$

Доказательство теоремы 1 состоит в непосредственной проверке справедливости соотношений (3) и (4) путем вычисления значений функций принадлежности участвующих в этих соотношениях нечетких множеств на основе определений, данных выше.

Тождества (3) и (4) назовем *законами де Моргана для нечетких множеств*. В отличие от классического случая соотношений (2), они состоят из четырех тождеств, одна пара которых относится к операциям объединения и пересечения, а вторая — к операциям произведения и суммы. Как и соотношение (2) в алгебре множеств, законы де Моргана в алгебре нечетких множеств позволяют преобразовывать выражения и формулы, в состав которых входят операции отрицания.

Дистрибутивный закон для нечетких множеств. Некоторые свойства операций над множествами не выполнены для нечетких множеств. Так, $A + A \neq A$, за исключением случая, когда A — «четкое» множество (т.е. функция принадлежности принимает только значения 0 и 1).

Верен ли дистрибутивный закон для нечетких множеств? В литературе иногда расплывчато утверждается, что «не всегда». Внесем полную ясность.

Теорема 2. Для любых нечетких множеств A, B и C:

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C). \tag{5}$$

В то же время равенство:

$$A(B + C) = AB + AC \tag{6}$$

справедливо тогда и только тогда, когда при всех $y \in Y$:

$$(\mu_A^2(y) - \mu_A(y))\mu_B(y)\mu_C(y) = 0.$$

Доказательство. Фиксируем произвольный элемент $y \in Y$. Для сокращения записи обозначим $a = \mu_A(y),\ b = \mu_B(y),\ c = \mu_C(y)$. Для доказательства тождества (5) необходимо показать, что

$$\min(a, \max(b, c)) = \max(\min(a, b), \min(a, c)). \qquad (7)$$

Рассмотрим различные упорядочения трех чисел *a, b, c*. Пусть сначала $a \le b \le c$. Тогда левая часть соотношения (7) есть $\min(a,c) = a$, а правая $\max(a,a) = a$, т.е. равенство (7) справедливо.

Пусть $b \le a \le c$. Тогда в соотношении (7) слева стоит $\min(a,c) = a$, а справа $\max(b,a) = a$, т.е. соотношение (7) опять является равенством.

Если $b \le c \le a$, то в соотношении (7) слева стоит $\min(a,c) = c$, а справа $\max(b,c) = c$, т.е. обе части снова совпадают.

Три остальные упорядочения чисел *a, b, c* разбирать нет необходимости, поскольку в соотношение (6) числа *b* и *c* входят симметрично. Тождество (5) доказано.

Второе утверждение теоремы 2 вытекает из того, что в соответствии с определениями операций над нечеткими множествами:

$$\mu_{A(B+C)}(y) = a(b + c - bc) = ab + ac - abc$$

и

$$\mu_{AB+AC}(y) = ab + ac - (ab)(ac) = ab + ac - a^2bc.$$

Эти два выражения совпадают тогда и только тогда, когда $a^2bc = abc$, что и требовалось доказать.

Определение 1. Носителем нечеткого множества *A* называется совокупность всех точек $y \in Y$, для которых $\mu_A(y) > 0$.

Следствие теоремы 2. Если носители нечетких множеств *B* и *C* совпадают с *У,* то равенство (6) *имеет место тогда и только тогда, когда А* — «четкое» (т.е. обычное, классическое, не нечеткое) множество.

Доказательство. По условию $\mu_B(y)\mu_C(y) \ne 0$ при всех $y \in Y$. Тогда из теоремы 2 следует, что $\mu_A^2(y) - \mu_A(y) = 0$, т.е. $\mu_A(y)$=1 или $\mu_A(y) = 0$, что и означает, что *A* — четкое множество.

Пример описания неопределенности с помощью нечеткого множества. Понятие «богатый» часто используется при обсуждении социально-экономических проблем, в том числе и в связи с подготовкой и принятием решений. Однако очевидно, что разные лица вкладывают в это понятие раз-

личное содержание. Сотрудники Института высоких статистических технологий и эконометрики провели небольшое пилотное (т.е. пробное) социологическое исследование представления различных слоёв населения о понятии «богатый человек».

Мини-анкета опроса выглядела так:

1. При каком месячном доходе (в десятках тыс. руб. на одного человека) Вы считали бы себя богатым человеком?

2. Оценив свой сегодняшний доход, к какой из категорий Вы себя относите:

а) богатые;

б) достаток выше среднего;

в) достаток ниже среднего;

г) бедные;

д) за чертой бедности?

(В дальнейшем вместо полного наименования категорий будем оперировать буквами, например «в» — категория, «б» — категория и т.д.)

3. Ваша профессия, специальность.

Всего было опрошено 74 человека, из них 40 — научные работники и преподаватели, 34 человека — не занятых в сфере науки и образования, в том числе 5 рабочих и 5 пенсионеров. Из всех опрошенных только один (!) считает себя богатым. Несколько типичных ответов научных работников и преподавателей приведено в табл. 1, а аналогичные сведения для работников коммерческой сферы — в табл. 2.

Таблица 1

Типичные ответы научных работников и преподавателей

Ответы на вопрос 3	**Ответы на вопрос 1, десятки тыс. руб./чел.**	**Ответы на вопрос 2**	**Пол**
Кандидат наук	6	д	ж
Преподаватель	6	в	ж
Доцент	6	б	ж
Учитель	60	в	м
Старший научный сотрудник	60	д	м
Инженер-физик	140	д	ж
Программист	150	г	м
Научный работник	270	г	м

Таблица 2

Типичные ответы работников коммерческой сферы

Ответы на вопрос 3	Ответы на вопрос 1	Ответы на вопрос 2	Пол
Вице-президент банка	600	а	ж
Зам. директора банка	300	б	ж
Начальник кредитного отдела	300	б	м
Начальник отдела ценных бумаг	60	б	м
Главный бухгалтер	120	д	ж
Бухгалтер	90	в	ж
Менеджер банка	66	б	м
Начальник отдела проектирования	60	в	ж

Разброс ответов на первый вопрос — от 60 тыс. до 6 млн руб. в месяц на человека. Результаты опроса показывают, что критерий богатства у финансовых работников в целом несколько выше, чем у научных работников и преподавателей (см. гистограммы на рис. 1 и рис. 2 ниже).

Опрос показал, что выявить какое-нибудь конкретное значение суммы, которая необходима «для полного счастья», пусть даже с небольшим разбросом, нельзя, что вполне естественно. Как видно из таблиц 1 и 2, денежный эквивалент богатства колеблется от 60 тыс. до 6 млн руб. в месяц. Подтвердилось мнение, что работники сферы образования в подавляющем большинстве причисляют свой достаток к категории «в» и ниже (81 % опрошенных), в том числе к категории «д» отнесли свой достаток 57 %.

Со служащими коммерческих структур и бюджетных организаций иная картина: «г» — категория 1 человек (4 %), «д» — категория 4 человека (17 %), «б» — категория — 46 % и 1 человек «а» — категория.

Пенсионеры, что не вызывает удивления, отнесли свой доход к категории «д» (4 человека), и лишь один человек указал «г» — категорию. Рабочие же ответили так: 4 человека — «в», и один человек — «б».

Для представления общей картины в табл. 3 приведены данные об ответах работников других профессий.

Таблица 3

Типичные ответы работников различных профессий

Ответы на вопрос 3	Ответы на вопрос 1	Ответы на вопрос 2	Пол
Работник торговли	6	б	ж
Дворник	12	в	ж
Водитель	60	в	м
Военнослужащий	60	в	м
Владелец бензоколонки	120	б	ж
Пенсионер	36	д	ж
Начальник фабрики	120	б	м
Хирург	30	в	м
Домохозяйка	60	в	ж
Слесарь-механик	150	в	м
Юрист	60	б	м
Оператор ЭВМ	120	д	м
Работник собеса	18	д	ж
Архитектор	150	б	ж

Прослеживается интересное явление: чем выше планка богатства для человека, тем к более низкой категории относительно этой планки он себя относит.

Для сводки данных естественно использовать гистограммы. Для этого необходимо сгруппировать ответы. Использовались 7 классов (интервалов):

1) до 300 тыс. руб. в месяц на человека (включительно);

2) от 300 до 600 тыс. руб.;

3) от 600 до 900 тыс. руб.;

4) от 900 до 1 200 тыс. руб.;

5) от 1 200 до 1 500 тыс. руб.;

6) от 1 500 до 1 800 тыс. руб.;

7) более 1 800 тыс. руб.

(Во всех интервалах левая граница исключена, а правая, наоборот — включена.)

Сводная информация представлена на рис. 1 (для научных работников и преподавателей) и рис. 2 (для всех остальных, т.е. для лиц, не занятых в сфере науки и образования — служащих иных бюджетных организаций, коммерческих структур, рабочих, пенсионеров).

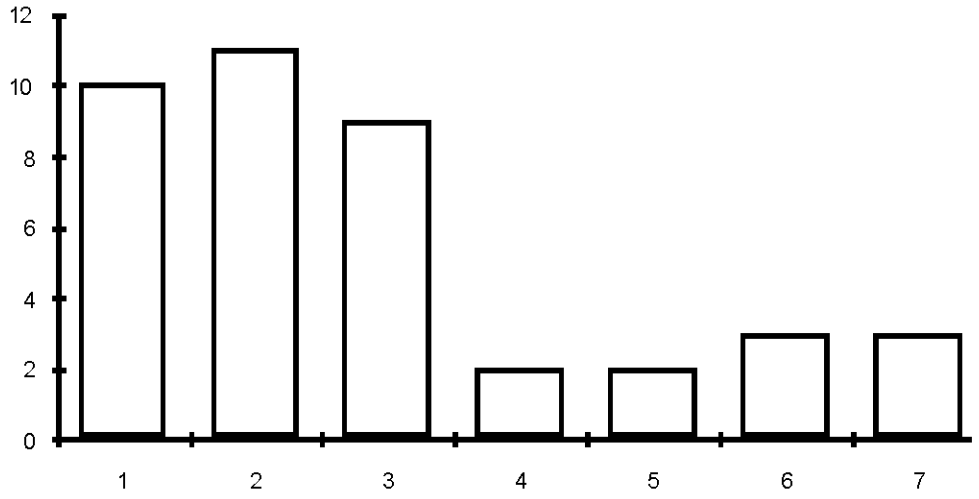


Рис. 1. Гистограмма ответов на вопрос 1
для научных работников и преподавателей (40 человек)

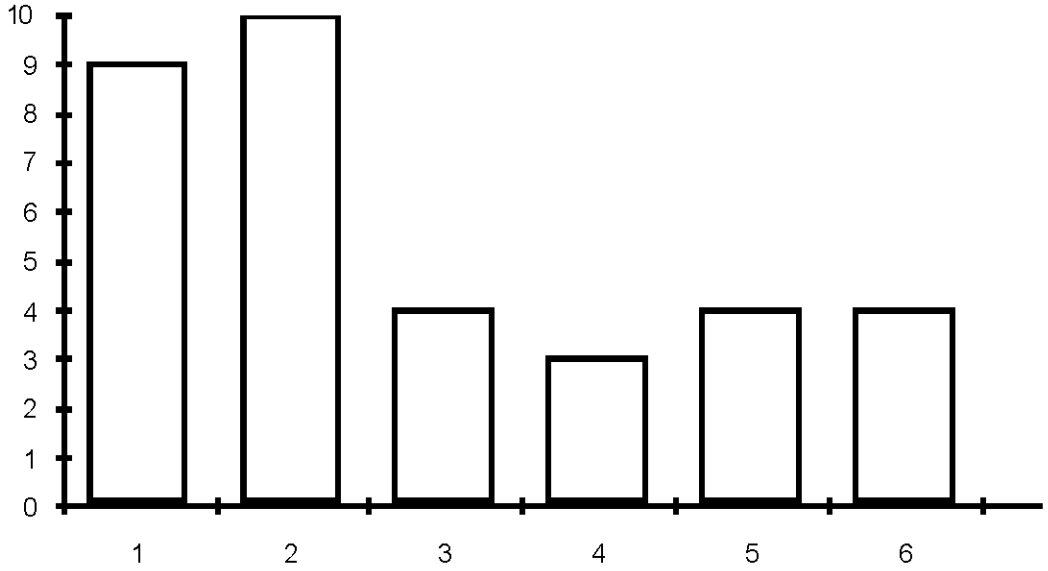


Рис. 2. Гистограмма ответов на вопрос 1 для лиц,
не занятых в сфере науки и образования (34 человека).

Для двух выделенных групп, а также для некоторых подгрупп второй группы рассчитаны сводные средние характеристики – выборочные средние арифметические, медианы, моды. При этом медиана группы — количество тыс. руб., названное центральным по порядковому номеру опрашиваемым в возрастающем ряду ответов на вопрос 1, а мода группы — интервал, на котором столбик гистограммы — самый высокий, т.е. в него «попало» максимальное количество опрашиваемых. Результаты приведены в табл. 4.

Таблица 4

Сводные средние характеристики ответов на вопрос 1 для различных групп (в десятках тыс. руб. в мес. на чел.)

Группа опрошенных	Среднее арифметическое	Медиана	Мода
Научные работники и преподаватели	70,0	43,5	(30; 60)
Лица, не занятые в сфере науки и образования	86,4	120,0	(30; 60)
Служащие коммерческих структур и бюджетных организаций	107,5	60,0	(30; 60)
Рабочие	90,0	78,0	–
Пенсионеры	61,8	60,0	–

Построим нечеткое множество, описывающее понятие «богатый человек» в соответствии с представлениями опрошенных. Для этого составим табл. 5 на основе рис. 1 и 2 с учетом размаха ответов на первый вопрос.

Таблица 5

Число ответов, попавших в интервалы

№	Номер интервала	0	1	2	3	4
1	Интервал, десятков тыс. руб. в месяц	(0; 6)	[6; 30]	(30; 60]	(60; 90]	(90; 120]
2	Число ответов в интервале	0	19	21	13	5
3	Доля ответов в интервале	0	0,257	0,284	0,176	0,068
4	Накопленное число ответов	0	19	40	53	58
5	Накопленная доля ответов	0	0,257	0,541	0,716	0,784

Продолжение табл. 5

№	Номер интервала	5	6	7	8
1	Интервал, десятков тыс. руб. в месяц	(120; 150]	(150; 180]	(180; 600)	[600; +∞)
2	Число ответов в интервале	6	7	2	1
3	Доля ответов в интервале	0,081	0,095	0,027	0,013
4	Накопленное число ответов	64	71	73	74
5	Накопленная доля ответов	0,865	0,960	0,987	1,000

Пятая строка табл. 5 задает функцию принадлежности нечеткого множества, выражающего понятие «богатый человек» в терминах его ежемесячного дохода. Это нечеткое множество является подмножеством множества из

9 интервалов, заданных в строке 2 табл. 5. Или множества из 9 условных номеров {0, 1, 2, …, 8}. Эмпирическая функция распределения, построенная по выборке из ответов 74 опрошенных на первый вопрос мини-анкеты, описывает понятие «богатый человек» как нечеткое подмножество положительной полуоси.

О разработке методики ценообразования на основе теории нечетких множеств. Для оценки значений показателей, не имеющих количественного выражения, можно использовать методы нечетких множеств. Например, в диссертации П. В. Битюкова [26] нечеткие множества применялись при моделировании задач ценообразования на электронные обучающие курсы, используемые при дистанционном обучении. Им было проведено исследование значений фактора «Уровень качества курса» с использованием нечетких множеств. В ходе практического использования предложенной П. В. Битюковым методики ценообразования значения ряда других факторов могут также определяться с использованием теории нечетких множеств. Например, ее можно использовать для расчета прогноза рейтинга специальности в вузе с помощью экспертов, а также значений других факторов, относящихся к группе «Особенности курса». Опишем подход П. В. Битюкова как пример практического использования теории нечетких множеств.

Значение оценки, присваиваемой каждому интервалу для фактора «Уровень качества курса», определяется на универсальной шкале [0; 1], где необходимо разместить значения лингвистической переменной «Уровень качества курса»: НИЗКИЙ, СРЕДНИЙ, ВЫСОКИЙ. Степень принадлежности некоторого значения вычисляется как отношение числа ответов, в которых оно встречалось в определенном интервале шкалы, к максимальному (для этого значения) числу ответов по всем интервалам.

Был проведен опрос экспертов о степени влияния уровня качества электронных курсов на их потребительную ценность. Каждому эксперту в процессе опроса предлагалось оценить с позиции потребителя ценность того или иного класса курсов в зависимости от уровня качества. Эксперты давали свою оценку для каждого класса курсов по 10-ти балльной шкале (где 1 — min, 10 — max). Для перехода к универсальной шкале [0; 1], все значения 10-ти балльной шкалы оценки ценности были разделены на максимальную оценку, т.е. на 10.

Используя свойства функции принадлежности, необходимо предварительно обработать данные с тем, чтобы уменьшить искажения, вносимые опросом. Естественными свойствами функций принадлежности являются наличие одного максимума и гладкие, затухающие до нуля фронты. Для об-

работки статистических данных можно воспользоваться так называемой матрицей подсказок. Предварительно удаляются явно ошибочные элементы. Критерием удаления служит наличие нескольких нулей в строке вокруг этого элемента.

Элементы матрицы подсказок вычисляются с использованием величин:

$$k_j = \sum_{i=1}^{n} b_{ij},\ j = \overline{1,n},$$

где b_{ij} — элемент таблицы с результатами анкетирования, сгруппированными по интервалам. Выбирается максимальный элемент: $k_{\max} = \max_j k_j$, и далее все элементы матрицы при $k_j \neq 0$ преобразуются по формуле:

$$c_{ij} = \frac{b_{ij} k_{\max}}{k_j},\ i = \overline{1,m},\ j = \overline{1,n}.$$

Для столбцов, где $k_j = 0$, применяется линейная аппроксимация:

$$c_{ij} = \frac{c_{ij-1} + c_{ij+1}}{2},\ i = \overline{1,m},\ j = \overline{1,n}.$$

Результаты расчетов сводятся в таблицу, на основании которой строятся функции принадлежности. Для этого находятся максимальные элементы по строкам:

$$c_{i\max} = \max_j c_{ij},\ \ i = 1,2,...,m,\ \ j = 1,2,...,n.$$

Функция принадлежности вычисляется по формуле: $\mu_{ij} = c_{ij} / c_{i\max}$. Результаты расчетов приведены в табл. 6.

Таблица 6

Значения функции принадлежности лингвистической переменной

μ_i	Интервал на универсальной шкале									
	0,1	**0,2**	**0,3**	**0,4**	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**	**1,0**
μ_1	0	0,2	1	1	0,89	0,67	0	0	0	0
μ_2	0	0	0	0	0	0,33	1	1	0	0
μ_3	0	0	0	0	0	0	0	0	1	1

На рис. 3 сплошными линиями показаны функции принадлежности значений лингвистической переменной «Уровень качества курса» после обработки таблицы, содержащей результаты опроса. Как видно из графика,

функции принадлежности удовлетворяют описанным выше свойствам. Для сравнения пунктирной линией показана функция принадлежности лингвистической переменной для значения НИЗКИЙ без обработки данных.

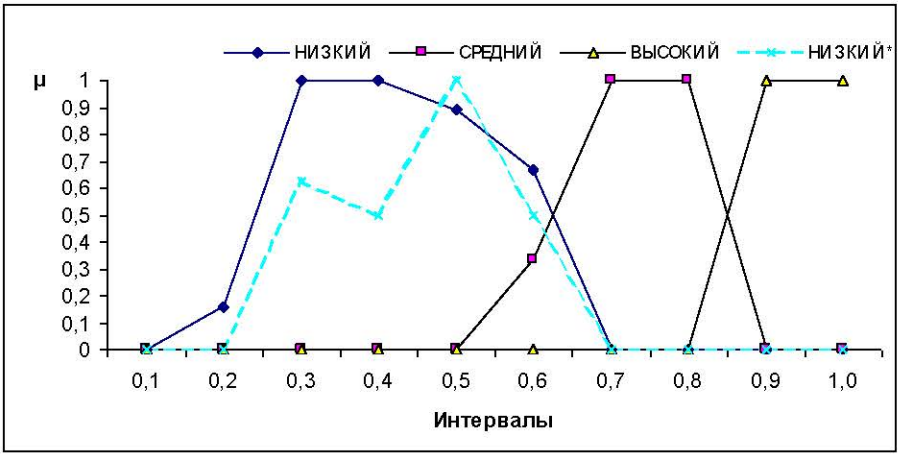


Рис. 3. График функций принадлежности значений лингвистической переменной «Уровень качества курса»

1.6. СВЕДЕНИЕ НЕЧЕТКИХ МНОЖЕСТВ К СЛУЧАЙНЫМ

Нечеткость и случайность. С самого начала появления современной теории нечеткости в 1960-е гг. (см. выше) началось обсуждение ее взаимоотношений с теорией вероятностей. Дело в том, что функция принадлежности нечеткого множества напоминает распределение вероятностей. Отличие только в том, что сумма вероятностей по всем возможным значениям случайной величины (или интеграл, если множество возможных значений несчетно) всегда равна 1, а сумма S значений функции принадлежности (в непрерывном случае — интеграл от функции принадлежности) может быть любым неотрицательным числом. Возникает искушение пронормировать функцию принадлежности, т.е. разделить все ее значения на S (при $S \neq 0$), чтобы свести ее к распределению вероятностей (или к плотности вероятности). Однако специалисты по нечеткости справедливо возражают против такого «примитивного» сведения, поскольку оно проводится отдельно для каждой размытости (нечеткого множества), и определения обычных операций над нечеткими множествами согласовать с ним нельзя. Последнее утверждение означает следующее. Пусть указанным образом преобразованы функции принадлежности нечетких множеств A и B. Как при этом преобразуются функции принадлежности $A \cap B, A \cup B, A + B, AB$? Установить это *невоз-*

можно в принципе. Последнее утверждение становится совершенно ясным после рассмотрения нескольких примеров пар нечетких множеств с одними и теми же суммами значений функций принадлежности, но различными результатами теоретико-множественных операций над ними. Причем и суммы значений соответствующих функций принадлежности для этих результатов теоретико-множественных операций, например, для пересечений множеств, также различны.

В работах по нечетким множествам время от времени утверждается, что теория нечеткости является самостоятельным разделом прикладной математики и не имеет отношения к теории вероятностей (см., например, обзор литературы в монографиях [2, 24]). Некоторые авторы, обсуждавшие взаимоотношения теории нечеткости и теории вероятностей, подчеркивали различие между этими областями теоретических и прикладных исследований. Обычно сопоставляют аксиоматику и сравнивают области приложений. Надо сразу отметить, что аргументы при втором типе сравнений не имеют доказательной силы, поскольку по поводу границ применимости даже такой давно выделившейся научной области, как вероятностно-статистические методы, имеются различные мнения. Более того, нет единства мнений об арифметике. Напомним, что итог рассуждений одного из наиболее известных французских математиков Анри Лебега по поводу границ применимости арифметики таков: «Арифметика применима тогда, когда она применима» (см. его монографию [27, с. 21–22]).

При сравнении различных аксиоматик теории нечеткости и теории вероятностей нетрудно увидеть, что списки аксиом различаются. Из этого, однако, отнюдь не следует, что между указанными теориями нельзя установить связь, типа известного сведения евклидовой геометрии на плоскости к арифметике (точнее к теории числовой системы R^2 — см., например, монографию [28]). Напомним, что эти две аксиоматики — евклидовой геометрии и арифметики — на первый взгляд весьма сильно различаются.

Можно понять желание энтузиастов теории нечеткости подчеркнуть принципиальную новизну своего научного аппарата. Однако не менее важно установить связи этого подхода с ранее известными.

Проекция случайного множества. Как оказалось, теория нечетких множеств тесно связана с теорией случайных множеств. Еще в 1975 г. в работе [29] было показано, что нечеткие множества естественно рассматривать как «проекции» случайных множеств. Рассмотрим этот метод сведения теории нечетких множеств к теории случайных множеств.

Определение 1. Пусть $A = A(\omega)$ — случайное подмножество конечного множества U. Нечеткое множество B, определенное на U, называется проекцией A и обозначается *Proj A*, если:

$$\mu_B(y) = P(y \in A) \tag{1}$$

при всех $y \in Y$.

Очевидно, каждому случайному множеству A можно поставить в соответствие с помощью формулы (1) нечеткое множество $B = Proj\ A$. Оказывается, верно и обратное.

Теорема 1. Для любого нечеткого подмножества B конечного множества U существует случайное подмножество A множества U такое, что $B = Proj\ A$.

Доказательство. Достаточно задать распределение случайного множества A. Пусть U_1 — носитель B (см. определение 1 в разделе 1.5 выше). Без ограничения общности можно считать, что $Y_1 = \{y_1, y_2, ..., y_m\}$ при некотором m и элементы U_1 занумерованы в таком порядке, что:

$$0 < \mu_B(y_1) \le \mu_B(y_2) \le ... \le \mu_B(y_m).$$

Введем множества:

$$Y(1) = Y_1, Y(2) = \{y_2, ..., y_m\}, ..., Y(t) = \{y_t, ..., y_m\}, ..., Y(m) = \{y_m\}.$$

Положим:

$$P(A = Y(1)) = \mu_B(y_1), \quad P(A = Y(2)) = \mu_B(y_2) - \mu_B(y_1), ...,$$
$$P(A = Y(t)) = \mu_B(y_t) - \mu_B(y_{t-1}), ..., P(A = Y(m)) = \mu_B(y_m) - \mu_B(y_{m-1}),$$
$$P(A = \varnothing) = 1 - \mu_B(y_m).$$

Для всех остальных подмножеств X множества U положим $P(A = X) = 0$. Поскольку элемент y_t входит во множества $Y(1)$, $Y(2)$, ..., $Y(t)$ и не входит во множества $Y(t + 1)$, ..., $Y(m)$, то из приведенных выше формул следует, что $P(y_t \in A) = \mu_B(y_t)$. Если $y \notin Y_1$, то, очевидно, $P(y \in A) = 0$. Теорема 1 доказана.

Распределение случайного множества с независимыми элементами, как показано выше (см. также главу 8 монографии [3]), полностью определяется

его проекцией. Для конечного случайного множества общего вида это не так. Для уточнения сказанного понадобится следующая теорема.

Теорема 2. Для случайного подмножества A множества Y из конечного числа элементов наборы чисел $P(A = X), X \subseteq Y$, и $P(X \subseteq A), X \subseteq Y$, выражаются один через другой.

Доказательство. Второй набор выражается через первый следующим образом:

$$P(X \subseteq A) = \sum_{X' : X \subseteq X'} P(A = X').$$

Элементы первого набора выразить через второй можно с помощью формулы включений и исключений из формальной логики, в соответствии с которой:

$$P(A = X) = P(X \subseteq A) - \sum P(X \cup \{y\} \subseteq A) + \sum P(X \cup \{y_1, y_2\} \subseteq A) - \ldots \pm P(Y \subseteq A).$$

В этой формуле в первой сумме y пробегает все элементы множества $Y \backslash X$, во второй сумме переменные суммирования y_1 и y_2 не совпадают и также пробегают это множество, и т.д. Ссылка на формулу включений и исключений завершает доказательство теоремы 2.

В соответствии с теоремой 2 случайное множество A можно характеризовать не только распределением, но и набором чисел $P(X \subseteq A), X \subseteq Y$. В этом наборе $P(\varnothing \subseteq A) = 1$, а других связей типа равенств нет. В этот набор входят числа $P(\{y\} \subseteq A) = P(y \in A)$, следовательно, фиксация проекции случайного множества эквивалентна фиксации $k = Card(Y)$ параметров из $(2^k - 1)$ параметров, задающих распределение случайного множества A в общем случае.

При обосновании возможности сведения теории нечетких множеств к теории случайных множеств будет применяться следующая теорема.

Теорема 3. Если $Proj\ A = B$, то $Proj\ \overline{A} = \overline{B}$.

Для доказательства достаточно воспользоваться тождеством из теории случайных множеств $P(\overline{A} = X) = P(A = \overline{X})$, формулой для вероятности накрытия $P(y \in A)$, определением отрицания нечеткого множества и тем, что сумма всех $P(A = X)$ равна 1. При этом под формулой для вероятности накрытия имеется в виду следующее утверждение: чтобы найти вероятность накрытия

фиксированного элемента q случайным подмножеством S конечного множества Q, достаточно вычислить:

$$P(q \in S) = \mathrm{P}(\{\omega : q \in S(\omega)\}) = \sum_{A: q \in A, A \subseteq 2^Q} P(S = A),$$

где суммирование идет по всем подмножествам A множества Q, содержащим q.

Пересечения и произведения нечетких и случайных множеств. Выясним, как операции над случайными множествами соотносятся с операциями над их проекциями. В силу законов де Моргана (теорема 1 в разделе 1.5) и теоремы 3 достаточно рассмотреть операцию пересечения случайных множеств.

Теорема 4. Если случайные подмножества A_1 и A_2 конечного множества Y независимы, то нечеткое множество $Proj\ (A_1 \bigcap A_2)$ является произведением нечетких множеств $Proj\ A_1$ и $Proj\ A_2$.

Доказательство. Надо показать, что для любого $y \in Y$:

$$P(y \in A_1 \cap A_2) = P(y \in A_1) P(y \in A_2). \tag{2}$$

По формуле для вероятности накрытия точки случайным множеством (см. выше):

$$P(y \in A_1 \cap A_2) = \sum_{X: y \in X} P((A_1 \cap A_2) = X). \tag{3}$$

Легко проверить, что распределение пересечения случайных множеств $A_1 \cap A_2$ можно выразить через их совместное распределение следующим образом:

$$P(A_1 \cap A_2 = X) = \sum_{X_1, X_2 : X_1 \cap X_2 = X} P(A_1 = X_1, A_2 = X_2). \tag{4}$$

Из соотношений (3) и (4) следует, что вероятность накрытия для пересечения случайных множеств можно представить в виде двойной суммы:

$$P(y \in A_1 \cap A_2) = \sum_{X: y \in X} \sum_{X_1, X_2 : X_1 \cap X_2 = X} P(A_1 = X_1, A_2 = X_2). \tag{5}$$

Заметим теперь, что правую часть формулы (5) можно переписать следующим образом:

$$\sum_{X_1, X_2 : y \in X_1, y \in X_2} P(A_1 = X_1,\ A_2 = X_2). \qquad (6)$$

Действительно, формула (5) отличается от формулы (6) лишь тем, что в ней сгруппированы члены, в которых пересечение переменных суммирования $X_1 \cap X_2$ принимает постоянное значение. Воспользовавшись определением независимости случайных множеств и правилом перемножения сумм, получаем, что из (5) и (6) вытекает равенство:

$$P(y \in A_1 \cap A_2) = \left(\sum_{X_1 : y \in X_1} P(A_1 = X_1) \right) \left(\sum_{X_2 : y \in X_2} P(A_2 = X_2) \right).$$

Для завершения доказательства теоремы 4 достаточно еще раз сослаться на формулу для вероятности накрытия точки случайным множеством.

Определение 2. Носителем случайного множества C называется совокупность всех тех элементов $y \in Y,$ для которых $P(y \in C) > 0.$

Теорема 5. Равенство:

$$Proj\,(A_1 \bigcap A_2) = Proj\,(A_1) \bigcap Proj\,(A_2)$$

верно тогда и только тогда, когда пересечение носителей случайных множеств $\overline{A_1} \cap A_2$ и $A_1 \cap \overline{A_2}$ пусто.

Доказательство. Необходимо выяснить условия, при которых:

$$P(y \in A_1 \cap A_2) = \min(P(y \in A_1), P(y \in A_2)). \qquad (7)$$

Положим:

$$p_1 = P(y \in A_1 \cap A_2),\ p_2 = P(y \in \overline{A_1} \cap A_2),\ p_3 = P(y \in A_1 \cap \overline{A_2}).$$

Тогда равенство (7) сводится к условию:

$$p_1 = \min(p_1 + p_2, p_1 + p_3). \qquad (8)$$

Ясно, что соотношение (8) выполнено тогда и только тогда, когда $p_2 p_3 = 0$ при всех $y \in Y$, т.е. не существует ни одного элемента $y_0 \in Y$ такого, что одновременно $P(y_0 \in \overline{A}_1 \cap A_2) > 0$ и $P(y_0 \in A_1 \cap \overline{A_2}) > 0$, а это эквивалентно пустоте пересечения носителей случайных множеств $\overline{A_1} \cap A_2$ и $A_1 \cap \overline{A_2}$. Теорема 5 доказана.

Сведение последовательности операций над нечеткими множествами к последовательности операций над случайными множествами. Выше получены некоторые связи между нечеткими и случайными множествами. Стоит отметить, что изучение этих связей в работе [29] началось с введения случайных множеств с целью развития и обобщения аппарата нечетких множеств Л. Заде. Дело в том, что математический аппарат нечетких множеств не позволяет в должной мере учитывать различные варианты зависимости между понятиями (объектами), моделируемыми с его помощью, т.е. не является достаточно гибким. Так, для описания «общей части» двух нечетких множеств есть лишь две операции — произведение и пересечение. Если применяется первая из них, то фактически предполагается, что множества ведут себя как проекции независимых случайных множеств (см. выше теорему 4). Операция пересечения также накладывает вполне определенные ограничения на вид зависимости между множествами (см. выше теорему 5), причем в этом случае найдены даже необходимые и достаточные условия. Желательно иметь более широкие возможности для моделирования зависимости между множествами (понятиями, объектами). Использование математического аппарата случайных множеств предоставляет такие возможности.

Цель сведения теории нечетких множеств к теории случайных множеств состоит в том, чтобы за любой конструкцией из нечетких множеств увидеть конструкцию из случайных множеств, определяющую свойства первой, аналогично тому, как за плотностью распределения вероятностей мы видим случайную величину. Рассмотрим результаты по сведению алгебры нечетких множеств к алгебре случайных множеств.

Определение 3. Вероятностное пространство $\{\Omega, G, P\}$ назовем делимым, если для любого измеримого множества $X \in G$ и любого положительного числа α, меньшего $P(X)$, можно указать измеримое множество $Y \subset X$ такое, что $P(Y) = \alpha$.

Пример. Пусть Ω — единичный куб конечномерного линейного пространства, G есть сигма-алгебра борелевских множеств, а P — мера Лебега. Тогда $\{\Omega, G, P\}$ — делимое вероятностное пространство.

Таким образом, делимое вероятностное пространство — это не экзотика. Обычный куб является примером такого пространства.

Доказательство сформулированного в примере утверждения проводится стандартными математическими приемами. Они основаны на том, что измеримое множество можно сколь угодно точно приблизить открытыми множествами, последние представляются в виде суммы не более чем счетного числа открытых шаров, а для шаров делимость проверяется непосредственно (от шара X тело объема $\alpha < P(X)$ отделяется соответствующей плоскостью).

Теорема 6. Пусть даны случайное множество A на делимом вероятностном пространстве $\{\Omega, G, P\}$ со значениями во множестве всех подмножеств множества U из конечного числа элементов, и нечеткое множество D на U. Тогда существуют случайные множества C_1, C_2, C_3, C_4 на том же вероятностном пространстве такие, что:

$$\Pr oj(A \cap C_1) = B \cap D, \quad \Pr oj(A \cap C_2) = BD, \quad \Pr oj(A \cup C_3) = B \cup D,$$

$$\Pr oj(A \cup C_4) = B + D, \quad \Pr ojC_i = D, \ i = 1,2,3,4,$$

где $B = Proj\ A$.

Доказательство. В силу справедливости законов де Моргана для нечетких (см. теорему 1 в разделе 1.5 выше) и для случайных множеств, а также теоремы 3 выше (об отрицаниях) достаточно доказать существование случайных множеств C_1 и C_2.

Рассмотрим распределение вероятностей на множестве всех подмножеств множества U, соответствующее случайному множеству C такому, что $Proj\ C = D$ (оно существует в силу теоремы 1). Построим случайное множество C_2 с указанным распределением, независимое от A. Тогда $\Pr oj(A \cap C_2) = BD$ по теореме 4.

Перейдем к построению случайного множества C_1. По теореме 7 необходимо и достаточно определить случайное множество $C_1(\omega)$ так, чтобы $Proj\ C_1 = D$ и пересечение носителей случайных множеств $A \cap \overline{C_1}$ и $\overline{A} \cap C_1$ было пусто, т.е.

$$p_3 = P(y \in A \cap \overline{C_1}) = 0$$

для $y \in Y_1 = \{y : \mu_B(y) \le \mu_D(y)\}$ и

$$p_2 = P(y \in \overline{A} \cap C_1) = 0$$

для $y \in Y_2 = \{y : \mu_B(y) \ge \mu_D(y)\}$.

Построим $C_1(\omega)$, исходя из заданного случайного множества $A(\omega)$. Пусть $y_1 \in Y_2$. Исключим элемент y_1 из $A(\omega)$ для стольких элементарных событий ω, чтобы для полученного случайного множества $A_1(\omega)$ было справедливо равенство:

$$P(y_1 \in A_1) = \mu_D(y_1)$$

(именно здесь используется делимость вероятностного пространства, на котором задано случайное множество $A(\omega)$). Для $y \neq y_1$, очевидно:

$$P(y \in A_1) = P(y \in A).$$

Аналогичным образом последовательно исключаем y из $A(\omega)$ для всех $y \in Y_2$ и добавляем y в $A(\omega)$ для всех $y \in Y_1$, меняя на каждом шагу $P(y \in A_i)$ только для $y = y_i$ так, чтобы:

$$P(y_i \in A_i) = \mu_D(y_i)$$

(ясно, что при рассмотрении $y_i \in Y_1 \cap Y_2$ случайное множество $A_i(\omega)$ не меняется). Перебрав все элементы $У$, получим случайное множество $A_k(\omega) = C_1(\omega)$, для которого выполнено требуемое. Теорема 6 доказана.

Основной результат о сведении теории нечетких множеств к теории случайных множеств дается следующей теоремой.

Теорема 7. Пусть $B_1, B_2, B_3, ..., B_t$ — некоторые нечеткие подмножества множества $У$ из конечного числа элементов. Рассмотрим результаты последовательного выполнения теоретико-множественных операций:

$$B^m = ((...((B_1 \circ B_2) \circ B_3) \circ ...) \circ B_{m-1}) \circ B_m, \quad m = 1,2,...,t,$$

где $\circ$ — символ одной из следующих теоретико-множественных операций над нечеткими множествами: пересечение, произведение, объединение, сумма (на разных местах могут стоять разные символы). Тогда существуют случайные подмножества $A_1, A_2, A_3, ..., A_t$ того же множества $У$ такие, что:

$$\Pr oj A_i = B_i, \quad i = 1,2,...,t,$$

и, кроме того, результаты теоретико-множественных операций связаны аналогичными соотношениями:

$$\Pr oj\{((...((A_1 \otimes A_2) \otimes A_3) \otimes ...) \otimes A_{m-1}) \otimes A_m\} = B^m,\ m = 1,2,...,t,$$

где знак $\otimes$ означает, что на рассматриваемом месте стоит символ пересечения $\cap$ случайных множеств, если в определении B^m стоит символ пересечения или символ произведения нечетких множеств, и соответственно символ объединения $\cup$ случайных множеств, если в B^m стоит символ объединения или символ суммы нечетких множеств.

Комментарий. Поясним содержание теоремы. Например, если:

$$B^5 = (((B_1 + B_2) \cap B_3)B_4) \cup B_5,$$

то

$$(((A_1 \otimes A_2) \otimes A_3) \otimes A_4) \otimes A_5 = (((A_1 \cup A_2) \cap A_3) \cap A_4) \cup A_5.$$

Как совместить справедливость дистрибутивного закона для случайных множеств (вытекающего из его справедливости для обычных множеств) с теоремой 2 раздела 1.5 выше, в которой показано, что для нечетких множеств, вообще говоря, $(B_1 + B_2)B_3 \neq B_1B_3 + B_2B_3$? Дело в том, что хотя в соответствии с теоремой 7 для любых трех нечетких множеств B_1, B_2 и B_3 можно указать три случайных множества A_1, A_2 и A_3 такие, что:

$$\Pr oj(A_i) = B_i,\ i = 1,2,3,\ \Pr oj(A_1 \cup A_2) = B_1 + B_2,\ \Pr oj((A_1 \cup A_2) \cap A_3) = B^3,$$

где

$$B^3 = (B_1 + B_2)B_3,$$

но при этом, вообще говоря,

$$\Pr oj(A_1 \cap A_3) \neq B_1B_3$$

и, кроме случаев, указанных в теореме 2 раздела 1.5,

$$\Pr oj((A_1 \cup A_2) \cap A_3) \neq B_1B_3 + B_2B_3.$$

Доказательство теоремы 7 проводится методом математической индукции. При t=1 распределение случайного множества строится с помощью теоремы 1. Затем конструируется само случайное множество A_1, определенное на делимом вероятностном пространстве (нетрудно проверить, что на делимом вероятностном пространстве можно построить случайное подмножество конечного множества с любым заданным распределением именно в силу делимости пространства). Далее случайные множества $A_2, A_3, ..., A_t$ строим по индукции с помощью теоремы 6. Теорема 7 доказана.

Замечание. Проведенное доказательство теоремы 7 проходит и в случае, когда при определении B^m используются отрицания, точнее, кроме B^m ранее введенного вида используются также последовательности результатов теоретико-множественных операций, очередной шаг в которых имеет вид:

$$B_1^m = \overline{B^{m-1}} \circ B_m, \quad B_2^m = B^{m-1} \circ \overline{B_m}, \quad B_3^m = \overline{B^{m-1}} \circ \overline{B_m}.$$

А именно, сначала при помощи законов де Моргана (теорема 1 раздела 1.5 выше) проводится преобразование, в результате которого в последовательности B^m остаются только отрицания отдельных подмножеств из совокупности $B_1, B_2, B_3, ..., B_t$, а затем с помощью теоремы 3 вообще удается избавиться от отрицаний и вернуться к условиям теоремы 7.

Итак, в настоящем разделе описаны связи между такими объектами нечисловой природы, как нечеткие и случайные множества, установленные в нашей стране в первой половине 1970-х гг. Через несколько лет, а именно, в начале 1980-х гг., близкие подходы стали развиваться и за рубежом. Одна из работ [30] носит примечательное название «Нечеткие множества как классы эквивалентности случайных множеств».

В нечисловой статистике разработан ряд методов статистического анализа нечетких данных. В том числе методы классификации, регрессии, проверки гипотез о совпадении функций принадлежности по опытным данным и т.д. При этом оказались полезными общие подходы статистики объектов нечисловой природы (см. главу 2 ниже). Методологические и прикладные вопросы теории нечеткости обсуждались и в научно-популярной литературе (см., например, статью [31], которая представляет интерес и в XXI в.).

1.7. ДАННЫЕ И РАССТОЯНИЯ В ПРОСТРАНСТВАХ ПРОИЗВОЛЬНОЙ ПРИРОДЫ

Как показано выше, исходные статистические данные могут иметь разнообразную математическую природу, являться элементами разнообразных

пространств — конечномерных, функциональных, бинарных отношений, множеств, нечетких множеств и т.д. Следовательно, центральной частью нечисловой статистики (и прикладной статистики в целом) является статистика в пространствах произвольной природы. Эта область прикладной статистики сама по себе не используется при анализе конкретных данных. Это очевидно, поскольку конкретные данные всегда имеют вполне определенную природу. Однако общие подходы, методы, результаты статистики в пространствах произвольной природы представляют собой научный инструментарий, готовый для применения в каждой конкретной области.

Статистика в пространствах произвольной природы. Много ли общего у статистических методов анализа данных различной природы? На этот естественный вопрос можно сразу же однозначно ответить — да, очень много. Такой ответ будет постоянно подтверждаться и конкретизироваться на протяжении всего учебника. Несколько примеров приведем сразу же.

Прежде всего отметим, что понятия случайного события, вероятности, независимости событий и случайных величин являются общими для любых конечных вероятностных пространств и любых конечных областей значений случайных величин (см., например, [32]). Поскольку все реальные явления и процессы можно описывать с помощью математических объектов, являющихся элементами конечных множеств, сказанное выше означает, что конечных вероятностных пространств и дискретных случайных величин (точнее, величин, принимающих значения в конечном множестве) вполне достаточно для всех практических применений. Переход к непрерывным моделям реальных явлений и процессов оправдан только тогда, когда этот переход облегчает проведение рассуждений и выкладок. Например, находить определенные интегралы зачастую проще, чем вычислять значения сумм. Не могу не отметить, что приведенные соображения о взаимном соотнесении дискретных и непрерывных математических моделей автор услышал более 30 лет назад от академика А. Н. Колмогорова (ясно, что за конкретную формулировку несет ответственность автор настоящего учебника).

Основные проблемы прикладной статистики — описание данных, оценивание, проверка гипотез — также в своей существенной части могут быть рассмотрены в рамках статистики в пространствах произвольной природы. Например, для описания данных могут быть использованы эмпирические и теоретические средние, плотности вероятностей и их непараметрические оценки, регрессионные зависимости. Правда, для этого пространства произвольной природы должны быть снабжены соответствующим математическим

инструментарием — расстояниями (показателями близости, мерами различия) между элементами рассматриваемых пространств.

Популярный в настоящее время метод оценивания параметров распределений — метод максимального правдоподобия — не накладывает каких-либо ограничений на конкретный вид элементов выборки. Они могут лежать в пространстве произвольной природы. Математические условия касаются только свойств плотностей вероятности и их производных по параметрам. Аналогично положение с методом одношаговых оценок, идущим на смену методу максимального правдоподобия (см. главу 2). Асимптотику решений экстремальных статистических задач достаточно изучить для пространств произвольной природы, а затем применять в каждом конкретном случае [33], когда задачу прикладной статистики удается представить в оптимизационном виде. Общая теория проверки статистических гипотез также не требует конкретизации математической природы рассматриваемых элементов выборок. Это относится, например, к лемме Неймана-Пирсона или теории статистических решений. Более того, естественная область построения теории статистик интегрального типа — это пространства произвольной природы (см. главу 2).

Совершенно ясно, что в конкретных областях прикладной статистики накоплено большое число результатов, относящимся именно к этим областям. Особенно это касается областей, исследования в которых ведутся сотни лет, в частности, статистики случайных величин (одномерной статистики). Однако принципиально важно указать на «ядро» прикладной статистики — статистику в пространствах произвольной природы. Если постоянно «держать в уме» это ядро, то становится ясно, что, например, многие методы непараметрической оценки плотности вероятности или кластер-анализа, использующие только расстояния между объектами и элементами выборки, относятся именно к статистике объектов произвольной природы, а не к статистике случайных величин или многомерному статистическому анализу. Следовательно, и применяться они могут во всех областях прикладной статистики, а не только в тех, в которых «родились».

Расстояния (метрики). В пространствах произвольной природы нет операции сложения, поэтому статистические процедуры не могут быть основаны на использовании сумм. Поэтому используется другой математический инструментарий, использующий понятия типа расстояния.

Как известно, расстоянием в пространстве X называется числовая функция двух переменных $d(x, y)$, $x \in X$, $y \in X$, определенная на этом про-

странстве, т.е. в стандартных обозначениях d: $X^2 \to R^1$, где R^1 — прямая, т.е. множество всех действительных чисел. Эта функция должна удовлетворять трем условиям (иногда их называют аксиомами):

1) неотрицательности: $d(x, y) \geq 0$, причем $d(x, x) = 0$, для любых значений $x \in X, y \in X$;

2) симметричности: $d(x, y) = d(y, x)$ для любых $x \in X, y \in X$;

3) неравенства треугольника: $d(x, y) + d(y, z) \geq d(x, z)$ для любых значений $x \in X, y \in X, z \in X$.

Для термина «расстояние» часто используются синонимы — «метрика» или «псевдометрика». Для метрики из $d(x, y) = 0$ следует $x = y$, для псевдометрики это не обязательно.

Пример 1. Если $d(x, x) = 0$ и $d(x, y) = 1$ при $x \neq y$ для любых значений $x \in X, y \in X$, то, как легко проверить, функция $d(x, y)$ — расстояние (метрика). Такое расстояние естественно использовать в пространстве X значений номинального признака: если два значения (например, названные двумя экспертами) совпадают, то расстояние равно 0, а если различны — то 1.

Пример 2. Расстояние, используемое в геометрии, очевидно, удовлетворяет трем приведенным выше аксиомам. Если X — это плоскость, а $x(1)$ и $x(2)$ — координаты точки $x \in X$ в некоторой прямоугольной системе координат, то эту точку естественно отождествить с двумерным вектором $(x(1), x(2))$. Тогда расстояние между точками $x = (x(1), x(2))$ и $y = (y(1), y(2))$ согласно известной формуле аналитической геометрии равно:

$$d(x, y) = \sqrt{(x(1) - y(1))^2 + (x(2) - y(2))^2}.$$

Пример 3. Евклидовым расстоянием в пространстве R^k векторов вида $x = (x(1), x(2), .., x(k))$ и $y = (y(1), y(2), ..., y(k))$ размерности k называется:

$$d(x, y) = \left(\sum_{j=1}^{k} (x(j) - y(j))^2 \right)^{1/2}.$$

В примере 2 рассмотрен частный случай примера 3 с $k = 2$.

Пример 4. В пространстве R^k векторов размерности k используют также так называемое «блочное расстояние», имеющее вид:

$$d(x, y) = \sum_{j=1}^{k} |x(j) - y(j)|.$$

Блочное расстояние соответствует передвижению по городу, разбитому на кварталы горизонтальными и вертикальными улицами. В результате можно передвигаться только параллельно одной из осей координат.

Пример 5. В пространстве функций, элементами которого являются функции $x = x(t)$, $y = y(t)$, $0 \le t \le 1$, часто используют расстояние Колмогорова:

$$d(x, y) = \sup_{0 \le t \le 1} |x(t) - y(t)|.$$

Пример 6. Пространство функций, элементами которого являются функции $x = x(t)$, $y = y(t)$, $0 \le t \le 1$, превращают в метрическое пространство (т.е. в пространство с метрикой), вводя расстояние:

$$d_p(x, y) = \left(\int_0^1 (x(t) - y(t))^p \, dt \right)^{1/p}.$$

Это пространство обычно обозначают L^p, где параметр $p \ge 1$ (при $p < 1$ не выполняются аксиомы метрического пространства, а именно, аксиома треугольника).

Пример 7. Рассмотрим пространство квадратных матриц порядка k. Как ввести расстояние между матрицами $A = \|a(i, j)\|$ и $B = \|b(i, j)\|$? Можно сложить расстояния между соответствующими элементами матриц:

$$d(A, B) = \sum_{i=1}^{k} \sum_{j=1}^{k} |a(i, j) - b(i, j)|.$$

Пример 8. Предыдущий пример наводит на мысль о следующем полезном свойстве расстояний. Если на некотором пространстве определены два или больше расстояний, то их сумма — также расстояние.

Пример 9. Пусть A и B — множества. Расстояние между множествами можно определить формулой:

$$d(A, B) = \mu(A \Delta B).$$

Здесь μ — мера на рассматриваемом пространстве множеств, Δ — символ симметрической разности множеств:

$$A \Delta B = (A \setminus B) \bigcup (B \setminus A).$$

Если мера — так называемая считающая, т.е. приписывающая единичный вес каждому элементу множества, то введенное расстояние есть число несовпадающих элементов в множествах A и B.

Пример 10. Между множествами можно ввести и другое расстояние:

$$d_1(A,B)=\frac{\mu(A\Delta B)}{\mu(A\cup B)}.$$

В ряде задач прикладной статистики используются функции двух переменных, для которых выполнены не все три аксиомы расстояния, а только некоторые. Их обычно называют показателями различия, поскольку чем больше различаются объекты, тем больше значение функции. Иногда в том же смысле используют термин «мера близости». Он менее удачен, поскольку большее значение функции соответствует меньшей близости.

Чаще всего отказываются от аксиомы, требующей выполнения неравенства треугольника, поскольку это требование не всегда находит обоснование в конкретной прикладной ситуации.

Пример 11. В конечномерном векторном пространстве показателем различия является:

$$d(x,y)=\sum_{j=1}^{k}(x(j)-y(j))^2$$

(сравните с примером 3).

Показателями различия, но не расстояниями являются такие популярные в прикладной статистике показатели, как дисперсия или средний квадрат ошибки при оценивании.

Иногда отказываются также и от аксиомы симметричности.

Пример 12. Показателем различия чисел x и y является:

$$d(x,y)=\left|\frac{x}{y}-1\right|.$$

Такой показатель различия используют в ряде процедур экспертного оценивания.

Что же касается первой аксиомы расстояния, то в различных постановках прикладной статистики ее обычно принимают. Вполне естественно, что наименьший показатель различия должен достигаться, причем именно на совпадающих объектах. Имеет ли смысл это наименьшее значение делать отличным от 0? Вряд ли, поскольку всегда можно добавить одну и ту же кон-

станту ко всем значениям показателя различия и тем самым добиться выполнения первой аксиомы.

В прикладной статистике используются самые разные расстояния и показатели различия, о них пойдет речь в соответствующих разделах учебника.

1.8. АКСИОМАТИЧЕСКОЕ ВВЕДЕНИЕ РАССТОЯНИЙ

В нечисловой статистике (и в прикладной статистике в целом) используют большое количество метрик и показателей различия (см. примеры в предыдущем разделе). Как обоснованно выбрать то или иное расстояние для использования в конкретной задаче? В 1959 г. американский статистик Джон Кемени предложил использовать аксиоматический подход, согласно которому следует сформулировать естественные для конкретной задачи аксиомы и вывести из них вид метрики. Этот подход получил большую популярность в нашей стране после выхода в 1972 г. перевода на русский язык книги Дж. Кемени и Дж. Снелла [34], в которой дана система аксиом для расстояния Кемени между упорядочениями. (Упорядочения, как и иные бинарные отношения, естественно представить в виде квадратных матриц из 0 и 1; тогда расстояние Кемени — это расстояние из примера 7 предыдущего раздела 1.6.) Последовала большая серия работ, в которых из тех или иных систем аксиом выводился вил метрики или показателя различия для различных видов данных, прежде всего для объектов нечисловой природы. Многие полученные результаты описаны в обзоре [35], содержащем 161 ссылку, в том числе 69 на русском языке. Рассмотрим некоторые задачи аксиоматического введения расстояний.

Аксиоматическое введение расстояния между толерантностями. Толерантность — это бинарное отношение, являющееся рефлексивным и симметричным. Его обычно используют для описания отношения сходства между реальными объектами, отношений знакомства или дружбы между людьми. От отношения эквивалентности отличается тем, что свойство транзитивности не предполагается обязательно выполненным. Действительно, Иванов может быть знаком с Петровым, Петров — с Сидоровым, но при этом ничего необычного нет в том, что Иванов и Сидоров не знакомы между собой.

Пусть множество X, на котором определено отношение толерантности, состоит из конечного числа элементов: $X = \{x_1, x_2, ..., x_k\}$. Тогда толерантность описывается квадратной матрицей $A = \|a(i, j)\|$, $i, j = 1, 2, ..., k$, такой,

что $a(i, j) = 1$, если x_i и x_j связаны отношением толерантности, и $a(i,j) = 0$ в противном случае. Матрица A симметрична: $a(i, j) = a(j, i)$, на главной диагонали стоят единицы: $a(i, i) = 1$. Любая матрица, удовлетворяющая приведенным в предыдущей фразе условиям, является матрицей, соответствующей некоторому отношению толерантности. Матрице A можно сопоставить неориентированный граф с вершинами в точках X: вершины x_i и x_j соединены ребром тогда и только тогда, когда $a(i, j) = 1$. Толерантности используются, в частности, при проведении экспертных исследований (см. раздел 3.7 ниже).

Будем говорить, что толерантность A_3 лежит между толерантностями A_1 и A_2, если при всех i, j число $a_3(i,j)$ лежит между числами $a_1(i, j)$ и $a_2(i, j)$, т.е. выполнены либо неравенства $a_1(i, j) \le a_3(i, j) \le a_2(i, j)$, либо неравенства $a_1(i, j) \ge a_3(i, j) \ge a_2(i, j)$.

Теорема 1 [2]. Пусть

(I) $d(A_1, A_2)$ — метрика в пространстве толерантностей, определенных на конечном множестве $X = \{x_1, x_2, ..., x_k\}$;

(II) $d(A_1, A_3) + d(A_3, A_2) = d(A_1, A_2)$ тогда и только тогда, когда A_3 лежит между A_1 и A_2;

(III) если отношения толерантности A_1 и A_2 отличаются только на одной паре элементов, т.е. $a_1(i, j) = a_2(i, j)$ при $(i, j) \ne (i_0, j_0)$, $i < j$, $i_0 < j_0$, и $a_1(i_0, j_0) \ne a_2(i_0, j_0)$, то $d(A_1, A_2) = 1$.

Тогда

$$d(A_1, A_2) = \sum_{1 \le i < j \le k} | a_1(i, j) - a_2(i, j) | = \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} | a_1(i, j) - a_2(i, j) |.$$

Таким образом, расстояние $d(A_1, A_2)$ только постоянным множителем 1/2 отличается от расстояния Кемени, введенного в пространстве всех бинарных отношений как расстояние Хемминга между описывающими отношения матрицами из 0 и 1 (см. пример 7 предыдущего раздела 1.6). Теорема 1 дает аксиоматическое введение расстояния в пространстве толерантностей. Оказалось, что оно является сужением расстояния Кемени на это пространство. Сам Дж. Кемени дал аналогичную систему аксиом для сужения на пространство упорядочений. Доказательство теоремы 1 вытекает из рассмотрений, связанных с аксиоматическим введением расстояний между множествами, и приводится ниже.

Мера симметрической разности как расстояние между множествами. Как известно, бинарное отношение можно рассматривать как подмножество декартова квадрата X^2 того множества X, на котором оно опреде-

лено. Поэтому теорему 1 можно рассматривать как аксиоматическое введение расстояния между множествами специального вида. Укажем систему аксиом для расстояния между множествами общего вида, описанного в примере 9 предыдущего раздела.

Определение 1. Множество B находится между множествами A и C, если $(A \cap C) \subseteq B \subseteq (A \cup C)$.

С помощью определения 1 в совокупности множеств вводятся геометрические соотношения, использование которых полезно для восприятия рассматриваемых ситуаций.

Расстояние между двумя точками в евклидовом пространстве не изменится, если обе точки сдвинуть на один и тот же вектор. Аналогичное свойство расстояния между множествами сформулируем в виде аксиомы 1. Оно соответствует аксиоме 3 Кемени и Снелла [34, с. 22] для расстояний между упорядочениями.

Аксиома 1. Если $A \cap C = B \cap C = \varnothing$, то $d(A,B) = d(A \cup C, B \cup C)$.

Определение 2. Непустая система множеств называется кольцом, если для любых двух входящих в нее множеств в эту систему входят их объединение, пересечение и разность. Множество X называется единицей системы множеств, если оно входит в эту систему, а все остальные множества системы являются подмножествами X. Кольцо множеств, содержащее единицу, называется алгеброй множеств [36, с. 38].

Теорема 2. Пусть W — алгебра множеств, d: $W^2 \to R^1$. Тогда аксиома 1 эквивалентна следующему условию: $d(A, B) = d(A \setminus B, B \setminus A)$ для любых $A, B \in W$.

Доказательство. Поскольку

$$(A \setminus B) \cap (A \cap B) = \varnothing, \ (B \setminus A) \cap (A \cap B) = \varnothing,$$

то равенство $d(A,B) = d(A \setminus B, B \setminus A)$ следует из аксиомы 1. Обратное утверждение вытекает из того, что в условиях аксиомы 1:

$$(A \cup C) \setminus (B \cup C) = A \setminus B, \ (B \cup C) \setminus (A \cup C) = B \setminus A.$$

Теорема 2 доказана.

С целью внести в алгебру множеств W связь расстояния и отношения «находиться между», аналогичную используемой при аксиоматическом введении расстояний в пространствах бинарных отношений (см. условие (II) в теореме 1), примем следующую аксиому.

Аксиома 2. Если B лежит между A и C, то $d(A, B)+d(B, C)= d(A, C)$.

Определение 3. Неотрицательная функция μ, определенная на алгебре множеств W, называется мерой, если для любых двух непересекающихся множеств A и B из W справедливо соотношение:

$$\mu(A \cup B) = \mu(A) + \mu(B).$$

Понятие меры — это обобщение понятий длины линии, площади фигуры, объема тела.

Теорема 3. Пусть W — алгебра множеств, аксиомы 1 и 2 выполнены для функции d: $W^2 \to [0; +\infty]$. Функция d симметрична: $d(A, B) = d(B, A)$ для любых A и B из W. Тогда существует, и притом единственная, мера μ на W такая, что:

$$d(A, B) = \mu(A \Delta B), \qquad (1)$$

при всех A и B из W, где $A \Delta B$ — симметрическая разность множеств A и B, т.е. $A \Delta B = (A \setminus B) \cup (B \setminus A)$.

Доказательство. Положим:

$$\mu(B) = d(\varnothing, B), \quad B \in W. \qquad (2)$$

Покажем, что определенная формулой (2) функция множества μ является мерой. Неотрицательность μ следует из неотрицательности d. Остается доказать аддитивность, т.е. что из $A \cap B = \varnothing$ следует, что:

$$\mu(A \cup B) = \mu(A) + \mu(B), \; A \in W, B \in W. \qquad (3)$$

Поскольку A всегда лежит между $\varnothing$ и $A \cup B$, то по аксиоме 2:

$$\mu(A \cup B) = d(\varnothing, A \cup B) = d(\varnothing, A) + d(A, A \cup B) = \mu(A) + d(A, A \cup B). \qquad (4)$$

Если $A \cap B = \varnothing$, то по аксиоме 1 $d(\varnothing, B) = d(A, A \cup B)$, откуда с учетом (4) и следует (3).

Докажем соотношение (1). Поскольку $A\backslash B$ и $B\backslash A$ имеют пустое пересечение, то согласно определению 1 пустое множество $\varnothing$ лежит между $A\backslash B$ и $B\backslash A$. Поэтому по аксиоме 2:

$$d(A \setminus B, B \setminus A) = d(A \setminus B, \varnothing) + d(\varnothing, B \setminus A).$$

Из симметричности и соотношения (2) следует, что:

$$d(A \setminus B, \varnothing) = d(\varnothing, A \setminus B) = \mu(A \setminus B),$$

откуда $d(A\backslash B, B\backslash A) = \mu(A\backslash B) + \mu(B\backslash A)$. Из соотношения (3) следует, что $\mu(A\backslash B) + \mu(B\backslash A) = \mu(A \Delta B)$. С другой стороны, по аксиоме 1:

$$d(A \setminus B, B \setminus A) = d((A \setminus B) \cup (A \cap B), (B \setminus A) \cup (A \cap B)) = d(A, B).$$

Из трех последних равенств вытекает справедливость равенства (1).

Остается доказать единственность меры μ в соотношении (1). Поскольку $A\Delta B = B$ при $A = \varnothing$, то из (1) следует (2), т.е. однозначность определения меры $\mu = \mu(d)$ по расстоянию d. Теорема 3 доказана.

Теорема 4 (обратная). Пусть μ — мера, определенная на алгебре множеств W. Тогда функция $d(A, B) = \mu(A\Delta B)$ является псевдометрикой, для нее выполнены аксиомы 1 и 2.

Доказательство. То, что функция $d(A, B)$ из (1) задает псевдометрику, хорошо известно (см., например, [37, с. 79]). Доказательство аксиомы 2 содержится в [38, с. 181–183]. Аксиома 1 следует из того, что условия $A \cap C = B \cap C = \varnothing$ обеспечивают справедливость соотношений:

$$(A \cup C)\Delta(B \cup C) = ((A \cup C) \setminus (B \cup C)) \cup ((B \cup C) \setminus (A \cup C)) = (A \setminus B) \cup (B \setminus A) = A\Delta B.$$

Замечание. Полагая в аксиоме 2 $A = B = C$, получаем, что $d(A,A) + d(A,A) = d(A,A)$, т.е. $d(A,A) = 0$. Согласно теоремам 3 и 4, из условий теоремы 3 следует неравенство треугольника. Таким образом, в теореме 3 действительно приведена система аксиом, определяющая семейство псевдометрик в пространстве множеств.

Обсудим независимость (друг от друга) условий теоремы 3. Отбрасывание неотрицательности функции d приводит к тому, что слово «мера» в теоремах 3 и 4 необходимо заменить на «заряд» [36, с. 328]. Этот термин обо-

значает аддитивную функцию множеств, не обладающую свойством неотрицательности. Заряд можно представить как разность двух мер.

Функция $d_1(A,B)=\sqrt{\mu(A\Delta B)}$ является псевдометрикой, для нее выполнена аксиома 1, но не выполнена аксиома 2, следовательно, ее нельзя представить в виде (1).

Приведем пример системы множеств W и метрики в ней, для которых верна аксиома 2, но не верна аксиома 1, а потому эту метрику нельзя представить в виде (1). Пусть W состоит из множеств $\varnothing, A, B, A\cup B$, причем $A\cap B=\varnothing$, а расстояния таковы:

$$d(\varnothing,A)=d(\varnothing,B)=1,\ \ d(A,A\cup B)=d(B,A\cup B)=d(A,B)=2,\ \ d(\varnothing,A\cup B)=3.$$

Если единица X алгебры множеств W конечна, т.е. $X=\{x_1, x_2, ..., x_k\}$, то расстояние (1) принимает вид:

$$d(A,B)=\sum_{i=1}^{k}\mu_i\,|\,\chi_A(x_i)-\chi_B(x_i)\,|, \qquad (5)$$

где χ_C — индикатор (индикаторная функция) множества C, т.е. $\chi_C(x)=1$, если $x\in C$, и $\chi_C(x)=0$ в противном случае. Как следует из теоремы 3, неотрицательный коэффициент μ_i — это мера одноэлементного множества $\{x_i\}$, а также расстояние этого множества от пустого множества, т.е.

$$\mu_i=\mu(\{x_i\})=d(\varnothing,\{x_i\}).$$

Если все коэффициенты μ_i положительны, то формула (5) определяет метрику, если хотя бы один равен 0, то — псевдометрику, поскольку в таком случае найдутся два различающиеся между собой множества A и B такие, что $d(A, B)=0$.

Расстояние определяется однозначно, если априори известны коэффициенты μ_i. В частности, равноправность объектов (элементов единицы алгебры множеств X) приводит к $\mu_i\equiv 1$. Требование равноправности содержится в аксиомах 2 и 4 Кемени [34, с. 21–22].

Применим полученные результаты к толерантностям и докажем теорему 1. Совокупность всех толерантностей, определенных на конечном множестве Y, естественным образом ассоциируется с совокупностью всех подмножеств множества $X=\{(y_i, y_j),\ 1\le i<j\le k\}$. Именно, пара (y_i, y_j) входит в под-

множество тогда и только тогда, когда y_i и y_j связаны отношением толерантности. Указанная совокупность подмножеств является алгеброй множеств с единицей X. Определение 1 понятия «находиться между» для множеств полностью соответствует ранее данному определению понятия «находиться между» для толерантностей.

Теорема 5. Пусть выполнены условия (I) и (II) теоремы 1 и аксиома 1. Тогда существуют числа $\mu_{ij} > 0$ такие, что:

$$d(A,B) = \sum_{1 \le i < j \le k} \mu_{ij} \, | \, a(i,j) - b(i,j) \, |. \qquad (6)$$

Для доказательства достаточно сослаться на теорему 3. Поскольку в условии (I) требуется, чтобы функция $d(A, B)$ являлась метрикой, то необходимо $\mu_{ij} > 0$.

Теорема 6. Пусть выполнены условия теоремы 1 и, кроме того, аксиома 1. Тогда верно заключение теоремы 1.

Доказательство. Рассмотрим толерантность A, для которой $a(i, j) = 1$ при $(i, j) = (i_0, j_0)$ и $a(i,j) = 0$ в противном случае. Согласно условию (III) теоремы 1 $d(\varnothing, A) = 1$, а согласно (6) имеем $d(\varnothing, A) = \mu_{i_0 j_0}$. Следовательно, коэффициент $\mu_{i_0 j_0} = 1$, что и требовалось доказать.

Для окончательного доказательства теоремы 1 осталось избавиться от требования справедливости аксиомы 1.

Доказательство теоремы 1. Рассмотрим две толерантности A и B такие, что при представлении их в виде множеств $A \subseteq B$. Это означает, что $a(i, j) \le b(i, j)$ при всех i, j. Поскольку X — конечное множество, то существует конечная последовательность толерантностей $A_1, A_2, \ldots, A_m, \ldots, A_t$ такая, что $A_1 = A$, $A_t = B$, $A_1 \subseteq A_2 \subseteq \ldots \subseteq A_m \subseteq \ldots \subseteq A_t$, причем A_{m+1} получается из A_m заменой ровно одного значения $a_m(i_m, j_m) = 0$ на $a_{m+1}(i_m, j_m) = 1$, для $(i, j) \ne (i_m, j_m)$ при этом $a_m(i, j) = a_{m+1}(i, j)$. Тогда A_m находится между A_{m-1} и A_{m+1}, следовательно, по условию (II):

$$d(A,B) = d(A_1, A_2) + d(A_2, A_3) + \ldots + d(A_m, A_{m+1}) + \ldots + d(A_{t-1}, A_t).$$

По условию (III) $d(A_m, A_{m+1}) = 1$ при всех m, а потому заключение теоремы 1 верно для любых A и B таких, что $A \subseteq B$.

Поскольку $A \cap B$ лежит между A и B, то по условию (II):

$$d(A, B) = d(A \cap B, A) + d(A \cap B, B).$$

При этом $A \cap B \subseteq A$ и $A \cap B \subseteq B$. Применяя результат предыдущего абзаца, получаем, заключение теоремы 1 верно всегда.

Замечание 1. Таким образом, условие (III) не только дает нормировку, но и заменяет аксиому 1.

Замечание 2. Условие (I) теоремы 1 не использовалось в доказательстве, но было приведено в первоначальной публикации [39], чтобы подчеркнуть цель рассуждения. По той же причине оно сохранено в формулировке теоремы 1, хотя в доказательстве удалось без него обойтись. Понадобилась только симметричность функции d.

Аксиоматическое введение метрики в пространстве неотрицательных суммируемых функций. Рассмотрим пространство $L(E, \mu)$ неотрицательных суммируемых функций на множестве E с мерой μ. Далее в настоящем разделе будем рассматривать только функции из пространства $L(E, \mu)$. Интегрирование всюду проводится по множеству (пространству) E и по мере μ. Будем писать $g = h$ или $g \le h$, если указанные соотношения справедливы почти всюду по μ на E (т.е. могут нарушаться лишь на множестве нулевой меры).

Аксиоматически введем расстояние в пространстве $L(E, \mu)$ (изложение следует работе [40]). Обозначим $M(g, h) = \max(g, h)$ и $m(g, h) = \min(g, h)$. Пусть функция D: $L(E, \mu) \times L(E, \mu) \to R^1$ — тот основной объект изучения, аксиомы для которого будут сейчас сформулированы.

Аксиома 1. Если $gh = 0$, $g + h \neq 0$, то $D(g, h) = 1$.

Аксиома 2. Если $h \le g$, то $D(g, h) = C \int (g - h)\, d\mu$, где множитель C не зависит от h, т.е. $C = C(g)$.

Лемма. Из аксиом 1,2 следует, что для $h \le g \neq 0$:

$$D(g,h) = \frac{\int (g-h)d\mu}{\int g d\mu}.$$

Для доказательства заметим, что по аксиоме 1 $D(g, 0) = 1$, а по аксиоме 2 $D(g, 0) = C\int g d\mu$, откуда $C = (\int g d\mu)^{-1}$. Подставляя это соотношение в аксиому 2, получаем заключение леммы.

Требование согласованности расстояния в пространстве $L(E, \mu)$ с отношением «находиться между» приводит, как и ранее для расстояния $d(A, B)$, к следующей аксиоме.

Аксиома 3. Для любых g и h справедливо равенство $D(g, h) = D(M(g, h), g) + D(M(g, h), h)$.

Замечание. В ряде реальных ситуаций естественно считать, что наибольшее расстояние между элементами пространства множеств (которое без ограничения общности можно положить равным 1), т.е. наибольшее несходство, соответствует множествам, не имеющим общих элементов. Расстояние, введенное в теореме 3 (формула (1)), этому условию не удовлетворяет. Поэтому в пространстве множеств была аксиоматически введена [35] так называемая D-метрика (от *dissimilarity* (англ.) — несходство), для которого это условие выполнено. Она имеет вид:

$$D(A,B) = \begin{cases} \dfrac{\mu(A \Delta B)}{\mu(A \cup B)}, & \mu(A \cup B) > 0, \\ 0, & \mu(A) = \mu(B) = 0. \end{cases} \tag{7}$$

Приведенные выше аксиомы являются обобщениями соответствующих аксиом для D-метрики в пространстве множеств.

Теорема 7. Из аксиом 1–3 следует, что

$$D(g,h) = \begin{cases} \dfrac{\int |g-h| \, d\mu}{\int M(g,h) d\mu}, & g + h \neq 0, \\ 0, & g = h = 0. \end{cases} \tag{8}$$

Доказательство. Поскольку

$$(M(g, h) - g) + (M(g, h) - h) = |g - h|,$$

то заключение теоремы 7 при $g + h \neq 0$ вытекает из леммы и аксиомы 3. Из аксиомы 2 при $g = 0$ следует, что $D(0, 0) = 0$. Легко видеть, что функция D, заданная формулой (8), удовлетворяет аксиомам 1–3 и, кроме того, $D(g, h) \leq 1$ при любых g и h.

Замечание. Если g и h — индикаторные функции множеств, то формула (8) переходит в формулу (7). Если g и h — функции принадлежности нечетких множеств, то формула (8) задает метрику в пространстве нечетких множеств, а именно, D-метрику в этом пространстве [35].

Теорема 8. Функция $D(g,h)$, определенная формулой (8), является метрикой в $L(E, \mu)$ (при отождествлении функций, отличающихся лишь на мно-

жестве нулевой меры), причем $D(g, f) + D(f, h) = D(g, h)$ тогда и только тогда, когда $f = g$, $f = h$ или $f = M(g, h)$.

Доказательство. Обратимся к определению метрики. Для рассматриваемой функции непосредственно очевидна справедливость условий неотрицательности и симметричности. Очевидна и эквивалентность условия $D(g, h) = 0$ равенству $g = h$. Остается доказать неравенство треугольника и установить, когда оно обращается в равенство.

Без ограничения общности можно считать, что рассматриваемые расстояния задаются верхней строкой формулы (8) и, кроме того,

$$R = \int M(g, f)\, d\mu - \int M(f, h)\, d\mu \geq 0$$

(частные случаи с использованием нижней строки формулы (8) рассматриваются элементарно, а справедливости последнего неравенства можно добиться заменой обозначений функций — элементов пространства $L(E, \mu)$).

Тогда

$$D(g, f) + D(f, h) \geq \frac{\int (|g - f| + |f - h|)\, d\mu}{\int M(g, f)\, d\mu}, \tag{9}$$

причем равенство имеет место тогда и только тогда, когда $R = 0$ или $f = h$.

Положим:

$$P = \int (|g - f| + |f - h| - |g - h|)\, d\mu, \quad Q = \int (M(g, f) - M(g, h))\, d\mu.$$

Ясно, что $P \geq 0$ и

$$\frac{\int (|g - f| + |f - h|)\, d\mu}{\int M(g, f)\, d\mu} = \frac{\int |g - h|\, d\mu + P}{\int M(g, h)\, d\mu + Q}. \tag{10}$$

Если $Q < 0$, то, очевидно, неравенство треугольника выполнено, причем неравенство является строгим. Рассмотрим случай $Q > 0$.

Воспользуемся следующим элементарным фактом: если $y \geq x$, $y > 0$, $P > Q > 0$, то

$$\frac{x + P}{y + Q} > \frac{x}{y}. \tag{11}$$

Из соотношений (10) и (11) вытекает, что для доказательства неравенства треугольника достаточно показать, что $P - Q > 0$.

Рассмотрим:

$$k = \{|g - f| + |f - h| - |g - h|\} - M(g, f) + M(g, h).$$

Применяя равенство $(M(g, h) - g) + (M(g, h) - h) = |g - h|$ к слагаемым, заключенным в фигурные скобки, получаем, что

$$k = M(f, h) + [M(g, f) + M(f, h) - M(g, h) - 2f].$$

Применяя соотношение:

$$M(g, h) = g + h - m(g, h) \tag{12}$$

к слагаемым, заключенным в квадратные скобки, получаем, что

$$k = M(f, h) - m(f, h) - m(g, f) + m(g, h).$$

Так как $M(f, h) - m(f, h) = |f - h|$, то

$$k = |f - h| - (m(g, f) - m(g, h)) \geq (f - h) - (m(g, f) - m(g, h)). \tag{13}$$

В соответствии с (12) правая часть (13) есть $M(g, f)$ — M(g, h), а потому

$$P - Q = \int k \, d\mu \geq Q > 0,$$

что завершает доказательство для случая $Q > 0$. При этом неравенство треугольника является строгим.

Осталось рассмотреть случай $Q = 0$. В силу соотношений (9) и (10) неравенство треугольника выполнено. Когда оно обращается в равенство? Тривиальные случаи: $f = g$ или $f = h$. Если же f отлично от g и h, то необходимо, чтобы $R = 0$ и $P = 0$. Как легко проверить, последнее условие эквивалентно неравенствам:

$$m(g, h) \leq f \leq M(g, h). \tag{14}$$

Из правого неравенства в (14) следует, что $M(g, f) \le M(g, M(g, h)) = M(g, h)$. Так как $Q = 0$, то $M(g, f) = M(g, h)$. Аналогичным образом из соотношений:

$$M(h, f) \le M(h, M(g, h)) = M(g, h) = M(g, f)$$

и $R = 0$ следует, что $M(f, h) = M(g, h)$.

Рассмотрим измеримое множество $X = \{x \in E: h(x) < g(x)\}$.

Тогда $M(g, h)(x) = M(f, h)(x) = g(x) > h(x)$, т.е. $h(x) < f(x) = M(g, h)(x)$ для почти всех $x \in X$. Для почти всех $y \in \{x \in E: h(x) > g(x)\}$ точно так же получаем $f(y) = M(g, h)(y)$. Для почти всех $z \in \{x \in E: h(x) = g(x)\}$ в силу (14) $f(z) = M(g, h)(z)$, что и завершает доказательство теоремы.

Замечание. Назовем функции g и h подобными, если существует число $b > 0$ такое, что $g = bh$. Тогда при $0 < b \le 1$ имеем $D(g, h) = 1 - b$, т.е. расстояние между подобными функциями линейно зависит от коэффициента подобия. Далее, пусть $a > 0$, тогда $D(ag, ah) = D(g, h)$. Таким образом, метрика (8) инвариантна по отношению к преобразованиям подобия, которые образуют группу допустимых преобразований в шкале отношений. Это дает основания именовать метрику (8) метрикой подобия [40].

Многообразие объектов нечисловой природы рассмотрено в [41]. Вероятностным моделям порождения нечисловых данных посвящена работа [42]. Сведение теории нечетких множеств к теории случайных множеств осуществлено в [43], тем самым продемонстрировано, что теория нечетких множеств — часть теории вероятностей. Статья [44] описывает расстояния в пространствах статистических данных и системы аксиом, порождающих такие расстояния. Естественные показатели различия изучены в работе [45].

Темы докладов, рефератов, исследовательских работ

1. Содержание первого сочинения по прикладной статистике — книги «Числа» в Библии.

2. Свойства основных шкал измерения.

3. Взаимосвязи различных классов объектов нечисловой природы между собой.

4. Вероятностные модели бинарных отношений.

5. Вероятностные модели парных сравнений.

6. Опишите с помощью нечеткого подмножества временной шкалы понятие «молодой человек» (на основе опроса 10–20 экспертов).

7. Опишите с помощью теории нечеткости понятие «куча зерен» (на основе опроса 10–20 экспертов).

8. Обсудите суждение: «Мы мыслим нечетко» (см. [32]). Почему нечеткость мышления помогает взаимопониманию?

9. Взаимосвязь теории нечеткости и теории вероятностей.

10. Методы оценивания функции принадлежности.

11. Теория нечеткости и интервальная математика.

12. Описание данных для выборок, элементы которых — нечеткие множества.

13. Регрессионный анализ нечетких переменных (согласно [24]).

14. Кластерный анализ нечетких данных.

15. Непараметрические оценки плотности распределения вероятностей в пространстве нечетких множеств (согласно подходу раздела 2.5).

16. Центральная роль статистики объектов произвольной природы в прикладной статистике.

17. Расстояния в пространствах функций.

18. Докажите, что аксиоматически введенный в разделе 1.7 показатель различия между множествами $d(A, B) = \mu(A \Delta B)$ удовлетворяет неравенству треугольника.

Контрольные вопросы и задачи

1. Приведите примеры практического использования количественных и категорированных данных.

2. Как соотносятся группы допустимых преобразований для различных шкал измерения?

3. Почему анализ нечисловых данных занимает одно из центральных мест в прикладной статистике?

4. Какая математическая модель используется для описания случайного множества?

5. В каких случаях целесообразно применение нечетких множеств?

6. Справедливо ли для нечетких множеств равенство $(A+B)C = AC + BC$? А равенство $(AB)C = (AC)(BC)$?

7. Как с точки зрения нечетких множеств можно интерпретировать вероятность накрытия определенной точки случайным множеством?

8. На множестве $Y = \{y_1, y_2, y_3\}$ задано нечеткое множество B с функцией принадлежности $\mu_B(y)$, причем $\mu_B(y_1) = 0{,}1$, $\mu_B(y_2) = 0{,}2$, $\mu_B(y_3) = 0{,}3$. Постройте случайное множество A так, чтобы $Proj\ A = B$.

9. На множестве $Y = \{y_1, y_2, y_3\}$ задано нечеткое множество B с функцией принадлежности $\mu_B(y)$, причем $\mu_B(y_1) = 0{,}2$, $\mu_B(y_2) = 0{,}1$, $\mu_B(y_3) = 0{,}5$. Постройте случайное множество A так, чтобы $Proj\ A = B$.

10. На множестве $Y = \{y_1, y_2, y_3\}$ задано нечеткое множество B с функцией принадлежности $\mu_B(y)$, причем $\mu_B(y_1) = 0{,}5$, $\mu_B(y_2) = 0{,}4$, $\mu_B(y_3) = 0{,}7$. Постройте случайное множество A так, чтобы $Proj\ A = B$.

11. На множестве $Y = \{y_1, y_2, y_3\}$ задано нечеткое множество B с функцией принадлежности $\mu_B(y)$, причем $\mu_B(y_1) = 0{,}3$, $\mu_B(y_2) = 0{,}2$, $\mu_B(y_3) = 0{,}1$. Постройте случайное множество A так, чтобы $Proj\ A = B$.

12. Докажите, что для блочного расстояния (пример 4 из раздела 1.6) справедливо неравенство треугольника.

13. Расскажите о многообразии расстояний в различных пространствах статистических данных.

14. Докажите, что если $d(x, y)$ — расстояние в некотором пространстве, то $\sqrt{d(x,y)}$ — также расстояние в этом пространстве.

Литература

1. *Суппес, П.* Основы теории измерений / П. Суппес, Дж. Зинес // Психологические измерения. — Москва : Мир, 1967. — С. 9–110.

2. *Орлов, А. И.* Устойчивость в социально-экономических моделях / А. И. Орлов. — Москва : Наука, 1979. — 296 с.

3. *Орлов, А. И.* Эконометрика : учебник для вузов / А. И. Орлов. — 3-е изд., испр. и доп. — Москва : Экзамен, 2004. — 576 с.

4. *Носовский, Г. В.* Империя. Русь, Турция, Китай, Европа, Египет. Новая математическая хронология древности / Г. В. Носовский, А. Т. Фоменко. — Москва : Факториал, 1996. — 752 с.

5. *Шубкин, В. П.* Социологические опыты / В. П. Шубкин. — Москва : Мысль, 1970. — 256 с.

6. *Щукина, Г. И.* Проблема познавательного интереса в педагогике / Г. И. Щукина. — Москва : Педагогика, 1971. — 352 с.

7. *Орлов, А. И.* Статистика объектов нечисловой природы (Обзор) / А. И. Орлов // Заводская лаборатория. — 1990. — Т. 56. — № 3. — С. 76–83.

8. *Орлов, А. И.* Объекты нечисловой природы // Заводская лаборатория. — 1995. — Т. 61. — № 3. — С. 43–52.

9. *Кендэл, М.* Ранговые корреляции / М. Кендэл. — Москва : Статистика, 1975. — 216 с.

10. *Беляев, Ю. К.* Вероятностные методы выборочного контроля / Ю. К. Беляев. — Москва : Наука, 1975. — 408 с.

11. *Лумельский, Я. П.* Статистические оценки результатов контроля качества / Я. П. Лумельский. — Москва : Изд-во стандартов, 1979. — 200 с.

12. *Дэвид, Г.* Метод парных сравнений / Г. Дэвид. — Москва : Статистика, 1978. — 144 с.

13. Организация и планирование машиностроительного производства (производственный менеджмент) : учебник / К. А. Грачева, М. К. Захарова, Л. А. Одинцова [и др.] ; под редакцией Ю. В. Скворцова, Л. А. Некрасова. — Москва : Высшая школа, 2003. — 470 с.

14. *Кендалл, М. Дж.* Статистические выводы и связи / М. Дж. Кендалл, А. Стьюарт. — Москва : Наука, 1973. — 900 с.

15. *Себер, Дж.* Линейный регрессионный анализ / Дж. Себер. — Москва : Мир, 1980. — 456 с.

16. *Орлов, А. И.* Асимптотика некоторых оценок размерности модели в регрессии // Прикладная статистика. Ученые записки по статистике. — Т. 45. — Москва : Наука, 1983. — С. 260–265.

17. *Борель, Э.* Вероятность и достоверность / Э. Борель. — Москва : ГИФМЛ, 1961. — 120 с.

18. *Орлов, А. И.* Вероятностные модели конкретных видов объектов нечисловой природы / А. И. Орлов // Заводская лаборатория. — 1995. — Т. 61. — № 5. — С. 43–51.

19. Вероятность и математическая статистика : энциклопедия / главный редактор Ю. В. Прохоров. — Москва : Большая Российская энциклопедия, 1999. — 910 с.

20. *Орлов, А. И.* Логистическое распределение / А. И. Орлов // Математическая энциклопедия. — Т. 3. — Москва : Советская энциклопедия, 1982. — С. 414.

21. *Тюрин, Ю. Н.* Статистические модели ранжирования / Ю. Н. Тюрин, А. П. Василевич, П. Ф. Андрукович // Статистические методы анализа экспертных оценок. — Москва : Наука, 1977. — С. 30–58.

22. *Орлов, А. И.* Случайные множества с независимыми элементами (люсианы) и их применения / А. И. Орлов // Алгоритмическое и программное

обеспечение прикладного статистического анализа. Ученые записки по статистике. — Т. 36. — Москва : Наука, 1980. — С. 287–308.

23. *Орлов, А. И.* Парные сравнения в асимптотике Колмогорова / А. И. Орлов // Экспертные оценки в задачах управления. — Москва : Изд-во Института проблем управления АН СССР, 1982. — С. 58–66.

24. *Орлов, А. И.* Задачи оптимизации и нечеткие переменные / А. И. Орлов. — Москва : Знание, 1980. — 64 с.

25. *Прохоров, Ю. В.* Теория вероятностей. (Основные понятия. Предельные теоремы. Случайные процессы) / Ю. В. Прохоров, Ю. А. Розанов. — Москва : Наука, 1973. — 496 с.

26. *Битюков, П. В.* Моделирование задач ценообразования на электронные обучающие курсы в области дистанционного обучения : специальность 08.00.13 «Математические и инструментальные методы экономики» : автореферат диссертации на соискание ученой степени кандидата экономических наук / Битюков Петр Вадимович ; Московский государственный университет экономики, статистики и информатики. — Москва : Изд-во МГУЭСИ, 2002. — 24 с.

27. *Лебег, А.* Об измерении величин / А. Лебег. — Москва : Учпедгиз, 1960. — 204 с.

28. *Ефимов, Н. В.* Высшая геометрия / Н. В. Ефимов. — Москва : ГИФМЛ, 1961. — 580 с.

29. *Орлов, А. И.* Основания теории нечетких множеств (обобщение аппарата Заде). Случайные толерантности / А. И. Орлов // Алгоритмы многомерного статистического анализа и их применения. — Москва : ЦЭМИ АН СССР, 1975. — С. 169–175.

30. *Гудмэн, И.* Нечеткие множества как классы эквивалентности случайных множеств / И. Гудмэн // Нечеткие множества и теория возможностей. Последние достижения. — Москва : Радио и связь, 1986. — С. 241–264. — Перевод издания: *Goodman, I. R.* Fuzzy sets as equivalence classes of random sets / I. R. Goodman // Fuzzy Set and Possibility Theory: Recent Developments / editor R. Yager Ronald. — Oxford : Pergamon Press, 1982. — P. 327–343.

31. *Орлов, А. И.* Математика нечеткости / А. И. Орлов // Наука и жизнь. — 1982. — № 7. — С. 60–67.

32. *Орлов, А. И.* Математика случая. Вероятность и статистика — основные факты / А. И. Орлов. — Москва : Экзамен, 2008.

33. *Орлов, А. И.* Асимптотика решений экстремальных статистических задач / А. И. Орлов // Анализ нечисловых данных в системных исследованиях :

сборник трудов. — Вып. 10. — Москва : Всесоюзный научно-исследовательский институт системных исследований, 1982. — С. 4–12.

34. *Кемени, Дж.* Кибернетическое моделирование. Некоторые приложения / Дж. Кемени, Дж. Снелл. — Москва : Советское радио, 1972. — 192 с.

35. *Раушенбах, Г. В.* Меры близости и сходства / Г. В. Раушенбах // Анализ нечисловой информации в социологических исследованиях. — Москва : Наука, 1986. — С. 169–203.

36. *Колмогоров, А. Н.* Элементы теории функций и функционального анализа / А. Н. Колмогоров, С. В. Фомин. — Москва : Наука, 1972. — 496 с.

37. *Окстоби, Дж.* Мера и категория / Дж. Окстоби. — Москва : Мир, 1974. — 158 с.

38. *Льюс, Р.* Психофизические шкалы / Р. Льюс, Е. Галантер // Психологические измерения. — Москва : Мир, 1967. — С. 111–195.

39. *Орлов, А. И.* Связь между нечеткими и случайными множествами: Нечеткие толерантности / А. И. Орлов // Исследования по вероятностно-статистическому моделированию реальных систем. — Москва : ЦЭМИ АН СССР, 1977. — С. 140–148.

40. *Орлов, А. И.* Метрика подобия: аксиоматическое введение, асимптотическая нормальность / А. И. Орлов, Г. В. Раушенбах // Статистические методы оценивания и проверки гипотез : межвузовский сборник научных трудов. — Пермь : Изд-во ПГНИУ, 1986. — С. 148–157.

41. *Орлов, А. И.* Многообразие объектов нечисловой природы / А. И. Орлов // Научный журнал КубГАУ. — 2014. — № 102. — С. 32–63.

42. *Орлов, А. И.* Вероятностные модели порождения нечисловых данных / А. И. Орлов // Научный журнал КубГАУ. — 2015. — № 105. — С. 39–66.

43. *Орлов, А. И.* Теория нечетких множеств — часть теории вероятностей / А. И. Орлов // Научный журнал КубГАУ. — 2013. — № 92. — С. 51–60.

44. *Орлов, А. И.* Расстояния в пространствах статистических данных / А. И. Орлов // Научный журнал КубГАУ. — 2014. — № 101. — С. 227–252.

45. *Орлов, А. И.* Естественные показатели различия / А. И. Орлов // Научный журнал КубГАУ. — 2020. — № 163. — С. 248–264.

ГЛАВА 2. СТАТИСТИЧЕСКИЕ МЕТОДЫ В ПРОСТРАНСТВАХ ПРОИЗВОЛЬНОЙ ПРИРОДЫ

2.1. ЭМПИРИЧЕСКИЕ И ТЕОРЕТИЧЕСКИЕ СРЕДНИЕ

Одна из основных статистических процедур — вычисление средних величин для тех или иных совокупностей данных. Законы больших чисел состоят в том, что эмпирические средние сходятся к теоретическим. В классическом варианте: выборочное среднее арифметическое при определенных условиях сходится по вероятности при росте числа слагаемых к математическому ожиданию. На основе законов больших чисел обычно доказывают состоятельность различных статистических оценок. В целом эта тематика занимает заметное место в теории вероятностей и математической статистике.

Однако математический аппарат при этом основан на свойствах сумм случайных величин (векторов, элементов линейных пространств). Следовательно, он не пригоден для изучения вероятностных и статистических проблем, связанных со случайными объектами нечисловой природы. Это такие объекты, как бинарные отношения, нечеткие множества, вообще элементы пространств без векторной структуры. Объекты нечисловой природы все чаще встречаются в прикладных исследованиях. Много конкретных примеров приведено выше в главе 1. Поэтому необходимо научиться усреднять различные нечисловые данные, т.е. определять эмпирические и теоретические средние в пространствах произвольной природы. Кроме того, представляется полезным получение законов больших чисел в пространствах нечисловой природы.

Для осуществления описанной научной программы необходимо решить следующие задачи.

А. Определить понятие эмпирического среднего.

Б. Определить понятие теоретического среднего.

В. Ввести понятие сходимости эмпирических средних к теоретическому.

Г. Доказать при тех или иных комплексах условий сходимость эмпирических средних к теоретическому.

Д. Обобщив это доказательство, получить метод обоснования состоятельности различных статистических оценок.

Е. Дать применения полученных результатов при решении конкретных задач.

Ввиду принципиальной важности рассматриваемых результатов приводим в настоящей главе доказательство закона больших чисел, а также результаты компьютерного анализа множества эмпирических средних.

Определения средних величин. Пусть X — пространство произвольной природы, $x_1, x_2, x_3, \ldots, x_n$ — его элементы. Чтобы ввести эмпирическое среднее для $x_1, x_2, x_3, \ldots, x_n$, будем использовать действительнозначную (т.е. с числовыми значениями) функцию $f(x,y)$ двух переменных со значениями в X. В стандартных математических обозначениях: $f : X^2 \to R^1$. Величина $f(x, y)$ интерпретируется как показатель различия между x и y: чем $f(x, y)$ больше, тем x и y сильнее различаются. В качестве f можно использовать расстояние в X, квадрат расстояния и т.п.

Определение 1. Средней величиной для совокупности $x_1, x_2, x_3, \ldots, x_n$ (относительно меры различия f), обозначаемой любым из трех способов:

$$x_{cp} = E_n(f) = E_n(x_1, x_2, x_3, \ldots, x_n; f),$$

называем решение оптимизационной задачи:

$$\sum_{i=1}^{n} f(x_i, y) \to \min, \qquad y \in X. \tag{1}$$

Это определение согласуется с классическими определениями средних величин. Если $X = R^1$, $f(x, y) = (x - y)^2$, то x_{cp} — выборочное среднее арифметическое. Если же $X = R^1$, $f(x,y) = |x - y|$, то при $n = 2k + 1$ имеем $x_{cp} = x(k + 1)$, при $n = 2k$ эмпирическое среднее является отрезком $[x(k), x(k + 1)]$. Здесь через $x(i)$ обозначен i-й член вариационного ряда, построенного по $x_1, x_2, x_3, \ldots, x_n$, т.е. i-я порядковая статистика. Таким образом, при $X = R^1$, $f(x, y) = |x - y|$ решение задачи (1) дает естественное определение выборочной медианы. Правда, несколько отличающееся от определения, обычно предлагаемого в курсах «Общей теории статистики», в котором при $n = 2k$ медианой называют полусумму двух центральных членов вариационного ряда $(x(k) + x(k + 1))/2$. Иногда $x(k)$ называют левой медианой, а $x(k + 1)$ — правой медианой [1].

Решением задачи (1) является множество $E_n(f)$, которое может быть пустым, состоять из одного или многих элементов. Выше приведен пример, когда решением является отрезок. Если $X = R^1 \setminus \{x_0\}$, $f(x,y) = (x - y)^2$, а среднее арифметическое выборки равно x_0, то $E_n(f)$ пусто.

При моделировании реальных ситуаций часто можно принять, что X состоит из конечного числа элементов. Тогда множество $E_n(f)$ непусто — минимум на конечном множестве всегда достигается.

Понятия случайного элемента $x = x(\omega)$ со значениями в X, его распределения, независимости случайных элементов используем согласно определениям главы 1, т.е. каноническому справочнику Ю.В. Прохорова и Ю.А. Розанова [2]. Будем считать, что функция f измерима относительно σ-алгебры, участвующей в определении случайного элемента $x = x(\omega)$. Тогда $f(x(\omega), y)$ при фиксированном y является действительнозначной случайной величиной. Предположим, что она имеет математическое ожидание.

Определение 2. Теоретическим средним $E(x,f)$ (другими словами, математическим ожиданием) случайного элемента $x = x(\omega)$ относительно меры различия f называется решение оптимизационной задачи:

$$Mf(x(\omega), y) \to \min, \ y \in X. \tag{2}$$

Это определение, как и для эмпирических средних, согласуется с классическим. Если $X = R^1$, $f(x, y) = (x - y)^2$, то $E(x, f) = M(x(\omega))$ — обычное математическое ожидание. При этом $Mf(x(\omega), y)$ — дисперсия случайной величины $x = x(\omega)$. Если же $X = R^1$, $f(x, y) = |x - y|$, то $E(x, f) = [a, b]$, где $a = \sup\{t: F(t) \le \le 0{,}5\}$, $b = \inf\{t: F(t) \ge 0{,}5\}$, где $F(t)$ — функция распределения случайной величины $x = x(\omega)$. Если график $F(t)$ имеет плоский участок на уровне $F(t) = 0{,}5$, то медиана — теоретическое среднее в смысле определения 2 — является отрезком. В классическом случае обычно говорят, что каждый элемент отрезка $[a; b]$ является одним из возможных значений медианы. Поскольку наличие указанного плоского участка — исключительный случай, то обычно решением задачи (2) является множество из одного элемента $a = b$ — классическая медиана распределения случайной величины $x = x(\omega)$.

Теоретическое среднее $E(x, f)$ можно определить лишь тогда, когда $Mf(x(\omega), y)$ существует при всех $y \in X$. Оно может быть пустым множеством, например, если $X = R^1 \backslash \{x_0\}$, $f(x, y) = (x - y)^2$, $x_0 = M(x(\omega))$. И то, и другое исключается, если X конечно. Однако и для конечных X теоретическое среднее может состоять не из одного, а из многих элементов. Отметим, однако, что в множестве всех распределений вероятностей на X подмножество тех распределений, для которых $E(x, f)$ состоит более чем из одного элемента, имеет ко-

размерность 1, поэтому основной является ситуация, когда множество $E(x,f)$ содержит единственный элемент [1].

Существование средних величин. Под существованием средних величин будем понимать непустоту множеств решений соответствующих оптимизационных задач.

Если X состоит из конечного числа элементов, то минимум в задачах (1) и (2) берется по конечному множеству. А потому, как уже отмечалось, эмпирические и теоретические средние существуют.

Ввиду важности обсуждаемой темы приведем доказательства. Для строгого математического изложения нам понадобятся термины из раздела математики под названием «общая топология». Топологические термины и результаты будем использовать в соответствии с классической монографией [3]. Так, топологическое пространство называется бикомпактным в том и только в том случае, когда из каждого его открытого покрытия можно выбрать конечное подпокрытие [3, с. 183].

Теорема 1. Пусть X — бикомпактное пространство, функция f непрерывна на X^2 (в топологии произведения). Тогда эмпирическое и теоретическое средние существуют.

Доказательство. Функция $f(x_i, y)$ от y непрерывна, сумма непрерывных функций непрерывна, непрерывная функция на бикомпакте достигает своего минимума, откуда и следует заключение теоремы относительно эмпирического среднего.

Перейдем к теоретическому среднему. По теореме Тихонова [3, с. 194] из бикомпактности X вытекает бикомпактность X^2. Для каждой точки (x, y) из X^2 рассмотрим $\varepsilon/2$ — окрестность в X^2 в смысле показателя различия f, т.е. множество:

$$U(x, y) = \{(x', y') : | f(x, y) - f(x', y') | < \varepsilon / 2\}.$$

Поскольку f непрерывна, то множества $U(x, y)$ открыты в рассматриваемой топологии в X^2. По теореме Уоллеса [3, с. 193] существуют открытые (в X) множества $V(x)$ и $W(y)$, содержащие x и y соответственно и такие, что их декартово произведение $V(x) \times W(y)$ целиком содержится внутри $U(x, y)$.

Рассмотрим покрытие X^2 открытыми множествами $V(x) \times W(y)$. Из бикомпактности X^2 вытекает существование конечного подпокрытия $\{V(x_i) \times \times W(y_i),\ i = 1, 2, ..., m\}$. Для каждого x из X рассмотрим все декартовы произведения $V(x_i) \times W(y_i)$, куда входит точка (x, y) при каком-либо y. Таких декар-

товых произведений и их первых множителей $V(x_i)$ конечное число. Возьмем пересечение таких первых множителей $V(x_i)$ и обозначим его $Z(x)$. Это пересечение открыто, как пересечение конечного числа открытых множеств, и содержит точку x. Из покрытия бикомпактного пространства X открытыми множествами $Z(x)$ выберем открытое подпокрытие $Z_1, Z_2, ..., Z_k$.

Покажем, что если x_1' и x_2' принадлежат одному и тому же Z_j при некотором j, то

$$\sup\{|f(x_1', y) - f(x_2', y)|, y \in X\} < \varepsilon. \qquad (3)$$

Пусть $Zj = Z(x_0)$ при некотором x_0. Пусть $V(x_i) \times W(y_i)$, $i \in I$, — совокупность всех тех исходных декартовых произведений из системы $\{V(x_i) \times W(y_i)$, $i = 1, 2, ..., m\}$, куда входят точки (x_0, y) при различных y. Покажем, что их объединение содержит также точки (x_1', y) и (x_2', y) при всех y. Действительно, если (x_0, y) входит в $V(x_i) \times W(y_i)$, то y входит в $W(y_i)$, а x_1' и x_2' вместе с x_0 входят в $V(x_i)$, поскольку x_1', x_2' и x_0 входят в $Z(x_0)$. Таким образом, (x_1', y) и (x_2', y) принадлежат $V(x_i) \times W(y_i)$, а потому согласно определению $V(x_i) \times W(y_i)$:

$$|f(x_1', y) - f(x_i, y_i)| < \varepsilon / 2, \quad |f(x_2', y) - f(x_i, y_i)| < \varepsilon / 2,$$

откуда и следует неравенство (3).

Поскольку X^2 — бикомпактное пространство, то функция f ограничена на X^2, а потому существует математическое ожидание $Mf(x(\omega), y)$ для любого случайного элемента $x(\omega)$, удовлетворяющего приведенным выше условиям согласования топологии, связанной с f, и измеримости, связанной с $x(\omega)$. Если x_1 и x_2 принадлежат одному открытому множеству Z_j, то:

$$|Mf(x_1, y) - Mf(x_2, y)| < \varepsilon,$$

а потому функция:

$$g(y) = Mf(x(\omega), y) \qquad (4)$$

непрерывна на X. Поскольку непрерывная функция на бикомпактном множестве достигает своего минимума, т.е. существуют такие точки z, на которых $g(z) = \inf\{g(y), y \in X\}$, то теорема 1 доказана.

В ряде интересных для приложений ситуаций X не является бикомпактным пространством. Например, если $X = R^1$. В этих случаях приходится наложить на показатель различия f некоторые ограничения, например, так, как это сделано в теореме 2.

Теорема 2. Пусть X — топологическое пространство, непрерывная (в топологии произведения) функция $f: X^2 \to R^1$ неотрицательна, симметрична (т.е. $f(x, y) = f(y, x)$ для любых x и y из X), существует число $D > 0$ такое, что при всех x, y, z из X:

$$f(x, y) \le D\{f(x, z) + f(z, y)\}. \tag{5}$$

Пусть в X существует точка x_0 такая, что при любом положительном R множество $\{x: f(x, x_0) \le R\}$ является бикомпактным. Пусть для случайного элемента $x(\omega)$, согласованного с топологией в рассмотренном выше смысле, существует $g(x_0) = Mf(x(\omega), x_0)$.

Тогда существуют (т.е. непусты) математическое ожидание $E(x, f)$ и эмпирические средние $E_n(f)$.

Замечание. Условие (5) — некоторое обобщение неравенства треугольника. Например, если g — метрика в X, а $f = g^p$ при некотором натуральном p, то для f выполнено соотношение (5) с $D = 2^p$.

Доказательство. Рассмотрим функцию $g(y)$, определенную формулой (4). Имеем:

$$f(x(\omega), y) \le D\ \{f(x(\omega), x_0) + f(x_0, y)\}. \tag{6}$$

Поскольку по условию теоремы $g(x_0)$ существует, а потому конечно, то из оценки (6) следует существование и конечность $g(y)$ при всех y из X. Докажем непрерывность этой функции.

Рассмотрим шар (в смысле меры различия f) радиуса R с центром в x_0:

$$K(R) = \{x : f(x, x_0) \le R\},\ R > 0.$$

В соответствии с условием теоремы $K(R)$ как подпространство топологического пространства X является бикомпактным. Рассмотрим произвольную точку x из X. Справедливо разложение:

$$f(x(\omega), y) = f(x(\omega), y)\chi(x(\omega) \in K(R)) + f(x(\omega), y)\chi(x(\omega) \notin K(R)),$$

где $\chi(C)$ — индикатор множества C. Следовательно,

$$g(y) = Mf(x(\omega), y)\chi(x(\omega) \in K(R)) + Mf(x(\omega), y)\chi(x(\omega) \notin K(R)). \qquad (7)$$

Рассмотрим второе слагаемое в (7). В силу (5):

$$f(x(\omega), y)\chi(x(\omega) \notin K(R)) \le D\{f(x(\omega), x_0)\chi(x(\omega) \notin K(R)) + f(x_0, y)\chi(x(\omega) \notin K(R))\}. \qquad (8)$$

Возьмем математическое ожидание от обеих частей (8):

$$Mf(x(\omega), y)\chi(x(\omega) \in K(R)) \le D\int_R^{+\infty} tdP\{f(x(\omega), x_0) \le t\} + Df(x_0, y)P(x(\omega) \in K(R)). \qquad (9)$$

В правой части (9) оба слагаемых стремятся к 0 при безграничном возрастании R: первое — в силу того, что

$$g(x_0) = Mf(x(\omega), x_0) = \int_0^{+\infty} tdP(f(x(\omega), x_0) \le t) < \infty,$$

второе — в силу того, что распределение случайного элемента $x(\omega)$ сосредоточено на X и

$$X \setminus \bigcup_{R>0} K(R) = \varnothing.$$

Пусть $U(x)$ — такая окрестность x (т.е. открытое множество, содержащее x), для которой

$$\sup\{f(y, x), y \in U(x)\} < +\infty.$$

Имеем:

$$f(y, x_0) \le D(f(x_0, x) + f(x, y)). \tag{10}$$

В силу (9) и (10) при безграничном возрастании R:

$$Mf(x(\omega), y)\chi(x(\omega) \notin K(R)) \to 0 \tag{11}$$

равномерно по $y \in U(x)$. Пусть $R(0)$ таково, что левая часть (11) меньше $\varepsilon > 0$ при $R > R(0)$ и, кроме того, $y \in U(x) \subseteq K(R(0))$. Тогда при $R > R(0)$:

$$|g(y) - g(x)| \le |Mf(x(\omega), y)\chi(x(\omega) \in K(R)) - Mf(x(\omega), x)\chi(x(\omega) \in K(R))| + 2\varepsilon. \tag{12}$$

Нас интересует поведение выражения в правой части формулы (12) при $y \in U(x)$. Рассмотрим f_1 — сужение функции f на замыкание декартова произведения множеств $U(x) \times K(R)$, и случайный элемент $x_1(\omega) = x(\omega)\chi(x(\omega) \in K(R))$.

Тогда

$$Mf(x(\omega), y)\chi(x(\omega) \in K(R)) = Mf_1(x_1(\omega), y)$$

при $y \in U(x)$, а непрерывность функции $g_1(y) = Mf_1(x_1(\omega), y)$ была доказана в теореме 1. Последнее означает, что существует окрестность $U_1(x)$ точки x такая, что

$$|Mf_1(x_1(\omega), y) - Mf_1(x_1(\omega), x)| < \varepsilon \tag{13}$$

при $y \in U_1(x)$. Из (12) и (13) вытекает, что при $y \in U(x) \bigcap U_1(x)$:

$$|g(y) - g(x)| < 3\varepsilon,$$

что и доказывает непрерывность функции $g(x)$.

Докажем существование математического ожидания $E(x,f)$. Пусть $R(0)$ таково, что

$$P(x(\omega) \in K(R(0))) > 1/2. \tag{14}$$

Пусть H — некоторая константа, значение которой будет выбрано позже. Рассмотрим точку x из множества $K(HR(0))^C$ — дополнения $K(HR(0))$, т.е. из внешности шара радиуса $HR(0)$ с центром в x_0. Пусть $x(\omega) \in K(R(0))$. Тогда имеем:

$$f(x_0, x) \le D\{f(x_0, x(\omega)) + f(x(\omega), x)\},$$

откуда

$$f(x(\omega), x) \ge \frac{1}{D} f(x_0, x) - f(x_0, x(\omega)) \ge \frac{HR(0)}{D} - R(0). \qquad (15)$$

Выбирая H достаточно большим, получим с учетом условия (14), что при $x \in K(HR(0))^C$ справедливо неравенство:

$$Mf(x(\omega), x) \ge \frac{1}{2}\left(\frac{HR(0)}{D} - R(0)\right). \qquad (16)$$

Можно выбрать H так, чтобы правая часть (16) превосходила $g(x_0) = Mf(x(\omega), x_0)$.

Сказанное означает, что Argmin $g(x)$ достаточно искать внутри бикомпактного множества $K(HR(0))$. Из непрерывности функции g вытекает, что ее минимум достигается на указанном бикомпактном множестве, а потому — и на всем X. Существование (непустота) теоретического среднего $E(x, f)$ доказана.

Докажем существование эмпирического среднего $E_n(f)$. Есть искушение проводить его дословно так же, как и доказательство существования математического ожидания $E(x, f)$, лишь с заменой 1/2 в формуле (16) на частоту попадания элементов выборки x_i в шар $K(R(0))$. Эта частота, очевидно, стремится к вероятности попадания случайного элемента $x = x(\omega)$ в $K(R(0))$, большей 1/2 в соответствии с (14). Однако это рассуждение показывает лишь, что вероятность непустоты $E_n(f)$ стремится к 1 при безграничном росте объема выборки. Точнее, оно показывает, что

$$\lim_{n\to\infty} P\{E_n(f) \ne \varnothing \wedge E_n(f) \subseteq K(HR(0))\} = 1.$$

Поэтому пойдем другим путем, не опирающимся к тому же на вероятностную модель выборки. Положим:

$$R(1) = \max\{f(x_i, x_0), i = 1,2,\ldots,n\}. \qquad (17)$$

Если x входит в дополнение шара $K(HR(1))$, то аналогично (15) имеем:

$$f(x_i, x_0) \geq \frac{HR(1)}{D} - R(1). \tag{18}$$

При достаточно большом H из (17) и (18) следует, что

$$\sum_{i=1}^{n} f(x_i, x_0) \leq nR(1) < \sum_{i=1}^{n} f(x_i, x), \quad x \in \{K(HR(1))\}^C.$$

Следовательно, Argmin достаточно искать на $K(HR(1))$. Заключение теоремы 2 следует из того, что на бикомпактном пространстве $K(HR(1))$ минимизируется непрерывная функция.

Теорема 2 полностью доказана. Перейдем к законам больших чисел.

2.2. ЗАКОНЫ БОЛЬШИХ ЧИСЕЛ

О формулировках законов больших чисел. Пусть $x, x_1, x_2, x_3, ..., x_n$ — независимые одинаково распределенные случайные элементы со значениями в X. Закон больших чисел — это утверждение о сходимости эмпирических средних к теоретическому среднему (математическому ожиданию) при росте объема выборки n, т.е. утверждение о том, что

$$E_n(f) = E_n(x_1, x_2, x_3, ..., x_n; f) \to E(x, f) \tag{1}$$

при $n \to \infty$. Однако и слева, и справа в формуле (1) стоят, вообще говоря, множества. Поэтому понятие сходимости в (1) требует обсуждения и определения.

В силу классического закона больших чисел при $n \to \infty$:

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i, y) \to Mf(x, y) \tag{2}$$

в смысле сходимости по вероятности, если правая часть существует (теорема А. Я. Хинчина, 1923 г.).

Если пространство X состоит из конечного числа элементов, то из соотношения (2) легко вытекает (см., например, [1, с. 192–193]), что

$$\lim_{n\to\infty} P\{E_n(f) \subseteq E(x, f)\} = 1. \tag{3}$$

Другими словами, $E_n(f)$ является состоятельной оценкой $E(x, f)$.

Если $E(x, f)$ состоит из одного элемента, $E(x, f) = \{x_0\}$, то соотношение (3) переходит в следующее:

$$\lim_{n\to\infty} P\{E_n(f) = \{x_0\}\} = 1. \qquad (4)$$

Однако с прикладной точки зрения доказательство соотношений (3)–(4) не дает достаточно уверенности в возможности использования $E_n(f)$ в качестве оценки $E(x, f)$. Причина в том, что в процессе доказательства объем выборки предполагается настолько большим, что при всех $y \in X$ одновременно левые части соотношений (2) сосредотачиваются в непересекающихся окрестностях правых частей.

Замечание. Если в соотношении (2) рассмотреть сходимость с вероятностью 1, то аналогично (3) получим так называемый усиленный закон больших чисел [1, с. 193–194]. Согласно этой теореме с вероятностью 1 эмпирическое среднее $E_n(f)$ входит в теоретическое среднее $E(x, f)$, начиная с некоторого объема выборки n, вообще говоря, случайного, $n = n(\omega)$. Мы не будем останавливаться на сходимости с вероятностью 1, поскольку в соответствующих постановках, подробно разобранных в монографии [1], нет принципиальных отличий от случая сходимости по вероятности.

Если X не является конечным, например, $X = R^1$, то соотношения (3) и (4) неверны. Поэтому необходимо искать иные формулировки закона больших чисел. В классическом случае сходимости выборочного среднего арифметического к математическому ожиданию, т.е. $\bar{x} \to M(x)$, можно записать закон больших чисел так: для любого $\varepsilon > 0$ справедливо предельное соотношение:

$$\lim_{n\to\infty} P\{\bar{x} \in (M(x) - \varepsilon;\ M(x) + \varepsilon)\} = 1. \qquad (5)$$

В этом соотношении в отличие от (3) речь идет о попадании эмпирического среднего $E_n(f) = \bar{x}$ не непосредственно внутрь теоретического среднего $E(x, f)$, а в некоторую *окрестность* теоретического среднего.

Обобщим эту формулировку. Как задать окрестность теоретического среднего в пространстве произвольной природы? Естественно взять его окрестность, определенную с помощью какой-либо метрики. Однако полезно обеспечить на ее дополнении до X *отделенность* множества значений $Mf(x(\omega), y)$ как функции y от минимума этой функции на всем X.

Поэтому мы сочли целесообразным определить такую окрестность с помощью самой функции $Mf(x(\omega),y)$.

Определение 1. Для любого $\varepsilon > 0$ назовем ε-пяткой функции $g(x)$ множество:

$$K_\varepsilon(g) = \{x: g(x) < \inf\{ g(y), y \in X\} + \varepsilon, x \in X\}.$$

Таким образом, в ε-пятку входят все те x, для которых значение $g(x)$ либо минимально, либо отличается от минимального (или от инфимума — точной нижней грани) не более чем на ε. Так, для $X = R^1$ и функции $g(x) = x^2$ минимум равен 0, а ε-пятка имеет вид интервала $(-\sqrt{\varepsilon}; \sqrt{\varepsilon})$. В формулировке (5) классического закона больших чисел утверждается, что при любом $\varepsilon > 0$ вероятность попадания среднего арифметического в ε^2-пятку математического ожидания стремится к 1. Поскольку $\varepsilon > 0$ произвольно, то вместо ε^2-пятки можно говорить о ε-пятке, т.е. перейти от (5) к эквивалентной записи:

$$\lim_{n\to\infty} P\left\{x \in K_\varepsilon(M(x(\omega) - x)^2)\right\} = 1. \qquad (6)$$

Соотношение (6) допускает непосредственное обобщение на общий случай пространств произвольной природы.

Схема закона больших чисел. Пусть $x, x_1, x_2, x_3, ..., x_n$ — независимые одинаково распределенные случайные элементы со значениями в пространстве произвольной природы X с показателем различия f: $X^2 \to R^1$. *Пусть выполнены некоторые математические условия регулярности.* Тогда для любого $\varepsilon > 0$ справедливо предельное соотношение:

$$\lim_{n\to\infty} P\{E_n(f) \subseteq K_\varepsilon(E(x, f))\} = 1. \qquad (7)$$

Аналогичным образом может быть сформулирована и общая идея усиленного закона больших чисел. Ниже приведены две конкретные формулировки «условий регулярности».

Законы больших чисел. Начнем с рассмотрения естественного обобщения конечного множества — бикомпактного пространства X.

Теорема 1. В условиях теоремы 1 раздела 2.1 справедливо соотношение (7).

Доказательство. Воспользуемся построенным при доказательстве теоремы 1 раздела 2.1 конечным открытым покрытием $\{Z_1, Z_2, ..., Z_k\}$ пространства X таким, что для него выполнено соотношение (3) раздела 2.1. Построим на его основе разбиение X на непересекающиеся множества $W_1, W_2, ..., W_m$ (объединение элементов разбиения $W_1, W_2, ..., W_m$ составляет X). Это можно сделать итеративно. На первом шаге из Z_1 следует вычесть $Z_2, ..., Z_k$ — это и будет W_1. Затем в качестве нового пространства надо рассмотреть разность X и W_1, а покрытием его будет $\{Z_2, ..., Z_k\}$. И так до k-го шага, когда последнее из рассмотренных покрытий будет состоять из единственного открытого множества Z_k. Остается из построенной последовательности $W_1, W_2, ..., W_k$ вычеркнуть пустые множества, которые могли быть получены при осуществлении описанной процедуры (поэтому, вообще говоря, m может быть меньше k).

В каждом из элементов разбиения $W_1, W_2, ..., W_m$ выберем по одной точке, которые назовем центрами разбиения и соответственно обозначим $w_1, w_2, ..., w_m$. Это и есть то конечное множество, которым можно аппроксимировать бикомпактное пространство X. Пусть y входит в W_j. Тогда из соотношения (3) раздела 2.1 вытекает, что:

$$\left| \frac{1}{n} \sum_{i=1}^{n} f(x_i, y) - \frac{1}{n} \sum_{i=1}^{n} f(x_i, w_j) \right| < \varepsilon. \tag{8}$$

Перейдем к доказательству соотношения (7). Возьмем произвольное $\delta > 0$. Рассмотрим некоторую точку b из $E(x, f)$. Доказательство будет основано на том, что с вероятностью, стремящейся к 1, для любого y вне $K_\delta(E(x, f))$ выполнено неравенство:

$$\frac{1}{n} \sum_{i=1}^{n} f(x_i, y) > \frac{1}{n} \sum_{i=1}^{n} f(x_i, b). \tag{9}$$

Для обоснования этого неравенства рассмотрим все элементы разбиения $W_1, W_2, ..., W_m$, имеющие непустое пересечение с внешностью δ-пятки $K_\delta(E(x, f))$. Из неравенства (8) следует, что для любого y вне $K_\delta(E(x, f))$ левая часть неравенства (9) не меньше:

$$\min_j \left(\frac{1}{n} \sum_{i=1}^{n} f(x_i, w_j) \right) - \varepsilon, \tag{10}$$

где минимум берется по центрам всех элементов разбиения, имеющим непустое пересечение с внешностью δ-пятки. Возьмем теперь в каждом таком разбиении точку v_i, лежащую вне δ-пятки $K_\delta(E(x,f))$. Тогда из неравенств (3) раздела 2.1 и (10) следует, что левая часть неравенства (9) не меньше:

$$\min_j \left(\frac{1}{n} \sum_{i=1}^{n} f(x_i, v_j) \right) - 2\varepsilon. \qquad (11)$$

В силу закона больших чисел для действительнозначных случайных величин каждая из участвующих в соотношениях (9) и (11) средних арифметических имеет своими пределами соответствующие математические ожидания, причем в соотношении (11) эти пределы не менее:

$$Mf(x(\omega), b) + \delta - 2\varepsilon,$$

поскольку точки v_i лежат вне δ-пятки $K_\delta(E(x,f))$. Следовательно, при

$$\delta - 2\varepsilon > 0$$

и достаточно большом n, обеспечивающем необходимую близость рассматриваемого конечного числа средних арифметических к их математическим ожиданиям, справедливо неравенство (9).

Из неравенства (9) следует, что пересечение $E_n(f)$ с внешностью $K_\delta(E(x,f))$ пусто. При этом точка b может входить в $E_n(f)$, а может и не входить. Во втором случае $E_n(f)$ состоит из иных точек, входящих в $K_\delta(E(x,f))$. Теорема 1 доказана.

Если X не является бикомпактным пространством, то необходимо суметь оценить рассматриваемые суммы «на периферии», вне бикомпактного ядра, которое обычно выделяется естественным путем. Один из возможных комплексов условий сформулирован выше в теореме 2 раздела 2.1.

Теорема 2. В условиях теоремы 2 раздела 2.1 справедлив закон больших чисел, т.е. соотношение (25).

Доказательство. Будем использовать обозначения, введенные в теореме 2 раздела 2.1 и при ее доказательстве. Пусть r и R, $r < R$ — положительные числа. Рассмотрим точку x в шаре $K(r)$ и точку y вне шара $K(R)$. Поскольку:

$$f(x_0, y) \le D\{f(x_0, x) + f(x, y)\},$$

то

$$f(x,y) \geq \frac{1}{D} f(x_0,y) - f(x_0,x) \geq \frac{R}{D} - r. \qquad (12)$$

Положим:

$$g_n(x) = g_n(x,\omega) = \frac{1}{n}\sum_{i=1}^{n} f(x_i,x).$$

Сравним $g_n(x_0)$ и $g_n(y)$. Выборку $x_1, x_2, x_3, \ldots, x_n$ разобьем на две части. В первую часть включим те элементы выборки, которые входят в $K(r)$, во вторую — все остальные (т.е. лежащие вне $K(r)$). Множество индексов элементов первой части обозначим $I = I(n, r)$. Тогда в силу неотрицательности f имеем:

$$g_n(y) \geq \frac{1}{n}\sum_{i\in I} f(x_i,y),$$

а в силу неравенства (12):

$$\sum_{i\in I} f(x_i,y) \geq \left(\frac{R}{D} - r\right) CardI(n,r),$$

где $Card\ I(n, r)$ — число элементов в множестве индексов $I(n,r)$. Следовательно,

$$g_n(y) \geq \frac{1}{n}\left(\frac{R}{D} - r\right) J, \qquad (13)$$

где $J = Card\ I(n,r)$ — биномиальная случайная величина $B(n,p)$ с вероятностью успеха $p = P\{x_i(\omega) \in K(r)\}$. По теореме Хинчина для $g_n(x_0)$ справедлив (классический) закон больших чисел. Пусть $\varepsilon > 0$. Выберем $n_1 = n_1(\varepsilon)$ так, чтобы при $n > n_1$ было выполнено соотношение:

$$P\{g_n(x_0) - g(x_0) > \varepsilon\} < \varepsilon, \qquad (14)$$

где $g(x_0) = Mf(x_1,x_0)$. Выберем r так, чтобы вероятность успеха $p > 0{,}6$. По теореме Бернулли можно выбрать $n_2 = n_2(\varepsilon)$ так, чтобы при $n > n_2$

$$P\{J > 0{,}5n\} > 1 - \varepsilon. \qquad (15)$$

Выберем R так, чтобы

$$\frac{1}{2}\left(\frac{R}{D}-r\right)>g(x_0)+\varepsilon.$$

Тогда

$$K_\varepsilon(g)\subseteq K(R) \tag{16}$$

и согласно (13), (14) и (15) при $n>n_3=\max(n_1,n_2)$ с вероятностью не менее $1-\varepsilon$ имеем:

$$g_n(y)>g_n(x_0) \tag{17}$$

для любого y вне $K(R)$. Из (16) следует, что минимизировать g_n достаточно внутри бикомпактного шара $K(R)$, при этом $E_n(f)$ не пусто и

$$E_n(f)\subseteq K(R) \tag{18}$$

с вероятностью не менее $1-2\varepsilon$.

Пусть $g_n^{'}$ и $g^{'}$ — сужения g_n и $g(x) = Mf(x(\omega), x)$ соответственно на $K(R)$ как функций от x. В силу (16) справедливо равенство $K_\varepsilon(g^{'})=K_\varepsilon(f)$. Согласно доказанной выше теореме 1 найдется $n_4=n_4(\omega)$ такое, что при $n>n_4$:

$$P(K_0(g_n^{'})\subseteq K_\varepsilon(g))>1-\varepsilon.$$

Согласно (18) с вероятностью не менее $1-2\varepsilon$:

$$K_0(g_n^{'})=E_n(f)$$

при $n>n_3$. Следовательно, при $n>n_5(\varepsilon)=\max(n_3,n_4)$ имеем:

$$P(E_n(f)\subseteq K_\varepsilon(g))>1-3\varepsilon,$$

что и завершает доказательство теоремы 2.

Справедливы и иные варианты законов больших чисел, полученные, в частности, в статье [4]. Разберем важный для прикладных исследований пример.

Медиана Кемени и экспертные оценки. Рассмотрим на основе развитой выше теории частный случай пространств нечисловой природы — пространство бинарных отношений на конечном множестве $Q = \{q_1, q_2, ..., q_k\}$ и его подпространства. Как известно, каждое бинарное отношение A можно описать матрицей $||a(i,j)||$ из 0 и 1, причем $a(i,j) = 1$ тогда и только тогда q_i и q_j находятся в отношении A, и $a(i,j) = 0$ в противном случае.

Определение 2. Расстоянием Кемени между бинарными отношениями A и B, описываемыми матрицами $||a(i, j)||$ и $||b(i, j)||$ соответственно, называется:

$$d(A, B) = \sum_{i,j=1}^{k} | a(i, j) - b(i, j) | .$$

Замечание. Иногда в определение расстояния Кемени вводят множитель, зависящий от k.

Определение 3. Медианой Кемени для выборки, состоящей из бинарных отношений, называется эмпирическое среднее, построенное с помощью расстояния Кемени (минимум берется по соответствующему подпространству).

Поскольку число бинарных отношений на конечном множестве конечно, то эмпирические и теоретические средние для произвольных показателей различия существуют и справедливы законы больших чисел, описанные формулами (3) и (4) выше.

Бинарные отношения, в частности, упорядочения, часто используются для описания мнений экспертов. Тогда расстояние Кемени измеряет близость мнений экспертов, а медиана Кемени позволяет находить итоговое усредненное мнение комиссии экспертов. Расчет медианы Кемени обычно включают в информационное обеспечение систем принятия решений с использованием оценок экспертов. Речь идет, например, о математическом обеспечении автоматизированного рабочего места «Математика в экспертизе» (АРМ «МАТЭК»), предназначенного, в частности, для использования при проведении экспертиз в задачах экологического страхования. Поэтому представляет большой практический интерес численное изучение свойств медианы Кемени

при конечном объеме выборки. Такое изучение дополняет описанную выше асимптотическую теорию, в которой объем выборки предполагается безгранично возрастающим ($n \to \infty$).

Компьютерное изучение свойств медианы Кемени при конечных объемах выборок. С помощью специально разработанной программной системы В. Н. Жихарев провел ряд серий численных экспериментов по изучению свойств выборочных медиан Кемени (в пространстве ранжировок без связей). Представление о полученных результатах дается таблицей 1, взятой из статьи [5]. В каждой серии методом статистических испытаний определенное число раз моделировался случайный и независимый выбор экспертных ранжировок, а затем находились все медианы Кемени для смоделированного набора мнений экспертов. При этом в сериях 1–5 распределение ответа эксперта предполагалось равномерным на множестве всех ранжировок. В серии 6 это распределение являлось монотонным относительно расстояния Кемени с некоторым центром (о понятии монотонности см. главу 1), т.е. вероятность выбора определенной ранжировки убывала с увеличением расстояния Кемени этой ранжировки от центра. Таким образом, серии 1–5 соответствуют ситуации, когда у экспертов нет почвы для согласия, нет группировки их мнений относительно некоторого единого среднего группового мнения, в то время как в серии 6 есть единое мнение — описанный выше центр, к которому тяготеют ответы экспертов.

Результаты, приведенные в табл. 1, можно комментировать разными способами. Неожиданным явилось большое число элементов в выборочной медиане Кемени — как среднее, так и особенно максимальное. Одновременно обращает на себя внимание убывание этих чисел при росте числа экспертов и особенно при переходе к ситуации реального существования группового мнения (серия 6). Достаточно часто один из ответов экспертов входит в медиану Кемени (т.е. пересечение множества ответов экспертов и медианы Кемени непусто), а диаметр медианы как множества в пространстве ранжировок заметно меньше диаметра множества ответов экспертов. По этим показателям — наилучшее положение в серии 6. Грубо говоря, всяческие «патологии» в поведении медианы Кемени наиболее резко проявляются в ситуации, когда ее применение не имеет содержательного обоснования, т.е. когда у экспертов нет основы для согласия, их ответы равномерно распределены на множестве ранжировок.

Таблица 1

Вычислительный эксперимент
по изучению медианы Кемени

Номер серии	1	2	3	4	5	6
Число испытаний	100	1 000	50	50	1 000	1 000
Количество объектов	5	5	7	7	5	5
Количество экспертов	10	30	10	30	10	10
Частота непустого пересечения	0,85	0,58	0,52	0,2	0,786	0,911
Среднее отношение диаметров	0.283	0,124	0,191	0,0892	0,202	0.0437
Средняя мощность медианы	5,04	2,41	6,4	2,88	3,51	1,35
Максимальная мощность медианы	30	14	19	11	40	12

Увеличение числа испытаний в 10 раз при переходе от серии 1 к серии 5 не очень сильно повлияло на приведенные в таблице характеристики, поэтому представляется, что суть дела выявляется при числе испытаний (в методе Монте-Карло), равном 100 или даже 50. Увеличение числа объектов или экспертов увеличивает число элементов в рассматриваемых пространствах ранжировок, а потому уменьшается частота попадания какого-либо из мнений экспертов внутрь медианы Кемени. А также отношение диаметра медианы к диаметру множества экспертов и число элементов медианы Кемени (среднее и максимальное). Можно сказать так: увеличение числа объектов или экспертов уменьшает степень дискретности задачи, приближает ее к непрерывному случаю, а потому уменьшает выраженность различных «патологий».

Есть много интересных направлений исследований, которые здесь не рассматриваем. Они связаны, в частности, со сравнением медианы Кемени с другими методами усреднения мнений экспертов, например, с нахождением итогового упорядочения по методу средних рангов [6]. А также с использованием малых окрестностей ответов экспертов для поиска входящих в медиану ранжировок (с целью сокращения расчетов). Или с построением теоретических и численных оценок скорости сходимости в законах больших чисел.

2.3. ЭКСТРЕМАЛЬНЫЕ СТАТИСТИЧЕСКИЕ ЗАДАЧИ

Если проанализировать приведенные выше в разделах 2.1 и 2.2 постановки и результаты, касающиеся эмпирических и теоретических средних и

законов больших чисел, то становится очевидной возможность их обобщения. Так, доказательства теорем практически не меняются, если считать, что функция $f(x, y)$ определена на декартовом произведении бикомпактных пространств X и Y, а не на X^2. Тогда можно считать, что элементы выборки лежат в X, а Y — пространство параметров, подлежащих оценке.

Обобщения законов больших чисел. Пусть, например, выборка $x_1 = x_1(\omega)$, $x_2 = x_2(\omega)$, ..., $x_n = x_n(\omega)$ взята из распределения с плотностью $p(x, y)$, где y — неизвестный параметр. Если положить:

$$f(x, y) = -\ln p(x,y),$$

то задача нахождения эмпирического среднего:

$$f_n(\omega, y) = \frac{1}{n}\sum_{k=1}^{n} f(x_k(\omega), y) \to \min$$

переходит в задачу оценивания неизвестного параметра y методом максимального правдоподобия:

$$\sum_{k=1}^{n} \ln\, p(x_k(\omega), y) \to \max\,.$$

Соответственно законы больших чисел переходят в утверждения о состоятельности этих оценок в случае пространств X и Y общего вида. При такой интерпретации функция $f(x, y)$ уже не является расстоянием или показателем различия. Однако для доказательства сходимости оценок к соответствующим значениям параметров это и не требуется. Достаточно непрерывности этой функции на декартовом произведении бикомпактных пространств X и Y.

В случае функции $f(x, y)$ общего вида можно говорить об определении в пространствах произвольной природы аналогов оценок минимального контраста, достаточно хорошо изученных в классической математической статистике, и о состоятельности таких оценок. Пусть при каждом конкретном значении параметра y справедливо предельное соотношение:

$$f_n(\omega, y) = \frac{1}{n}\sum_{k=1}^{n} f(x_k(\omega), y) \to Mf(x_1(\omega), y) = g(y),$$

где f — функция контраста. Тогда состоятельность оценок минимального контраста вытекает из справедливости предельного перехода:

$$Arg\min\left\{\frac{1}{n}\sum_{k=1}^{n} f(x_k(\omega), y)\right\} \to Arg\min\{Mf(x_1(\omega), y)\}.$$

Частными случаями оценок минимального контраста являются, устойчивые (робастные) оценки Тьюки — Хубера [1, 6–9], а также оценки параметров в задачах аппроксимации (параметрической регрессии) в пространствах произвольной природы (см. ниже раздел 2.7).

Можно пойти и дальше в обобщении законов больших чисел. Пусть известно, что при каждом конкретном y при безграничном росте n имеет быть сходимость по вероятности:

$$f_n(\omega, y) \to f(y),$$

где $f_n(\omega, y)$ — последовательность случайных функций на пространстве Y, а $f(y)$ — некоторая функция на Y. В каких случаях и в каком смысле имеет место сходимость:

$$Argmin\ \{f_n(\omega, y), y \in X\} \to Argmin\ \{f(y), y \in X\}?$$

Другими словами, когда из поточечной сходимости функций вытекает сходимость точек минимума?

Причем здесь можно под n понимать натуральное число. А можно рассматривать сходимость по направленному множеству (см. прил. 1), или же, что практически то же самое — «сходимость по фильтру» в смысле Картана и Бурбаки [3, с. 118]. В частности, можно описывать ситуацию вектором, координаты которого — объемы нескольких выборок, и все они безгранично растут. В классической математической статистике такие постановки рассматривать не любят, поскольку без использования понятия направленного множества трудно строго описать подобный предельный переход.

Поскольку, как хорошо известно, основные задачи прикладной статистики можно представить в виде оптимизационных задач, то ответ на поставленный вопрос о сходимости точек минимума дает возможность единообразного подхода к изучению асимптотики решений разнообразных экстремальных ста-

тистических задач. Одна из возможных формулировок, основанная на бикомпактности пространств X и Y и нацеленная на изучение оценок минимального контраста, дана и обоснована выше. Другой подход развит в работе [4]. Он основан на использовании понятий асимптотической равномерной разбиваемости и координатной асимптотической равномерной разбиваемости пространств. С помощью указанных подходов удается стандартным образом обосновывать состоятельность оценок характеристик и параметров в основных задачах прикладной статистики.

Рассматриваемую тематику можно развивать дальше, в частности, рассматривать аналоги законов больших чисел в случае пространств, не являющихся бикомпактными, а также изучать *скорость* сходимости $Argmin\{f_n(x(\omega), y), y \in X\}$ к $Argmin\{f(y), y \in X\}$.

Примеры применения результатов о предельном поведении точек минимума приведены ниже. В частности, экстремальный вид имеют параметрические задачи восстановления зависимостей, в том числе задачи оценивания информативных подмножеств признаков (раздел 2.7). Ряд методов классификации основан на решении оптимизационных задач, например, так ищут оптимальное разбиение пространства и «центры» кластеров (раздел 2.8). При снижении размерности пространства с целью сжатия информации, в частности, методами главных компонент, метрического и неметрического многомерного шкалирования необходимо решать экстремальные статистические задачи рассмотренного выше вида (раздел 2.9).

2.4. ОДНОШАГОВЫЕ ОЦЕНКИ

В прикладной статистике используются разнообразные параметрические модели. Термин «параметрический» означает, что вероятностно-статистическая модель полностью описывается конечномерным вектором фиксированной размерности. Причем эта размерность не зависит от объема выборки.

Рассмотрим выборку $x_1, x_2, ..., x_n$ из распределения с плотностью $f(x; \theta_0)$, где $f(x; \theta_0)$ — элемент параметрического семейства плотностей распределения вероятностей $\{f(x; \theta), \theta \in \Theta\}$. Здесь Θ — заранее известное k-мерное пространство параметров, являющееся подмножеством евклидова пространства R^k, а конкретное значение параметра θ_0 статистику неизвестно. Обычно в прикладной статистике применяются параметрические семейства с $k = 1, 2, 3$ (см. главу 1.2 в [54]). В статистике нечисловых данных вместо плотности часто

рассматриваются вероятности попадания в точки. Напомним, что в параметрических задачах оценивания принимают вероятностную модель, согласно которой результаты наблюдений $x_1, x_2, ..., x_n$ рассматривают как реализации n независимых случайных величин (векторов, элементов произвольных пространств).

Задача оценивания состоит в том, чтобы оценить неизвестное статистику значение параметра θ_0 наилучшим (в каком-либо смысле) образом.

Выбор «наилучших» в каком-либо смысле оценок в определенной параметрической модели прикладной статистики — научно-исследовательская работа, растянутая во времени. Выделим два этапа. *Этап асимптотики*: оценки строятся и сравниваются по их свойствам при безграничном росте объема выборки. На этом этапе рассматривают такие характеристики оценок, как состоятельность, асимптотическая эффективность и др. *Этап конечных объемов выборки:* оценки сравниваются, скажем, при $n = 10$. Ясно, что исследование начинается с этапа асимптотики: чтобы сравнивать оценки, надо сначала их построить и быть уверенными, что они не являются абсурдными (такую уверенность дает доказательство состоятельности).

С какой оценки начинать? Одним из наиболее известных и простых в употреблении методов является метод моментов. Название связано с тем, что этот метод опирается на использование выборочных моментов. Они приравниваются теоретическим моментам, выраженным в виде гладких функций от параметров. Решением этой системы уравнений является вектор оценок метода моментов, координаты которого являются функциями от выборочных моментов. Обычно оценки метода моментов легко вычисляются. Однако они, как правило, не являются наилучшими. Обычно существуют другие оценки, дисперсия которых при любых значениях параметров меньше, чем для оценок метода моментов. Таковы одношаговые оценки и оценки максимального правдоподобия. Рассмотрим их.

Оценки максимального правдоподобия. В работах, предназначенных для первоначального знакомства с математической статистикой, обычно рассматривают оценки максимального правдоподобия (сокращенно ОМП):

$$\theta_0(n) = \theta_0(n; x_1, x_2, ..., x_n) = Arg \min_{\theta \in \Theta} \prod_{i=1}^{n} f(x_i; \theta). \qquad (1)$$

Таким образом, сначала строится плотность распределения вероятностей, соответствующая выборке. Поскольку элементы выборки независимы,

то эта плотность представляется в виде произведения плотностей для отдельных элементов выборки. Совместная плотность рассматривается в точке, соответствующей наблюденным значениям. Это выражение как функция от параметра (при заданных элементах выборки) называется функцией правдоподобия. Затем тем или иным способом ищется значение параметра, при котором значение совместной плотности максимально. Это и есть оценка максимального правдоподобия.

Хорошо известно, что оценки максимального правдоподобия входят в класс наилучших асимптотически нормальных оценок (определение дано ниже). Однако при конечных объемах выборки в ряде задач ОМП недопустимы, т.к. они хуже (дисперсия и средний квадрат ошибки больше), чем другие оценки, в частности, несмещенные [10]. Именно поэтому в ГОСТ 11.010-81 (в настоящее время отменен как нормативный документ, но может использоваться как научная публикация) для оценивания параметров отрицательного биномиального распределения используются несмещенные оценки, а не ОМП [11]. Из сказанного следует, что априорно предпочитать ОМП другим видам оценок можно — если можно — лишь на этапе изучения асимптотического поведения оценок.

В отдельных случаях ОМП находятся явно, в виде конкретных формул, пригодных для вычисления.

Пример 1. Найдем ОМП для выборки из нормального распределения, каждый элемент которой имеет плотность:

$$f(x;m,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\}.$$

Таким образом, надо оценить двумерный параметр (m, σ^2).

Произведение плотностей вероятностей для элементов выборки, т.е. функция правдоподобия, имеет вид:

$$H(m;\sigma^2)=\sigma^{-n}(2\pi)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-m)^2\right\}. \qquad (2)$$

Требуется решить задачу оптимизации:

$$H(m;\sigma^2)\to\max.$$

Как и во многих иных случаях, задача оптимизации проще решается, если прологарифмировать функцию правдоподобия, т.е. перейти к функции:

$$h(m;\sigma^2) = \ln H(m;\sigma^2),$$

называемой логарифмической функцией правдоподобия. Для выборки из нормального распределения:

$$h(m;\sigma^2) = (-n)\ln\sigma + \left(-\frac{n}{2}\right)\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2. \qquad (3)$$

Необходимым условием максимума является равенство 0 частных производных от логарифмической функции правдоподобия по параметрам, т.е.

$$\frac{\partial h(m,\sigma^2)}{\partial m} = 0, \quad \frac{\partial h(m,\sigma^2)}{\partial(\sigma^2)} = 0. \qquad (4)$$

Система (4) называется системой уравнений максимального правдоподобия. В общем случае число уравнений равно числу неизвестных параметров, а каждое из уравнений выписывается путем приравнивания 0 частной производной логарифмической функции правдоподобия по тому или иному параметру.

При дифференцировании по m первые два слагаемых в правой части формулы (3) обращаются в 0, а последнее слагаемое дает уравнение:

$$\frac{\partial}{\partial m}\sum_{i=1}^{n}(x_i - m)^2 = \sum_{i=1}^{n}2(x_i - m)(-1) = 0, \quad \sum_{i=1}^{n}x_i = nm\ .$$

Следовательно, оценкой m^* максимального правдоподобия параметра m является выборочное среднее арифметическое,

$$m^* = \bar{x}\ .$$

Для нахождения оценки дисперсии необходимо решить уравнение:

$$\frac{\partial}{\partial(\sigma^2)}h(m;\sigma^2) = \frac{\partial}{\partial(\sigma^2)}(-n)\ln\sqrt{\sigma^2} - \frac{\partial}{\partial(\sigma^2)}\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2 = 0.$$

Легко видеть, что

$$\frac{\partial}{\partial(\sigma^2)}(-n)\ln\sqrt{\sigma^2}=\frac{(-n)}{2\sigma^2},\ -\frac{\partial}{\partial(\sigma^2)}\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-m)^2=\frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-m)^2.$$

Следовательно, оценкой $(\sigma^2)^*$ максимального правдоподобия для дисперсии σ^2 с учетом найденной ранее оценки для параметра m является выборочная дисперсия,

$$(\sigma^2)^*=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2.$$

Итак, система уравнений максимального правдоподобия решена аналитически, ОМП для математического ожидания и дисперсии нормального распределения — это выборочное среднее арифметическое и выборочная дисперсия. Отметим, что последняя оценка является смещенной.

Отметим, что в условиях примера 1 оценки метода максимального правдоподобия совпадают с оценками метода моментов. Причем вид оценок метода моментов очевиден и не требует проведения каких-либо рассуждений.

В большинстве случаев аналитических решений не существует, для нахождения ОМП необходимо применять численные методы. Так обстоит дело, например, с выборками из гамма-распределения или распределения Вейбулла — Гнеденко. Во многих работах по нечисловой статистике каким-либо итерационным методом решают систему уравнений максимального правдоподобия ([12] и др.) или впрямую максимизируют функцию правдоподобия типа (8) (см. [13] и др.).

Однако применение численных методов порождает многочисленные проблемы. Сходимость итерационных методов требует обоснования. В ряде примеров функция правдоподобия имеет много локальных максимумов, а потому естественные итерационные процедуры не сходятся [14]. Для данных ВНИИ железнодорожного транспорта по усталостным испытаниям стальным образцов уравнение максимального правдоподобия имеет 11 корней [15]. Какой из одиннадцати использовать в качестве оценки параметра?

Как следствие осознания указанных трудностей, стали появляться работы по доказательству сходимости алгоритмов расчета ОМП для конкрет-

ных вероятностных моделей и конкретных алгоритмов. Примером является статья [16], посвященная одному из разделов нечисловой статистики.

Однако теоретическое доказательство сходимости итерационного алгоритма — это еще не всё. Возникает вопрос об обоснованном выборе момента прекращения вычислений в связи с достижением требуемой точности. В большинстве случаев он не решен.

Но и это не все. Точность вычислений необходимо увязывать с объемом выборки — чем он больше, тем точнее надо находить оценки параметров, в противном случае нельзя говорить о состоятельности метода оценивания. Более того, при увеличении объема выборки необходимо увеличивать и количество используемых в компьютере разрядов, например, переходить от одинарной точности расчетов к двойной, — опять-таки ради достижения состоятельности оценок.

Таким образом, при отсутствии явных формул для оценок максимального правдоподобия нахождение ОМП натыкается на ряд проблем, так сказать, вычислительного характера. Специалисты по математической статистике обычно позволяют себе игнорировать все эти проблемы, рассуждая об ОМП в теоретическом плане. Однако прикладная статистика не может их игнорировать. Отмеченные проблемы ставят под вопрос целесообразность практического использования ОМП.

Нет необходимости абсолютизировать ОМП. Кроме них, существуют другие виды оценок, обладающих хорошими статистическими свойствами. Примером являются одношаговые оценки (ОШ-оценки).

В прикладной статистике разработано много видов оценок. Упомянем квантильные оценки. Они основаны на идее, аналогичной методу моментов, но только вместо выборочных и теоретических моментов приравниваются выборочные и теоретические квантили. Другая группа оценок базируется на идее минимизации расстояния (показателя различия) между эмпирическими данными и элементом параметрического семейства. В простейшем случае минимизируется евклидово расстояние между эмпирическими и теоретическими гистограммами, а точнее, векторами, составленными из высот столбиков гистограмм.

Одношаговые оценки. Одношаговые оценки (ОШ-оценки, ОШО) имеют столь же хорошие асимптотические свойства, что и оценки максимального правдоподобия, при тех же условиях регулярности, что и ОМП. Грубо говоря, они представляют собой результат первой итерации при решении системы уравнений максимального правдоподобия по методу Нью-

тона — Ватсона. Одношаговые оценки выписываются в виде явных формул, а потому требуют существенно меньше машинного времени, а также могут применяться при ручном счете (на калькуляторах). Снимаются вопросы о сходимости алгоритмов, о выборе момента прекращения вычислений, о влиянии округлений при вычислениях на окончательный результат. ОШ-оценки были использованы нами при разработке ГОСТ 11.011-83 (в настоящее время отменен как нормативный документ, но может использоваться как научная публикация) [17] вместо ОМП.

Как и раньше, рассмотрим выборку $x_1, x_2, ..., x_n$ из распределения с плотностью $f(x; \theta_0)$, где $f(x; \theta_0)$ — элемент параметрического семейства плотностей распределения вероятностей $\{f(x;\theta), \theta \in \Theta\}$. Здесь Θ — известное статистику k-мерное пространство параметров, являющееся подмножеством евклидова пространства R^k, а конкретное значение параметра θ_0 неизвестно. Его и будем оценивать.

Обозначим $\theta = (\theta^1, \theta^2, \ldots, \theta^k)$. Рассмотрим вектор-столбец частных производных логарифма плотности вероятности:

$$s(x;\theta) = \left\| \frac{\partial}{\partial \theta^\alpha} \ln f(x;\theta), \quad \alpha = 1,2,...,k \right\|$$

и матрицу частных производных второго порядка для той же функции:

$$b(x;\theta) = \left\| \frac{\partial^2}{\partial \theta^\alpha \partial \theta^\beta} \ln f(x;\theta), \quad \alpha, \beta = 1,2,...,k \right\|.$$

Положим:

$$s_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} s(x_i, \theta), \quad b_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} b(x_i, \theta).$$

Пусть матрица информации Фишера $I(\theta_0) = M[-b_n(\theta_0)]$ положительно определена.

Определение 1 [14, с. 269]. Оценку $\theta(n)$ параметра θ_0 называют наилучшей асимптотически нормальной оценкой (сокращенно НАН-оценкой), если распределение случайного вектора $\sqrt{n}(\theta(n) - \theta_0)$ сходится при $n \to \infty$ к нормальному распределению с нулевым математическим ожиданием и ковариационной матрицей $I^{-1}(\theta_0)$.

Определение 1 корректно: $I^{-1}(\theta_0)$ является нижней асимптотической границей для ковариационной матрицы случайного вектора $\sqrt{n}(\theta^*(n)-\theta_0)$, где $\theta^*(n)$ — произвольная оценка. ОМП являются НАН-оценками (см. [14] и др.). Некоторые другие оценки также являются НАН-оценками, например, байесовские. Сказанное об ОМП и байесовских оценках справедливо при некоторых внутриматематических условиях регулярности (см., например, [18]). В ряде случаев несмещенные оценки являются НАН-оценками, более того, они лучше, чем ОМП (их дисперсия меньше), при конечных объемах выборки [10].

Для анализа реальных данных естественно рекомендовать какую-либо из НАН-оценок. (Это утверждение всегда верно на этапе асимптотики при изучении конкретной задачи прикладной статистики. Теоретически можно предположить, что при тщательном изучении для конкретных конечных объемов выборки наилучшей окажется какая-либо оценка, не являющаяся НАН-оценкой. Однако такие ситуации нам пока не известны.)

Пусть $\theta_1(n)$ и I_n^{-1} — некоторые оценки θ_0 и $I^{-1}(\theta_0)$ соответственно.

Определение 2. Одношаговой оценкой (ОШ-оценкой, или ОШО) называется оценка:

$$\theta_2(n)=\theta_1(n)+I_n^{-1}s_n(\theta_1(n)).$$

Теорема 1 [19]. Пусть выполнены следующие условия.

(I) Распределение $\sqrt{n}s_n(\theta_0)$ сходится при $n\to\infty$ к нормальному распределению с математическим ожиданием 0 и ковариационной матрицей $I(\theta_0)$ и, кроме того, существует $Mb_n(\theta_0)b_n'(\theta_0)$.

(II) При некотором $\varepsilon>0$ и $n\to\infty$

$$\sup_{\theta:0<|\theta-\theta_0|<\varepsilon}\frac{|s_n(\theta)-s_n(\theta_0)-b_n(\theta_0)(\theta-\theta_0)|}{|\theta-\theta_0|^2}=O_p(1).$$

(III) Для любого $\varepsilon>0$

$$\lim_{n\to\infty}P\{n^{1/4}(|\theta_1(n)-\theta_0|+\|I_n^{-1}-I^{-1}(\theta_0)\|)>\varepsilon\}=0.$$

Тогда ОШ-оценка является НАН-оценкой.

Доказательство. Рассмотрим тождество:

$$\sqrt{n}(\theta_2(n)-\theta_0)=\sqrt{n}(\theta_1(n)-\theta_0)+\sqrt{n}I_n^{-1}s_n(\theta_1(n)).$$

В силу условия (II) теоремы:

$$\sqrt{n}I_n^{-1}s_n(\theta_1(n)) = \sqrt{n}I_n^{-1}s_n(\theta_0) + \sqrt{n}I_n^{-1}b_n(\theta_0)(\theta_1(n)-\theta_0) + \sqrt{n}I_n^{-1}O_p(|\,\theta_1(n)-\theta_0\,|^2). \quad (5)$$

Из условия (I) теоремы следует, что первое слагаемое в правой части формулы (5) сходится при $n \to \infty$ по распределению к многомерному нормальному закону с математическим ожиданием 0 и ковариационной матрицей $I^1(\theta_0)$. Согласно условию (III):

$$\sqrt{n}\,|\,\theta_1(n)-\theta_0\,|^2 \to 0$$

по вероятности. Кроме того, согласно тому же условию последовательность матриц I_n^{-1} ограничена по вероятности. Поэтому третье слагаемое в правой части формулы (5) сходится к 0 по вероятности. Для завершения доказательства теоремы осталось показать, что:

$$\sqrt{n}(\theta_1(n)-\theta_0) + \sqrt{n}I_n^{-1}b_n(\theta_0)(\theta_1(n)-\theta_0) \to 0 \quad (6)$$

по вероятности. Левая часть формулы (6) преобразуется к виду:

$$(E + I_n^{-1}b_n(\theta_0))\sqrt{n}(\theta_1(n)-\theta_0), \quad (7)$$

где E — единичная матрица. Поскольку из условия (I) теоремы следует, что для $b_n(\theta_0)$ справедлива (многомерная) центральная предельная теорема, то

$$b_n(\theta_0) = -I(\theta_0) + O_p(n^{-1/2}).$$

С учетом условия (III) теоремы заключаем, что

$$E + I_n^{-1}b_n(\theta_0) = o_p(n^{-1/4}). \quad (8)$$

Из соотношений (7), (8) и условия (III) теоремы вытекает справедливость формулы (6). Теорема доказана.

Прокомментируем условия теоремы. Условия (I) и (II) обычно предполагаются справедливыми при рассмотрении оценок максимального правдоподобия [14]. Эти условия можно выразить в виде требований, наложенных

непосредственно на плотность $f(x; \theta)$ из параметрического семейства, как это сделано, например, в [18]. Условие (III) теоремы, наложенное на исходные оценки, весьма слабое. Обычно используемые оценки $\theta_1(n)$ и I_n^{-1} являются не $n^{-1/4}$ — состоятельными, а $\sqrt{n}$ — состоятельными, т.е. условие (III) заведомо выполняется.

Какие оценки годятся в качестве начальных? В классических областях прикладной статистики в качестве $\theta_1(n)$ можно использовать оценки метода моментов, как это сделано в ГОСТ 11.011-83 (в настоящее время отменен, но может использоваться как научная публикация) [17], или, например, квантильные. В качестве I_n^{-1} в теоретической работе [14] предлагается использовать простейшую оценку:

$$I_n^{-1} = -b_n^{-1}(\theta_1(n)). \qquad (9)$$

Для гамма-распределения с неизвестными параметрами формы, масштаба и сдвига ОШ-оценки применены в [17]. При этом оценка (9) оказалась непрактичной, поскольку с точностью до погрешностей измерений и вычислений $\det(b_n) = 0$ для реальных данных о наработке резцов до предельного состояния, приведенных в [17]. Поскольку $\det(b_n) = 0$, то обратная матрица не существует, вычисления по формуле (9) невозможны. Поэтому в [17] в качестве ОШ-оценки была применена непосредственно первая итерация метода Ньютона — Рафсона решения системы уравнений максимального правдоподобия, т.е. была использована оценка:

$$I_n^{-1} = I^{-1}(\theta_1(n)). \qquad (10)$$

В формуле (10) непосредственно используется явный вид зависимости матрицы информации Фишера от неизвестных параметров распределения.

В других случаях выбор тех или иных начальных оценок, в частности, выбор между (9) и (10), может определяться, например, простотой вычислений. Можно использовать также устойчивые аналоги [1] перечисленных выше оценок.

Полезно отметить, что еще в 1925 г., т.е. непосредственно при разработке метода максимального правдоподобия, его создатель Р. Фишер считал, что первая итерация по методу Ньютона — Рафсона дает хорошую оценку вектору неизвестных параметров [14, с. 298]. Он, однако, рассматривал эту оценку как аппроксимацию ОМП. А. А. Боровков воспринимает ОШ-оценки как способ «приближенного вычисления оценок максимального правдоподо-

бия» [20, с. 225] и показывает асимптотическую эквивалентность ОШ-оценок и ОМП (в более сильных предположениях, чем в теореме 1; другими словами, теорема 1 обобщает результаты А. А. Боровкова относительно ОШ-оценок). Мы же полагаем, что ОШ-оценки имеют самостоятельную ценность, причем не меньшую, а в большинстве реальных задач большую, чем ОМП. По нашему мнению, ОМП целесообразно применять (на этапе асимптотики) только тогда, когда они находятся явно. Во всех остальных случаях следует использовать на этом этапе ОШ-оценки (или какие-либо иные, выбранные из дополнительных соображений).

С чем связана популярность оценок максимального правдоподобия? Из всех НАН-оценок они наиболее просто вводятся, ранее других предложены. Поэтому среди математиков сложилась устойчивая традиция рассматривать ОМП в курсах математической статистики. Однако при этом игнорируются вычислительные вопросы, а также отодвигаются в сторону многочисленные иные НАН оценки.

В прикладной статистике — иные приоритеты. На первом месте — ОШ-оценки, все остальные НАН-оценки, в том числе ОМП, рассматриваются в качестве дополнительных возможностей.

В задачах нечисловой статистики вместо оценок метода моментов используют иные начальные оценки, свои для каждого конкретного вида нечисловых данных [12, 13, 16].

Одношаговые оценки для параметров гамма-распределения расписаны в стандарте [11] и статье [21]. Алгоритмическое и программное обеспечение ОШ-оценок для распределения Вейбулла — Гнеденко и гамма-распределения рассмотрено в содержательной монографии [22]. История вопроса освещена в статье [19].

2.5. НЕПАРАМЕТРИЧЕСКИЕ ОЦЕНКИ ПЛОТНОСТИ

Эмпирическая функция распределения — это состоятельная непараметрическая оценка функции распределения числовой случайной величины. А как оценить плотность? Если формально продифференцировать эмпирическую функцию распределения, то получим бесконечности в точках, соответствующих элементам выборки, и 0 во всех остальных. Ясно, что это не оценка плотности.

Как же действовать? Каждому элементу выборки соответствует в эмпирическом распределении вероятность $1/n$, где n — объем выборки. Целесо-

образно эту вероятность не помещать в одну точку, а «размазать» вокруг нее, построив «холмик». Если «холмики» налегают друг на друга, то получаем положительную плотность на всей прямой. Чтобы получить состоятельную оценку плотности, необходимо выбирать ширину «холмика» в зависимости от объема выборки. При этом число «холмиков», покрывающих фиксированную точку, должно безгранично расти. Но одновременно доле таких «холмиков» следует убывать, поскольку покрывающие «холмики» должны быть порождены лишь ближайшими членами вариационного ряда.

Реализация описанной идеи привела к различным вариантам непараметрических оценок плотности. Основополагающей является работа Н. В. Смирнова 1951 г. [23]. Вначале рассматривались непараметрические оценки плотности распределения числовых случайных величин и конечномерных случайных векторов. В 1980-х гг. удалось сконструировать такие оценки в пространствах произвольной природы [24], а затем и для конкретных видов нечисловых данных [25].

О гистограммах. При описании числовых данных часто используют гистограммы. При этом область изменения случайной переменной разбивают на интервалы равной длины, подсчитывают число попаданий в каждый интервал и строят соответствующую столбиковую диаграмму. Она напоминает график плотности. И действительно, Н. В. Смирнов показал в работе [23], что последовательность гистограмм при определенных условиях сходится к плотности.

Процедура построения гистограммы зависит от субъективного мнения статистика. Не существует научно обоснованных правил выбора числа интервалов и их длины. Рекомендации по этому поводу, приводимые в различных изданиях, отражают лишь традицию и/или субъективное мнение авторов.

К настоящему времени разработано много методов оценивания плотности распределения. О некоторых из них речь пойдет ниже. Что же касается гистограмм, то с научной точки зрения их надо отнести к истории статистики.

Непараметрические оценки плотности в пространствах произвольной природы. Сначала рассмотрим непараметрические оценки плотности в наиболее общей ситуации. Напомним, что в нечисловой статистике выделяют общую теорию и статистику в конкретных пространствах нечисловой природы (например, статистику ранжировок). В общей теории есть два основных сюжета. Один связан со средними величинами и асимптотическим

поведением решений экстремальных статистических задач, второй — с непараметрическими оценками плотности. Первый сюжет рассмотрен выше, второму посвящен настоящий раздел.

Понятие плотности в пространстве произвольной природы X требует специального обсуждения. В пространстве X должна быть выделена некоторая специальная мера μ, относительно которой будут рассматриваться плотности, соответствующие другим мерам, например, мере ν, задающей распределение вероятностей некоторого случайного элемента ξ. В таком случае $\nu(A) = P(\xi \in A)$ для любого случайного события A. Плотность $f(x)$, соответствующая мере ν — это такая функция, что:

$$\nu(A) = \int_A f(x) d\mu$$

для любого случайного события A. Для случайных величин и векторов мера μ — это объем множества A, в математических терминах — мера Лебега. Для дискретных случайных величин и элементов со значениями в конечном множестве X в качестве меры μ естественно использовать считающую меру, которая событию A ставит в соответствие число его элементов. Используют также нормированную случайную меру, когда число точек в множестве A делят на число точек во всем пространстве X. В случае считающей меры значение плотности в точке x совпадает с вероятностью попасть в точку x, т.е. $f(x) = P(\xi = x)$. Таким образом, с рассматриваемой точки зрения стирается грань между понятиями «плотность вероятности» и «вероятность (попасть в точку)».

Как могут быть использованы непараметрические оценки плотности распределения вероятностей в пространствах нечисловой природы? Например, для решения задач классификации (диагностики, распознавания образов — см. раздел 2.8). Зная плотности распределения классов, можно решать основные задачи диагностики — как задачи выделения кластеров, так и задачи отнесения вновь поступающего объекта к одному из диагностических классов. В задачах кластер-анализа можно находить моды плотности и принимать их за центры кластеров или за начальные точки итерационных методов типа k-средних или динамических сгущений. В задачах собственно диагностики (дискриминации, распознавания образов с учителем) можно принимать решения о диагностике объектов на основе отношения плотностей, соответствующих классам. При неизвестных плотностях представляется естественным использовать их состоятельные оценки.

Методы оценивания плотности вероятности в пространствах общего вида предложены и первоначально изучены в работе [24]. В частности, в задачах диагностики объектов нечисловой природы предлагаем использовать непараметрические ядерные оценки плотности типа Парзена — Розенблатта (этот вид оценок и его название впервые были введены в статье [24]). Они имеют вид:

$$f_n(x) = \frac{1}{\eta_n(h_n, x)} \sum_{1 \le i \le n} K\left(\frac{d(x_i, x)}{h_n}\right),$$

где K: $R^1_+ \to R^1$ — так называемая ядерная функция, $x_1, x_2, \ldots, x_n \in X$, — выборка, по которой оценивается плотность, $d(x_i, x)$ — показатель различия (метрика, расстояние, мера близости) между элементом выборки x_i и точкой x, в которой оценивается плотность, последовательность h_n показателей размытости такова, что $h_n \to 0$ и $nh_n \to \infty$ при $n \to \infty$, а $\eta_n(h_n, x)$ — нормирующий множитель, обеспечивающий выполнение условия нормировки (интеграл по всему пространству от непараметрической оценки плотности $f_n(x)$ по мере μ должен равняться 1). Ранее американские исследователи Е. Парзен и М. Розенблатт использовали подобные статистики в случае $X = R^1$ с $d(x_i, x) = |x_i - x|$.

Введенные описанным образом ядерные оценки плотности — частный случай так называемых линейных оценок, также впервые предложенных в работе [24]. В теоретическом плане они выделяются тем, что удается получать результаты такого же типа, что в классическом одномерном случае, но, разумеется, с помощью совсем иного математического аппарата.

Свойства непараметрических ядерных оценок плотности. Рассмотрим выборку со значениями в некотором пространстве произвольного вида. В этом пространстве предполагаются заданными показатель различия d и мера μ. Одна из основных идей рассматриваемого подхода состоит в том, чтобы согласовать их между собой. А именно, на их основе построим новый показатель различия d_1, так называемый «естественный», в терминах которого проще формулируются свойства непараметрической оценки плотности. Для этого рассмотрим шары $L_t(x) = \{y \in X : d(y, x) \le t\}$ радиуса $t \ge 0$ и их меры $F_x(t) = \mu(L_t(x))$. Предположим, что $F_x(t)$ как функция t при фиксированном x непрерывна и строго возрастает. Введем функцию $d_1(x, y) = F_x(d(x, y))$. Это — монотонное преобразование показателя различия или расстояния, а потому $d_1(x, y)$ — также показатель различия (даже если d — *метрика*, для d_1 неравенство треугольника

может быть не выполнено). Другими словами, $d_1(x, y)$, как и $d(x, y)$, можно рассматривать как показатель различия (меру близости) между x и y.

Для вновь введенного показателя различия $d_1(x,y)$ введем соответствующие шары $L_{1t}(x)=\{y\in X: d_1(y,x)\le t\}$. Поскольку обратная функция $F^{-1}{}_x(t)$ определена однозначно, то

$$L_{1t}(x)=\{y\in X: d_1(y,x)\le F_x^{-1}(t)\}=L_T(x),$$

где $T = F^{-1}{}_x(t)$. Следовательно, справедлива цепочка равенств $F^{1}{}_x(t) = \mu(L_{1t}(x)) = \mu(L_T(x)) = F_x(F^{-1}{}_x(t)) = t$ (для всех тех значений параметра t, для которых определены все участвующие в записи математические объекты).

Переход от d к d_1 напоминает классическое преобразование, использованное Н. В. Смирновым при изучении непараметрических критериев согласия и однородности, а именно, преобразование $\eta = F(\xi)$, переводящее случайную величину ξ с непрерывной функцией распределения $F(x)$ в случайную величину η, равномерно распределенную на отрезке [0, 1]. Оба рассматриваемых преобразования существенно упрощают дальнейшие рассмотрения. Преобразование $d_1 = F_x(d)$ зависит от точки x, что не влияет на дальнейшие рассуждения, поскольку ограничиваемся изучением сходимости в отдельно взятой точке.

Функцию $d_1(x, y)$, для которой мера шара радиуса t равна t, называем в соответствии с работой [24] «естественным показателем различия» или «естественной метрикой». В случае конечномерного пространства R^k и евклидовой метрики d имеем $d_1(x,y) = c_k d^k(x,y)$, где c_k — объем шара единичного радиуса в R^k.

Поскольку можно записать, что:

$$K\left(\frac{d(x_i,x)}{h_n}\right)=K_1\left(\frac{d_1(x_i,x)}{h_n}\right),$$

где

$$K_1(u)=K\left(\frac{F_x^{-1}(uh_n)}{h_n}\right),$$

то переход от одного показателя различия к другому, т.е. от d к d_1, соответствует переходу от одной ядерной функции к другой, т.е. от K к K_1. Выгода от такого перехода заключается в том, что утверждения о поведении непараметрических оценок плотности приобретают более простую формулировку.

Теорема 1. Пусть d — естественная метрика, плотность f непрерывна в точке x и ограничена на всем пространстве X, причем $f(x) > 0$, ядерная функция $K(u)$ удовлетворяет простым условиям регулярности:

$$\int_0^1 K(u)du = 1, \int_0^\infty (|K(u)| + K^2(u))du < \infty.$$

Тогда $\eta_n(h_n, x) = nh_n$, оценка $f_n(x)$ является состоятельной, т.е. $f_n(x) \to f(x)$ по вероятности при $n \to \infty$ и, кроме того,

$$\lim_{n\to\infty} (nh_n Df_n(x)) = f(x)\int_0^{+\infty} K^2(u)du.$$

Теорема 1 доказывается методами, развитыми в работе [24]. Однако остается открытым вопрос о скорости сходимости ядерных оценок, в частности, о поведении величины $\alpha_n = M(f_n(x) - f(x))^2$ — среднего квадрата ошибки, и об оптимальном выборе показателей размытости h_n. Для того, чтобы продвинуться в решении этого вопроса, введем новые понятия. Для случайного элемента $X(\omega)$ со значениями в X рассмотрим так называемое круговое распределение $G(x, t) = P\{d(X(\omega), x) \le t\}$ и круговую плотность $g(x, t) = G'_t(x, t)$.

Теорема 2. Пусть ядерная функция $K(u)$ непрерывна и финитна, т.е. существует число E такое, что $K(u)=0$ при $u>E$. Пусть круговая плотность является достаточно гладкой, т.е. допускает разложение:

$$g(x,t) = f(x) + tg'_t(x,0) + \frac{t^2}{2}g''_{tt}(x,0) + \frac{t^3}{3!}g'''_{ttt}(x,0) + \ldots + \frac{t^k}{k!}g^{(k)}_{t^{(k)}}(x,0) + o(h_n^k)$$

при некотором натуральном k, причем остаточный член равномерно ограничен на $[0,hE]$. Пусть:

$$\int_0^E u^i K(u)du = 0, i = 1,2,\ldots,k-1.$$

Тогда

$$\alpha_n = [Mf_n(x) - f(x)]^2 + Df_n(x) =$$

$$= h_n^{2k}\left(\int_0^E u^k K(u)du\right)^2 (g^k_{t^{(k)}}(x,0))^2 + \frac{f(x)}{nh_n}\int_0^E K^2(u)du + o\left(h_n^{2k} + \frac{1}{nh_n}\right).$$

Доказательство теоремы 2 проводится с помощью разработанной в нечисловой статистике математической техники, образцы которой представлены, в частности, в работе [24]. Если коэффициенты при основных членах в правой части последней формулы не равны 0, то величина α_n достигает минимума, равного $\alpha_n = O\left(n^{-1+\frac{1}{2k+1}}\right)$, при $h_n = n^{-\frac{1}{2k+1}}$. Эти выводы совпадают с классическими результатами, полученными ранее рядом авторов для весьма частного случая прямой $X = R^1$ (см., например, монографию [18, с. 316]). Заметим, что для уменьшения смещения оценки приходится применять знакопеременные ядра $K(u)$.

Непараметрические оценки плотности в конечных пространствах [25]. В случае пространств из конечного числа элементов естественных метрик не существует. Однако можно получить аналоги теорем 1 и 2, переходя к пределу не только по объему выборки n, но и по новому параметру дискретности m.

Рассмотрим некоторую последовательность X_m, $m = 1, 2, \ldots$, конечных пространств. Пусть в X_m заданы показатели различия d_m. Будем использовать нормированные считающие меры μ_m, ставящие в соответствие каждому подмножеству A долю элементов всего пространства X_m, входящих в A. Как и ранее, рассмотрим как функцию t объем шара радиуса t, т.е.

$$F_{mx}(t) = \mu_m(\{y \in X_m : d_m(x, y) \le t\}).$$

Введем аналог естественного показателя различия $d_{1m}(x, y) = F_{mx}(d_m(x, y))$. Наконец, рассмотрим аналоги преобразования Смирнова $F^1_{mx}(t) = \mu_m(\{y \in X_m : d_{1m}(x, y) \le t\})$. Функции $F^1_{mx}(t)$, в отличие от ситуации предыдущего подраздела, уже не совпадают тождественно с t, они кусочно-постоянны и имеют скачки в некоторых точках t_i, $i = 1, 2, \ldots$, причем в этих точках $F^1_{mx}(t_i) = t_i$.

Теорема 3. Пусть точки скачков равномерно сближаются, т.е. $\max(t_i - t_{i-1}) \to 0$ при $m \to \infty$ (другими словами, $\sup |F^1_{mx}(t) - t| \to 0$ при $m \to \infty$).

Тогда существует последовательность параметров дискретности m_n такая, что при предельном переходе $n \to \infty, m \to \infty, m \geq m_n$ справедливы заключения теорем 1 и 2.

Пример 1. Пространство $X_m = 2^{\sigma(m)}$ всех подмножеств конечного множества $\sigma(m)$ из m элементов допускает (см. главу 1 или монографию [1]) аксиоматическое введение метрики $d(A,B) = card(A \Delta B)/2^m$, где Δ — символ симметрической разности множеств. Рассмотрим непараметрическую ядерную оценку плотности типа Парзена — Розенблатта:

$$f_{nm}(A) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{1}{h_n} \Phi\left(\frac{2card(A \Delta X_i) - m}{\sqrt{m}} \right) \right),$$

где $\Phi(\cdot)$ — функция нормального стандартного распределения. Можно показать, что эта оценка удовлетворяет условиям теоремы 3 с $m_n = (\ln n)^6$.

Пример 2. Рассмотрим пространство функций $f : Y_r \to Z_q$, определенных на конечном множестве $Y_r = \{1/r,\ 2/r, ..., (r-1)/r,\ 1\}$, со значениями в конечном множестве $Z_q = \{0,\ 1/q,\ 2/q, ...,\ ^{q-1}/_{q},\ 1\}$. Это пространство можно интерпретировать как пространство нечетких множеств (см. главу 1), а именно, Y_r — носитель нечеткого множества, а Z_q — множество значений функции принадлежности. Очевидно, число элементов пространства X_m равно $(q + 1)^r$. Будем использовать расстояние $d(f,g) = \sup |f(y) - g(y)|$ в этом пространстве. Непараметрическая оценка плотности имеет вид:

$$f_{nm}(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{[2 \sup\limits_{y} |x(y) - x_i(y)| + 1/q]^r}{h_n (1 + 1/q)^r} \right).$$

Если $r = n^{\alpha},\ q = n^{\beta}$, то при $\beta > \alpha$ выполнены условия теоремы 7, а потому справедливы теоремы 1 и 2.

Пример 3. Рассматривая пространства ранжировок m объектов, в качестве расстояния $d(A, B)$ между ранжировками A и B примем минимальное число инверсий, необходимых для перехода от A к B. Тогда $\max(t_i - t_{i-1})$ не стремится к 0 при $m \to \infty$, условия теоремы 3 не выполнены.

Пример 4. В прикладных работах наиболее распространенный пример объектов нечисловой природы — вектор разнотипных данных: реальный

объект описывается вектором, часть координат которого — значения количественных признаков, а часть — качественных (номинальных и порядковых). Для пространств разнотипных признаков, т.е. декартовых произведений непрерывных и дискретных пространств, возможны различные постановки. Пусть, например, число градаций качественных признаков остается постоянным. Тогда непараметрическая оценка плотности сводится к произведению двух величин — частоты попадания в точку в пространстве качественных признаков и классической оценки типа Парзена — Розенблатта в пространстве количественных переменных. В общем случае расстояние $d(x, y)$ можно, например, рассматривать как сумму трех расстояний. А именно, евклидова расстояния d_1 между количественными факторами, расстояния d_2 между номинальными признаками ($d_2(x, y) = 0$, если $x = y$, и $d_2(x, y) = 1$, если $x \neq y$) и расстояния d_3 между порядковыми переменными (если x и y — номера градаций, то $d_3(x, y) = |x - y|$). Наличие количественных факторов приводит к непрерывности и строгому возрастанию функции $F_{mx}(t)$, а потому для непараметрических оценок плотности в пространствах разнотипных признаков верны теоремы 1 и 2.

Программная реализация описания данных с помощью непараметрических оценок плотности включена в ряд программных продуктов по прикладной статистике, в частности, в пакет программ анализа данных ППАНД [26].

2.6. СТАТИСТИКИ ИНТЕГРАЛЬНОГО ТИПА

В прикладной статистике широко используются статистики типа омега-квадрат и типа Колмогорова — Смирнова [6, 54]. Они применяются для проверки согласия с фиксированным распределением или семейством распределений, для проверки однородности двух выборок, симметрии распределения относительно 0, их обобщения — при оценивании условной плотности и регрессии в пространствах произвольной природы и т.д.

Статистики интегрального типа и их асимптотика. Рассмотрим статистики интегрального типа:

$$\xi_\alpha = \xi(f_\alpha, F_\alpha) = \int_X f_\alpha(x,\omega) dF_\alpha(x,\omega), \qquad (1)$$

где X — некоторое пространство, по которому происходит интегрирование (например, $X = [0; 1]$, $X = R^1$ или $X = R^k$). Здесь $\{\alpha\}$ — направленное множество,

переход к пределу по которому обозначен как $\alpha \to \infty$ (см. приложение 1). Случайные функции f_α: $X \times \Omega \to Y$ обычно принимают значения, являющиеся числами. Но иногда рассматривают и постановки, в которых $Y = R^k$ или Y — банахово пространство (т.е. полное нормированное пространство [27]). Наконец, $F_\alpha(x, \omega)$ — случайная функция распределения или случайная вероятностная мера; в последнем случае используют также обозначение $dF_\alpha(x, \omega) = F_\alpha(dx, \omega)$.

Предполагаются выполненными необходимые для корректности изложения внутриматематические предположения измеримости, например, сформулированные в [28, 29].

Пример 1. Рассмотрим критерий Лемана — Розенблатта, т.е. критерий типа омега-квадрат для проверки однородности двух независимых выборок [6]. Его статистика имеет вид:

$$LR = \frac{mn}{m+n} \int_{-\infty}^{\infty} (F_m(x) - G_n(x))^2 dH_{m+n}(x),$$

где $F_m(x)$ — эмпирическая функция распределения, построенная по первой выборке объема m, $G_n(x)$) — эмпирическая функция распределения, построенная по второй выборке объема n, а $H_{m+n}(x)$ — эмпирическая функция распределения, построенная по объединенной выборке объема $m + n$. Легко видеть, что:

$$H_{m+n}(x) = \frac{m}{m+n} F_m(x) + \frac{n}{m+n} G_n(x).$$

Ясно, что статистика LR имеет вид (1). При этом x — действительное число, $X = Y = R^1$, в роли α выступает пара (m, n), и $\alpha \to \infty$ означает, что $\min(m, n) \to \infty$. Далее,

$$f_\alpha(x,\omega) = \frac{mn}{m+n} (F_m(x) - G_n(x))^2.$$

Наконец, $F_\alpha(x, \omega) = H_{m+n}(x)$.

Теперь обсудим асимптотическое поведение функций $f_\alpha(x, \omega)$ и $F_\alpha(x, \omega)$, с помощью которых определяется статистика Лемана — Розенблатта LR. Ограничимся случаем, когда справедлива гипотеза однородности, т.е. совпадают функции распределения, соответствующие генеральным совокупностям, из которых взяты выборки. Их общую функцию распределения обозна-

чим $F(x)$. Она предполагается непрерывной. Введем в рассмотрение выборочные процессы:

$$\xi_m(x) = \sqrt{m}(F_m(x) - F(x)), \quad \eta_n(x) = \sqrt{n}(G_n(x) - F(x)).$$

Нетрудно проверить, что

$$f_\alpha(x,\omega) = \left(\sqrt{\frac{n}{m+n}}\xi_m(x) - \sqrt{\frac{m}{m+n}}\eta_n(x) \right)^2.$$

Сделаем замену переменной $t = F(x)$. Тогда выборочные процессы переходят в соответствующие эмпирические процессы (см. приложение 1):

$$f_\alpha(F^{-1}(t),\omega) = \left(\sqrt{\frac{n}{m+n}}\xi_m^0(t) - \sqrt{\frac{m}{m+n}}\eta_n^0(t) \right)^2, \quad 0 \le t \le 1,$$

где $\xi_m^0(t) = \xi_m(F^{-1}(t)), \quad \eta_n^0(t) = \eta_n(F^{-1}(t))$.

Конечномерные распределения этого процесса, т.е. распределения случайных векторов:

$$(f_\alpha(F^{-1}(t_1),\omega), f_\alpha(F^{-1}(t_2),\omega), \ldots, f_\alpha(F^{-1}(t_k),\omega))$$

для всех возможных наборов (t_1, t_2, ..., t_k), сходятся к конечномерным распределениям квадрата броуновского моста $\xi^2(t)$. В соответствии с разделом П-5 приложения 1 рассматриваемая сходимость по распределению обозначается так:

$$f_\alpha(F^{-1}(t),\omega) \Rightarrow \xi^2(t), \quad 0 \le t \le 1. \tag{2}$$

Нетрудно видеть, что при любом x:

$$F_\alpha(x,\omega) = H_{m+n}(x) \to F(x)$$

при $\alpha \to \infty$ (сходимость по вероятности). С помощью замены переменной $t = F(x)$ получаем, что

$$F_\alpha(F^{-1}(t),\omega) = H_{m+n}(F^{-1}(t)) \to t \tag{3}$$

при $\alpha \to \infty$. Из соотношений (2) и (3) хотелось бы сделать вывод, что в случае статистики Лемана — Розенблатта типа омега-квадрат:

$$\xi_\alpha = \int_X f_\alpha(x,\omega)dF_\alpha(x,\omega) = LR \Rightarrow \int_0^1 \xi^2(t)dt\,,$$

т.е. предельным распределением этой статистики является классическое распределение [30], найденное как предельное для одновыборочной статистики критерия согласия омега-квадрат, известного также как критерий Крамера — Мизеса — Смирнова.

Действительно, сформулированное утверждение справедливо. Однако доказательство нетривиально.

Так, может показаться очевидным следующее утверждение.

Утверждение 1. Пусть f: [0; 1] $\to R^1$ — ограниченная функция, $G_n(x)$ и $G(x)$ — функции распределения, $G_n(0) = G(0) = 0$, $G_n(1) = G(1) = 1$, причем $G_n(x) \to G(x)$ при всех x. Тогда:

$$\lim_{n\to\infty} \int_0^1 f(x)d(G_n(x) - G(x)) = 0\,. \qquad (4)$$

Это утверждение неверно (ср. [31, с. 42]). Действительно, пусть $f(x) = 1$, если x рационально, и $f(x) = 0$, если x иррационально, $G(x) = x$, а $G_n(x)$ имеет скачки величиной 2^{-n} в точках $m/2^n$, $m = 1, 2, ..., 2^n$ при всех $n = 1, 2, \ldots$ Тогда $G_n(x) \to G(x)$ при всех x, однако,

$$\int_0^1 f(x)dG_n(x) = 1,\ \int_0^1 f(x)dG(x) = 0$$

при всех $n = 1, 2, \ldots$ Следовательно, вопреки сформулированному выше утверждению 1,

$$\int_0^1 f(x)d(G_n(x) - G(x)) = 1\,,$$

т.е. соотношение (4) неверно.

Итак, сформулируем проблему. Пусть известно, что последовательность случайных функций $f_\alpha(x, \omega)$ сходится по распределению при $\alpha \to \infty$ к случайной

функции $f(x, \omega)$. Пусть последовательность случайных мер $F_\alpha(A, \omega)$, определенных на множествах A из достаточно обширного семейства, сходится по распределению к вероятностной мере $F(A)$ при $\alpha \to \infty$. Если речь идет о конечномерном пространстве и меры задаются функциями распределения, то сходимость $F_\alpha(x,\omega)$ к $F(x)$ должна иметь место во всех точках непрерывности $F(x)$. В каких случаях можно утверждать, что при $\alpha \to \infty$ справедлив предельный переход:

$$\xi_\alpha = \xi(f_\alpha, F_\alpha) = \int_X f_\alpha(x,\omega) dF_\alpha(x,\omega) \Rightarrow \xi = \xi(f,F) = \int_X f(x,\omega) dF(x) \, ?$$

Выше показано, что, например, ограниченности $f_\alpha(x, \omega)$ для этого недостаточно.

Метод аппроксимации ступенчатыми функциями. Рассмотрим общий метод, позволяющий получить предельные распределения не только для статистик интегрального типа, но и для других статистических критериев, например, для критериев типа Колмогорова. Пусть $T = \{C_1, C_2, ..., C_k\}$ — разбиение пространства X на непересекающиеся подмножества. Пусть в каждом элементе C_j разбиения T выделена точка x_j, $j = 1, 2, ..., k$. На множестве функций $f: X \to Y$ введем оператор A_T: если $x \in C_j$, то:

$$A_T f(x) = f(x_j), j = 1, 2, \ldots . k. \tag{5}$$

Тогда $A_T f$ — аппроксимация функции f ступенчатыми (кусочно-постоянными) функциями.

Пусть $f_\alpha(x, \omega)$ — последовательность случайных функций на X, а $K(\bullet)$ — функционал на множестве всех возможных их траекторий как функций от x. Для изучения распределения $K(f_\alpha)$ методом аппроксимации ступенчатыми функциями используют разложение:

$$K(f_\alpha) = K(A_T f_\alpha) + \{K(f_\alpha) - K(A_T f_\alpha)\}. \tag{6}$$

Согласно (5) распределение первого слагаемого в (6) определяется конечномерным распределением случайного элемента, а именно, распределением вектора:

$$(f_\alpha(x_1, \omega), f_\alpha(x_2, \omega), \ldots f_\alpha(x_k, \omega)). \tag{7}$$

В обычных постановках предельной теории классических непараметрических критериев распределение вектора (7) сходится при $\alpha \to \infty$ к соответствующему конечномерному распределению предельной случайной функции $f(x, \omega)$, т.е. к распределению случайного вектора:

$$(f(x_1, \omega), f(x_2, \omega), \ldots f(x_k, \omega)). \tag{8}$$

В соответствии с теорией наследования сходимости (прил. 1) при слабых условиях на функционал $K(\cdot)$ из сходимости по распределению вектора (7) к вектору (8) следует сходимость по распределению $K(A_T f_\alpha)$ к $K(A_T f)$.

Используя аналогичное (6) разложение:

$$K(f) = K(A_T f) + \{K(f) - K(A_T f)\}, \tag{9}$$

можно устанавливать сходимость по распределению $K(f_\alpha)$ к $K(f)$ при $\alpha \to \infty$ в два этапа: сначала выбрать разбиение T так, чтобы вторые слагаемые в правых частях соотношений (6) и (9) были малы, а затем при фиксированном операторе A_T воспользоваться сходимостью по распределению $K(A_T f_\alpha)$ к $K(A_T f)$.

Рассмотрим простой пример применения метода аппроксимации ступенчатыми функциями.

Обобщение теоремы Хелли. Пусть f: $[0; 1] \to R^1$ — измеримая функция, $F_n(x)$ — функции распределений, сосредоточенных на отрезке $[0; 1]$. Пусть $F_n(x)$ сходятся в основном к функции распределения $F(x)$, т.е.

$$\lim_{n \to \infty} F_n(x) = F(x) \tag{10}$$

для всех x, являющихся точками непрерывности $F(x)$.

Утверждение 2. Если $f(x)$ — непрерывная функция, то:

$$\lim_{n \to \infty} \int_0^1 f(x) dF_n(x) = \int_0^1 f(x) dF(x) \tag{11}$$

(рассматриваются интегралы Лебега — Стилтьеса).

Утверждение 2 известно в литературе как первая теорема Хелли [27, с. 344–346], вторая теорема Хелли [32, с. 174–175], лемма Хелли — Брея [33, с. 193–194].

Естественно поставить вопрос: при каких f из (10) следует (11)? Необходимо ввести условия и на F_n: если $F_n \equiv F$, то соотношение (11) верно для любой измеримой функции f, для которой интеграл в (11) существует. Поэтому рассмотрим следующую постановку.

Постановка 1. Пусть функция f такова, что для *любой* последовательности F_n, удовлетворяющей (10), справедливо (11). Что можно сказать о функции f?

В работах [28, 29] найдены следующие необходимые и достаточные условия на функцию f.

Теорема 1. Пусть ограниченная на [0; 1] функция f интегрируема по Риману — Стилтьесу по функции распределения $F(x)$. Тогда для *любой* последовательности функций распределения F_n, сходящейся в основном к F, имеет место предельный переход (11).

Теорема 2. Пусть функция f не интегрируема по Риману — Стилтьесу по функции распределения $F(x)$. Тогда *существует* последовательность функций распределения F_n, сходящаяся в основном к F, для которой соотношение (11) не выполнено.

Теоремы 1 и 2 в совокупности дают необходимые и достаточные условия для f в постановке 1. А именно, необходимо и достаточно, чтобы ограниченная на [0; 1] функция f была интегрируема по Риману — Стилтьесу по F.

Напомним определение интегрируемости функции f по Риману — Стилтьесу по функции распределения F [27, с. 341]. Рассмотрим разбиение $T = \{C_1, C_2, \ldots, C_m\}$, где

$$C_i = [y_{i-1}, y_i),\ i = 1, 2, \ldots, m-1,\ C_m = [y_{m-1}, y_m], \qquad (12)$$
$$0 = y_0 < y_1 < y_2 < \ldots < y_m = 1.$$

Выберем в C_i произвольную точку x_i, $i = 1, 2, \ldots, m$, и составим сумму:

$$S(T) = \sum_{i=1}^{m} f(x_i)[F(y_i) - F(y_{i-1})].$$

Если при $\max(y_i - y_{i-1}) \to 0$ эти суммы стремятся к некоторому пределу (не зависящему ни от способа дробления отрезка [0; 1], ни от выбора точек x_i в каждом из элементов разбиения), то этот предел называется интегралом Римана — Стилтьеса от функции f по функции F по отрезку [0; 1] и обозначается символом, приведенным в правой части равенства (11).

Рассмотрим суммы Дарбу — Стилтьеса:

$$S_H(T)=\sum_{i=1}^{m} m_i[F(y_i)-F(y_{i-1})],\quad S_B(T)=\sum_{i=1}^{m} M_i[F(y_i)-F(y_{i-1})],$$

где

$$m_i=\inf\{f(x), x\in X_i\},\quad M_i=\sup f\{f(x), x\in X_i\}.$$

Ясно, что:

$$S_H(T)\le S(T)\le S_B(T).$$

Необходимым и достаточным условием интегрируемости по Риману-Стилтьесу является следующее: для любой последовательности разбиений T_k, $k = 1, 2, 3, \ldots$ вида (12) такой, что $\max(y_i - y_{i-1}) \to 0$ при $k\to\infty$, имеем:

$$\lim_{k\to\infty}[S_B(T_k)-S_H(T_k)]=0. \tag{13}$$

Напомним, что согласно разделу П-3 приложения 1 колебанием $\delta(f, B)$ функции f на множестве B называется $\delta(f, B) = \sup\{|f(x) - f(y)|, x\in B, y\in B\}$. Поскольку:

$$\delta(f, C_i) = M_i - m_i,$$

то условие (13) можно записать в виде:

$$\lim_{k\to\infty}\sum_{C\in T_k}\delta(f,C)F(C)=0. \tag{14}$$

Условие (14), допускающее обобщение с $X = [0; 1]$ и f: $[0; 1] \to R^1$ на X и f более общего вида, и будем использовать при доказательстве теорем 1 и 2.

Доказательство теоремы 1. Согласно методу аппроксимации ступенчатыми функциями рассмотрим оператор A_T. Как легко проверить, имеет место разложение:

$$\begin{aligned}\beta_n &= \int_0^1 f(x)dF_n(x)-\int_0^1 f(x)dF(x)=\int_0^1\{f(x)-A_Tf(x)\}dF_n(x)+\\ &+\int_0^1\{A_Tf(x)-f(x)\}dF(x)+\left\{\int_0^1 A_Tf(x)dF_n(x)-\int_0^1 A_Tf(x)dF(x)\right\}.\end{aligned} \tag{15}$$

Поскольку:

$$|f(x) - A_T f(x)| \le \delta(f, X_i), \ x \in C_i,$$

то первое слагаемое в правой части (15) не превосходит:

$$\sum_{C \in T} \delta(f, C) F_n(C), \tag{16}$$

а второе не превосходит:

$$\sum_{C \in T} \delta(f, C) F(C).$$

Согласно определению оператора A_T третье слагаемое в (15) имеет вид:

$$\sum_{i=1}^{m} f(x_i)(F_n(C_i) - F(C_i)).$$

Очевидно, оно не превосходит по модулю:

$$\sup_{x \in X} |f(x)| \sum_{C \in T} |F_n(C) - F(C)|$$

(здесь используется ограниченность f на X).

Согласно (16) первое слагаемое в правой части (15) не превосходит:

$$\sum_{C \in T} \delta(f, C) F(C) + \sum_{C \in T} \delta(f, C) |F_n(C) - F(C)|.$$

Поскольку:

$$\delta(f, C) \le 2 \sup_{x \in X} |f(x)|,$$

то первое слагаемое в правой части (15) не превосходит:

$$\sum_{C \in T} \delta(f, C) F(C) + 2 \sup_{x \in X} |f(x)| \sum_{C \in T} |F_n(C) - F(C)|.$$

Из оценок, относящихся к трем слагаемым в разложении (15), следует, что

$$|\beta_n| \le 2\sum_{C\in T}\delta(f,C)F(C) + 3\sup_{x\in X}|f(x)|\sum_{C\in T}|F_n(C)-F(C)|. \qquad (17)$$

Используя оценку (17), докажем, что $\beta_n \to 0$ при $n \to \infty$. Пусть дано $\varepsilon > 0$. Согласно условию интегрируемости функции f по Риману — Стилтьесу, т.е. условию (14), можно указать разбиение $T = T(\varepsilon)$ такое, что

$$\sum_{C\in T(\varepsilon)}\delta(f,C)F(C) < \frac{\varepsilon}{4}, \qquad (18)$$

и в точках y_i, $i = 1, 2, \ldots, m-1$ (см. (12)), функция F непрерывна.

Поскольку

$$F_n(X_i) = F_n(y_i) - F_n(y_{i-1}),$$

то из (10) следует, что существует число $n = n(\varepsilon)$ такое, что при $n > n(\varepsilon)$ справедливо неравенство:

$$\sum_{C\in T(\varepsilon)}|F_n(C)-F(C)| < \frac{\varepsilon}{6}\left(\sup_{x\in X}|f(x)|\right)^{-1}. \qquad (19)$$

Из (17), (18) и (19) следует, что при $n > n(\varepsilon)$ справедливо неравенство:

$$\left|\int_0^1 f(x)dF_n(x) - \int_0^1 f(x)dF(x)\right| < \varepsilon,$$

что и требовалось доказать.

Обсудим условие ограниченности f. Если оно не выполнено, то из (10) не всегда следует (11).

Пример 2. Пусть $f(x) = 1/x$ при $x > 0$ и $f(0) = 0$. Пусть $F(0{,}5) = 0$, т.е. предельное распределение сосредоточено на [1/2; 1]. Пусть распределение F_n на [0; 1/2) имеет единственный атом в точке $x = 1/n$ величиной $n^{-1/2}$, а на [1/2; 1]

справедливо (10). Тогда по причинам, изложенным при доказательстве теоремы 1,

$$\lim_{n\to\infty}\int_{1/2}^{1} f(x)dF_n(x)=\int_{1/2}^{1} f(x)dF(x),$$

однако

$$\int_{0}^{1/2} f(x)dF_n(x)=\sqrt{n},\quad \int_{0}^{1/2} f(x)dF(x)=0,$$

т.е. соотношение (11) не выполнено.

Условие ограниченности подынтегральной функции f можно заменить, как это сделано, например, в [28], на условие строгого возрастания функции распределения F.

Лемма. Пусть функция распределения F всюду строго возрастает, т.е. из $x_1 < x_2$ вытекает $F(x_1) < F(x_2)$. Пусть функция f интегрируема по Риману — Стилтьесу по F, т.е. выполнено (14). Тогда функция f ограничена.

Доказательство. Рассмотрим точки $0 = y_0 < y_1 < y_2 < \ldots < y_{2m} = 1$ и два разбиения:

$$T_1=\{[0;y_1),[y_1;y_3),[y_3;y_5),\ldots,[y_{2m-1};1]\},\ T_2=\{[0;y_2),[y_2;y_4),[y_4;y_6),\ldots,[y_{2m-2};1]\}.$$

Тогда для любых двух точек x и x' можно указать конечную последовательность точек $x_1 = x$, x_2, x_3, …, x_s, $x_{s+1} = x'$ такую, что любые две соседние точки x_i, x_{i+1}, $i = 1, 2, \ldots, s$, одновременно принадлежат некоторому элементу C_i разбиения T_1 или разбиения T_2, причем $C_i \neq C_j$ при $i \neq j$. Действительно, пусть $x\in[y_p;y_{p+1}), x'\in[y_q;y_{q+1})$. Пусть для определенности $q > p$. Тогда можно положить $x_2 = y_{p+1}$, $x_3 = y_{p+2}$, …, $x_s = y_q$. Поскольку среди элементов разбиений T_1 и T_2 есть $C_1 = [y_p; y_{p+2})$, то $x = x_1 \in C_1, x_2 = y_{p+1} \in C_1$. Далее, $x_2 \in [y_{p+1}, y_{p+3}) = C_2, x_3 \in C_2$, и т.д.

Из указанных выше свойств последовательности $x_1 = x, x_2, x_3, \ldots, x_s, x_{s+1} = x'$ следует, что:

$$|f(x)-f(x')|\le\sum_{i=1}^{s}|f(x_{i+1})-f(x_i)|\le\sum_{C\in T_1}\delta(f,C)+\sum_{C\in T_2}\delta(f,C).$$

Пусть теперь число $\max(y_i - y_{i-2})$ настолько мало, что согласно (14):

$$\sum_{C\in T_1}\delta(f,C)F(C)<1,\ \sum_{C\in T_2}\delta(f,C)F(C)<1.$$

Тогда согласно двум последним соотношениям:

$$|f(x)-f(x')|\le 2[\min\{F(C): C\in T_1\bigcup T_2\}]^{-1},$$

что и доказывает лемму.

Доказательство теоремы 2. Пусть условие (14) не выполнено, т.е. существуют число $\gamma > 0$ и последовательность разбиений T_n, $n = 1, 2, \ldots$, такие, что $\max(y_i - y_{i-1}) \to 0$ при $n\to\infty$ и при всех n:

$$\sum_{C\in T_n}\delta(f,C)F(C)\ge\gamma. \qquad (20)$$

Для доказательства теоремы построим две последовательности функций распределения F_{1n} и F_{2n}, $n = 1, 2, \ldots$, для которых выполнено (10), но последовательность:

$$\delta_n=\int_0^1 f(x)dF_{1n}(x)-\int_0^1 f(x)dF_{2n}(x)$$

не стремится к 0 при $n \to \infty$. Тогда (11) не выполнено хотя бы для одной из последовательностей F_{1n} и F_{2n}.

Для любого C — элемента некоторого разбиения T — можно указать, как вытекает из определения $\delta(f, C)$, точки $x_1(C)$ и $x_2(C)$ такие, что:

$$f(x_1(C)) - f(x_2(C)) > 1/2\ \delta(f, C). \qquad (21)$$

Построим F_{1n} и F_{2n} следующим образом. Пусть $F_{1n}(C) = F_{2n}(C) = F(C)$ для любого C из T_n. При этом F_{1n} имеет в C один атом в точке $x_1(C)$ величиной $F(C)$, а F_{2n} имеет в C также один атом в точке $x_2(C)$ той же величины $F(C)$. Другими словами, распределение F_{1n} в C сосредоточено в одной точке, а именно, в $x_1(C)$, а распределение F_{2n} сосредоточено в $x_2(C)$. Тогда:

$$\delta_n=\sum_{C\in T_n}(f(x_1(C))-f(x_2(C)))F(C). \qquad (22)$$

Из (20), (21) и (22) следует, что

$$\delta_n \geq \frac{1}{2}\sum_{C\in T_n}\delta(f,C)F(C) \geq \frac{\gamma}{2}.$$

Остается показать, что для последовательностей функций распределения F_{1n} и F_{2n} выполнено (10). Пусть x — точка непрерывности F. Пусть:

$$y_1(x, T) = \max\{y_{kn}: y_{kn} < x\}, y_2(x, T) = \min\{ y_{kn}/ y_{kn} > x\},$$

где y_{kn} — точки, определяющие разбиения T_n согласно (12). В соответствии с определением F_{in}:

$$F_{in}(y_j(x, T_n))= F(y_j(x, T_n)), i = 1, 2, j = 1, 2,$$

а потому

$$|F_{in}(\mathrm{x}) - F(x)| \leq F(y_2(x, T_n)) - F(y_1(x, T_n)), i = 1, 2.$$

В силу условия $\max(y_{kn} - y_{(k-1)n}) \to 0$ и непрерывности F в точке x правая часть последнего соотношения стремится к 0 при $n \to \infty$, что и заканчивает доказательство теоремы 2.

Теоремы 1 и 2 демонстрируют основные идеи предельной теории статистик интегрального типа и непараметрических критериев в целом. Как показывают эти теоремы, основную роль в рассматриваемой теории играет предельное соотношение (14). Отметим, что если $\delta(f, T_n) \to 0$ при $n \to \infty$, то (14) справедливо, но, вообще говоря, не наоборот. Естественно возникает еще ряд постановок. Пусть (14) выполнено для f_1 и f_2. При каких функциях h это соотношение выполнено для $h(x, f_1(x), f_2(x))$? В прикладной статистике вместо $f(x)$ рассматривают $f_\alpha(x, \omega)$ и $f(x, \omega)$, а вместо интегрирования по функциям распределения $F_n(x)$ — интегрирование по случайным мерам $F_\alpha(\omega)$. Как меняются формулировки в связи с такой заменой? В связи со слабой сходимостью (т.е. сходимостью по распределению) $A_T f_\alpha$ к A_T и переходом от $f_\alpha(x, \omega)$ к $h_\alpha(x, f_{1\alpha}(x, \omega), f_{2\alpha}(x, \omega))$ возникает следующая постановка. Пусть κ_α слабо сходится к к при $\alpha\to\infty$. Когда распределения $g_\alpha(\kappa_\alpha)$ сближаются с распределениями $g_\alpha(\kappa)$? Полным ответом на последний вопрос являются необходимые и

достаточные условия наследования сходимости. Они приведены в приложении 1.

Основные результаты. Наиболее общая теорема типа теоремы 1 выглядит так [29].

Теорема 3. Пусть существует последовательность разбиений T_n, $n = 1, 2, ...$, такая, что при $n \to \infty$ и $\alpha \to \infty$:

$$\Delta(f_\alpha, T_n) = \sum_{C \in T_n} \delta(f_\alpha, C) F(C) \to 0. \tag{23}$$

Пусть для любого C, входящего хотя бы в одно из разбиений T_n,

$$F_\alpha(C, \omega) \to F(C) \tag{24}$$

при $\alpha \to \infty$ (сходимость по вероятности). Пусть f_α асимптотически ограничены по вероятности при $\alpha \to \infty$. Тогда

$$\xi(f_\alpha, F_\alpha) - \xi(f_\alpha, F) \to 0 \tag{25}$$

при $\alpha \to \infty$ (сходимость по вероятности).

Как известно, полное сепарабельное метрическое пространство называется польским. Это понятие понадобится для формулировки аналога теоремы 2.

Теорема 4. Пусть X — польское пространство, Y конечномерно, существует измельчающаяся последовательность T_n разбиений, для которой соотношение (23) не выполнено. Тогда существует удовлетворяющая (24) последовательность F_α, для которой соотношение (25) неверно, хотя F_α слабо сходится к F при $\alpha \to \infty$.

Условие (23) естественно назвать условием римановости, поскольку в случае, рассмотренном в теореме 1, оно является условием интегрируемости по Риману — Стилтьесу. Рассмотрим *наследуемость римановости* при переходе от $f_{1\alpha}(x, \omega)$ со значениями в Y_1 и $f_{2\alpha}(x, \omega)$ со значениями в Y_2, удовлетворяющих (23), к $h_\alpha(x, f_{1\alpha}(x, \omega), f_{2\alpha}(x, \omega))$ со значениями в Y_3.

Положим:

$$Y_k(a, \varepsilon) = \{(y, y') : y \in Y_k, y' \in Y_k, \| y \|_k < a, \| y' \|_k < a, \| y - y' \|_k < \varepsilon\}, k = 1,2,$$

где $\|\cdot\|_k$ — норма (т.е. длина вектора) в пространстве Y_k, $k = 1, 2$. Рассмотрим также множества:

$$A(C,a,\varepsilon)=\{(x,x',y_1,y_1^*,y_2,y_2^*): x,x'\in C,(y_k,y_k^*)\in Y_k(a,\varepsilon),k=1,2\}$$

и функции:

$$q_\alpha(x,x',y_1,y_1^*,y_2,y_2^*)=h_\alpha(x,y_1,y_2)-h_\alpha(x',y_1^*,y_2^*).$$

Наконец, понадобится измеритель колеблемости:

$$c(h_\alpha,T,a,\varepsilon)=\sum_{C\in T}\sup_{A(C,a,\varepsilon)}\|q_\alpha\|_3 F(C)$$

и множество:

$$Z(a)=X\times\{y_1:\|y_1\|_1<a\}\times\{y_2:\|y_2\|_2<a\}.$$

Теорема 5. Пусть функции h_α асимптотически (при $\alpha\to\infty$) ограничены на множестве $Z(a)$ при любом положительном a. Пусть функции $f_{1\alpha}$ и $f_{2\alpha}$ асимптотически ограничены по вероятности и удовлетворяют условию (23). Пусть для участвующей в (23) последовательности T_n:

$$c(h_\alpha, T_n, a, \varepsilon)\to 0 \qquad (26)$$

при $\alpha\to\infty$, $n\to\infty$, $\varepsilon\to 0$ и любом положительном a. Тогда функции $f_{3\alpha}(x, \omega) = h_\alpha(x, f_{1\alpha}(x, \omega), f_{2\alpha}(x, \omega))$ удовлетворяют условию (23) и асимптотически ограничены по вероятности.

Теорема 6. Пусть условие (26) не выполнено для h_α. Тогда существуют детерминированные ограниченные функции $f_{1\alpha}$ и $f_{2\alpha}$ такие, что соотношение (23) выполнено для $f_{1\alpha}$ и $f_{2\alpha}$ и не выполнено для $f_{3\alpha}$.

Пример 3. Пусть $X = [0; 1]^k$, пространства Y_1 и Y_2 конечномерны, функция $h_\alpha \equiv h(x, y_1, y_2)$ непрерывна. Тогда условие (26) выполнено.

С помощью теорем 3 и 5 и результатов о наследовании сходимости можно изучить асимптотическое поведение статистик интегрального типа:

$$\xi_\alpha=\int_X h_\alpha(x,f_{1\alpha}(x,\omega),f_{2\alpha}(x,\omega))F_\alpha(dx,\omega)$$

со значениями в банаховом пространстве Y.

Теорема 7. Пусть для некоторой последовательности T_n разбиений X справедливы соотношения (23) для $f_{1\alpha}$ и $f_{2\alpha}$ и (24) для F_α. Пусть последовательность функций h_α удовлетворяет условию в теореме 5, конечномерные распределения $(f_{1\alpha}(x, \omega), f_{2\alpha}(x, \omega))$ слабо сходятся к конечномерным распределениям $(f_1(x, \omega), f_2(x, \omega))$, причем для f_1 и f_2 справедливо соотношение (23). Тогда:

$$\lim_{\alpha\to\infty} L(\xi_\alpha, \eta_\alpha) = 0,$$

где L — расстояние Прохорова (см. раздел П-3 прил. 1),

$$\eta_\alpha = \int_X h_\alpha(x, f_1(x,\omega), f_2(x,\omega))F(dx).$$

Теорема 7 дает общий метод получения асимптотических распределений статистик интегрального типа. Важно, что соотношение (23) выполнено для эмпирического процесса и для процессов, связанных с оцениванием параметров при проверке согласия [28].

Один из выводов общей теории состоит в том, что в качестве F_α можно использовать практически любую состоятельную оценку истинной функции распределения. Этот вывод использовался при построении критерия типа омега-квадрат для проверки симметрии распределения относительно 0 и обнаружения различий в связанных выборках (см. ниже).

Асимптотическое поведение критериев типа Колмогорова может быть получено с помощью описанного выше метода аппроксимации ступенчатыми функциями. Этот метод не требует обращения к теории сходимости вероятностных мер в функциональных пространствах. Для критериев Колмогорова и Смирнова достаточно использовать лишь свойства эмпирического процесса и броуновского моста. В случае проверки согласия добавляется необходимость изучения еще одного случайного процесса. Он является разностью между двумя функциями распределения. Одна — функция распределения элементов выборки. Вторая — случайный элемент параметрического семейства распределений, полученный путем подстановки оценок параметров вместо их истинных значений.

Статистика интегрального типа для проверки симметрии распределения. В прикладной статистике часто возникает необходимость проверки гипотезы о симметрии распределения относительно 0. Так, при проверке однородности связанных выборок необходимость проверки этой гипотезы основана

на следующем факте [6]. Если случайные величины X и Y независимы и одинаково распределены, то для функции распределения $H(x)=P(Z \le x)$ случайной величины $Z = X - Y$ выполнено, как нетрудно видеть, соотношение:

$$H(-x)=1 - H(x).$$

Это соотношение означает симметрию функции распределения относительно 0. Плотность такой функции распределения является четной функцией, ее значения в точках x и $(-x)$ совпадают. Проверка гипотезы однородности связанных выборок в наиболее общем случае сводится к проверке симметрии функции распределения разности $Z = X - Y$ относительно 0.

Рассмотрим методы проверки этой гипотезы. Сначала обсудим, какого типа отклонения от гипотезы симметрии можно ожидать при альтернативных гипотезах?

Рассмотрим сначала альтернативу сдвига:

$$H_{11} : G(x) = F(x + a).$$

В этом случае распределение Z при альтернативе отличается сдвигом от симметричного относительно 0. Для проверки гипотезы однородности может быть использован критерий знаковых рангов, разработанный Вилкоксоном (см., например, справочник [34, с. 46–53]).

Альтернативная гипотеза общего вида записывается как:

$$H_{12} : H(-x_0) \neq 1 - H(x_0)$$

при некотором x_0 . Таким образом, проверке подлежит гипотеза симметрии относительно 0, которую можно переписать в виде:

$$H(x) + H(-x) - 1 = 0 .$$

Для построенной по выборке $Z_j = X_j - Y_j$, $j = 1, 2, ..., n$, эмпирической функции распределения $H_n(x)$ последнее соотношение выполнено лишь приближенно:

$$H_n(x) + H_n(-x) - 1 \approx 0.$$

Как измерять отличие от 0? По тем же соображениям, что и в [6, п. 4.6], [54, п. 8.4], целесообразно использовать статистику типа омега-квадрат. Соответствующий критерий был предложен в работе [35]. Он имеет вид:

$$\omega_n^2 = \sum_{j=1}^{n} (H_n(Z_j) + H_n(-Z_j) - 1)^2. \qquad (27)$$

Представим эту статистику в интегральном виде. Рассмотрим выборочный процесс:

$$\xi_n(x) = \sqrt{n}(H_n(x) - H(x)).$$

При справедливости нулевой гипотезы:

$$(H_n(Z_j) + H_n(-Z_j) - 1)^2 = \frac{1}{n}(\xi_n(x) + \xi_n(-x))^2 = \frac{1}{n} f_n(x, \omega).$$

Положим:

$$F_n(x, \omega) = H_n(x).$$

Тогда, как легко видеть, статистика, заданная формулой (27), представляется в виде:

$$\omega_n^2 = \int_{-\infty}^{+\infty} f_n(x, \omega) dF_n(x, \omega).$$

Таким образом, асимптотическое поведение этой статистики может быть изучено с помощью описанной выше предельной теории статистик интегрального типа. Исторически ход мысли был обратным — сначала была построена и изучена статистика (27), а потом путем обобщения разработанных при анализе конкретной статистики методов исследования была построена общая теория, включающая в себя ряд необходимых и достаточных условий.

Критерий проверки гипотезы симметрии распределения относительно 0 с помощью статистики (27) является состоятельным, т.е. если функция рас-

пределения элементов выборки не удовлетворяет рассматриваемой гипотезе, то вероятность отклонения гипотезы стремится к 1 при росте объема выборки.

В работе [35] найдено предельное распределение этой статистики:

$$\lim_{n\to\infty} P(\omega_n^2 < x) = S_0(x).$$

В табл. 1 приведены критические значения статистики типа омега-квадрат для проверки симметрии распределения (и тем самым для проверки однородности связанных выборок), соответствующие наиболее распространенным значениям уровней значимости (расчеты проведены Г. В. Мартыновым; см. также [31]).

Таблица 1

Критические значения статистики ω_n^2 для проверки симметрии распределения

Значение функции распределения $S_0(x)$	**Уровень значимости** $\alpha = 1 - S_0(x)$	**Критическое значение** ***x*** **статистики** ω_n^2
0,90	0,10	1,20
0,95	0,05	1,66
0,99	0,01	2,80

Как следует из табл. 1, правило принятия решений при проверке симметрии распределения (или однородности связанных выборок) в наиболее общей постановке и при уровне значимости 5 % формулируется так. Вычислить статистику ω_n^2. Если $\omega_n^2 \le 1{,}66$, то принять гипотезу однородности. В противном случае — отвергнуть.

Пример. Пусть величины Z_j, $j = 1, 2, ..., 20$, таковы:

20, 18, (–2), 34, 25, (–17), 24, 42, 16, 26,
13, (–23), 35, 21, 19, 8, 27, 11, (–5), 7.

Соответствующий вариационный ряд $Z(1) < Z(2) < ... < Z(20)$ имеет вид:

$$(-23) < (-17) < (-5) < (-2) < 7 < 8 < 11 < 13 < 16 < 18 <$$
$$< 19 < 20 < 21 < 24 < 25 < 26 < 27 < 34 < 35 < 42.$$

Для расчета значения статистики ω_n^2 построим табл. 2 из 7 столбцов и 20 строк, не считая заголовков столбцов (сказуемого таблицы). В первом столбце указаны номера (ранги) членов вариационного ряда, во втором — сами эти члены, в третьем — значения эмпирической функции распределения при значениях аргумента, совпадающих с членами вариационного ряда. В следующем столбце приведены члены вариационного ряда с обратным знаком, а затем указываются соответствующие значения эмпирической функции распределения. Например, поскольку минимальное наблюдаемое значение равно (–23), то $H_n(x) = 0$ при $x < -23$, а потому для членов вариационного ряда с 14-го по 20-й в пятом столбце стоит 0. В качестве другого примера рассмотрим минимальный член вариационного ряда, т.е. (–23). Меняя знак, получаем 23. Это число стоит между 13-м и 14-м членами вариационного ряда, $21 < 23 < 24$. На этом интервале эмпирическая функция распределения совпадает со своим значением в левом конце, поэтому следует записать в пятом столбце значение 0,65. Остальные ячейки пятого столбца заполняются аналогично. На основе третьего и пятого столбцов элементарно заполняется шестой столбец, а затем и седьмой. Остается найти сумму значений, стоящих в седьмом столбце. Подобная таблица удобна как для ручного счета, так и при использовании электронных таблиц типа *Excel*.

Таблица 2

Расчет значения статистики ω_n^2 для проверки симметрии распределения

j	$Z(j)$	$H_n(Z(j))$	$-Z(j)$	$H_n(-Z(j))$	$H_n(Z(j)) + H_n(-Z(j)) - 1$	$(H_n(Z(j)) + H_n(-Z(j)) - 1)^2$
1	–23	0,05	23	0,65	–0,30	0,09
2	–17	0,10	17	0,45	–0,45	0,2025
3	–5	0,15	5	0,20	–0,65	0,4225
4	–2	0,20	2	0,20	–0,60	0,36
5	7	0,25	–7	0,10	–0,65	0,4225
6	8	0,30	–8	0,10	–0,60	0,36
7	11	0.35	–11	0,10	–0,55	0,3025
8	13	0,40	–13	0,10	–0,50	0,25
9	16	0,45	–16	0,10	–0,45	0,2025
10	18	0,50	–18	0,05	–0,45	0,2025
11	19	0,55	–19	0,05	–0,40	0,16
12	20	0,60	–20	0,05	–0,35	0,1225
13	21	0,65	–21	0,05	–0,30	0,09

j	$Z(j)$	$H_n(Z(j))$	$-Z(j)$	$H_n(-Z(j))$	$H_n(Z(j)) + H_n(-Z(j)) - 1$	$(H_n(Z(j)) + H_n(-Z(j)) - 1)^2$
14	24	0,70	–24	0	–0,30	0,09
15	25	0,75	–25	0	–0,25	0,0625
16	26	0,80	–26	0	–0,20	0,04
17	27	0,85	–27	0	–0,15	0,0225
18	34	0,90	–34	0	–0,10	0,01
19	35	0,95	–35	0	–0,05	0,0025
20	42	1,00	–42	0	0	0

Результаты расчетов (суммирование значений по седьмому столбцу табл. 2) показывают, что значение статистики $\omega_n^2 = 3{,}055$. В соответствии с табл. 1 это означает, что на любом используемом в прикладных статистических исследованиях уровнях значимости отклоняется гипотеза симметрии распределения относительно 0 (а потому и гипотеза однородности в связанных выборках).

2.7. МЕТОДЫ ВОССТАНОВЛЕНИЯ ЗАВИСИМОСТЕЙ

Сначала рассмотрим параметрические постановки задач регрессионного анализа (восстановления зависимостей) в пространствах произвольной природы, затем — непараметрические, после чего перейдем к оцениванию нечисловых параметров в классической ситуации, когда отклик и факторы принимают числовые значения.

Задача аппроксимации зависимости (параметрической регрессии). Пусть X и Y — некоторые пространства. Пусть имеются статистические данные — n пар (x_k, y_k), где $x_k \in X$, $y_k \in Y$, $k = 1, 2, \ldots, n$. Задано параметрическое пространство Θ произвольной природы и семейство функций $g(x, \theta)$: $X \times \Theta \to Y$. Требуется подобрать параметр $\theta \in \Theta$ так. чтобы $g(x_k, \theta)$ наилучшим образом приближали y_k, $k = 1, 2, \ldots, n$. Пусть f_k — последовательность показателей различия в Y. При сделанных предположениях параметр θ естественно оценивать путем решения экстремальной задачи:

$$\theta_n = Arg \min_{\theta \in \Theta} \sum_{k=1}^{n} f_k(g(x_k, \theta), y_k). \qquad (1)$$

Часто, но не всегда, все f_k совпадают. В классической постановке, когда $X = R^k$, $Y = R^1$, функции f_k различны при неравноточных наблюдениях, например, когда число опытов меняется от одной точки x проведения опытов к другой.

Если $f_k(y_1, y_2) = f(y_1, y_2) = (y_1 - y_2)^2$, то получаем общую постановку метода наименьших квадратов (см. подробности, например, в [6, гл. 5]):

$$\theta_n = Arg \min_{\theta \in \Theta} \sum_{k=1}^{n} (g(x_k, \theta) - y_k)^2 .$$

В рамках детерминированного анализа данных остается единственный теоретический вопрос — о существовании θ_n. Если все участвующие в формулировке задачи (1) функции непрерывны, а минимум берется по бикомпакту, то θ_n существует. Есть и иные условия существования θ_n [4, 36, 37].

При появлении нового наблюдения x в соответствии с методологией восстановления зависимости рекомендуется выбирать оценку соответствующего y по правилу:

$$y^* = g(x, \theta_n).$$

Обосновать такую рекомендацию в рамках детерминированного анализа данных невозможно. Это можно сделать только в вероятностной теории, равно как и изучить асимптотическое поведение θ_n, доказать состоятельность этой оценки.

Как и в классическом случае, вероятностную теорию целесообразно строить для трех различных постановок.

1. Переменная x — детерминированная (например, время), переменная y — случайная, ее распределение зависит от x.

2. Совокупность (x_k, y_k), $k = 1, 2, \ldots, n$, — выборка из распределения случайного элемента со значениями в $X \times Y$.

3. Имеется детерминированный набор пар (x_{k0}, y_{k0}), $k = 1, 2, \ldots, n$, результат наблюдения (x_k, y_k) является случайным элементом, распределение которого зависит от (x_{k0}, y_{k0}). Это — постановка так называемого конфлюэнтного анализа.

Во всех трех случаях:

$$f_n(\omega, \theta) = \sum_{k=1}^{n} f_k(g(x_k, \theta), y_k),$$

однако случайность входит в правую часть по-разному в зависимости от постановки, от которой зависит и определение предельной функции $f(\theta)$.

Проще всего выглядит $f(\theta)$ в случае второй постановки при $f_k \equiv f$:

$$f(\theta) = Mf(g(x_1, \theta), y).$$

В случае первой постановки:

$$f(\theta) = \lim_{n\to\infty} \sum_{k=1}^{n} Mf_k(g(x_k,\theta), y_k(\omega))$$

в предположении существования указанного предела. Ситуация усложняется для третьей постановки:

$$f(\theta) = \lim_{n\to\infty} \sum_{k=1}^{n} Mf_k(g(x_k(\omega),\theta), y_k(\omega)) .$$

Во всех трех случаях на основе общих результатов о поведении решений экстремальных статистических задач можно изучить [4, 36, 37] асимптотику оценок θ_n. При выполнении соответствующих внутриматематических условий регулярности оценки оказываются состоятельными, т.е. удается восстановить зависимость.

Аппроксимация и регрессия. Соотношение (1) дает решение задачи аппроксимации. Поясним, как эта задача соотносится с нахождением регрессии. Согласно [38] для случайной величины (ξ, η) со значениями в $X\times Y$ регрессией η на ξ относительно меры близости f естественно назвать решение задачи:

$$Mf(g(\xi), \eta) \to \min_{g} , \qquad (2)$$

где f: $Y \times Y \to R^1$, g: $X \to Y$, минимум берется по множеству всех измеримых функций.

Можно исходить и из формально другого определения. Для каждого $x\in X$ рассмотрим случайную величину $\eta(x)$, распределение которой является условным распределением η при условии $\xi = x$. В соответствии с определением математического ожидания в пространстве общей природы назовем условным математическим ожиданием решение экстремальной задачи:

$$M(\eta \mid \xi = x) = Arg\min\{ Mf(y, \eta(x)), y \in Y\} .$$

Оказывается, при обычных предположениях измеримости решение задачи (2) совпадает с $M(\eta \mid \xi = x)$. (Внутриматематические уточнения типа «равенство имеет место почти всюду» здесь опущены.)

Если заранее известно, что условное математическое ожидание $M(\eta \mid \xi = x)$ принадлежит некоторому параметрическому семейству $g(x, \theta)$, то задача

нахождения регрессии сводится к оцениванию параметра θ в соответствии с рассмотренной выше второй постановкой вероятностной теории параметрической регрессии.

Если же нет оснований считать, что регрессия принадлежит некоторому параметрическому семейству, то можно использовать непараметрические оценки регрессии. Они строятся с помощью непараметрических оценок плотности (см. раздел 2.5).

Непараметрические методы восстановления зависимости. Пусть ν_1 — мера в X, ν_2 — мера в Y, а их прямое произведение $\nu = \nu_1 \times \nu_2$ — мера в $X \times Y$. Пусть $g(x, y)$ — плотность случайного элемента (ξ, η) по мере ν. Тогда условная плотность $g(y \mid x)$ распределения η при условии $\xi = x$ имеет вид:

$$g(y \mid x) = \frac{g(x, y)}{\int\limits_Y g(x, y)\nu_2(dy)} \tag{3}$$

(в предположении, что интеграл в знаменателе отличен от 0). Следовательно,

$$Mf(y, \eta(x)) = \int\limits_Y f(y, a)g(a \mid x)\nu_2(da),$$

а потому:

$$M(\eta \mid \xi = x) = Arg\min_{y \in Y} Mf(y, \eta(x)) = Arg\min_{y \in Y} \int\limits_Y f(y, a)g(a \mid x)\nu_2(da).$$

Заменяя $g(x, y)$ в (3) непараметрической оценкой плотности $g_n(x, y)$, получаем оценку условной плотности:

$$g_n(y \mid x) = \frac{g_n(x, y)}{\int\limits_Y g_n(x, y)\nu_2(dy)}. \tag{4}$$

Если $g_n(x, y)$ — состоятельная оценка $g(x, y)$, то числитель (4) сходится к числителю (3). Сходимость знаменателя (4) к знаменателю (3) обосновывается с помощью предельной теории статистик интегрального типа (см. раздел 2.6). В итоге получаем утверждение о состоятельности непараметрической оценки (4) условной плотности (3).

Непараметрическая оценка регрессии ищется как:

$$M_n(\eta \mid \xi = x) = Arg \min_{y \in Y} \int_Y f(y,a) g_n(a \mid x) \nu_2(da).$$

Состоятельность этой оценки следует из приведенных выше общих результатов об асимптотическом поведении решений экстремальных статистических задач.

Оценивание объектов нечисловой природы в классических постановках регрессионного анализа. Нечисловая статистика тесно связана с классическими областями прикладной статистики. Ряд трудностей в классических постановках удается понять и разрешить лишь с помощью общих результатов прикладной статистики. В частности, это касается оценивания параметров, когда параметр имеет нечисловую природу.

Рассмотрим типовую прикладную постановку задачи восстановления регрессионной зависимости, линейной по параметрам (см. также [6, глава 5.1]). Исходные данные имеют вид $(x_i, y_i) \in R^2$, $i = 1,2,...,n$. Цель состоит в том, чтобы с достаточной точностью описать y как многочлен (полином) от x, т.е. модель имеет вид:

$$y_i = \sum_{k=0}^{m} a_k x_i^k + \varepsilon_i, \quad i = 1,2,...,n, \tag{5}$$

где m — неизвестная степень полинома; $a_0, a_1, a_2, ..., a_m$ — неизвестные коэффициенты многочлена; ε_i, $i = 1,2,...,n$ — погрешности, которые для простоты примем независимыми и имеющими одно и то же нормальное распределение с нулевым математическим ожиданием и дисперсией σ^2.

Замечание. Здесь наглядно проявляется одна из причин живучести вероятностно-статистических моделей на основе нормального распределения. Такие модели, как правило, неадекватны реальной ситуации [6, глава 4.1]. Однако с математической точки зрения они позволяют глубже проникнуть в суть изучаемого явления. Поэтому такие модели полезны для первоначального анализа ситуации. В ходе дальнейших исследований необходимо снять нереалистическое предположение нормальности и перейти к непараметрическим моделям.

В прикладной статистике часто используют следующую технологию анализа данных. Сначала пытаются применить модель (5) для линейной функции (m = 1), при неудаче (неадекватности модели) переходят к многочлену второго порядка (m = 2), если снова неудача, то берут модель (2) с m = 3 и т.д. Адекватность модели обычно проверяют по F-критерию Фишера.

Обсудим свойства этой процедуры. Если степень полинома задана ($m = m_0$), то его коэффициенты оценивают методом наименьших квадратов, свойства этих оценок хорошо известны. Однако в рассматриваемой постановке m тоже является неизвестным параметром и подлежит оценке. Таким образом, требуется оценить объект (m, a_0, a_1, a_2, ..., a_m), множество значений которого можно описать как $R^1 \cup R^2 \cup R^3 \cup ...$. Это — объект нечисловой природы, обычные методы оценивания для него неприменимы. Разработанные к настоящему времени методы оценивания степени полинома носят в основном эвристический характер (см., например, гл. 12 монографии [39]). Рассмотрим некоторые из них.

Оценивание степени полинома. Полезно рассмотреть основной показатель качества регрессионной модели (5). Одни и те же данные можно обрабатывать различными способами. На первый взгляд, показателем отклонений данных от модели может служить остаточная сумма квадратов SS. Чем этот показатель меньше, тем приближение лучше, значит, и модель лучше описывает реальные данные. Однако это рассуждение годится только для моделей с одинаковым числом параметров. Ведь если добавляется новый параметр, по которому можно минимизировать, то и минимум, как правило, оказывается меньше.

В качестве основного показателя качества регрессионной модели используют следующую оценку остаточной дисперсии:

$$\hat{\sigma}^2(m) = \frac{SS}{n-m-1}.$$

Таким образом, вводят корректировку на число параметров, оцениваемых по наблюдаемым данным. Корректировка состоит в уменьшении знаменателя на указанное число. В модели (5) это число равно (m+1). В случае задачи восстановления линейной функции одной переменной оценка остаточной дисперсии имеет вид:

$$\hat{\sigma}^2 = \frac{SS}{n-2},$$

поскольку число оцениваемых параметров $m + 1 = 2$.

Еще раз — почему *при подборе вида модели* знаменатель дроби, оценивающей остаточную дисперсию, приходится корректировать на число параметров? Если этого не делать, то придется заключить, что всегда многочлен второй степени лучше соответствует данным, чем линейная функция, многочлен третьей степени лучше приближает исходные данные, чем многочлен

второй степени, и т.д. В конце концов доходим до многочлена степени (n-1) с n коэффициентами, который проходит через все заданные точки. Но его прогностические возможности, скорее всего, существенно меньше, чем даже у линейной функции. *Излишнее усложнение статистических моделей вредно.* Типовое поведение скорректированной оценки остаточной дисперсии:

$$v(m) = \hat{\sigma}^2(m)$$

в случае расширяющейся системы моделей (т.е. при возрастании натурального параметра m) выглядит так. Сначала наблюдаем заметное убывание. Затем оценка остаточной дисперсии колеблется около некоторой константы (дисперсии погрешности).

Поясним ситуацию на примере модели восстановления зависимости, выраженной многочленом:

$$x(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + \ldots + a_m t^m.$$

Пусть эта модель справедлива при $m = m_0$. При $m < m_0$ в скорректированной оценке остаточной дисперсии учитываются не только погрешности измерений, но и соответствующие (старшие) члены многочлена (предполагаем, что коэффициенты при них отличны от 0). При $m \geq m_0$ имеем:

$$\lim_{n\to\infty} v(m) = \sigma^2.$$

Следовательно, скорректированная оценка остаточной дисперсии будет колебаться около указанного предела. Поэтому представляется естественным, что в качестве оценки неизвестной статистику степени многочлена (полинома) можно использовать первый локальный минимум скорректированной оценки остаточной дисперсии, т.е.

$$m^* = \min\{\, m : v(m-1) > v(m), \quad v(m) \leq v(m+1)\}.$$

В работе [40] найдено предельное распределение этой оценки степени многочлена.

Теорема. При справедливости некоторых условий регулярности:

$$\lim_{n\to\infty} P(m^* < m_0) = 0, \quad \lim_{n\to\infty} P(m^* = m_0 + u) = \lambda(1-\lambda)^u, \quad u = 0,1,2,\ldots,$$

где

$$\lambda = \Phi(1) - \Phi(-1) = \frac{1}{\sqrt{2\pi}} \int_{-1}^{1} \exp\left\{-\frac{x^2}{2}\right\} dx \approx 0{,}68268.$$

Таким образом, предельное распределение оценки m^* степени многочлена (полинома) является геометрическим. Это означает, в частности, что оценка не является состоятельной. При этом вероятность получить меньшее значение, чем истинное, исчезающе мала. Далее имеем:

$$P(m^* = m_0) \to 0{,}68268, \quad P(m^* = m_0 + 1) \to 0{,}68268(1 - 0{,}68268) = 0{,}21663,$$

$$P(m^* = m_0 + 2) \to 0{,}68268(1 - 0{,}68268)^2 = 0{,}068744,$$

$$P(m^* = m_0 + 3) \to 0{,}68268(1 - 0{,}68268)^3 = 0{,}021814...$$

Разработаны и иные методы оценивания неизвестной степени многочлена, например, путем многократного применения процедуры проверки адекватности регрессионной зависимости с помощью критерия Фишера. Предельное поведение таких оценок — таково же, как в приведенной выше теореме, только значение параметра λ иное. Отметим, что для степени многочлена давно предложены состоятельные оценки [41]. Для этого достаточно уровень значимости (при проверке адекватности регрессионной зависимости с помощью критерия Фишера) сделать убывающим при росте объема выборки.

Построение информативного подмножества признаков. В более общем случае многомерной линейной регрессии данные имеют вид (y_i, X_i), $i = 1,2,...,n$, где $X_i = (x_{i1}, x_{i2}, ..., x_{iN}) \in R^N$ — вектор предикторов (факторов, объясняющих переменных), а модель такова:

$$y_i = \sum_{j \in K} a_j x_{ij} + \varepsilon_i, \quad i = 1,2,...,n \qquad (6)$$

(здесь K — некоторое подмножество множества $\{1, 2, ..., n\}$; ε_i — те же, что и в модели (5); a_j — неизвестные коэффициенты при предикторах с номерами из K). Множество K называют *информативным подмножеством признаков*, поскольку согласно формуле (6) остальные признаки можно отбросить без потери информации. Проблема состоит в том, что при анализе реальных дан-

ных неизвестно, какие признаки входят в K, а какие нет. Ясна важность оценивания информативного подмножества признаков.

Модель (5) сводится к модели (6), если:

$$x_{i1} = 1,\ x_{i2} = x_i,\ x_{i3} = x_i^2,\ x_{i4} = x_i^3, \ldots, x_{ij} = x_i^{j-1}, \ldots$$

В модели (5) есть естественный порядок ввода предикторов в рассмотрение — в соответствии с возрастанием степени многочлена, а в модели (6) естественного порядка нет, поэтому здесь приходится рассматривать произвольное подмножество множества предикторов. Есть только частичный порядок — чем мощность подмножества меньше, тем лучше. Модель (6) особенно актуальна в технических исследованиях (см. многочисленные примеры в журнале «Заводская лаборатория»). Она применяется в задачах управления качеством продукции и других технико-экономических исследованиях, в медицине, экономике, маркетинге и социологии, когда из большого числа факторов, предположительно влияющих на изучаемую переменную, надо отобрать по возможности наименьшее число значимых факторов и с их помощью сконструировать прогнозирующую формулу (6).

Задача оценивания модели (6) разбивается на две последовательные задачи: оценивание множества K — подмножества множества всех предикторов, а затем — неизвестных параметров a_j. Методы решения второй задачи хорошо известны и подробно изучены (обычно используют метод наименьших квадратов). Гораздо хуже обстоит дело с оцениванием объекта нечисловой природы K. Как уже отмечалось, существующие методы — в основном эвристические, они зачастую не являются даже состоятельными. Даже само понятие состоятельности в данном случае требует специального определения.

Определение. Пусть K_0 — истинное подмножество предикторов, т.е. подмножество, для которого справедлива модель (6), а подмножество предикторов K_n — его оценка. Оценка K_n называется состоятельной, если:

$$\lim_{n \to \infty} Card(K_n \Delta K_0) = 0,$$

где Δ — символ симметрической разности множеств; $Card(K)$ означает число элементов множества K, а предел понимается в смысле сходимости по вероятности.

Задача оценивания в моделях регрессии, таким образом, разбивается на две — оценивание структуры модели и оценивание параметров при за-

данной структуре. В модели (5) структура описывается неотрицательным целым числом m, в модели (6) — множеством K. Структура — объект нечисловой природы. Задача ее оценивания сложна, в то время как задача оценивания численных параметров при заданной структуре хорошо изучена, разработаны эффективные (в смысле прикладной математической статистики) методы.

Такова же ситуация и в других методах многомерного статистического анализа — в факторном анализе (включая метод главных компонент) и в многомерном шкалировании, в иных оптимизационных постановках проблем прикладного многомерного статистического анализа.

Множество K и параметры a_j линейной зависимости можно оценивать путем решения задачи оптимизации:

$$\sum_{i=1}^{n}\left(y_i - \sum_{j\in K} a_j x_{ij}\right)^2 \to \min\ , \qquad (7)$$

в которой минимум берется по K, a_j, $j \in K$. Математическая природа множества, по которому проводится минимизация, весьма сложна. Это и объясняет тот факт, что к настоящему времени разработано много эвристических методов оценивания информативного множества параметров K, свойства которых плохо изучены. На основе общих результатов нечисловой статистики об асимптотическом поведении решений экстремальных статистических задач удалось показать, что оценки, полученные путем решения задачи (7), являются состоятельными [42].

2.8. МЕТОДЫ КЛАССИФИКАЦИИ

Как известно, математический аппарат нечисловой статистики базируется на использовании расстояний (мер близости, показателей различия) в пространствах таких объектов. Это вызвано отсутствием в таких пространствах операций суммирования, на которых основано большинство методов других областей статистики. Любые методы, использующие только расстояния (меры близости, показатели различия) между объектами, следует относить к нечисловой статистике, поскольку такие методы могут работать с объектами произвольного пространства, если в нем задана метрика или ее анало-

ги. Таким образом, весьма многие методы классической прикладной статистики следует включать в нечисловую статистику.

В настоящем разделе рассматривается важное направление прикладной статистики — математические методы классификации. Значительную их часть следует отнести к нечисловой статистике, а именно, методы классификации, основанные на расстояниях между объектами.

Основные направления в математической теории классификации. Какие научные исследования относить к этой теории? Исходя из потребностей специалиста, применяющего математические методы классификации, целесообразно принять, что сюда входят исследования, во-первых, отнесенные самими авторами к этой теории; во вторых, связанные с ней общностью тематики, хотя бы их авторы и не упоминали термин «классификация».

В литературных источниках наряду с термином «классификация» в близких смыслах используются термины «группировка», «распознавание образов», «диагностика», «дискриминация», «сортировка», «типология», «систематика», «районирование», «сегментирование» и др. Терминологический разнобой связан прежде всего с традициями научных кланов, к которым относятся авторы публикаций, а также с внутренним делением самой теории классификации.

В научных исследованиях по современной теории классификации можно выделить два относительно самостоятельных направления. Одно из них опирается на опыт таких наук, как биология, география, геология, и таких прикладных областей, как ведение классификаторов продукции и библиотечное дело. Типичные объекты рассмотрения — классификация химических элементов (таблица Д. И. Менделеева), биологическая систематика, универсальная десятичная классификация публикаций (УДК), классификатор товаров на основе штрих-кодов.

Другое направление опирается на опыт технических исследований, экономики, маркетинговых исследований, социологии, медицины. Типичные задачи — техническая и медицинская диагностика, а также, например, разбиение на группы отраслей промышленности, тесно связанных между собой, выделение групп однородной продукции. Обычно используются такие термины, как «распознавание образов», «кластер-анализ» или «дискриминантный анализ». Это направление обычно опирается на математические модели; для проведения расчетов интенсивно используется компьютеры.

В 60-х гг. XX в. внутри статистических методов достаточно четко оформилась область, посвященная методам классификации. Несколько мо-

дифицируя формулировки М. Дж. Кендалла и А. Стьюарта 1966 г. (см. русский перевод [43, с. 437]), в теории классификации выделим три подобласти: дискриминация (дискриминантный анализ), кластеризация (кластер-анализ), группировка. Опишем эти подобласти.

В дискриминантном анализе классы предполагаются заданными — плотностями вероятностей или обучающими выборками. Задача состоит в том, чтобы вновь поступающий объект отнести в один из этих классов. У понятия «дискриминация» имеется много синонимов: диагностика, распознавание образов с учителем, автоматическая классификация с учителем, статистическая классификация и т.д.

При кластеризации и группировке целью является выявление и выделение классов. Синонимы: построение классификации, распознавание образов без учителя, автоматическая классификация без учителя, типология, таксономия и др. При этом задача кластер-анализа состоит в выяснении по эмпирическим данным, насколько элементы «группируются» или распадаются на изолированные «скопления», «кластеры» (от *cluster* (англ.) — гроздь, скопление). Иными словами, задача — выявление естественного разбиения на классы, свободного от субъективизма исследователя, а цель — выделение групп однородных объектов, сходных между собой, при резком отличии этих групп друг от друга.

При группировке, наоборот, «мы хотим разбить элементы на группы независимо от того, естественны ли границы разбиения или нет» [43, с. 437]. Цель по-прежнему состоит в выявлении групп однородных объектов, сходных между собой (как в кластер-анализе), однако «соседние» группы могут не иметь резких различий (в отличие от кластер-анализа). Границы между группами условны, не являются естественными, зависят от субъективизма исследователя. Аналогично при лесоустройстве проведение просек (границ участков) зависит от специалистов лесного ведомства, а не от свойств леса.

Задачи кластеризации и группировки принципиально различны, хотя для их решения могут применяться одни и те же алгоритмы. Важная для практической деятельности проблема состоит в том, чтобы понять, разрешима ли задача кластер-анализа для конкретных данных или возможна только их группировка, поскольку совокупность объектов достаточно однородна и не разбивается на резко разделяющиеся между собой кластеры.

Как правило, в математических задачах кластеризации и группировки основное — выбор метрики, расстояния между объектами, меры близости, сходства, различия. Хорошо известно, что для любого заданного разбиения

объектов на группы и любого положительного числа $\varepsilon > 0$ можно указать метрику такую, что расстояния между объектами из одной группы будут меньше ε, а между объектами из разных групп — больше $1/\varepsilon$. Тогда любой разумный алгоритм кластеризации даст именно заданное разбиение.

Понимание и обсуждение постановок задач осложняется использованием одного и того же термина в разных смыслах. Термином «классификация» (и сходными терминами, такими, как «диагностика») обозначают по крайней мере три разные сущности. Во-первых, процедуру построения классификации (и выделение классов, используемых при диагностике). Во-вторых, построенную классификацию (систему выделенных классов). В-третьих, процедуру ее использования (правила отнесения вновь поступающего объекта к одному из ранее выделенных классов). Другими словами, имеем естественную триаду: построение — изучение — использование классификации.

Как уже отмечалось, для построения системы диагностических классов используют разнообразные методы кластерного анализа и группировки объектов. Наименее известен второй член триады (отсутствующий у Кендалла и Стьюарта [43]) — изучение отношений эквивалентности, полученных в результате построения системы диагностических классов. Статистический анализ полученных, в частности экспертами, отношений эквивалентности — часть статистики бинарных отношений и тем самым — нечисловой статистики (см. главу 3).

Диагностика в узком смысле слова (процедура использования классификации, т.е. отнесения вновь поступающего объекта к одному из выделенных ранее классов) — предмет дискриминантного анализа. Отметим, что с точки зрения нечисловой статистики дискриминантный анализ является частным случаем общей схемы регрессионного анализа, соответствующим ситуации, когда зависимая переменная принимает конечное число значений, а именно — номера классов, а вместо квадрата разности стоит функция потерь от неправильной классификации. Однако есть ряд специфических постановок и методов принятия решений, выделяющих задачи диагностики среди всех регрессионных задач.

О построении диагностических правил. Начнем с краткого обсуждения одного распространенного заблуждения. Предположим, что исходные статистические данные явно неоднородны. Иногда рекомендуют сначала построить систему диагностических классов (т.е. кластеров), а потом в каждом диагностическом классе отдельно проводить регрессионный анализ или применять иные статистические методы.. Однако обычно забывают, что при

этом нельзя опираться на вероятностную модель многомерного нормального распределения, так как распределение результатов наблюдений, попавших в определенный кластер, будет отнюдь не нормальным, а усеченным нормальным (усечение определяется фиксированными границами кластера) или более сложным (если границы кластера случайны). Одна из возможных рекомендаций [44] — применять в таких случаях робастные методы восстановления зависимостей.

Состоятельная оценка числа классов. Процедуры построения диагностических правил делятся на вероятностные и детерминированные. К первым относятся так называемые задачи расщепления смесей. В них предполагается, что распределение вновь поступающего случайного элемента является смесью вероятностных законов, соответствующих диагностическим классам. Как и при выборе степени полинома в регрессии (см. предыдущий раздел), при анализе реальных статистических данных (технических, социально-экономических, медицинских и др.) встает вопрос об оценке числа элементов смеси, т.е. числа диагностических классов. Были изучены результаты применения обычно рекомендуемого критерия Уилкса для оценки числа элементов смеси. Оказалось (см. статью [44]), что оценка с помощью известного критерия Уилкса не является состоятельной, асимптотическое распределение этой оценки — геометрическое, как и в случае задачи восстановления зависимости в регрессионном анализе. Итак, в [44] продемонстрирована несостоятельность обычно используемых оценок. Для получения состоятельных оценок достаточно связать уровень значимости в критерии Уилкса с объемом выборки, как это было предложено и для задач регрессии [40, 41]. Таким образом, разработаны состоятельные оценки такого нечислового объекта, как число классов.

Как уже отмечалось, задачи построения системы диагностических классов целесообразно разбить на два типа — с четко разделенными кластерами (задачи кластер-анализа) и с условными границами, непрерывно переходящими друг в друга классами (задачи группировки). Такое деление полезно, хотя в обоих случаях могут применяться одинаковые алгоритмы.

Сколько существует алгоритмов построения системы диагностических правил? Иногда называют то или иное число. На самом же деле их бесконечно много, в чем нетрудно убедиться. Действительно, рассмотрим один определенный алгоритм — алгоритм средней связи. Он основан на использовании некоторой меры близости $d(x,y)$ между объектами x и y. Как он работает? На первом шаге каждый объект рассматривается как отдельный кластер. На каждом следующем шаге объединяются две ближайших класте-

ра. Расстояние между объектами рассчитывается как средняя связь (отсюда и название алгоритма), т.е. как среднее арифметическое расстояний между парами объектов, один из которых входит в первый кластер, а другой — во второй. В конце концов все объекты объединяются вместе, и результат работы алгоритма представляет собой дерево последовательных объединений (в терминах теории графов), или «Дендрограмму». Из нее можно выделить кластеры разными способами. Один подход — исходя из заданного числа кластеров. Второй — из соображений предметной области. Третий — исходя из устойчивости (если разбиение долго не менялось при возрастании порога объединения — значит, оно отражает реальность) и т.д.

К алгоритму средней связи естественно сразу добавить алгоритм ближайшего соседа. В этом алгоритме расстоянием между кластерами называется минимальное из расстояний между парами объектов, один из которых входит в первый кластер, а другой — во второй. А также и алгоритм дальнего соседа (когда расстоянием между кластерами называется максимальное из расстояний между парами объектов, один из которых входит в первый кластер, а другой — во второй).

Каждый из трех описанных алгоритмов (средней связи, ближайшего соседа, дальнего соседа), как легко проверить, порождает бесконечное (континуальное) семейство алгоритмов кластер-анализа. Дело в том, что величина $d^a(x, y)$, $a > 0$, также является мерой близости между x и y и порождает новый алгоритм. Если параметр a пробегает отрезок, то получается бесконечно много алгоритмов классификации.

Устойчивость — критерий естественности классификации. Каким из них пользоваться при обработке данных? Дело осложняется тем, что практически в любом пространстве статистических данных существует весьма много мер близости различных видов. Именно в связи с обсуждаемой проблемой следует указать на принципиальное различие между кластер-анализом и задачами группировки.

Если классы реальны, естественны, существуют на самом деле, четко отделены друг от друга, то любой алгоритм кластер-анализа их выделит. Следовательно, *в качестве критерия естественности классификации следует рассматривать устойчивость результата классификации относительно выбора алгоритма кластер-анализа.*

Проверить устойчивость можно, применив к одним и тем же данным несколько подходов, например, столь непохожие алгоритмы, как «ближнего соседа» и «дальнего соседа». Если полученные результаты содержательно

близки, то они адекватны действительности. В противном случае следует предположить, что естественной классификации не существует, задача кластер-анализа не имеет решения, и можно проводить только группировку.

Как уже отмечалось, часто применяется так называемый агломеративный иерархический алгоритм «Дендрограмма», в котором вначале все элементы рассматриваются как отдельные кластеры, а затем на каждом шагу объединяются два наиболее близких кластера. Для работы «Дендрограммы» необходимо задать правило вычисления расстояния между кластерами. Оно вычисляется через расстояние $d(x, y)$ между элементами x и y. Поскольку $d^a(x, y)$ при $0 < a < 1$ также расстояние, то, как правило, существует бесконечно много различных вариантов этого алгоритма. Представим себе, что они применяются для обработки одних и тех же реальных данных. Если при всех a получается одинаковое разбиение элементов на кластеры, т.е. результат работы алгоритма устойчив по отношению к изменению a (в смысле общей схемы устойчивости, введенной в [1]), то имеем «естественную» классификацию. В противном случае результат зависит от субъективно выбранного исследователем параметра a, т.е. задача кластер-анализа неразрешима (предполагаем, что выбор a нельзя специально обосновать). Задача группировки в этой ситуации имеет много решений. Из них можно выбрать одно по дополнительным критериям.

Следовательно, получаем эвристический критерий: если решение задачи кластер-анализа существует, то оно находится с помощью любого алгоритма. Целесообразно использовать наиболее простой.

Продолжим обсуждение в более широком контексте.

Проблема поиска естественной классификации. Существуют различные точки зрения на эту проблему. Естественная классификация обычно противопоставляется искусственной. На Всесоюзной школе-семинаре «Использование математических методов в задачах классификации» (г. Пущино, 1986 г.), в частности, были высказаны мнения, что естественная классификация:

- закон природы;

- основана на глубоких закономерностях, тогда как искусственная классификация — на неглубоких;

- для конкретного индивида та, которая наиболее быстро вытекает из его тезауруса;

- удовлетворяет многим целям; цель искусственной классификации задает человек;

- классификация с точки зрения потребителя продукции;
- классификация, позволяющая делать прогнозы;
- имеет критерием устойчивость.

Приведенные высказывания уже дают представление о больших расхождениях в понимании «естественной классификации». Этот термин следует признать нечетким, как, впрочем, и многие другие термины, и профессиональные — социально-экономические, научно-технические, и используемые в обыденном языке. Нетрудно подробно обосновать нечеткость естественного языка и тот факт, что «мы мыслим нечетко», что, однако, не слишком мешает нам решать производственные и жизненные проблемы. Кажущееся рациональным требование выработать сначала строгие определения, а потом развивать науку — невыполнимо. Следовать ему — значит отвлекать силы от реальных задач. При системном подходе к теории классификации становится ясно, что строгие определения можно надеяться получить на последних этапах построения теории. Мы же сейчас находимся на первых этапах. Поэтому, не давая строгого определения понятиям «естественная классификация» и «естественная диагностика», обсудим, как проверить на «естественность» полученную расчетным путем классификацию (набор диагностических классов).

Можно выделить два критерия «естественности», по поводу которых имеется относительное согласие:

А. Естественная классификация должна быть реальной, соответствующей действительному миру, лишенной внесенного исследователем субъективизма.

Б. Естественная классификация должна быть важной или с научной точки зрения (давать возможность прогноза, предсказания новых свойств, сжатия информации и т.д.), или с практической.

Пусть классификация проводится на основе информации об объектах, представленной в виде матрицы «объект — признак» или матрицы попарных расстояний (мер близости). Пусть алгоритм классификации дал разбиение на кластеры. Как можно получить доводы в пользу естественности этой классификации? Например, уверенность в том, что она — закон природы, может появиться только в результате ее длительного изучения и практического применения. Это соображение относится и к другим из перечисленных выше критериев, в частности к Б (важности). Сосредоточимся на критерии А (реальности).

Понятие «реальности» кластера требует специального обсуждения (оно начато в работе [44]). Рассмотрим существо различий между понятиями

«классификация» (как результат кластер-анализа) и «группировка». Пусть, к примеру, необходимо деревья, растущие в определенной местности, разбить на группы находящихся рядом друг с другом. Ясна интуитивная разница между несколькими отдельными рощами, далеко отстоящими друг от друга и разделенными полями, и сплошным лесом, разбитым просеками на квадраты с целью лесоустройства.

Однако формально определить эту разницу столь же сложно, как определить понятие «куча зерен», чем занимались еще в Древней Греции. Ясно, что одно зерно не составляет кучи, два зерна не составляют кучи... Если к тому, что не составляет кучи, добавить еще одно зерно, то куча не получится. Значит — по принципу математической индукции — никакое количество зерен не составляет кучи. Но ясно, что миллиард зерен — большая куча зерен — подсчитайте объем! (Этот пример французского математика Э. Бореля (1871–1956 гг.) относится к теории нечеткости.)

Переформулируем сказанное в терминах «кластер-анализа» и «методов группировки». Выделенные с помощью первого подхода кластеры реальны, а потому могут рассматриваться как кандидаты в «естественные». Группировка дает «искусственные» классы, которые не могут быть «естественными».

Выборку из унимодального распределения можно, видимо, рассматривать как «естественный», «реальный» кластер. Применим к ней какой-либо алгоритм классификации («средней связи», «ближайшего соседа» и т.п.). Он даст какое-то разбиение на классы, которые, разумеется, не являются «реальными», поскольку отражают прежде всего свойства алгоритма, а не исходных данных. Как отличить такую ситуацию от противоположной, когда имеются реальные кластеры и алгоритм классификации более или менее точно их выделяет? Как известно, «критерий истины — практика», но слишком много времени необходимо для применения подобного критерия. Поэтому представляет интерес критерий, оценивающий «реальность» выделяемых с помощью алгоритма классификации кластеров одновременно с применением этого алгоритма.

Такой показатель существует — это критерий устойчивости. Устойчивость — понятие широкое. Общая схема формулирования и изучения проблем устойчивости рассмотрена в [1]. В частности, поскольку значения признаков всегда измеряются с погрешностями, то «реальное» разбиение должно быть устойчиво (т.е. не меняться или меняться слабо) при малых отклонениях исходных данных. Алгоритмов классификации существует бесконечно много, и «реальное» разбиение должно быть устойчиво по отношению к переходу к другому алгоритму. Другими словами, если «реальное» разбиение

на классы возможно, то оно находится с помощью любого алгоритма автоматической классификация. Следовательно, критерием естественности классификации может служить совпадение результатов работы двух достаточно различающихся алгоритмов, например «ближайшего соседа» и «дальнего соседа».

Критерии «естественности» кластеров и классификаций. Выше рассмотрены два типа «глобальных» критериев «естественности классификации», касающихся разбиения в целом. «Локальные» критерии относятся к отдельным кластерам. Простейшая постановка такова: достаточно ли однородны два кластера (две совокупности) для их объединения? Если объединение возможно, то кластеры не являются «естественными». Преимущество этой постановки в том, что она допускает применение статистических критериев однородности двух выборок. В одномерном случае (классификация по одному признаку) разработано большое число подобных критериев — Крамера-Уэлча, Смирнова, омега-квадрат (Лемана — Розенблатта), Вилкоксона, Вандер — Вардена, Лорда, Стьюдента и др. [6, 30]. Имеются критерии и для многомерных данных. Для одного из видов объектов нечисловой природы — люсианов — статистические методы выделения «реальных» кластеров развиты в разделе 3.4.

Что касается глобальных критериев, то для изучения устойчивости по отношению к малым отклонениям исходных данных естественно использовать метод статистических испытаний и проводить расчеты по «возмущенным» данным. Некоторые теоретические утверждения, касающиеся влияния «возмущений» на кластеры различных типов, получены в работе [44].

Опишем практический опыт реализации анализа устойчивости. Несколько алгоритмов классификации были применены к данным, полученным при проведении маркетинга образовательных услуг и приведенным в работе [45]. Для анализа данных были использованы широко известные алгоритмы «ближайшего соседа», «дальнего соседа» и алгоритм кластер-анализа из работы [46]. С содержательной точки зрения полученные разбиения отличались мало. Поэтому есть основания считать, что с помощью этих алгоритмов действительно выявлена «реальная» структура данных.

Идея устойчивости как критерия «реальности» иногда реализуется неадекватно. Так, для однопараметрических алгоритмов иногда предлагают выделять разбиения, которым соответствуют наибольшие интервалы устойчивости по параметру, т.е. наибольшие приращения параметра между очередными объединениями кластеров. Для данных работы [45] это предложение не дало полезных результатов — были получены различные разбиения:

три алгоритма — три разбиения. И с теоретической точки зрения такое предложение несостоятельно. Покажем это.

Действительно, рассмотрим алгоритм «ближайшего соседа», использующий меру близости $d(x, y)$, и однопараметрическое семейство алгоритмов с мерой близости $d^a(x, y)$, $a > 0$, также являющихся алгоритмами «ближайшего соседа». Тогда дендрограммы, полученные с помощью этих алгоритмов, совпадают при всех a, поскольку при их реализации происходит лишь сравнение мер близости между объектами. Другими словами, дендрограмма, полученная с помощью алгоритма «ближайшего соседа», является адекватной в порядковой шкале (измерения меры близости $d(x, y)$), т.е. сохраняется при любом строго возрастающем преобразовании этой меры. Однако выделенные по обсуждаемому методу «устойчивые разбиения» меняются, другими словами, не являются адекватными в порядковой шкале. В частности, при достаточно большом a «наиболее объективным» в соответствии с рассматриваемым предложением будет, как нетрудно показать, разбиение на два кластера! Таким образом, разбиение, выдвинутое в соответствии с рассматриваемым предложением как «устойчивое», на самом деле оказывается весьма неустойчивым.

Рассмотрим с позиций нечисловой статистики несколько конкретных вопросов теории классификации.

Вероятностная теория кластер-анализа. Как и для прочих статистических методов, свойства алгоритмов кластер-анализа необходимо изучать на вероятностных моделях. Это касается, например, условий естественного объединения двух кластеров.

Вероятностные постановки нужно применять, в частности, при перенесении результатов, полученных по выборке, на генеральную совокупность. Вероятностная теория кластер-анализа и методов группировки различна для исходных данных типа таблиц «объект × признак» и матриц сходства. Для первых параметрическая вероятностно-статистическая теория называется «расщеплением смесей». Непараметрическая теория основана на непараметрических оценках плотностей вероятностей и их мод. Основные результаты, связанные с непараметрическими оценками плотности, обсуждались в разделе 2.5.

Если исходные данные — матрица сходства $\|d(x, y)\|$, то необходимо признать, что развитой вероятностно-статистической теории пока нет. Подходы к ее построению намечены в работе [44]. Одна из основных проблем — проверка «реальности» кластера, его объективного существования независимо от расчетов исследователя. Проблема «реальности» кластера давно об-

суждается специалистами различных областей. Типичное рассуждение, напомним, таково. Предположим, что результаты наблюдений можно рассматривать как выборку из некоторого распределения с монотонно убывающей плотностью при увеличении расстояния от некоторого центра. Примененный к подобным данным какой-либо алгоритм кластер-анализа порождает некоторое разбиение. Ясно, что оно — чисто формальное, поскольку выделенным таксонам (кластерам) не соответствуют никакие «реальные» классы. Другими словами, задача кластер-анализа не имеет решения, а алгоритм дает лишь группировку. При обработке реальных данных мы не знаем вида плотности. Проблема состоит в том, чтобы определить, каков результат работы алгоритма (реальные кластеры или формальные группы).

Частный случай этой проблемы — проверка обоснованности объединения двух кластеров, которые мы рассматриваем как два множества объектов, а именно, множества $\{a_1, a_2, ..., a_k\}$ и $\{b_1, b_2, ..., b_m\}$. Пусть, например, используется алгоритм типа «Дендрограмма». Естественной представляется следующая идея. Пусть есть две совокупности мер близости. Одна — меры близости между объектами, лежащими внутри одного кластера, т.е. $d(a_i, a_j)$, $1 \le i < j \le k$, $d(b_\alpha, b_\beta)$, $1 \le \alpha < \beta \le m$. Другая совокупность — меры близости между объектами, лежащими в разных кластерах, т.е. $d(a_i, b_\alpha)$, $1 \le i \le k$, $1 \le \alpha \le m$. Эти две совокупности мер близости предлагается рассматривать как независимые выборки и проверять гипотезу о совпадении их функций распределения. Если гипотеза не отвергается, объединение кластеров считается обоснованным; в противном случае — объединять нельзя, алгоритм прекращает работу.

В рассматриваемом подходе есть две некорректности (см. также работу [44, разд. 4]). Во-первых, меры близости не являются независимыми случайными величинами. Во-вторых, не учитывается, что объединяются не заранее фиксированные кластеры (с детерминированным составом), а полученные в результате работы некоторого алгоритма, и их состав (в частности, количество элементов) оказывается случайным. От первой из этих некорректностей можно частично избавиться. Справедливо следующее утверждение [47].

Теорема 1. Пусть $a_1, a_2, ..., a_k, b_1, b_2, ..., b_m$ — независимые одинаково распределенные случайные величины (со значениями в произвольном пространстве). Пусть случайная величина $d(a_1,a_2)$ имеет все моменты. Тогда при $k, m \to \infty$ распределение статистики:

$$\frac{8\sqrt{3}U - 3(k+m)(k+m-1)(k(k+1)+m(m+1))}{2(k+m)\sqrt{km(k^2+m^2)}}$$

сходится к стандартному нормальному распределению с математическим ожиданием 0 и дисперсией 1. Здесь U — сумма рангов элементов первой выборки в объединенной выборке; первая выборка составлена из внутрикластерных расстояний (мер близости) $d(a_i, a_j)$, $1 \le i < j \le k$, и $d(b_\alpha, b_\beta)$, $1 \le \alpha < \beta \le m$, а вторая — из межкластерных расстояний $d(a_i, b_\alpha)$, $1 \le i \le k$, $1 \le \alpha \le m$.

На основе теоремы 1 очевидным образом формулируется правило проверки обоснованности объединения двух кластеров. Другими словами, мы проверяем статистическую гипотезу, согласно которой объединение двух кластеров образует однородную совокупность. Если величина U слишком мала, статистическая гипотеза однородности отклоняется (на заданном уровне значимости), и возможность объединения отбрасывается. Таким образом, хотя расстояния между объектами в кластерах зависимы, но эта зависимость слаба, и доказана математическая теорема о допустимости применения критерия Вилкоксона для проверки возможности объединения кластеров.

О вычислительной сходимости алгоритмов кластер-анализа. Алгоритмы кластер-анализа и группировки зачастую являются итерационными. Например, формулируется правило улучшения решения задачи кластер-анализа шаг за шагом, но момент остановки вычислений не обсуждается. Примером является известный алгоритм «Форель», в котором постепенно улучшается положение центра кластера. В этом алгоритме на каждом шаге строится шар определенного заранее радиуса, выделяются элементы кластеризуемой совокупности, попадающие в этот шар, и новый центр кластера строится как центр тяжести выделенных элементов. При анализе алгоритма «Форель» возникает проблема: завершится ли процесс улучшения положения центра кластера через конечное число шагов или же он может быть бесконечным. Она получила название «проблема остановки». Для широкого класса так называемых «эталонных алгоритмов» проблема остановки была решена в работе [44]: процесс улучшения остановится через конечное число шагов.

Отметим, что алгоритмы кластер-анализа могут быть модифицированы разнообразными способами. Например, описывая алгоритм «Форель» в стиле статистики объектов нечисловой природы, заметим, что вычисление центра тяжести для совокупности многомерных точек — это нахождение эмпирического среднего для меры близости, равной квадрату евклидова расстояния. Если взять другую естественную меру близости — блочное расстояние (п. 1.7), то получим алгоритм кластер-анализа «Медиана», отличающийся от «Форели» тем, что новый центр кластера строится не с по-

мощью средних арифметических координат элементов, попавших в кластер, а с помощью медиан.

Проблема остановки возникает не только при построении диагностических классов. Она принципиально важна, в частности, и при оценивании параметров вероятностных распределений методом максимального правдоподобия. Обычно не представляет большого труда выписать систему уравнений максимального правдоподобия и предложить решать ее каким-либо численным методом. Однако когда остановиться, сколько итераций сделать, какая точность оценивания будет при этом достигнута? Общий ответ, видимо, невозможно найти, но обычно нет ответа и для конкретных семейств распределения вероятностей. Именно поэтому нет оснований рекомендовать решать системы уравнений максимального правдоподобия. Вместо них целесообразно использовать так называемые одношаговые оценки (подробнее об этих оценках см. раздел 2.4). Эти оценки задаются конечными формулами, но асимптотически столь же хороши (на профессиональном языке — эффективны), как и оценки максимального правдоподобия.

О сравнении алгоритмов диагностики по результатам обработки реальных данных. Перейдем к этапу применения диагностических правил, когда классы, к одному из которых нужно отнести вновь поступающий объект, уже выделены.

В прикладных исследованиях применяют различные методы дискриминантного анализа, основанные на вероятностно-статистических моделях, а также с ними не связанные, т.е. эвристические, использующие детерминированные методы анализа данных. Независимо от «происхождения», каждый подобный алгоритм должен быть исследован как на параметрических и непараметрических вероятностно-статистических моделях порождения данных, так и на различных массивах реальных данных. Цель исследования — выбор наилучшего алгоритма в определенной области применения, включение его в стандартные программные продукты, методические материалы, учебные программы и пособия. Но для этого надо уметь сравнивать алгоритмы по качеству. Как это делать?

Часто используют такой показатель качества алгоритма диагностики, как «вероятность правильной классификации» (при обработке конкретных данных — «частота правильной классификации»). Чуть ниже мы покажем, что этот показатель качества некорректен, а потому пользоваться им не рекомендуется. Целесообразно применять другой показатель качества алгоритма диагностики — оценку специального вида так называемые «расстояния

Махаланобиса» между классами. Изложение проведем на примере разработки программного продукта для специалистов по диагностике материалов. Прообразом является диалоговая система «АРМ материаловеда», разработанная Институтом высоких статистических технологий и эконометрики для ВНИИ эластомерных материалов (Москва).

При построении информационно-исследовательской системы диагностики материалов (ИИСДМ) возникает задача сравнения прогностических правил «по силе». Прогностическое правило — это алгоритм, позволяющий по характеристикам материала прогнозировать его свойства. Если прогноз дихотомичен («есть» или «нет»), то правило является алгоритмом диагностики, при котором материал относится к одному из двух классов. Ясно, что случай нескольких классов может быть сведен к конечной последовательности выбора между двумя классами.

Прогностические правила могут быть извлечены из научно-технической литературы и практики. Каждое из них обычно формулируется в терминах небольшого числа признаков, но наборы признаков сильно меняются от правила к правилу. Поскольку в ИИСДМ должно фиксироваться лишь ограниченное число признаков, то возникает проблема их отбора. Естественно отбирать лишь те их них, которые входят в наборы, дающие наиболее «надежные» прогнозы. Для придания точного смысла термину «надежный» необходимо иметь способ сравнения алгоритмов диагностики по прогностической «силе».

Результаты обработки реальных данных с помощью некоторого алгоритма диагностики в рассматриваемом случае двух классов описываются долями: правильной диагностики в первом классе κ; правильной диагностики во втором классе λ; долями классов в объединенной совокупности π_i, $i = 1.2$; $\pi_1 + \pi_2 = 1$.

При изучении качества алгоритмов классификации их сравнивают по результатам дискриминации вновь поступающей контрольной выборки. А именно, по контрольной выборке определяются величины $\kappa, \lambda, \pi_1, \pi_2$. Однако иногда вместо контрольной используют обучающую выборку. Т.е. указанные величины определяются ретроспективно, в результате анализа уже имеющихся данных. Обычно это связано с трудоемкостью получения данных. Тогда κ и λ зависимы. Однако в случае, когда решающее правило основано на использовании дискриминантной поверхности, параметры которой оцениваются по обучающим выборкам, величины κ и λ асимптотически (при безграничном росте объемов выборок) независимы [44], поскольку тогда положение дискри-

минантной поверхности стабилизируется (стремится к пределу). Это позволяет использовать приводимые ниже результаты и в этом случае.

Нецелесообразность применения «доли правильной диагностики». Нередко как показатель качества алгоритма диагностики (прогностической «силы») используют долю правильной диагностики (классификации):

$$\mu = \pi_1 \kappa + \pi_2 \lambda.$$

Однако показатель μ определяется, в частности, через характеристики π_1 и π_2, частично заданные исследователем (например, на них влияет тактика отбора образцов для изучения). В аналогичной медицинской задаче величина μ оказалась больше для тривиального прогноза, согласно которому у всех больных течение заболевания будет благоприятно. Тривиальный прогноз сравнивался с алгоритмом выделения больных с прогнозируемым тяжелым течением заболевания. Он был разработан группой исследователей (врачей и математиков) под руководством академика АН СССР И. М. Гельфанда. Применение этого алгоритма с медицинской точки зрения вполне оправдано [48], поскольку гипердиагностика тяжелого течения заболевания позволяет снизить частоту летального исхода.

Другими словами, по доле правильной классификации алгоритм академика И. М. Гельфанда оказался хуже тривиального — объявить всех больных легкими, не требующими специального наблюдения. Этот вывод (о преимуществе тривиального прогноза), очевидно, нелеп. И причина появления нелепости вполне понятна. Хотя доля тяжелых больных невелика, но смертельные исходы сосредоточены именно в этой группе больных. Поэтому целесообразна гипердиагностика — рациональнее часть легких больных объявить тяжелыми, чем сделать ошибку в противоположную сторону. Применение теории статистических решений в рассматриваемой постановке вряд ли возможно, поскольку оценить количественно потери от смерти больного нельзя по этическим соображениям. Поэтому, на наш взгляд, долю правильной диагностики μ нецелесообразно использовать как показатель качества алгоритма диагностики.

Поясним сказанное. Применение теории статистических решений требует знания потерь от ошибочной диагностики, а в большинстве научно-технических и экономических задач определить потери, как уже отмечалось, сложно. В частности, из-за необходимости оценивать человеческую жизнь в денежных единицах. По этическим соображениям это, на наш взгляд, недо-

пустимо. Сказанное не означает отрицания пользы страхования, но, очевидно, страховые выплаты следует рассматривать лишь как способ первоначального смягчения потерь от утраты близких.

Метод нормального усреднения. Для выявления информативного набора признаков целесообразно использовать *метод нормального усреднения*. Его называют также *методом пересчета на модель линейного дискриминантного анализа*. Согласно этому методу статистической оценкой прогностической «силы» является:

$$\delta^* = \Phi(d^*/2), \quad d^* = \Phi^{-1}(\kappa) + \Phi^{-1}(\lambda),$$

где $\Phi(x)$ — функция стандартного нормального распределения вероятностей с математическим ожиданием 0 и дисперсией 1, а $\Phi^{-1}(y)$ — обратная ей функция. Таким образом, прогностическая сила δ^* является средним по Колмогорову (см. раздел 3.1), рассчитанным по долям правильной классификации κ и λ, причем для усреднения используется функция $F(x) = \Phi^{-1}(x)$, участвующая в определении среднего по Колмогорову.

Пример 1. Если доли правильной классификации $\kappa = 0{,}90$ и $\lambda = 0{,}80$, то $\Phi^{-1}(\kappa) = 1{,}28$ и $\Phi^{-1}(\lambda) = 0{,}84$, откуда $d^* = 2{,}12$ и прогностическая сила $\delta^* = \Phi^{-1}(1{,}06) = 0{,}86$. При этом доля правильной классификации μ может принимать любые значения между 0,80 и 0,90, в зависимости от доли элементов того или иного класса среди анализируемых данных.

Если классы описываются выборками из многомерных нормальных совокупностей с одинаковыми матрицами ковариаций, а для классификации применяется классический линейный дискриминантный анализ Р. Фишера, то величина d^* представляет собой состоятельную статистическую оценку так называемого расстояния Махаланобиса между рассматриваемыми двумя совокупностями (конкретный вид этого расстояния сейчас не имеет значения), независимо от порогового значения, определяющего конкретное решающее правило. В общем случае показатель δ^* вводится как эвристический.

Пусть алгоритм классификации применялся к совокупности, состоящей из m объектов первого класса и n объектов второго класса.

Теорема 2 [47]. Пусть $m, n \to \infty$. Тогда для всех x:

$$P\left\{\frac{\delta^* - \delta}{A(\kappa, \lambda)} < x\right\} \to \Phi(x),$$

где δ — истинная «прогностическая сила» алгоритма диагностики; δ^* — ее эмпирическая оценка,

$$A^2(\kappa,\lambda)=\frac{1}{4}\left\{\left[\frac{\varphi(d^*/2)}{\varphi(\Phi^{-1}(\kappa))}\right]^2\frac{\kappa(1-\kappa)}{m}+\left[\frac{\varphi(d^*/2)}{\varphi(\Phi^{-1}(\lambda))}\right]^2\frac{\lambda(1-\lambda)}{n}\right\};$$

$\varphi(x)=\Phi'(x)$ — плотность стандартного нормального распределения вероятностей с математическим ожиданием 0 и дисперсией 1.

С помощью теоремы 2 по κ и λ обычным образом определяют доверительные границы для «прогностической силы» δ.

Пример 2. В условиях примера 1 при $m = n = 100$ найдем асимптотическое среднее квадратическое отклонение $A(0{,}90;\ 0{,}80)$.

Поскольку $\varphi(\Phi^{-1}(\kappa)) = \varphi(1{,}28) = 0{,}176$, $\varphi(\Phi^{-1}(\lambda)) = \varphi(0{,}84) = 0{,}280$, $\varphi(d^*/2) = \varphi(1{,}06) = 0{,}227$, то подставляя в выражение для A^2 численные значения, получаем, что

$$A^2(0{,}90;\ 0{,}80)=\frac{0{,}0372}{m}+\frac{0{,}0265}{n}$$

(численные значения плотности стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1 и функции, обратной к функции этого распределения, можно было взять, например, из справочника [30]).

При $m = n = 100$ имеем $A(0{,}90;\ 0{,}80) = 0{,}0252$. При доверительной вероятности $\gamma = 0{,}95$ имеем $u(0{,}95) = \Phi^{-1}(1{,}0{,}975) = 1{,}96$, а потому нижняя доверительная граница для прогностической силы δ есть $\delta_H = 0{,}86 - 1{,}96$ Ч $0{,}0252 = 0{,}81$, а верхняя доверительная граница такова: $\delta_B = 0{,}86 + 1{,}96$ Ч $0{,}0252 = 0{,}91$. Аналогичный расчет при $m = n = 1000$ дает $\delta_H = 0{,}845$, $\delta_B = 0{,}875$.

Как проверить обоснованность пересчета на модель линейного дискриминантного анализа? Допустим, что классификация состоит в вычислении некоторого прогностического индекса y и сравнении его с заданным порогом c. Объект относят к первому классу, если $y \le c$, ко второму, если $y > c$. Прогностический индекс — это обычно линейная функция от характеристик рассматриваемых объектов. Другими словами, от координат векторов, описывающих объекты.

Возьмем два значения порога c_1 и c_2. Если пересчет на модель линейного дискриминантного анализа обоснован, то, как можно показать, «прогностические силы» для обоих правил совпадают: $\delta(c_1) = \delta(c_2)$. Выполнение этого равенства можно проверить как статистическую гипотезу.

Пусть κ_1 — доля объектов первого класса, для которых $y \le c_1$, а κ_2 — доля объектов первого класса, для которых $c_1 < y \le c_2$. Аналогично пусть λ_2 — доля объектов второго класса, для которых $c_1 < y \le c_2$, а λ_3 — доля объектов второго класса, для которых $y > c_2$. Тогда можно рассчитать две оценки одного и того же расстояния Махаланобиса. Они имеют вид:

$$d^*(c_1) = \Phi^{-1}(\kappa_1) + \Phi^{-1}(\lambda_2 + \lambda_3), \quad d^*(c_2) = \Phi^{-1}(\kappa_1 + \kappa_2) + \Phi^{-1}(\lambda_3).$$

Теорема 3 [47]. Если истинные прогностические силы двух правил диагностики совпадают, $\delta(c_1) = \delta(c_2)$, то при $m \to \infty, n \to \infty$ при всех x:

$$P\left\{\frac{d^*(c_1) - d^*(c_2)}{B} < x\right\} \to \Phi(x),$$

где

$$B^2 = \frac{1}{m}T(\kappa_1;\kappa_2) + \frac{1}{n}T(\lambda_3;\lambda_2);$$

$$T(x;y) = \frac{x(1-x)}{\varphi^2(\Phi^{-1}(x))} + \frac{(x+y)(1-x-y)}{\varphi^2(\Phi^{-1}(x+y))} - \frac{2x(1-x-y)}{\varphi(\Phi^{-1}(x))\varphi(\Phi^{-1}(x+y))}.$$

Из теоремы 3 вытекает метод проверки рассматриваемой гипотезы: при выполнении неравенства:

$$\left|\frac{d^*(c_1) - d^*(c_2)}{B}\right| \le \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

она принимается на уровне значимости, асимптотически равном α, в противном случае — отвергается.

Пример 3. Пусть данные примеров 1 и 2 соответствуют порогу c_1. Пусть порогу c_2 соответствуют $\kappa' = 0{,}95$ и $\lambda' = 0{,}70$. Тогда в обозначениях теоремы 3 $\kappa_1 = 0{,}90$, $\kappa_2 = 0{,}05$, $\lambda_2 = 0{,}10$, $\lambda_3 = 0{,}70$. Далее $d^*(c_1) = 2{,}12$ (пример 1), $d^*(c_2) = 2{,}17$, $T(\kappa_1, \kappa_2) = 2{,}22$, $T(\lambda_3, \lambda_2) = 0{,}89$. Гипотеза о совпадении прогностических сил на двух порогах принимается на уровне значимости $\alpha = 0{,}05$ тогда и только тогда, когда:

$$\frac{0{,}05^2}{\frac{2{,}22}{m} + \frac{0{,}89}{n}} \le 1{,}96^2,$$

т.е. когда

$$\frac{2,22}{m}+\frac{0,89}{n}\geq 0,00065.$$

Так, гипотеза принимается при $m = n = 1000$ и отвергается при $m = n = 5000$.

Подходы к построению прогностических правил. Для решения задач диагностики используют два подхода — параметрический и непараметрический. Первый из них обычно основан на использовании того или иного индекса и сравнения его с порогом. Индекс может быть построен по статистическим данным, например, как в уже упомянутом линейном дискриминантном анализе Фишера. Часто индекс представляет собой линейную функцию от характеристик, выбранных специалистами предметной области, коэффициенты которой подбирают эмпирически. При этом качественные характеристики тем или иным способом «оцифровываются», их градациям искусственно приписывают численные значения.

Непараметрический подход связан с леммой Неймана — Пирсона в математической статистике и с теорией статистических решений. Он опирается на использование непараметрических оценок плотностей распределений вероятностей, описывающих диагностические классы.

Обсудим ситуацию подробнее. Математические методы диагностики, как и статистические методы в целом, делятся на параметрические и непараметрические. Первые основаны на предположении, что классы описываются распределениями из некоторых параметрических семейств. Обычно рассматривают многомерные нормальные распределения, при этом зачастую принимают гипотезу о том, что ковариационные матрицы для различных классов совпадают. Именно в таких предположениях сформулирован классический дискриминантный анализ Фишера. Как известно, обычно нет оснований считать, что наблюдения извлечены из нормального распределения [6, раздел 4.1].

Поэтому с прикладной точки зрения более корректными, чем параметрические, являются непараметрические методы диагностики. Исходная идея таких методов основана на лемме Неймана — Пирсона, входящей в стандартный курс математической статистики. Согласно этой лемме решение об отнесении вновь поступающего объекта (сигнала, наблюдения и др.) к одному из двух классов принимается на основе отношения плотностей $f(x)/g(x)$, где $f(x)$ — плотность распределения, соответствующая первому классу, а $g(x)$ — плотность распределения, соответствующая второму

классу. Если плотности распределения неизвестны, то применяют их непараметрические оценки, построенные по обучающим выборкам. Пусть обучающая выборка объектов из первого класса состоит из n элементов, а обучающая выборка для второго класса — из m объектов. Тогда рассчитывают значения непараметрических оценок плотностей $f_n(x)$ и $g_m(x)$ для первого и второго классов соответственно, а диагностическое решение принимают по их отношению. Таким образом, для решения задачи диагностики достаточно научиться строить непараметрические оценки плотности для выборок объектов произвольной природы.

Методы построения непараметрических оценок плотности распределения вероятностей в пространствах произвольной природы рассмотрены в разделе 2.5.

Методы классификации — одна из наиболее обширных частей нечисловой статистики как области прикладной статистики. В настоящем разделе рассмотрены лишь наиболее важные вопросы, относящиеся к этим методам.

2.9. МЕТОДЫ ШКАЛИРОВАНИЯ

В прикладной статистике каждый элемент выборки описывается тем или иным математическим объектом, например, нечетким множеством или вектором. Естественным является желание наглядно представить себе имеющиеся статистические данные. Однако человек может непосредственно воспринимать лишь числовые данные или точки на плоскости. Анализировать скопления точек в трехмерном пространстве уже гораздо труднее. Непосредственное (визуальное) восприятие данных более высокой размерности невозможно. Поэтому вполне естественным является желание перейти от многомерной выборки или выборки, состоящей из объектов нечисловой природы, к данным небольшой размерности, чтобы «на них можно было посмотреть». Статистические технологии такого перехода объединяют термином «методы шкалирования». Если исходные данные — многомерные вектора, то говорят о «методах снижения размерности».

Кроме стремления к наглядности, есть и другие мотивы для шкалирования и снижения размерности. Те факторы, от которых интересующая исследователя переменная не зависит, лишь мешают статистическому анализу. Во-первых, на сбор информации о них расходуются ресурсы. Во-вторых, как можно доказать, их включение в анализ ухудшает свойства статистических процедур (в частности, увеличивает дисперсию оценок па-

раметров и характеристик распределений). Поэтому желательно избавиться от таких факторов.

При анализе данных обычно рассматривают не одну, а множество задач, в частности, по-разному выбирая независимые и зависимые переменные. Поэтому рассмотрим задачу шкалирования (снижения размерности) в следующей формулировке. Дана выборка. Требуется перейти от нее к совокупности векторов малой размерности, максимально сохранив структуру исходных данных, по возможности не теряя информации, содержащейся в данных. Задача конкретизируется в рамках каждого конкретного метода шкалирования (снижения размерности).

Метод главных компонент является одним из наиболее часто используемых методов снижения размерности. Основная его идея состоит в последовательном выявлении направлений, в которых данные имеют наибольший разброс. Пусть выборка состоит из n-мерных векторов, одинаково распределенных с вектором $X = (x(1), x(2), \ldots, x(n))$. Рассмотрим линейные комбинации:

$$Y(\lambda(1), \lambda(2), \ldots, \lambda(n)) = \lambda(1)x(1) + \lambda(2)x(2) + \ldots + \lambda(n)x(n),$$

где

$$\lambda^2(1) + \lambda^2(2) + \ldots + \lambda^2(n) = 1.$$

Здесь вектор $\lambda = (\lambda(1), \lambda(2), \ldots, \lambda(n))$ лежит на единичной сфере в n-мерном пространстве.

В методе главных компонент прежде всего находят направление максимального разброса, т.е. такой вектор λ, при котором достигает максимума дисперсия случайной величины $Y(\lambda) = Y(\lambda(1), \lambda(2), ..., \lambda(n))$. Тогда вектор λ задает первую главную компоненту, а величина $Y(\lambda)$ является проекцией случайного вектора X на ось первой главной компоненты.

Затем, выражаясь терминами линейной алгебры, рассматривают гиперплоскость в n-мерном пространстве, перпендикулярную первой главной компоненте, и проектируют на эту гиперплоскость все элементы выборки. Размерность гиперплоскости на 1 меньше, чем размерность исходного пространства.

В рассматриваемой гиперплоскости процедура повторяется. В ней находят направление наибольшего разброса, т.е. вторую главную компоненту. Затем выделяют гиперплоскость, перпендикулярную первым двум глав-

ным компонентам. Ее размерность на 2 меньше, чем размерность исходного пространства. Далее — следующая итерация.

С точки зрения линейной алгебры речь идет о построении нового базиса в n-мерном пространстве, ортами которого служат главные компоненты.

Дисперсия, соответствующая каждой новой главной компоненте, меньше (точнее, не больше), чем для предыдущей. Обычно останавливаются, когда она меньше заданного порога. Если отобрано k главных компонент, то это означает, что от n-мерного пространства удалось перейти к k-мерному, т.е. сократить размерность с n-до k, практически не исказив структуру исходных данных.

Для визуального анализа данных часто используют проекции исходных векторов на плоскость первых двух главных компонент. Обычно хорошо видна структура данных, выделяются компактные кластеры объектов и отдельно выделяющиеся элементы выборки.

Метод главных компонент является одним из методов **факторного анализа** [49]. Различные алгоритмы факторного анализа объединены тем, что во всех них происходит переход к новому базису в исходном n-мерном пространстве. Важным является понятие «нагрузка фактора», применяемое для описания роли исходного фактора (переменной) в формировании определенного вектора из нового базиса.

Новая идея по сравнению с методом главных компонент состоит в том, что на основе нагрузок происходит разбиение факторов на группы. В одну группу объединяются факторы, имеющие сходное влияние на элементы нового базиса. Затем из каждой группы рекомендуется оставить одного представителя. Иногда вместо выбора представителя расчетным путем формируется новый фактор, являющийся центральным для рассматриваемой группы. Снижение размерности происходит при переходе к системе факторов, являющихся представителями групп. Остальные факторы отбрасываются.

Описанная процедура может быть осуществлена не только с помощью факторного анализа. Речь идет о кластер-анализе признаков (факторов, переменных). Для разбиения признаков на группы можно применять различные алгоритмы кластер-анализа. Достаточно ввести расстояние (меру близости, показатель различия) между признаками. Пусть X и Y — два признака. Различие $d(X, Y)$ между ними можно измерять с помощью выборочных коэффициентов корреляции:

$$d_1(X, Y) = 1 - r_n(X, Y), \quad d_2(X, Y) = 1 - \rho_n(X, Y),$$

где $r_n(X, Y)$ — выборочный линейный коэффициент корреляции Пирсона, $\rho_n(X, Y)$ — выборочный коэффициент ранговой корреляции Спирмена.

Процедуры шкалирования. На использовании расстояний (мер близости, показателей различия) $d(X, Y)$ между объектами (или признаками) X и Y основан обширный класс методов многомерного шкалирования [50, 51]. Основная идея этого класса методов состоит в представлении каждого объекта точкой геометрического пространства (обычно размерности 1, 2 или 3), координатами которой служат значения скрытых (латентных) факторов, в совокупности достаточно адекватно описывающих объект. При этом отношения между объектами заменяются отношениями между точками — их представителями. Так, данные о сходстве объектов — расстояниями между точками, данные о превосходстве — взаимным расположением точек [52].

В практике используется ряд различных моделей шкалирования. Во всех них встает проблема оценки истинной размерности факторного пространства. Рассмотрим эту проблему на примере обработки данных о сходстве объектов с помощью так называемого метрического шкалирования.

Пусть имеется n объектов $O(1)$, $O(2)$, ..., $O(n)$, для каждой пары объектов $O(i)$, $O(j)$ задана мера их сходства $s(i,j)$. Считаем, что всегда $s(i, j) = s(j, i)$. Происхождение чисел $s(i, j)$ не имеет значения для описания работы алгоритма. Они могли быть получены либо непосредственным измерением, либо с использованием экспертов, либо путем вычисления по совокупности описательных характеристик, либо как-то иначе.

В евклидовом пространстве рассматриваемые n объектов должны быть представлены конфигурацией n точек, причем в качестве меры близости точек-представителей выступает евклидово расстояние $d(i, j)$ между соответствующими точками. Степень соответствия между совокупностью объектов и совокупностью представляющих их точек определяется путем сопоставления матриц сходства $\|s(i, j)\|$ и расстояний $\|d(i, j)\|$. Метрический функционал сходства имеет вид:

$$S = \sum_{i<j} | s(i, j) - d(i, j) |^2 .$$

Геометрическую конфигурацию надо выбирать так, чтобы функционал S достигал своего наименьшего значения [52].

Замечание. В неметрическом шкалировании вместо близости самих мер близости и расстояний рассматривается близость упорядочений на множестве мер близости и множестве соответствующих расстояний. Вместо функ-

ционала S используются аналоги ранговых коэффициентов корреляции Спирмена и Кендалла. Другими словами, неметрическое шкалирование исходит из предположения, что меры близости измерены в порядковой шкале.
Пусть евклидово пространство имеет размерность m. Рассмотрим минимум среднего квадрата ошибки:

$$\alpha_m = \frac{2}{n(n-1)} \min S,$$

где минимум берется по всем возможным конфигурациям n точек в m-мерном евклидовом пространстве.

Исходя из общих результатов об асимптотическом поведении решений экстремальных статистических задач (раздел 2.3), можно показать, что в задачах метрического и неметрического шкалирования рассматриваемые минимумы достигаются на некоторых конфигурациях.

Ясно, что при росте m величина α_m монотонно убывает (точнее, не возрастает). Можно показать, что при $m \geq n - 1$ она равна 0 (если $s(i, j)$ — метрика). Для увеличения возможностей содержательной интерпретации желательно действовать в пространстве возможно меньшей размерности. При этом, однако, размерность необходимо выбрать так, чтобы точки представляли объекты без больших искажений. Возникает вопрос: как рационально выбирать размерность, т.е. натуральное число m?

В рамках детерминированного анализа данных обоснованного ответа на этот вопрос, видимо, нет. Следовательно, необходимо изучить поведение α_m в тех или иных вероятностных моделях. Если меры близости $s(i, j)$ являются случайными величинами, распределение которых зависит от «истинной размерности» m_0 (и, возможно, от каких-либо еще параметров), то можно в классическом математико-статистическом стиле ставить задачу оценки m_0, искать состоятельные оценки и т.д.

Начнем строить вероятностные модели. Примем, что объекты моделируются точками в евклидовом пространстве размерности k, где k достаточно велико. То, что «истинная размерность» равна m_0, означает, что все эти точки лежат на гиперплоскости размерности m_0. Примем для определенности, что совокупность рассматриваемых точек представляет собой выборку из кругового нормального распределения с дисперсией $\sigma^2(0)$. Это означает, что объекты $O(1)$, $O(2)$, …, $O(n)$ являются независимыми в совокупности случайными векторами, каждый из которых строится как $\zeta(1)e(1) + \zeta(2)e(2) + \ldots + \zeta(m_0)e(m_0)$, где $e(1)$,

$e(2)$, … , $e(m_0)$ — ортонормальный базис в подпространстве размерности m_0, в котором лежат рассматриваемые точки, а $\zeta(1)$, $\zeta(2)$, … , $\zeta(m_0)$ — независимые в совокупности одномерные нормальные случайные величины с математическим ожиданием 0 и дисперсией $\sigma^2(0)$.

Рассмотрим две модели получения мер близости $s(i, j)$. В первой из них $s(i, j)$ отличаются от евклидова расстояния между соответствующими точками из-за того, что точки известны с искажениями. Пусть $c(1)$, $c(2)$, ..., $c(n)$ — рассматриваемые точки. Тогда:

$$s(i, j) = d(c(i) + \varepsilon(i), c(j) + \varepsilon(j)), \;\; i, j = 1, 2, ..., n,$$

где d — евклидово расстояние между точками в k-мерном пространстве, вектора $\varepsilon(1)$, $\varepsilon(2)$, ..., $\varepsilon(n)$ представляют собой выборку из кругового нормального распределения в k-мерном пространстве с нулевым математическим ожиданием и ковариационной матрицей $\sigma^2(1)I$, где I — единичная матрица. Другими словами, $\varepsilon(i) = \eta(i, 1)e(1) + \eta(i, 2)e(2) + \ldots + \eta(i, k)e(k)$, где $e(1)$, $e(2)$, ..., $e(k)$ — ортонормальный базис в k-мерном пространстве, а $\{\eta(i, t)$, $i = 1, 2, ..., n$, $t = 1, 2, ..., k\}$ — совокупность независимых в совокупности одномерных случайных величин с нулевым математическим ожиданием и дисперсией $\sigma^2(1)$.

Во второй модели искажения наложены непосредственно на сами расстояния:

$$s(i, j) = d(c(i), c(j)) + \varepsilon(i, j), \;\; i,j = 1, 2, \ldots, n, i \neq j,$$

где $\{\varepsilon(i, j), i, j = 1, 2, ..., n\}$ — независимые в совокупности нормальные случайные величины с математическим ожиданием 0 и дисперсией $\sigma^2(1)$.

В работе [53] показано, что для обеих сформулированных моделей минимум среднего квадрата ошибки α_m при $n \to \infty$ сходится по вероятности к:

$$f(m) = f_1(m) + \sigma^2(1)(k - m), \;\; m = 1, 2, ..., k,$$

где

$$f_1(m) = \begin{cases} \sigma^2(0)(m_0 - m), & m < m_0, \\ 0, & m \geq m_0. \end{cases}$$

Таким образом, функция $f(m)$ линейна на интервалах (1; m_0) и (m_0; k), причем на первом интервале она убывает быстрее, чем на втором. Отсюда следует, что статистика:

$$m^* = Arg \min_m \{\alpha_{m+1} - 2\alpha_m + \alpha_{m-1})$$

является состоятельной оценкой истинной размерности m_0 (сопоставьте с рассмотренными выше оценками истинной размерности модели в задачах восстановления зависимости (раздел 2.7) и расщепления смесей (раздел 2.8)).

Итак, из вероятностной теории вытекает рекомендация — в качестве оценки размерности факторного пространства использовать m^*. Отметим, что подобная рекомендация была сформулировано как эвристическая одним из основателей многомерного шкалирования Дж. Краскалом [50]. Он исходил из опыта практического использования многомерного шкалирования и вычислительных экспериментов. Вероятностная теория нечисловой статистики позволила обосновать эту давнюю эвристическую рекомендацию.

Применение общих результатов нечисловой статистики к методу главных компонент. Напомним, что исходные данные — набор векторов ξ_1, ξ_2, ..., ξ_n, лежащих в евклидовом пространстве R^k размерности k. Цель состоит в снижении размерности, т.е. в уменьшении числа рассматриваемых показателей. Для этого берут всевозможные линейные ортогональные нормированные центрированные комбинации исходных показателей, получают k новых показателей, из них берут первые m, где $m < k$ (подробности см. выше). Матрицу преобразования C выбирают так, чтобы максимизировать информационный функционал:

$$I_n(C) = \frac{s^2(z(1)) + s^2(z(2)) + ... + s^2(z(m))}{s^2(x(1)) + s^2(x(2)) + ... + s^2(x(k))}, \qquad (1)$$

где $x(i)$, i = 1, 2, … , k, — исходные показатели; исходные данные имеют вид $\xi_j = (x_j(1), x_j(2), \dots, x_j(k))$, j = 1, 2, … , n; при этом $z(\alpha)$, α = 1, 2, … , m, — комбинации исходных показателей, полученные с помощью матрицы C. Наконец, $s^2(z(\alpha))$, α = 1, 2, … , m, $s^2(x(i))$, i = 1, 2, … , k, — выборочные дисперсии переменных, указанных в скобках.

Укажем подробнее, как новые показатели (главные компоненты) $z(\alpha)$ строятся по исходным показателям $x(i)$ с помощью матрицы C:

$$z_j(\alpha) = \sum_{\beta=1}^{k} c_{\alpha\beta}(x_j(\beta) - \overline{x(\beta)}), \quad \alpha = 1,2,...,m, \quad j = 1,2,...,n,$$

где

$$\overline{x(\beta)} = \frac{1}{n}\sum_{j=1}^{n} x_j(\beta).$$

Матрица $C = \|c_{\alpha\beta}\|$ порядка $m \times k$ такова, что:

$$\sum_{\beta=1}^{k} c_{\alpha\beta}^2 = 1, \ \alpha = 1,2,...,m \tag{2}$$

(нормированность),

$$\sum_{\beta=1}^{k} c_{\alpha\beta} c_{\gamma\beta} = 0, \ \alpha,\gamma = 1,2,...,m, \ \alpha \neq \gamma \tag{3}$$

(ортогональность).

Решением основной задачи метода главных компонент является:

$$C_n = Arg\min_{C} (-I_n(C)),$$

где минимизируемая функция определена формулой (1), а минимизация проводится по всем матрицам C, удовлетворяющим условиям (2) и (3).

Вычисление матрицы C_n — задача детерминированного анализа данных. Однако, как и в иных случаях, например, для медианы Кемени, возникает вопрос об асимптотическом поведении C_n. Является ли решение основной задачи метода главных компонент устойчивым, т.е. существует ли предел C_n при $n \to \infty$? Чему равен этот предел?

Ответ, как обычно, может быть дан только в вероятностной теории. Пусть $\xi_1, \xi_2, \ldots, \xi_n$ — независимые одинаково распределенные случайные вектора. Положим:

$$z_\infty(\alpha) = \sum_{\beta=1}^{k} c_{\alpha\beta}(x_1(\beta) - Mx_1(\beta)), \ \alpha = 1,2,...,m,$$

где матрица $C = \|c_{\alpha\beta}\|$ удовлетворяет условиям (6) и (7). Введем функцию от матрицы:

$$I(C) = \frac{D(z_\infty(1)) + D(z_\infty(2)) + ... + D(z_\infty(m))}{D(x(1)) + D(x(2)) + ... + D(x(k))}.$$

Легко видеть, что при $n \to \infty$ и любом C:

$$I_n(C) \to I(C).$$

Рассмотрим решение предельной экстремальной задачи:

$$C_\infty = \underset{C}{Arg\min}\,(-I(C)).$$

Естественно ожидать, что:

$$\lim_{n\to\infty} C_n = C_\infty.$$

Действительно, это соотношение вытекает из приведенных выше общих результатов об асимптотическом поведении решений экстремальных статистических задач.

Таким образом, теория, развитая для пространств произвольной природы, позволяет единообразным образом изучать конкретные процедуры прикладной статистики.

Введению теоретических и эмпирических средних величин и законам больших чисел в пространствах произвольной природы посвящена работа [55]. Асимптотика решений экстремальных статистических задач изучена в [56]. Показано, что при оценивании параметров одношаговые оценки предпочтительнее оценок максимального правдоподобия [57]. Оценки плотности распределения вероятностей в пространствах произвольной природы, прежде всего ядерные оценки, в том числе в дискретных пространствах, рассмотрены в статьях [58–60]. Предельная теория непараметрических статистик, в том числе статистик интегрального типа, построена в [61]. Многообразие моделей регрессионного анализа, вероятностно-статистические модели корреляции и регрессии — предмет работ [62, 63]. Базовые результаты математической теории классификации приведены в [64]. Установлено, что прогностическая сила — наилучший показатель качества алгоритма диагностики [65]. На примере задач классификации сформулированы основные требования к методам анализа данных [66]. Методы снижения размерности пространства статистических данных и оценивания размерности вероятностно-статистической модели развиты в [67, 68].

Темы докладов, рефератов, исследовательских работ

1. Средние величины в теории и практике анализа статистических данных.
2. Средние и законы больших чисел в пространстве упорядочений.
3. Оптимизационные постановки основных задач прикладной статистики.
4. Минимизация расстояния как способ построения оценок параметров.
5. Примеры одношаговых оценок.
6. С помощью метода аппроксимации ступенчатыми функциями найдите асимптотическое распределение статистики Колмогорова.
7. Непараметрические оценки плотности в непрерывных и дискретных пространствах.
8. Критерии качества регрессионной модели.
9. Использование непараметрических оценок плотности для восстановления зависимости.
10. Классификация методов классификации.
11. Сравнительный анализ методов метрического и неметрического шкалирования.
12. Основные алгоритмы факторного анализа.
13. Состоятельные оценки размерности модели в задачах восстановления зависимости, классификации (расщепления смесей), многомерного шкалирования.
14. Применение общих результатов нечисловой статистики в конкретных областях прикладной статистики.

Контрольные вопросы и задачи

1. Как соотносятся эмпирические и теоретические средние величины для числовых данных и в пространствах произвольной природы?
2. Как соотносятся законы больших чисел для числовых случайных величин и в пространствах произвольной природы?
3. Какие экстремальные статистические задачи Вы знаете?
4. Как связаны законы больших чисел в пространствах произвольной природы и утверждения об асимптотическом поведении решений экстремальных статистических задач?
5. Почему одношаговые оценки предпочтительнее оценок максимального правдоподобия?
6. Почему описание числовых данных с помощью непараметрических оценок плотности предпочтительнее их описания с помощью гистограмм?

7. Можно ли строить непараметрические оценки плотности для результатов наблюдений из дискретных пространств?

8. Какие статистики интегрального типа Вы знаете?

9. Какую роль играет условие интегрируемости по Риману — Стилтьесу в предельной теории статистик интегрального типа?

10. Как соотносятся параметрическая регрессия и непараметрическая регрессия?

11. Как влияет предварительное выделение однородных групп на проведение регрессионного анализа?

12. Как соотносятся задачи группировки и задачи кластер-анализа?

13. В таблице приведены попарные расстояния между десятью социально-психологическими признаками способных к математике школьников [45]. Примените к этим данным алгоритмы ближнего соседа, средней связи и дальнего соседа. Для каждого из трех алгоритмов выделите наиболее устойчивые разбиения на кластеры.

Таблица к задаче 13

Попарные расстояния между признаками

	1	**2**	**3**	**4**	**5**	**6**	**7**	**9**	**10**
2	1 028	–	–	–	–	–	–	–	–
3	1 028	608	–	–	–	–	–	–	–
4	1 050	688	610	–	–	–	–	–	–
5	1 012	686	636	634	–	–	–	–	–
6	1 006	566	538	616	562	–	–	–	–
7	1 012	1 026	748	692	774	732	–	–	–
8	960	1 088	1 144	1 122	1 120	1 130	1 110	–	–
9	1 026	878	874	830	836	802	904	1 040	–
10	990	744	674	744	718	580	814	1 090	830

14. Какие Вам известны методы наглядного представления данных, основанные на идеях шкалирования и снижения размерности?

Литература

1. *Орлов, А. И.* Устойчивость в социально-экономических моделях / А. И. Орлов. — Москва : Наука, 1979. — 296 с.

2. *Прохоров, Ю. В.* Теория вероятностей. (Основные понятия. Предельные теоремы. Случайные процессы.) / Ю. В. Прохоров, Ю. А. Розанов. — Москва : Наука, 1973. — 496 с.

3. *Келли, Дж.* Общая топология / Дж. Келли. — Москва : Наука, 1968. — 384 с.

4. *Орлов, А. И.* Асимптотика решений экстремальных статистических задач / А. И. Орлов // Анализ нечисловых данных в системных исследованиях : сборник трудов. — Вып. 10. — Москва : Всесоюзный научно-исследовательский институт системных исследований, 1982. — С. 4–12.

5. *Жихарев, В. Н.* Законы больших чисел и состоятельность статистических оценок в пространствах произвольной природы / В. Н. Жихарев, А. И. Орлов // Статистические методы оценивания и проверки гипотез : межвузовский сборник научных трудов. — Пермь : Изд-во ПГНИУ, 1998. — С. 65–84.

6. *Орлов, А. И.* Эконометрика : учебник для вузов / А. И. Орлов. — 3-е изд., испр. и доп. — Москва : Экзамен, 2004. — 576 с.

7. *Смоляк, С. А.* Устойчивые методы оценивания: статистическая обработка неоднородных совокупностей / С. А. Смоляк, Б. П. Титаренко. — Москва : Статистика, 1980. — 208 с.

8. *Хьюбер, П.* Робастность в статистике / П. Хьюбер. — Москва : Мир, 1984. — 304 с.

9. Робастность в статистике. Подход на основе функций влияния / Ф. Хампель, Э. Рончетти, П. Рауссеу, В. Штаэль. — Москва : Мир, 1989. — 512 с.

10. *Лумельский, Я. П.* К вопросу сравнения несмещенных и других оценок / Я. П. Лумельский // Прикладная статистика. — Москва : Наука, 1983. — С. 316–319.

11. ГОСТ 11.010-81. Прикладная статистика. Правила определения оценок параметров и доверительных границ для биномиального и отрицательного биномиального распределений = Applied statistics. Point and interval estimators for parameters of binomial and negative binomial distribution : государтсвенный стандарт Союза ССР : утвержден и введен в действие Постановлением Государственного комитета СССР по стандартам от 30 марта 1981 г. № 1666 : дата введения 01 января 1982 г. — Москва : Изд-во стандартов, 1982. — 32 с. (В настоящее время отменен как нормативный документ, но может использоваться как научная публикация.)

12. *Сатаров, Г. А.* Новая статистическая модель парных сравнений // Экспертные оценки в задачах управления / Г. А. Сатаров, Д. С. Шмерлинг. — Москва : Изд-во Института проблем управления АН СССР, 1982. — С. 67–79.

13. *Лапига, А. Г.* Многокритериальные задачи управления качеством: построение прогноза качества в балльной шкале / А. Г. Лапига // Заводская лаборатория. — 1983. — Т. 49. — № 7. — С. 55–59.

14. *Закс, Ш.* Теория статистических выводов / Ш. Закс. — Москва : Мир, 1975. — 776 с.

15. *Бахмутов, В. О.* Использование метода максимального правдоподобия для оценки однородности результатов усталостных испытаний / В. О. Бахмутов, Л. Н. Косарев // Заводская лаборатория. — 1986. — Т. 52. — № 5. — С. 52–57.

16. *Резникова, А. Я.* Оценивание параметров вероятностных моделей парных и множественных сравнений / А. Я. Резникова, Д. С. Шмерлинг // Статистические методы оценивания и проверки гипотез : межвузовский сборник научных трудов. — Пермь : Изд-во ПГНИУ, 1984. — С. 110–120.

17. ГОСТ 11.011-83. Прикладная статистика. Правила определения оценок и доверительных границ для параметров гамма-распределения = Applied statistics. Regulations for determinations of estimates and confidence limits for parameters of gamma distribution : государственный стандарт Союза ССР : издание официальное : утвержден Постановлением Государственного комитета СССР по стандартам от 27 июня 1983 г. № 2684 : введен впервые : дата введения 1 января 1985 г. — Москва : Изд-во стандартов, 1984. — 53 с. (В настоящее время отменен как нормативный документ, но может использоваться как научная публикация.)

18. *Ибрагимов, И. А.* Асимптотическая теория оценивания / И. А. Ибрагимов, Р. З. Хасьминский. — Москва : Наука, 1979. — 528 с.

19. *Орлов, А. И.* О нецелесообразности использования итеративных процедур нахождения оценок максимального правдоподобия / А. И. Орлов // Заводская лаборатория. — 1986. — Т. 52. — № 5. — С. 67–69.

20. *Боровков, А. А.* Математическая статистика : учебное пособие для вузов / А. А. Боровков. — Москва : Наука, 1984. — 472 с.

21. *Орлов, А. И.* Одношаговые оценки для параметров гамма-распределения / А. И. Орлов, Н. Г. Миронова // Надежность и контроль качества. — 1988. — № 9. — С. 18–22.

22. *Петрович, М. Л.* Статистическое оценивание и проверка гипотез на ЭВМ / М. Л. Петрович, М. И. Давидович. — Москва : Финансы и статистика, 1989. — 191 с.

23. *Смирнов, Н. В.* О приближении плотностей распределения случайных величин / Н. В. Смирнов // Ученые записки МГПИ им. В. П. Потемкина. — 1951. — Т. XVI. — Вып. 3. — С. 69–96.

24. *Орлов, А. И.* Непараметрические оценки плотности в топологических пространствах / А. И. Орлов // Прикладная статистика. Ученые записки по статистике. — Т. 45. — Москва : Наука, 1983. — С. 12–40.

25 *Орлов, А. И.* Ядерные оценки плотности в пространствах произвольной природы / А. И. Орлов // Статистические методы оценивания и проверки гипотез : межвузовский сборник научных трудов. — Пермь : Изд-во ПГНИУ, 1996. — С. 68–75.

26. Пакет программ анализа данных «ППАНД» : учебное пособие / А. И. Орлов, И. Л. Легостаева [и др.]. — Москва : Сотрудничающий центр ВОЗ по профессиональной гигиене, 1990. — 93 с.

27. *Колмогоров, А. Н.* Элементы теории функций и функционального анализа : учебник / А. Н. Колмогоров, С. В. Фомин. — Москва : Наука, 1972. — 496 с.

28. *Орлов, А. И.* Асимптотическое поведение статистик интегрального типа / А. И. Орлов // Доклады АН СССР. — 1974. — Т. 219. — № 4. — С. 808–811.

29. *Орлов, А. И.* Асимптотическое поведение статистик интегрального типа / А. И. Орлов // Вероятностные процессы и их приложения : межвузовский сборник научных трудов. — Москва : МИЭМ, 1989. — С. 118–123.

30. *Большев, Л. Н.* Таблицы математической статистики / Л. Н. Большев, Н. В. Смирнов. — Москва : Наука, 1983. — 416 с.

31. *Мартынов, Г. В.* Критерии омега-квадрат / Г. В. Мартынов. — Москва : Наука, 1978. — 80 с.

32. *Гнеденко, Б. В.* Курс теории вероятностей : учебник / Б. В. Гнеденко. — 7-е изд., испр. — Москва : Эдиториал УРСС, 2001. — 320 с.

33. *Лоэв, М.* Теория вероятностей / М. Лоэв. — Москва : Издательство иностранной литературы, 1962. — 720 с.

34. *Холлендер, М.* Непараметрические методы статистики / М. Холлендер, Д. Вульф. — Москва : Финансы и статистика, 1983. — 518 с.

35. *Орлов, А. И.* О проверке симметрии распределения / А. И. Орлов. — Теория вероятностей и ее применения. — 1972. — Т. 17. — № 2. — С. 372–377.

36. *Орлов, А. И.* Общий взгляд на статистику объектов нечисловой природы / А. И. Орлов // Анализ нечисловой информации в социологических исследованиях. — Москва : Наука, 1985. — С. 58–92.

37. *Орлов, А. И.* Некоторые неклассические постановки в регрессионном анализе и теории классификации / А. И. Орлов // Программно-

алгоритмическое обеспечение анализа данных в медико-биологических исследованиях. — Москва : Наука, 1987. — С. 27–40.

38. *Орлов, А. И.* Статистика объектов нечисловой природы и экспертные оценки / А. И. Орлов // Экспертные оценки. Вопросы кибернетики. — Вып. 58. — Москва : Научный Совет АН СССР по комплексной проблеме «Кибернетика», 1979. — С. 17–33.

39. *Себер, Дж.* Линейный регрессионный анализ / Дж. Себер. — Москва : Мир, 1980. — 456 с.

40. *Орлов, А. И.* Асимптотика некоторых оценок размерности модели в регрессии / А. И. Орлов // Прикладная статистика. Ученые записки по статистике. Т. 45. — Москва : Наука, 1983. — С. 260–265.

41. *Орлов, А. И.* Об оценивании регрессионного полинома / А. И. Орлов // Заводская лаборатория. — 1994. — Т. 60. — № 5. — С. 43–47.

42. *Орлов, А. И.* Методы поиска наиболее информативных множеств признаков в регрессионном анализе / А. И. Орлов // Заводская лаборатория. — 1995. — Т. 61. — № 1. — С. 56–58.

43. *Кендалл, М. Дж.* Многомерный статистический анализ и временные ряды / М. Дж. Кендалл, А. Стьюарт. — Москва : Наука, 1976. — 736 с.

44. *Орлов, А. И.* Некоторые вероятностные вопросы теории классификации / А. И. Орлов // Прикладная статистика. Ученые записки по статистике. — Т. 45. — Москва : Наука, 1983. — С. 166–179.

45. *Орлов, А. И.* Математические методы в изучении способных к математике школьников / А. И. Орлов, Г. А. Гусейнов // Исследования по вероятностно-статистическому моделированию реальных систем. — Москва : ЦЭМИ АН СССР, 1977. — С. 80–93.

46. *Куперштох, В. Л.* Сумма внутренних связей как показатель качества классификации / В. Л. Куперштох, Б. Г. Миркин, В. А. Трофимов // Автоматика и телемеханика. — 1976. — № 3. — С. 91–98.

47. *Орлов, А. И.* Математические методы исследования и диагностика материалов (обобщающая статья) / А. И. Орлов // Заводская лаборатория. — 2003. — Т. 69. — № 3. — С. 53–64.

48. Прогнозирование исхода инфаркта миокарда с помощью программы «Кора-3» / И. М. Гельфанд, М. А. Алексеевская, Ш. А. Губерман [и др.] // Кардиология. — 1977. — Т. 17. — № 6. — С. 19–23.

49. *Харман, Г.* Современный факторный анализ / Г. Харман. — Москва : Статистика, 1972. — 488 с.

50. *Терехина, А. Ю.* Анализ данных методами многомерного шкалирования / А. Ю. Терехина. — Москва : Наука, 1986. — 168 с.

51. *Перекрест, В. Т.* Нелинейный типологический анализ социально-экономической информации: математические и вычислительные методы / В. Т. Перекрест. — Ленинград : Наука, 1983. — 176 с.

52. Анализ нечисловой информации / Ю. Н. Тюрин, Б. Г. Литвак, А. И. Орлов [и др.]. — Москва : Научный Совет АН СССР по комплексной проблеме «Кибернетика», 1981. — 80 с.

53. *Орлов, А. И.* Методы снижения размерности / А. И. Орлов // Толстова, Ю. Н. Основы многомерного шкалирования. — Москва : Издательство КДУ, 2006. — 160 с.

54. *Орлов, А. И.* Прикладная статистика / А. И. Орлов. — 2-е изд., испр. и доп. — Москва : Экзамен, 2007. — 672 с.

55. *Орлов, А. И.* Средние величины и законы больших чисел в пространствах произвольной природы / А. И. Орлов // Научный журнал КубГАУ. — 2013. — № 89. — С. 556–586.

56. *Орлов, А. И.* Предельная теория решений экстремальных статистических задач / А. И. Орлов // Научный журнал КубГАУ. — 2017. — № 133. — С. 579–600.

57. *Орлов, А. И.* Оценивание параметров: одношаговые оценки предпочтительнее оценок максимального правдоподобия / А. И. Орлов // Научный журнал КубГАУ. — 2015. — № 109. — С. 208–237.

58. *Орлов, А. И.* Оценки плотности распределения вероятностей в пространствах произвольной природы / А. И. Орлов // Научный журнал КубГАУ. — 2014. — № 99. — С. 15–32.

59. *Орлов, А. И.* Предельные теоремы для ядерных оценок плотности в пространствах произвольной природы / А. И. Орлов // Научный журнал КубГАУ. — 2015. — № 108. — С. 316–333.

60. *Орлов, А. И.* Непараметрические ядерные оценки плотности вероятности в дискретных пространствах / А. И. Орлов // Научный журнал КубГАУ. — 2016. — № 122. — С. 833–855.

61. *Орлов, А. И.* Предельная теория непараметрических статистик / А. И. Орлов // Научный журнал КубГАУ. — 2014. — № 100. — С. 31–52.

62. *Орлов, А. И.* Многообразие моделей регрессионного анализа (обобщающая статья) / А. И. Орлов // Заводская лаборатория. Диагностика материалов. — 2018. — Т. 84. — № 5. — С. 63–73.

63. *Орлов, А. И.* Вероятностно-статистические модели корреляции и регрессии / А. И. Орлов // Научный журнал КубГАУ. — 2020. — № 160. — С. 130–162.

64. *Орлов, А. И.* Базовые результаты математической теории классификации / А. И. Орлов // Научный журнал КубГАУ. — 2015. — № 110. — С. 219–239.

65. *Орлов, А. И.* Прогностическая сила — наилучший показатель качества алгоритма диагностики / А. И. Орлов // Научный журнал КубГАУ. — 2014. — № 99. — С. 33–49.

66. *Орлов, А. И.* Основные требования к методам анализа данных (на примере задач классификации) / А. И. Орлов // Научный журнал КубГАУ. — 2020. — № 159. — С. 239–267.

67. *Луценко, Е. В.* Методы снижения размерности пространства статистических данных / Е. В. Луценко, А. И. Орлов // Научный журнал КубГАУ. — 2016. — № 119. — С. 92–107.

68. *Орлов, А. И.* Оценивание размерности вероятностно-статистической модели / А. И. Орлов // Научный журнал КубГАУ. — 2020. — № 162. — С. 1–36.

ГЛАВА 3. СТАТИСТИКА НЕЧИСЛОВЫХ ДАННЫХ КОНКРЕТНЫХ ВИДОВ

От статистики в пространствах произвольной природы перейдем к обсуждению проблем анализа конкретных видов нечисловых данных. Начнем с задач обработки результатов измерений в шкалах, отличных от абсолютных.

3.1. ИНВАРИАНТНЫЕ АЛГОРИТМЫ И СРЕДНИЕ ВЕЛИЧИНЫ

Инвариантные алгоритмы и средние величины. Основное требование к алгоритмам анализа данных формулируется в теории измерений (см. главу 1) так: *выводы, сделанные на основе данных, измеренных в шкале определенного типа, не должны меняться при допустимом преобразовании шкалы измерения этих данных.* Другими словами, выводы должны быть *инвариантны* по отношению к допустимым преобразованиям шкалы.

Таким образом, одна из основных целей теории измерений — борьба с субъективизмом исследователя при приписывании численных значений реальным объектам. Так, расстояния можно измерять в аршинах, метрах, микронах, милях, парсеках и других единицах измерения. Массу (вес) — в пудах, килограммах, фунтах и других единицах измерения. Цены на товары и услуги можно указывать в юанях, рублях, тенге, гривнах, латах, кронах, марках, долларах США и иных валютах (при фиксированных курсах пересчета). Подчеркнем очень важное, хотя и вполне очевидное обстоятельство: выбор единиц измерения зависит от конкретного исследователя, или от соглашения, к которому пришла группа лиц, т.е. субъективен. *Статистические выводы могут быть адекватны реальности только тогда, когда они не зависят от того, какую единицу измерения предпочтет исследователь, т.е. когда они инвариантны относительно допустимого преобразования шкалы.*

Оказывается, сформулированное условие является достаточно сильным. Из многих алгоритмов анализа статистических данных ему удовлетворяют лишь некоторые. Покажем это на примере сравнения средних величин.

Пусть $X_1, X_2, \ldots, X_n$ — выборка объема n. В статистике часто используют выборочное среднее арифметическое:

$$X_{ср} = \frac{X_1 + X_2 + \ldots + X_n}{n}.$$

Использование среднего арифметического настолько привычно, что второе слово в термине часто опускают. И говорят о средней зарплате, среднем доходе и других средних для конкретных экономических данных, подразумевая под «средним» среднее арифметическое. Такая традиция может приводить к ошибочным выводам. Покажем это на примере расчета средней заработной платы (среднего дохода) работников условного предприятия (табл. 1).

Таблица 1

Численность работников различных категорий, их заработная плата и суммарные доходы (в условных единицах)

№ п/п	Категория работников	Число работников	Заработная плата	Суммарные доходы
1	Низкоквалифицированные рабочие	40	100	4 000
2	Высококвалифицированные рабочие	30	200	6 000
3	Инженеры и служащие	25	300	7 500
4	Менеджеры	4	1 000	4 000
5	Генеральный директор (владелец)	1	18 500	18 500
6	*Всего*	100	–	40 000

Первые три строки в табл. 1 вряд ли требуют пояснений. Менеджеры — это директора (управляющие) по направлениям. А именно, по производству (главный инженер), по финансам, по маркетингу и сбыту, по персоналу (по кадрам). Владелец сам руководит предприятием в качестве генерального директора. В столбце «заработная плата» указаны доходы одного работника соответствующей категории, а в столбце «суммарные доходы» — доходы всех работников соответствующей категории.

Фонд оплаты труда составляет 40 000 условных единиц, работников всего 100, следовательно, средняя заработная плата составляет 40 000 / 100 = 400 единиц. Однако эта средняя арифметическая величина явно не соответствует интуитивному представлению о «средней зарплате». Из 100 работников лишь 5 имеют заработную плату, ее превышающую, а зарплата остальных 95 % существенно меньше средней арифметической. Причина очевидна — заработная плата одного человека — генерального директора — превышает заработную плату 95 работников — низкоквалифицированных и высококвалифицированных рабочих, инженеров и служащих.

Ситуация напоминает ту, что описана в известном рассказе о больнице, в которой 10 больных, из них у 9 температура 40 °С, а один уже отмучился, лежит в морге с температурой 0 °С. Между тем средняя температура по больнице равна 36 °С — лучше не бывает!

Сказанное показывает, что среднее арифметическое можно использовать лишь для достаточно однородных совокупностей (без больших выбросов в ту или иную сторону). А какие виды средних величин целесообразно использовать для описания заработной платы? Вполне естественно использовать медиану. Для данных табл. 1 медиана — это среднее арифметическое 50-го и 51-го работника, если их заработные платы расположены в порядке неубывания. В вариационном ряду сначала идут зарплаты 40 низкоквалифицированных рабочих, а затем — с 41-го до 70-го работника — заработные платы высококвалифицированных рабочих. Следовательно, медиана попадает именно на них и равна 200. У 50-ти работников заработная плата не превосходит 200, и у 50-ти — не менее 200, поэтому медиана показывает «центр», около которого группируется основная масса исследуемых величин. Еще одна средняя величина — мода, наиболее часто встречающееся значение. В рассматриваемом случае мода — это заработная плата низкоквалифицированных рабочих, т.е. 100 условных единиц. Таким образом, для описания зарплаты имеем три средние величины — моду (100 единиц), медиану (200 единиц) и среднее арифметическое (400 единиц). Для наблюдающихся в реальной жизни распределений доходов и заработной платы справедлива та же закономерность: мода меньше медианы, а медиана меньше среднего арифметического.

Для чего в технических, экономических, медицинских и иных исследованиях используются средние величины? Обычно для того, чтобы заменить совокупность чисел одним числом, чтобы сравнивать совокупности с помощью средних.

Пусть, например, Y_1, Y_2, ..., Y_n — совокупность оценок экспертов, «выставленных» одному объекту экспертизы (например, одному из вариантов стратегического развития фирмы), Z_1, Z_2, ..., Z_n — второму (другому варианту такого развития). Как сравнивать эти совокупности? Очевидно, самый простой способ — по средним значениям.

А как вычислять средние? Известны различные виды средних величин: среднее арифметическое, медиана, мода, среднее геометрическое, среднее гармоническое, среднее квадратическое. Напомним, что общее понятие средней величины введено французским математиком первой половины XIX в. академиком О. Коши. Оно таково: средней величиной является любая функ-

ция $f(X_1, X_2, ..., X_n)$ такая, что при всех возможных значениях аргументов значение этой функции не меньше, чем минимальное из чисел $X_1, X_2, ..., X_n$, и не больше, чем максимальное из этих чисел. Все перечисленные выше виды средних величин являются средними по Коши.

При допустимом преобразовании шкалы значение средней величины, очевидно, меняется. Но выводы о том, для какой совокупности среднее больше, а для какой — меньше, не должны меняться (в соответствии с требованием инвариантности выводов, принятом как основное требование в теории измерений). Сформулируем соответствующую математическую задачу поиска вида средних величин, результат сравнения которых устойчив относительно допустимых преобразований шкалы.

Пусть $f(X_1, X_2, ..., X_n)$ — среднее по Коши. Пусть среднее по первой совокупности меньше среднего по второй совокупности:

$$f(Y_1, Y_2, ..., Y_n) < f(Z_1, Z_2, ..., Z_n).$$

Тогда согласно теории измерений для устойчивости результата сравнения средних необходимо, чтобы для любого допустимого преобразования g (из группы допустимых преобразований в соответствующей шкале) было справедливо также неравенство:

$$f(g(Y_1), g(Y_2), ..., g(Y_n)) < f(g(Z_1), g(Z_2), ..., g(Z_n)),$$

т.е. среднее преобразованных значений из первой совокупности также было меньше среднего преобразованных значений для второй совокупности. Причем сформулированное условие должно быть верно для любых двух совокупностей $Y_1, Y_2, ..., Y_n$ и $Z_1, Z_2, ..., Z_n$. И для любого допустимого преобразования. Средние величины, удовлетворяющие сформулированному условию, назовем *допустимыми* (в соответствующей шкале). Согласно теории измерений только допустимыми средними величинами можно пользоваться при анализе мнений экспертов и иных данных, измеренных в рассматриваемой шкале.

С помощью математической теории, развитой в монографии [1], удается описать вид допустимых средних величин в основных шкалах. Сразу ясно, что для данных, измеренных в шкале наименований, допустимых средних нет.

Средние величины в порядковой шкале. Рассмотрим обработку, для определенности, мнений экспертов, измеренных в порядковой шкале. Справедливо следующее утверждение.

Теорема 1. Из всех средних по Коши допустимыми средними в порядковой шкале являются только члены вариационного ряда (порядковые статистики).

Теорема 1 справедлива при условии, что среднее $f(X_1, X_2, ..., X_n)$ является непрерывной (по совокупности переменных) и симметрической функцией. Последнее означает, что при перестановке аргументов значение функции $f(X_1, X_2, ..., X_n)$ не меняется. Это условие является вполне естественным, ибо среднюю величину мы находим для *совокупности* (*множества*), а не для *последовательности*. Множество не меняется в зависимости от того, в какой последовательности мы перечисляем его элементы.

Согласно теореме 1 в качестве среднего для данных, измеренных в порядковой шкале, можно использовать, в частности, медиану (при нечетном объеме выборки). При четном же объеме следует применять один из двух центральных членов вариационного ряда — как их иногда называют, левую медиану или правую медиану. Моду тоже можно использовать — она всегда является членом вариационного ряда. Можно применять выборочные квартили, минимум и максимум, децили и т.п. Но никогда нельзя рассчитывать среднее арифметическое, среднее геометрическое и т.д.

Приведем численный пример, показывающий некорректность использования среднего арифметического $f(X_1, X_2) = (X_1 + X_2)/2$ в порядковой шкале. Пусть $Y_1 = 1$, $Y_2 = 11$, $Z_1 = 6$, $Z_2 = 8$. Тогда $f(Y_1, Y_2) = 6$, что меньше, чем $f(Z_1, Z_2) = 7$. Пусть строго возрастающее преобразование g таково, что $g(1) = 1$, $g(6) = 6$, $g(8) = 8$, $g(11) = 99$. Таких преобразований много. Например, можно положить $g(x) = x$ при x, не превосходящих 8, и $g(x) = 99(x - 8)/3 + 8$ для x, больших 8. Тогда $f(g(Y_1), g(Y_2)) = 50$, что больше, чем $f(g(Z_1), g(Z_2)) = 7$. Как видим, в результате допустимого, т.е. строго возрастающего преобразования шкалы упорядоченность средних величин изменилась.

Таким образом, теория измерений выносит жесткий приговор среднему арифметическому — использовать его в порядковой шкале нельзя. Однако же те, кто не знает теории измерений, используют его. Всегда ли они ошибаются? Оказывается, можно в какой-то мере (но отнюдь не полностью!) реабилитировать среднее арифметическое, если перейти к вероятностной постановке и к тому же удовлетвориться результатами для больших объемов выборок. В монографии [1] получено также следующее утверждение.

Теорема 2. Пусть Y_1, Y_2, ..., Y_m — независимые одинаково распределенные случайные величины с функцией распределения $F(x)$, а Z_1, Z_2, ..., Z_n — независимые одинаково распределенные случайные величины с функцией распределения $H(x)$, причем выборки Y_1, Y_2, ..., Y_m и Z_1, Z_2, ..., Z_n независимы между собой и $MY_1 > MZ_1$. Для того, чтобы вероятность события:

$$\left\{\omega : \frac{g(Y_1)+g(Y_2)+...+g(Y_m)}{m} > \frac{g(Z_1)+g(Z_2)+...+g(Z_n)}{n}\right\}$$

стремилась к 1 при $\min(m,n) \to \infty$ для любой строго возрастающей непрерывной функции g, удовлетворяющей условию:

$$\overline{\lim_{|x|\to\infty}} \left|\frac{g(x)}{x}\right| < \infty,$$

необходимо и достаточно, чтобы при всех x выполнялось неравенство $F(x) \le H(x)$, причем существовало число x_0, для которого $F(x_0) < H(x_0)$.

Примечание. Условие с верхним пределом носит чисто внутриматематический характер. Фактически функция g — произвольное допустимое преобразование в порядковой шкале.

Согласно теореме 2 средним арифметическим можно пользоваться и в порядковой шкале, если сравниваются выборки из двух распределений, удовлетворяющих приведенному в теореме неравенству. Проще говоря, одна из функций распределения должна всегда лежать над другой. Функции распределения не могут пересекаться, им разрешается только касаться друг друга. Это условие выполнено, например, если функции распределения отличаются только сдвигом, т.е.

$$F(x) = H(x + b)$$

при некотором b. Последнее условие выполняется, если два значения некоторой величины измеряются с помощью одного и того же средства измерения, у которого распределение погрешностей не меняется при переходе от измерения одного значения рассматриваемой величины к измерению другого.

Средние по Колмогорову. Естественная система аксиом (требований к средним величинам) приводит к так называемым ассоциативным средним. Их

общий вид нашел в 1930 г. А. Н. Колмогоров [2]. Теперь их называют «средними по Колмогорову».

Для чисел $X_1, X_2, ..., X_n$ среднее по Колмогорову вычисляется по формуле:

$$G\{(F(X_1) + F(X_2) + ... + F(X_n))/n\},$$

где F — строго монотонная функция (т.е. строго возрастающая или строго убывающая), G — функция, обратная к F. Среди средних по Колмогорову — много хорошо известных персонажей. Так, если $F(x) = x$, то среднее по Колмогорову — это среднее арифметическое, если $F(x) = \ln x$, то среднее геометрическое, если $F(x) = 1/x$, то среднее гармоническое, если $F(x) = x^2$, то среднее квадратическое, и т.д. (в последних трех случаях усредняются положительные величины). Среднее по Колмогорову — частный случай среднего по Коши. С другой стороны, такие популярные средние, как медиана и мода, нельзя представить в виде средних по Колмогорову. В монографии [1] доказаны следующие утверждения.

Теорема 3. При справедливости некоторых внутриматематических условий регулярности в шкале интервалов из всех средних по Колмогорову допустимым является только среднее арифметическое.

Таким образом, среднее геометрическое или среднее квадратическое температур (в шкале Цельсия), потенциальных энергий или координат точек не имеют смысла. В качестве среднего надо применять среднее арифметическое. А также можно использовать медиану или моду.

Теорема 4. При справедливости некоторых внутриматематических условий регулярности в шкале отношений из всех средних по Колмогорову допустимыми являются только степенные средние с $F(x) = x^c$, $c \neq 0$, и среднее геометрическое.

Есть ли средние по Колмогорову, которыми нельзя пользоваться в шкале отношений? Конечно, есть. Например, с $F(x) = e^x$.

Замечание 1. Среднее геометрическое является пределом степенных средних при $c \to 0$.

Замечание 2. Подробное описание «внутриматематических условий регулярности», упомянутых в формулировках теорем 3 и 4, можно найти в [1, 3].

Аналогично средним величинам могут быть изучены и другие статистические характеристики — показатели разброса, связи, расстояния и др. (см., например, [1]). Нетрудно показать, например, что коэффициент корреляции не меняется при любом допустимом преобразовании в шкале интерва-

лов, как и отношение дисперсий. Дисперсия не меняется в шкале разностей, коэффициент вариации — в шкале отношений, и т.д.

Приведенные выше результаты о средних величинах широко применяются, причем не только в экономике, менеджменте, теории экспертных оценок или социологии, но и в инженерном деле, например, для анализа методов агрегирования датчиков в АСУ ТП доменных печей. Велико прикладное значение теории измерений в задачах стандартизации и управления качеством, в частности, в квалиметрии. Здесь есть и интересные теоретические результаты. Так, например, любое изменение коэффициентов весомости единичных показателей качества продукции приводит к изменению упорядочения изделий по средневзвешенному показателю (эта теорема доказана проф. В. В. Подиновским).

При подготовке и принятии решений необходимо использовать только инвариантные алгоритмы обработки данных. В настоящем разделе показано, что требование инвариантности выделяет из многих алгоритмов усреднения лишь некоторые, соответствующие используемым шкалам измерения. Инвариантные алгоритмы в общем случае рассматриваются в математической теории измерений [4]. Нацеленное на прикладные исследования изложение теории измерений дается в монографиях [1, 5, 6].

3.2. ТЕОРИЯ СЛУЧАЙНЫХ ТОЛЕРАНТНОСТЕЙ

В прикладных исследованиях обычно используют три конкретных вида бинарных отношений — ранжировки, разбиения и толерантности. Статистические теории ранжировок [7] и разбиений [8] достаточно сложны с математической точки зрения. Поэтому продвинуться удается не очень далеко. Теория случайных ранжировок, в частности, изучает в основном равномерные распределения на множестве ранжировок. Теория случайных толерантностей позволяет рассмотреть принципиально более общие ситуации. Это объясняется, грубо говоря, тем, что для теории толерантностей оказываются полезными суммы некоторых независимых случайных величин, а для теорий ранжировок и разбиений аналогичные случайные величины зависимы, а потому изучение их сумм затруднено. Теория случайных толерантностей является частным случаем теории люсианов, рассматриваемой в разделе 3.4. Здесь приводим результаты, специфичные именно для толерантностей.

Пусть X — конечное множество из k элементов. Толерантность A на множестве X, как и любое бинарное отношение, однозначно описывается

матрицей $\|a(i, j)\|$, $1 \le i, j \le k$, где $a(i, j) = 1$, если элементы с номерами i и j связаны отношением толерантности, и $a(i, j) = 0$ в противном случае. Поскольку толерантность — это рефлексивное и симметричное бинарное отношение, то достаточно рассматривать часть матрицы, лежащую над главной диагональю: $\|a(i, j), 1 \le i < j \le k\|$. Между наборами $\|a(i, j), 1 \le i < j \le k\|$ из 0 и 1 и толерантностями на X имеется взаимно-однозначное соответствие.

Пусть $A = A(\omega)$ — случайная толерантность, равномерно распределенная на множестве всех толерантностей на X. Легко видеть, что в этом случае $a(i, j)$, $1 \le i < j \le k$, — независимые случайные величины, принимающие значения 0 и 1 с вероятностями 0,5. Этот факт, несмотря на свою математическую тривиальность, является решающим для построения базовой части теории толерантностей. Для аналогичных постановок в теории ранжировок и разбиений величины $a(i, j)$ оказываются зависимыми.

Следовательно, случайная величина:

$$B(A) = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} a(i, j)$$

имеет биномиальное распределение с параметрами $k(k - 1)/2$, 1/2 и асимптотически нормальна при $k \to \infty$.

Проверка гипотез о согласованности. Рассмотрим s независимых толерантностей $A_1, A_2, \ldots, A_s$, равномерно распределенных на множестве всех толерантностей на X. Рассмотрим вектор:

$$\xi_{ks} = \{d(A_p, A_q),\ 1 \le p < q \le s\} = \sum_{1 \le i < j \le k} \{|\, a_p(i, j) - a_q(i, j)\,|,\ 1 \le p < q \le s\}, \qquad (1)$$

где $d(A_p, A_q)$ — расстояние между толерантностями A_p и A_q, аксиоматически введенное в главе 1. В (1) предполагается, что пары (p, q), $p < q$, располагаются в раз навсегда установленном порядке, для определенности в лексикографическом (т.е. пары упорядочиваются в соответствии со значением p, а при одинаковых p — по значению q).

Вектор ξ_{ks} является суммой $k(k - 1)/2$ независимых одинаково распределенных случайных векторов, а потому асимптотически нормален при $k \to \infty$. Координаты этого вектора независимы, поскольку, как нетрудно видеть, координаты каждого слагаемого независимы (это свойство не сохраняется при отклонении от равномерности распределения). Распределения случайных ве-

личин $a_p(i, j)$ и $|a_p(i, j) - a_q(i, j)|$ совпадают, поэтому распределения $B(A)$ и $d(A_p, A_q)$ также совпадают.

В силу многомерной центральной предельной теоремы (прил. 1) распределение вектора:

$$\eta_{ks} = \sqrt{\frac{2}{k(k-1)}}\left(\xi_{rs} - \frac{k(k-1)}{2}\left(\frac{1}{2}, \frac{1}{2}, \ldots, \frac{1}{2}\right)\right)$$

сходится при $k \to \infty$ к распределению многомерного нормального вектора η_s, ковариационная матрица которого совпадает с ковариационной матрицей вектора η_{ks}, а математическое ожидание равно 0. Таким образом, координаты случайного вектора η_s независимы и имеют стандартное нормальное распределение с математическим ожиданием 0 и дисперсией 1. В соответствии с теоремами о наследовании сходимости (см. прил. 1) распределение $f(\eta_{ks})$ сходится при $k \to \infty$ к распределению $f(\eta_s)$ для достаточно широкого класса функций f, в частности, для всех непрерывных функций. В качестве примеров рассмотрим статистики:

$$W = \sum_{1 \le p < q \le s} d(A_p, A_q), \quad N = \sum_{1 \le p < q \le s} \left(d(A_p, A_q) - \frac{k(k-1)}{4}\right)^2.$$

При $k \to \infty$ распределения случайных величин:

$$\frac{8W - s(s-1)k(k-1)}{2\sqrt{s(s-1)k(k-1)}}, \quad \frac{8N}{k(k-1)}$$

сходятся соответственно к стандартному нормальному распределению с математическим ожиданием 0 и дисперсией 1 и распределению хи-квадрат с $s(s-1)/2$ степенями свободы. Статистики W и N могут быть использованы для проверки гипотезы о равномерности распределения толерантностей.

Как известно, в теории ранговой корреляции [7], т.е. в теории случайных ранжировок, в качестве единой выборочной меры связи нескольких признаков используется коэффициент согласованности $W(R)$, называемый также коэффициентом конкордации [9, табл. 6.10]. Его распределение затабулировано в предположении равномерности распределения на пространстве ранжировок (без связей). Непосредственным аналогом $W(R)$ в случае толерантностей является статистика W. Статистики W и N играют ту же роль для то-

лерантностей, что $W(R)$ для ранжировок, однако математико-статистическая теория в случае толерантностей гораздо проще, чем для ранжировок.

Обобщением равномерно распределенных толерантностей являются *толерантности с независимыми связями*. В этой постановке предполагается, что $a(i, j)$, $1 \le i < j \le k$, — независимые случайные величины, принимающие значения 0 и 1. Обозначим $P(a(i, j) = 1) = p(i, j)$. Тогда $P(a(i, j) = 0) = 1 - p(i, j)$. Таким образом, распределение толерантности с независимыми связями задается нечеткой толерантностью, т.е. вектором:

$$P = \{p(i, j), 1 \le i < j \le k\}.$$

Пусть имеется s независимых случайных толерантностей $A_1, A_2, \ldots, A_s$ с независимыми связями, распределения которых задаются векторами $P_1, P_2, \ldots, P_s$ соответственно. Рассмотрим проверку гипотезы согласованности:

$$H_0: P_1 = P_2 = \ldots = P_s.$$

Она является более слабой, чем гипотеза равномерности:

$$H_0^{'}: P_1 = P_2 = \ldots = P_s = (1/2, 1/2, \ldots, 1/2),$$

для проверки которой используют статистики W и N (см. выше).

Пусть сначала $s = 2$. Тогда

$$P\{|a_1(i, j) - a_2(i, j)| = 1\} = q(i, j), P\{|a_1(i, j) - a_2(i, j)| = 0\} = 1 - q(i, j),$$

где

$$q(i, j) = p_1(i, j)\,(1 - p_2(i, j)) + p_2(i, j)\,(1 - p_1(i, j)).$$

Следовательно, расстояние $d(A_1, A_2)$ между двумя случайными толерантностями с независимыми связями есть сумма $k(k - 1)/2$ независимых случайных величин, принимающих значения 0 и 1, причем математическое ожидание и дисперсия $d(A_1, A_2)$ таковы:

$$Md(A_1, A_2) = \sum_{1 \le i < j \le k} q(i, j), \quad Dd(A_1, A_2) = \sum_{1 \le i < j \le k} q(i, j)(1 - q(i, j)). \qquad (2)$$

Пусть $k \to \infty$. Если $Dd(A_1, A_2) \to \infty$, то условие Линденберга Центральной Предельной Теоремы теории вероятностей выполнено (см. прил. 1), и распределение нормированного расстояния:

$$\frac{d(A_1, A_2) - Md(A_1, A_2)}{\sqrt{Dd(A_1, A_2)}} \tag{3}$$

сходится к стандартному нормальному распределению с математическим ожиданием 0 и дисперсией 1. Если существует число $\delta > 0$ такое, что при всех k, i, j, $1 \le i < j \le k$, вероятности $p_1(i, j)$ и $p_2(i, j)$ лежат внутри интервала $(\delta; 1 - \delta)$, то $Dd(A_1, A_2) \to \infty$.

Соотношения (2), (3) и им подобные позволяют рассчитать мощность критериев, основанных на статистиках W и N, при $k \to \infty$, подобно тому, как это сделано в [1, глава 4.5]. Поскольку подобные расчеты не требуют новых идей, не будем приводить их здесь.

Обычно P_1 и P_2 неизвестны. Для проверки гипотезы $P_1 = P_2$ в некоторых случаях можно порекомендовать отвергать гипотезу на уровне значимости α, если $d(A_1, A_2) \ge d_0$, где d_0 есть $(1 - \alpha)$ — квантиль распределения расстояния между двумя независимыми равномерно распределенными случайными толерантностями, т.е. квантиль биномиального распределения $B(A)$. Укажем достаточные условия такой рекомендации.

Пусть

$$p = (p_1(i, j) + p_2(i, j))/2, \quad p_1(i, j) = p + \Delta,$$

тогда

$$p_2(i, j) = p - \Delta, \; q = q(i, j) = 2p(1 - p) + 2\Delta^2. \tag{4}$$

Если существует число $\delta > 0$ такое, что:

$$q - 1/2 > \delta > 0 \tag{5}$$

при всех k, i, j, то гипотеза $P_1 = P_2$ будет отвергаться с вероятностью, стремящейся к 1 при $k \to \infty$. Из (4) следует, что при фиксированном p существует Δ такое, что выполнено (5), тогда и только тогда, когда $0{,}25 < p < 0{,}75$.

Своеобразие постановки задачи проверки гипотезы состоит в том, что при росте k число неизвестных параметров, т.е. координат векторов P_i, растет пропорционально объему данных. Поэтому и столь далекая от оптимальности процедура, как описанная в двух предыдущих абзацах, представляет некоторый практический интерес. Для случая $s \geq 4$ в теории люсианов (раздел 3.4) разработаны методы проверки гипотезы согласованности H_0: $P_1 = P_2 = \ldots = P_s$.

Нахождение группового мнения. Пусть $A_1, A_2, \ldots, A_s$ — случайные толерантности, описывающие мнения s экспертов. Для нахождения группового мнения будем использовать медиану Кемени, т.е. эмпирическое среднее относительно расстояния Кемени, введенного в главе 1. Медианой Кемени является:

$$A_{cp} = Arg \min_{A} \sum_{p=1}^{s} d(A_p, A).$$

Легко видеть, что $A_{cp} = ||a_{cp}(i, j)||$ удовлетворяет условию: $a_{cp}(i, j) = 1$, если:

$$\sum_{p=1}^{s} a_p(i, j) > \frac{s}{2},$$

и $a_{cp}(i, j) = 0$, если:

$$\sum_{p=1}^{s} a_p(i, j) < \frac{s}{2}.$$

Следовательно, при нечетном s групповое мнение A_{cp} определяется однозначно. При четном s неоднозначность возникает в случае:

$$\sum_{p=1}^{s} a_p(i, j) = \frac{s}{2}.$$

Тогда медиана Кемени A_{cp} — не одна толерантность, а множество толерантностей, минимум суммы расстояний достигается и при $a_{cp}(i, j) = 1$, и при $a_{cp}(i, j) = 0$.

Асимптотическое поведение группового мнения (медианы Кемени для толерантностей) вытекает из общих результатов о законах больших чисел в

пространствах произвольной природы (глава 2), поэтому рассматривать его здесь нет необходимости.

Дихотомические (бинарные) признаки в классической асимптотике. Многое в предыдущем изложении определялось спецификой толерантностей. В частности, особая роль равномерности распределения на множестве всех толерантностей оправдывала специальное рассмотрение статистик W и N; аксиоматически введенное расстояние d между толерантностями играло важную роль в приведенных выше результатах. Однако модель толерантностей с независимыми связями уже меньше связана со спецификой толерантностей. В ней толерантности можно рассматривать просто как частный случай люсианов. Широко применяется следующая модель порождения данных.

Пусть $A_1, A_2, \ldots, A_s$ — независимые люсианы. Это значит, что статистические данные имеют вид:

$$(A_1, A_2, \ldots, A_s) = \|X_{ij},\ i = 1,2, \ldots, s;\ j = 1, 2, \ldots, k\|, \qquad (6)$$

где X_{ij} — независимые в совокупности испытания Бернулли с вероятностями успеха:

$$(P_1, P_2, \ldots, P_s) = \|p_{ij},\ ,\ i = 1,2, \ldots, s;\ j = 1, 2, \ldots, k\|, \qquad (7)$$

где P_i — вектор вероятностей, описывающий распределение люсиана A_i. Особое значение имеют одинаково распределенные люсианы, для которых $P_1 = P_2 = \ldots = P_s = P$, где символом P обозначен общий вектор вероятностей.

Как обычно в математической статистике, содержательные результаты при изучении модели (6) — (7) можно получить в асимптотических постановках. При этом есть два принципиально разных предельных перехода: $s \to \infty$ и $k \to \infty$. Первый из них — традиционный: число неизвестных параметров постоянно, объем выборки s растет. Во втором число параметров растет, объем выборки остается постоянным, но общий объем данных ks растет пропорционально числу неизвестных параметров. Аналогом является асимптотическое изучение коэффициентов ранговой корреляции Кендалла и Спирмена: число ранжировок, т.е. объем выборки, постоянно (и равно 2), а число ранжируемых объектов растет.

Вторая постановка изучается в разделе 3.4, посвященном люсианам. Некоторые задачи в первой постановке рассмотрим здесь.

Случайные толерантности используются, в частности, для оценки нечетких толерантностей [1]. Для описания результатов опроса группы экспертов о сходстве объектов строят нечеткую толерантность $M = ||\mu_{ij}||$, $\mu_{ij} = l_{ij}/n_{ij}$, где n_{ij} — число ответов о сходстве i-го и j-го объектов, а l_{ij} — число положительных ответов из них. Если эксперты действуют в соответствии с единым вектором параметров P, то M — состоятельная оценка для P. Следующий вопрос при таком подходе — верно ли, что две группы экспертов «думают одинаково», т.е. используют совпадающие вектора P? Рассмотрим эту постановку на более общем языке люсианов.

Пусть $A_1, A_2, ..., A_m$ и $B_1, B_2, ..., B_n$ — две группы независимых в совокупности люсианов, одинаково распределенные в каждой группе с параметрами $P(A)$ и $P(B)$ соответственно. Требуется проверить гипотезу $P(A) = P(B)$. Естественным является переход к пределу при $\min(m, n) \to \infty$.

Пусть гипотеза справедлива. Предположим, что $p_i = p_i(A) = p_i(B) \neq 0$ при всех $i = 1, 2, ..., k$. (Разбор последствий нарушений этого условия оставляем читателю.) Пусть s_i — число единиц на i-м месте в первой группе люсианов, а t_i — во второй. Рассмотрим случайные величины:

$$\xi_i = \sqrt{\frac{mn}{m+n}}\left(\frac{s_i}{m} - \frac{t_i}{n}\right)\frac{1}{\sqrt{p_i(1-p_i)}}. \qquad (8)$$

Они независимы в совокупности. В соответствии с приведенными в приложении 1 предельными теоремами распределения случайных величин ξ_i при $\min(m, n) \to \infty$ сходятся к стандартному нормальному распределению с математическим ожиданием 0 и дисперсией 1. Эти свойства сохраняются при замене p_i в (8) на состоятельные оценки, построенные по статистическим данным, соответствующим i-му месту. Будем использовать эффективную оценку [10, с. 529]:

$$p_i^* = \frac{s_i + t_i}{m+n}. \qquad (9)$$

Подставим (9) в (8), получим статистики:

$$\xi_i^* = \sqrt{\frac{mn(m+n)}{(s_i+t_i)(m+n-s_i-t_i)}}\left(\frac{s_i}{m} - \frac{t_i}{n}\right).$$

Полученные статистики можно использовать для проверки рассматриваемой гипотезы, например, с помощью критериев, основанных на статистиках:

$$W = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} a_i \xi_i^*, \quad T = \sum_{i=1}^{k} (\xi_i^*)^2, \quad \sum_{i=1}^{k} a_i^2 = 1.$$

С помощью результатов приложения 1 получаем, что W имеет в пределе при $\min(m, n) \to \infty$ стандартное нормальное распределение, а T — распределение хи-квадрат с k степенями свободы.

Рассмотрим распределение статистики W при альтернативных гипотезах. Положим:

$$\eta_{1m}^i = \frac{\sqrt{m}\left(\frac{s_i}{m} - p_i(A)\right)}{\sqrt{p_i(A)(1-p_i(A))}}, \quad \eta_{2n}^i = \frac{\sqrt{n}\left(\frac{t_i}{n} - p_i(B)\right)}{\sqrt{p_i(B)(1-p_i(B))}}.$$

Эти случайные величины независимы, распределение каждой из них при $\min(m, n) \to \infty$ сходится к стандартному нормальному распределению. Поскольку:

$$\frac{s_i}{m} = \frac{\eta_{1m}^i}{\sqrt{m}} \sqrt{p_i(A)(1-p_i(A))} + p_i(A), \quad \frac{t_i}{n} = \frac{\eta_{2n}^i}{\sqrt{n}} \sqrt{p_i(B)(1-p_i(B))} + p_i(B),$$

то

$$\sqrt{\frac{mn}{m+n}}\left(\frac{s_i}{m} - \frac{t_i}{n}\right) = F + G,$$

где

$$F = \sqrt{\frac{mn}{m+n}}\left(\frac{\eta_{1m}^i}{\sqrt{m}} \sqrt{p_i(A)(1-p_i(A))} - \frac{\eta_{2n}^i}{\sqrt{n}} \sqrt{p_i(B)(1-p_i(B))}\right)$$

и

$$G = \sqrt{\frac{mn}{m+n}}\left(p_i(A) - p_i(B)\right).$$

В силу результатов приложения 1 распределение F при $\min(m, n) \to \infty$ сближается с нормальным распределением, математическое ожидание которого равно 0, а дисперсия:

$$\frac{n}{m+n} p_i(A)(1-p_i(A)) + \frac{m}{m+n} p_i(B)(1-p_i(B)) \le \frac{1}{4}.$$

Поэтому, чтобы получить собственное (т.е. невырожденное) распределение W при альтернативах, естественно рассмотреть модель:

$$p_i(A) = p_i + \frac{\theta_i}{2}\sqrt{\frac{m+n}{mn}}\sqrt{p_i(1-p_i)},\ \ p_i(B) = p_i - \frac{\theta_i}{2}\sqrt{\frac{m+n}{mn}}\sqrt{p_i(1-p_i)},\ \ i = 1,2,...,k\ ,$$

где θ_i — некоторые фиксированные числа. Тогда при $\min(m, n) \to \infty$ оценки p_i^* из (9) сходятся к p_i и ξ_i^* являются независимыми асимптотически нормальными случайными величинами с математическими ожиданиями θ_i и единичными дисперсиями. Опираясь на результаты приложения 1, заключаем, что распределение статистики W сходится к нормальному распределению с математическим ожиданием:

$$\theta_0 = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} a_i \theta_i$$

и единичной дисперсией.

Если в последней формуле $\theta_0 = 0$, то асимптотическое распределение W таково же, как и в случае справедливости нулевой гипотезы. От указанного недостатка свободна статистика T. Тем же путем, как и для W, получаем, что при $\min(m, n) \to \infty$ распределение T сходится к нецентральному хи-квадрат распределению с k степенями свободы и параметром нецентральности:

$$\Theta = \sum_{i=1}^{k} \theta_i^2\ .$$

Можно рассматривать ряд других задач, например, проверку совпадения параметров для нескольких групп люсианов (аналог дисперсионного анализа), установление зависимости $P(B)$ от $P(A)$ (аналог регрессионного анализа), отнесение вновь поступающего люсиана к одной из групп (речь идет о задаче диагностики — аналоге дискриминантного анализа; она пред-

ставляет интерес, например, при применении тестов типа MMPI оценки психического состояния личности) и т.д. Однако принципиальных трудностей на пути развития соответствующих методов не видно, и мы не будем их здесь рассматривать. Создание соответствующих алгоритмов проводится специалистами по прикладной статистике в соответствии с непосредственными заказами пользователей.

3.3. МЕТОД ПРОВЕРКИ ГИПОТЕЗ ПО СОВОКУПНОСТИ МАЛЫХ ВЫБОРОК

Одна из областей применения прикладной статистики — статистические методы управления качеством продукции [5, гл. 13]. К ним относится статистический приемочный контроль, в котором по результатам испытаний элементов выборки делается вывод о качестве партии продукции. В простейшем варианте проводится контроль по альтернативному признаку, при котором возможны лишь два результата контроля конкретной единицы продукции — «соответствует требованиям» или «не соответствует требованиям», короче — «да» или «нет».

Рассмотрим статистический приемочный контроль по двум альтернативным признакам одновременно. В терминах теории люсианов обсудим проблему проверки независимости двух альтернативных признаков. Ее приходится проводить по совокупности малых выборок, т.е. в так называемой асимптотике А. Н. Колмогорова, когда число неизвестных параметров распределения не является постоянным, а растет пропорционально объему данных.

Испытания по двум альтернативным признакам. При статистическом контроле качества продукции, в частности, при сертификации, чаще всего используют контроль по альтернативным признакам. При этом устанавливается, соответствует ли контролируемый параметр единицы продукции (изделия, детали) заданным в нормативно-технической документации требованиям или не соответствует. Если соответствует — единица продукции признается годной. Примем для определенности, что в этом случае результат контроля кодируется символом 0. Если же не соответствует — единица продукции признается дефектной, а результат контроля кодируется символом 1.

Таким образом, в рассматриваемой нами математической модели контроля альтернативный признак — это функция $X = X(w)$, определенная на множестве единиц продукции $W = \{w\}$ и принимающая два значения 0 и 1. Причем

$X(w) = 0$ означает, что единица продукции w является годной, а $X(w) = 1$ — что она является дефектной.

Методы статистического контроля, в частности, включенные в государственные стандарты и иную нормативно-техническую документацию (НТД), как правило, используют контроль по одному признаку. В НТД указывают правила выбора планов контроля и расчета различных их характеристик, приводят графики оперативных характеристик и т.п.

Однако на производстве контроль нередко проводится по нескольким альтернативным признакам. Возникает проблема выбора плана контроля и расчета его характеристик.

Рассмотрим сначала контроль по двум альтернативным признакам $X(w)$ и $Y(w)$. В вероятностной модели $X(w)$ и $Y(w)$ — случайные величины, принимающие два значения — 0 и 1. Пусть, пользуясь стандартной (для статистических методов управления качеством) терминологией:

$$p_1 = P(X(w) = 1) —$$

входной уровень дефектности для первого признака, а

$$p_2 = P(Y(w) = 1) —$$

для второго. Вероятности результатов контроля по двум признакам одновременно описываются четырьмя числами:

$$P(X(w) = 0,\ Y(w) = 0) = p_{00},\ P(X(w) = 1,\ Y(w) = 0) = p_{10},$$
$$P(X(w) = 0,\ Y(w) = 1) = p_{01},\ P(X(w) = 1,\ Y(w) = 1) = p_{11}.$$

При этом справедливы соотношения:

$$p_{00} + p_{10} + p_{01} + p_{11} = 1,\ \ p_{10} + p_{11} = p_1,\ p_{01} + p_{11} = p_2.$$

С прикладной точки зрения наиболее интересна вероятность p_{00} того, что единица продукции является годной (по всем параметрам), и вероятность ее дефектности $(1 - p_{00})$, т.е. входной уровень дефектности для изделия в целом.

В табл. 1 сведены вместе введенные выше вероятности.

Таблица 1

Вероятности результатов испытаний при контроле по двум альтернативным признакам

Значения признаков	$X = 0$	$X = 1$	**Всего**
$Y = 0$	p_{00}	p_{10}	$1 - p_2$
$Y = 1$	p_{01}	p_{11}	p_2
Всего	$1 - p_1$	p_1	1

Есть три важных частных случая — поглощения, несовместности и независимости дефектов. Другими словами, поглощения, несовместности и независимости событий $\{w: X(w) = 1\}$ и $\{w: Y(w) = 1\}$. В случае поглощения одно из этих событий содержит другое, а потому:

$$p_{00} = 1 - \max(p_1 , p_2).$$

В случае несовместности:

$$p_{00} = 1 - p_1 - p_2.$$

В случае независимости:

$$p_{00} = (1 - p_1)(1 - p_2) = 1 - p_1 - p_2 + p_1 p_2.$$

Очевидно, что вероятность годности изделия всегда заключена между значениями, соответствующими случаям поглощения и несовместности. Кроме того, известно, что при большом числе признаков и малой вероятности дефектности по каждому из них случаи поглощения и независимости дают (в асимптотике) крайние значения для вероятности годности изделия, т.е. формулы, соответствующие независимости и несовместности, асимптотически совпадают. Причина этого явления состоит в том, что при малости p_1 и p_2 их произведение $p_1 p_2$ является бесконечно малой более высокого порядка по сравнения с p_1 и p_2.

Рассмотрим несколько примеров. Пусть некоторая продукция, скажем, гвозди, контролируются по двум альтернативным признакам, для определен-

ности, по весу и длине. Пусть результаты контроля 1 000 единиц продукции представлены в табл. 2.

Таблица 2

Результаты 1000 испытаний по двум альтернативным признакам (случай поглощения)

Значения признаков	$X = 0$	$X = 1$	**Всего**
$У = 0$	952	0	952
$У = 1$	0	48	48
Всего	952	48	1 000

Судя по данным табл. 2, дефекты всегда встречаются парами — если есть один, то есть и другой. Входной уровень дефектности как по каждому показателю, так и по обоим вместе — один и тот же, а именно, 0,048. Получив по результатам статистического наблюдения данные типа приведенных в табл. 2, целесообразно перейти к контролю только одного показателя, а не двух. Каково именно? Видимо, того, контроль которого дешевле.

Совсем иная ситуация в случае несовместности дефектов (табл. 3).

Таблица 3

Результаты 1000 испытаний по двум альтернативным признакам (случай несовместности)

Значения признаков	$X = 0$	$X = 1$	**Всего**
$У = 0$	904	48	952
$У = 1$	48	0	48
Всего	952	48	1 000

Судя по данным табл. 3, дефекты всегда встречаются поодиночке — если есть один, то другого нет. В результате входной уровень дефектности по каждому признаку по-прежнему равен 0,048, в то время как доля дефектных изделий (т.е. имеющих хотя бы один дефект) вдвое выше, т.е. входной уровень дефектности для изделия в целом равен 0,096.

Случай независимости результатов контроля по двум независимым признакам (табл. 4) лежит между крайними случаями поглощения и несовместности. Независимость альтернативных признаков обосновывается путем статистической проверки с помощью описанного ниже критерия $n^{1/2}V$.

Таблица 4

Результаты 1 000 испытаний по двум альтернативным признакам (случай независимости)

Значения признаков	$X = 0$	$X = 1$	Всего
$У = 0$	909	43	952
$У = 1$	43	5	48
Всего	952	48	1 000

Согласно данным табл. 4, входной уровень дефектности для каждого из двух альтернативных признаков по-прежнему равен 0,048, в то время как для изделий в целом он равен 0,091, т.е. на 5,2 % меньше, чем в случае несовместности, и на 89,6 % больше, чем в случае поглощения.

Проблема состоит в том, что таблицы и стандарты по статистическому приемочному контролю относятся обычно к случаю одного контролируемого параметра. А как быть, если контролируемых параметров несколько? Приведенные выше примеры показывают, что входной уровень дефектности изделия в целом не определяется однозначно по входным уровням дефектности отдельных его параметров.

Гипотеза независимости. Как должны соотноситься характеристики планов контроля по отдельным признакам с характеристиками плана контроля по двум (или многим) признакам одновременно? Рассмотрим распространенную рекомендацию — складывать уровни дефектности, т.е. считать, что уровень дефектности изделия в целом равен сумме уровней дефектности по отдельным его параметрам. Она, очевидно, опирается на гипотезу несовместности дефектов, а потому во многих случаях преувеличивает дефектность, следовательно, ведет к использованию излишне жестких планов контроля, что экономически невыгодно.

Зная специфику применяемых технологических процессов, в ряде конкретных случаев можно предположить, что дефекты по различным признакам возникают независимо друг от друга. Это предположение необходимо обосновывать по статистическим данным. Если же оно обосновано, следует рассчитывать входной уровень дефектности по формуле:

$$1 - p_{00} = p_1 + p_2 - p_1 p_2,$$

соотвтствующей независимости признаков.

Итак, необходимо уметь проверять по статистическим данным гипотезу независимости двух альтернативных признаков. Речь идет о статистической проверке нулевой гипотезы:

$$H_0: p_{11} = p_1 p_2 \qquad (1)$$

(что эквивалентно проверке равенства $p_{00} = (1 - p_1)(1 - p_2)$). Нетрудно проверить, что гипотеза о справедливости равенства (1) эквивалентна гипотезе:

$$H_0 : p_{00} p_{11} - p_{10} p_{01} = 0. \qquad (2)$$

В простейшем случае предполагается, что проведено n независимых испытаний (X_i, Y_i), $i = 1, 2, ..., n$, в каждом из которых проконтролированы два альтернативных признака, а вероятности результатов контроля не меняются от испытания к испытанию. Общий вид статистических данных приведен в табл. 5.

Таблица 5

Общий вид результатов контроля по двум альтернативным признакам

Значения признаков	$X = 0$	$X = 1$	**Всего**
$У = 0$	a	b	a + b
$У = 1$	c	d	c + d
Всего	a + c	b + d	n

В табл. 5 величина a — число испытаний, в которых $(X_i, Y_i) = (0, 0)$, величина b — число испытаний, в которых $(X_i, Y_i) = (1, 0)$, и т.д.

Случайный вектор (a, b, c, d) имеет мультиномиальное распределение с числом испытаний n и вектором вероятностей исходов $(p_{00}, p_{10}, p_{01}, p_{11})$. Состоятельными оценками этих вероятностей являются дроби a/n, b/n, c/n, d/n соответственно. Следовательно, критерий проверки гипотезы (2) может быть основан на статистике:

$$Z = ad — bc. \qquad (3)$$

Как вытекает из известной формулы для ковариаций мультиномиального вектора (см., например, формулу (6.3.5) в учебнике С.Уилкса [11] на с. 153):

$$M(Z) = n\,(p_{10}\,p_{01}\ - p_{00}\,p_{11}), \qquad (4)$$

что равно 0 при справедливости гипотезы независимости (2).

Связь между переменными X и Y обычно измеряется коэффициентом, отличающимся от Z нормирующим множителем:

$$V = (ad - bc)\{(a + b)(a + c)(b + d)(c + d)\}^{-1/2}$$

(см. классическую монографию М. Дж. Кендалла и А. Стьюарта [12, с. 723]).

При справедливости гипотезы H_0 и больших n случайная величина nV^2 имеет хи-квадрат распределение с одной степенью свободы, а $n^{1/2}V$ имеет стандартное нормальное распределение с математическим ожиданием 0 и дисперсией 1 (см. [12, с. 736]). Значение $n^{1/2}V$ для данных табл. 4 равно 1,866, т.е. на уровне значимости 0,05 гипотезу независимости следует принять.

Рассмотрим еще один пример. Пусть проведено 100 испытаний, результаты которых описаны в табл. 6. Тогда:

$$V = (50 \times 20 - 10 \times 20)\,(60 \times 70 \times 30 \times 40)^{-1/2} =$$
$$= (1000 - 200) \times 5940000^{-1/2} = 800 / 2245 = 0{,}35635,$$
$$n^{1/2}V = 3{,}5635\ .$$

Таблица 6

Результаты 100 испытаний
по двум альтернативным признакам

Значения признаков	$X = 0$	$X = 1$	**Всего**
$Y = 0$	50	10	60
$Y = 1$	20	20	40
Всего	70	30	100

Поскольку полученное значение $n^{1/2}V$ превышает критическое значение при любом применяемом в статистике уровне значимости, то гипотезу о независимости признаков необходимо отклонить.

Проверка гипотез по совокупности малых выборок. К сожалению, приведенный простой метод годится не всегда. При статистическом анализе реальных данных возникают проблемы, связанные с отсутствием достаточно больших однородных выборок, т.е. выборок, в которых постоянны параметры вероятностных распределений. Реально единицы продукции представляются на контроль партиями, из каждой партии контролируются лишь несколько изделий, т.е. малая выборка. При этом от партии к партии меняются параметры p_{00}, p_{10}, p_{01}, p_{11}, описывающие уровень дефектности. Поэтому необходимы статистические методы, позволяющие проверять гипотезу независимости признаков по совокупности малых выборок. Построим один из возможных методов.

Рассмотрим вероятностную модель совокупности k малых выборок объемов n_1 , n_2 , ..., n_k соответственно. Пусть j выборка (X_{jt}, Y_{jt}), $t = 1, 2, ..., n_j$, имеет распределение, задаваемое вектором параметров $(p_{00j}, p_{10j}, p_{01j}, p_{11j})$ в соответствии с ранее введенными обозначениями, $j = 1, 2, ..., k$. Будем проверять гипотезу:

$$H_0: \; p_{11j} = (p_{10j} + p_{11j})\,(p_{01j} + p_{11j}), j = 1, 2, ..., k, \tag{5}$$

или в эквивалентной формулировке:

$$H_0: \; p_{11j}\,p_{00j} - p_{10j}\,p_{01j}\,, j = 1, 2, ..., k\,. \tag{6}$$

Основная идея состоит в нахождении асимптотического распределения статистики типа $n^{1/2}V$ при росте числа k малых выборок. А именно, будем использовать статистику:

$$S = g_1\,Z_1 + \; g_2\,Z_2 + ... + \; g_k\,Z_k\,, \tag{7}$$

где Z_1, Z_2, ..., Z_k — статистики, рассчитанные по формуле (3) для каждой из k выборок, т.е. $Z_j = a_j d_j - b_j c_j$, $j = 1, 2, ..., k$, а g_1 , g_2 , ..., g_k — некоторые весовые коэффициенты, которые, в частности, могут совпадать. Поскольку:

$$M(S) = g_1\,M(Z_1) + \; g_2\,M(Z_2) + ... + \; g_k\,M(Z_k),$$

то при справедливости гипотезы независимости (5)–(6) имеем $M(S) = 0$, поскольку:

$$M(Z_j) = 0, j = 1, 2, ..., k,$$

при всех возможных значениях вектора параметров (p_{00j}, p_{10j}, p_{01j}, p_{11j}) согласно соотношению (4). Поскольку слагаемые в сумме (7) независимы, то при росте k случайная величина S в силу Центральной Предельной Теоремы является асимптотически нормальной. Дисперсия этой величины равна сумме дисперсий слагаемых:

$$D(S) = g_1^2 D(Z_1) + g_2^2 D(Z_2) + ... + g_k^2 D(Z_k). \quad (8)$$

Для оценивания дисперсии S необходимо использовать **несмещенные оценки** дисперсий в каждой из k выборок (и в этом одна из основных «изюминок» разбираемого метода). Предположим, что построены статистики T_j такие, что:

$$M(T_j) = D(Z_j), j = 1, 2, ..., k. \quad (9)$$

Тогда при некоторых математических «условиях регулярности», на которых нет необходимости здесь останавливаться, несмещенная оценка дисперсии статистики S, имеющая согласно формулам (8) и (9) вид:

$$L = g_1^2 T_1 + g_2^2 T_2 + ... + g_k^2 T_k,$$

в силу закона больших чисел такова, что дробь $D(S)/L$ приближается к 1 при росте числа выборок (сходимость по вероятности). Отсюда следует, что распределение случайной величины $Q = SL^{-1/2}$ приближается при росте числа выборок к стандартному нормальному распределению с математическим ожиданием 0 и дисперсией 1. Следовательно, критерий проверки гипотезы (5)–(6) независимости признаков, состоящий в том, что при $(-1{,}96) < Q < 1{,}96$ гипотеза принимается, а при Q , выходящих за пределы интервала (–1,96; 1,96), гипотеза отклоняется, имеет уровень значимости, приближающийся к 0,05 при росте числа выборок. Мощность этого критерия зависит от значения величины $M(S)D(S)^{-1/2}$ при альтернативной гипотезе.

Для реализации намеченного плана осталось научиться несмещенно оценивать $D(Z_j)$. К сожалению, в литературе по несмещенному оцениванию не рассматривают случай мультиномиального распределения, поэтому кратко опишем процедуру построения несмещенной оценки $D(Z_j)$. Поскольку согласно формулам (3) и (4):

$$\begin{aligned} D(Z_j) &= M(Z_j^2) - (M(Z_j))^2 = M(a_j^2 d_j^2) - 2M(a_j b_j c_j d_j) + \\ &+ M(b_j^2 c_j^2) + n_j^2 (p_{00j} p_{11j} - p_{01j} p_{10j})^2, \end{aligned} \quad (10)$$

то для вычисления $D(Z_j)$ достаточно найти входящие в правую часть формулы (10) начальные смешанные моменты мультиномиального распределения (четвертого порядка). Теоретически это просто — известен вид характеристической функции мультиномиального распределения (см., например, формулу (6.3.4) в монографии [11, с. 152]), а начальные смешанные моменты равны значениям ее соответствующих производных в 0, деленным на нужную степень мнимой единицы (формула (5.2.3) в монографии [11, с. 131]). Например, с помощью описанной процедуры после некоторых вычислений получаем, что (для упрощения записи здесь и далее опустим индекс j):

$$M(a^2d^2) = n(n-1)(n-2)(n-3)p_{11}^2p_{00}^2 + \\ + n(n-1)(n-2)(p_{11}^2p_{00} + p_{11}p_{00}^2) + n(n-1)p_{11}p_{00}. \qquad (11)$$

Формула (11) показывает, что начальные смешанные моменты мультиномиального распределения являются многочленами от параметров p_{11}, p_{00}, p_{10}, p_{01} этого распределения, однако конкретный вид этих многочленов достаточно громоздок, поэтому не будем их здесь выписывать, ограничившись формулой (11) в качестве образца.

Как вытекает из формул (10) и (11), для построения несмещенной оценки $D(Z_j)$ достаточно научиться несмещенно оценивать произведения типа $p_{11}^r p_{00}^m$, где целые неотрицательные числа r, m не превосходят 2. Эта задача решается, начиная с меньших степеней r и m. Известно, что для ковариации мультиномиального вектора:

$$M(ad) = -n\,p_{00}\,p_{11} \qquad (12)$$

(см., например, формулу (6.3.5) в монографии [11, с. 153]), а потому несмещенной оценкой для $p_{00}p_{11}$ является $(-ad/n)$. Далее, поскольку справедлива аналогичная (11) формула:

$$M(a^2d) = n(n-1)(n-2)p_{11}p_{00}^2 + n(n-1)p_{11}p_{00}, \qquad (13)$$

то с помощью формулы (12) преобразуем формулу (13) к виду:

$$M(a^2d + (n-1)ad) = n(n-1)(n-2)p_{11}p_{00}^2, \qquad (14)$$

т.е. несмещенной оценкой $p_{11}p_{00}^2$ является $ad(a+n-1)\{n(n-1)(n-2)\}^{-1}$.

Следующий шаг — аналогичным образом с помощью формул (12) и (14) получаем несмещенную оценку для $p_{11}^2 p_{00}^2$, а затем и для $D(Z_j)$. Промежуточные формулы опущены из-за громоздкости. Окончательный результат таков:

$$T_j = (b_j + d_j)(c_j + d_j)(a_j + c_j)(a_j + b_j)(n-1)^{-1}.$$

Как легко видеть,

$$\frac{Z_j}{\sqrt{T_j}} = V_j \sqrt{n_j - 1},$$

т.е. в случае одной выборки предлагаемый метод проверки независимости совпадает с классическим.

Таким образом, общая идея рассматриваемого метода проверки гипотез по совокупности малых выборок состоит в том, что подбирается статистика, математическое ожидание которой для каждой малой выборки равно 0 при справедливости проверяемой гипотезы. Затем для каждой выборки строится несмещенная оценка дисперсии этой статистики. Итоговая статистика критерия для проверки гипотезы — это сумма рассматриваемых статистик для всех малых выборок, деленная на квадратный корень из суммы всех несмещенных оценок дисперсий рассматриваемых статистик. При справедливости нулевой гипотезы эта итоговая статистика имеет в асимптотике стандартное нормальное распределение (при выполнении некоторых математических «условий регулярности», которые обычно выполняются при анализе реальных статистических данных).

Впервые такой способ проверки гипотез по совокупности малых выборок был предложен в монографии [1, раздел 4.5]. Нестандартность постановки состоит в том, что число неизвестных параметров растет пропорционально объему данных, т.е. имеет место так называемая «асимптотика Колмогорова», или асимптотика растущей размерности. Дальнейшее развитие применительно к данных типа «да» — «нет» (или «годен» — «дефектен») шло в рамках теории люсианов как части статистики объектов нечисловой природы (см. следующий раздел 3.4).

3.4. ТЕОРИЯ ЛЮСИАНОВ

Асимптотика растущей размерности и проверяемые гипотезы. Продолжим изучение модели порождения данных (6)–(7) раздела 3.2. Будем

использовать асимптотику $s = \text{const}$, $k \to \infty$. При этом число неизвестных параметров растет пропорционально объему данных.

В последние десятилетия (с начала 1970-х гг.) в прикладной статистике все большее распространение получают постановки, в которых число неизвестных параметров растет вместе с объемом выборки. Результаты, полученные в подобных постановках, называют найденными «в асимптотике растущей размерности» или «в асимптотике А. Н. Колмогорова» [13], перенося терминологию исследований по дискриминантному анализу на общий случай. Как известно, в задаче дискриминации в две совокупности (т.е. отнесения вновь появляющегося объекта к одному из двух классов) академик АН СССР А. Н. Колмогоров (1903–1987 гг.) предложил рассматривать асимптотику:

$$A \to \infty, \ N_i \to \infty, \ \frac{A}{N_i} \to \lambda_i > 0, \ i = 1,2 ,$$

где A — размерность пространства (число признаков), N_i — объемы обучающих выборок, λ_i — константы, $i = 1,2$. Эта асимптотика естественна при обработке многих видов технических, организационно-экономических, социологических, медицинских данных, поскольку число признаков, определяемых для каждого изучаемого объекта, респондента или пациента, обычно имеет тот же порядок, что и объем выборки.

Пусть $A_1, A_2, ..., A_s$ — независимые (между собой) люсианы с векторами параметров $P_1, P_2, ..., P_s$ соответственно. *Гипотезой согласованности* будем называть гипотезу:

$$P_1 = P_2 = ... = P_s. \qquad (1)$$

Для ранжировок и разбиений под согласованностью понимают более частную гипотезу, предполагающую отрицание равномерности распределений (т.е. одинаковой вероятности появления каждой возможной ранжировки или разбиения), что соответствует замене проверки гипотезы (1) на проверку гипотезы:

$$P_1 = P_2 = ... = P_s = (1/2, 1/2, ..., 1/2). \qquad (2)$$

Как разъяснено в [1, 14], гипотеза (1) более адекватна конкретным задачам обработки реальных данных, например, экспертных оценок, чем (2).

Поэтому полученные от экспертов данные, содержащие противоречия, целесообразно рассматривать как люсианы и проверять гипотезу (1), а не подбирать ближайшие ранжировки или разбиения , после чего проверять согласованность методами теории случайных ранжировок или разбиений, как иногда рекомендуется.

Пусть $A_1, A_2, ..., A_m$ и $B_1, B_2, ..., B_n$ — независимые в совокупности люсианы длины k, одинаково распределенные в каждой группе с параметрами $P(A)$ и $P(B)$ соответственно. *Гипотезой однородности* называется гипотеза:

$$P(A) = P(B).$$

В асимптотике растущей размерности принимаем, что m и n постоянны, а $k \to \infty$.

Пусть (A_i, B_i), $i = 1. 2, ..., s$ — последовательность (фиксированной длины) пар люсианов. Пары предполагаются независимыми между собой. Требуется проверить гипотезу независимости A_i и B_i, т.е. внутри пар. В ранее введенных обозначениях *гипотеза независимости* — это гипотеза:

$$P(X_{ij}(A) = 1, X_{ij}(B) = 1) = P(X_{ij}(A) = 1)P(X_{ij}(B) = 1),$$
$$i = 1, 2, ..., s; j = 1, 2, ..., k,$$

проверяемая в предположении:

$$P_1(A) = P_2(A) = ... = P_s(A), P_1(B) = P_2(B) = ... = P_s(B).$$

В настоящем разделе излагается метод проверки гипотез о люсианах в асимптотике растущей размерности на примере гипотезы согласованности. Эти результаты получены в работах [1, 14, 15]. Дальнейшее изучение проведено Г. В. Рыдановой, Т. Н. Дылько, Г. В. Раушенбахом, О. В. Филипповым, А. М. Никифоровым и др. Гипотеза однородности рассмотрена, например, в [15]. Методы проверки гипотезы независимости люсианов развиты и изучены Г. В. Рыдановой [16] на основе описанного ниже подхода. Она помимо доказательства предельных теорем провела подробное изучение скорости сходимости методом статистических испытаний.

Методы проверки согласованности люсианов нашли практическое применение, в частности, в медицине. Они были использованы в кардиологии при анализе данных кинетотопографии [15, 17, 18]. Эти методы включе-

ны в методические рекомендации Академии медицинских наук СССР и Ученого Медицинского Совета Минздрава СССР по управлению научными медицинскими исследованиями [19].

Метод проверки гипотез о люсианах в асимптотике растущей размерности. Будем использовать дальнейшее развитие метода, описанного в предыдущем разделе 3.3. Почему нельзя использовать иные подходы, имеющиеся в математической статистике? Поскольку число неизвестных параметров растет вместе с объемом выборки и пропорционально ему, эти параметры не являются мешающими (в том смысле, как этот термин понимается в теории математической статистики). Отметим, что согласно [20] равномерно наиболее мощных критериев не существует, поскольку параметров много. Не останавливаясь на других подходах математической статистики, констатируем необходимость применения метода проверки гипотез по совокупности малых выборок.

Пусть имеются k выборок, независимых между собой. Пусть при справедливости нулевой гипотезы по каждой из выборок можно построить несмещенную оценку $\xi_i \in R^p$ векторного нуля $0 \in R^p$, где $p \le 1$, $i = 1, 2, ..., k$. Другими словами, пусть распределение i-й выборки описывается параметром θ_i, лежащим в произвольном пространстве, а нулевая гипотеза, очевидно, состоит в том, что $\theta_i \in \Theta_{0i}$, где Θ_{0i} — собственное подмножество множества $\{\theta_i\}$. Предполагается, что можно по i-й выборке вычислить статистику ξ_i такую, что:

$$M\xi_i = 0 \tag{3}$$

при всех $\theta_i \in \Theta_{0i}$. Очевидно, $\xi_i \equiv 0$ удовлетворяют (1). Однако для рассматриваемого метода необходимо, чтобы при всех $\theta_i \in \Theta_{0i}$ ковариационная матрица вектора ξ_i была ненулевой:

$$Cov(\xi_i) = M(\xi_i^T \xi_i) \ne 0\,. \tag{4}$$

В теории математической статистики иногда используют понятие полноты параметрического семейства распределений. Если рассматриваемое семейство является полным — а так и есть для люсианов, — то не существует достаточной статистики, удовлетворяющей одновременно условиям (1) и (2) (см., например, [21, § 2.12–2.14]). Поэтому будем использовать статистики, не являющиеся достаточными.

Следующее предположение — ковариационные матрицы статистик ξ_i, т.е. $Cov(\xi_i)$, также допускают несмещенные оценки S_i по тем же выборкам:

$$M(S_i) = Cov(\xi_i) \tag{5}$$

при всех $\theta_i \in \Theta_{0i}$.

Рассматриваемый метод основан на том, что поскольку случайные вектора ξ_i определяются по независимым между собой выборкам, то ξ_i независимы в совокупности, а потому случайный вектор:

$$\xi = \sum_{i=1}^{k} \xi_i \tag{6}$$

является суммой независимых случайных векторов, имеет в силу (3) нулевое математическое ожидание, а его ковариационная матрица равна:

$$C_k = \sum_{i=1}^{k} Cov(\xi_i).$$

При справедливости многомерной центральной предельной теоремы (простейшее условие справедливости этой теоремы для ξ_i в случае люсианов — отделенность от 0 и 1 всех элементов матриц P_j, равномерная по s и k) вектор ξ является асимптотически нормальным, т.е. при $k \to \infty$ распределение ξ сближается (в смысле, раскрытом в приложении 1) с многомерным нормальным распределением $N(0; C_k)$.

Однако эту сходимость нельзя непосредственно использовать для проверки исходной гипотезы, поскольку матрица C_k неизвестна статистику. Необходимо оценить эту матрицу по статистическим данным. В силу (5) в качестве оценки C_k естественно использовать:

$$C_k^* = \sum_{i=1}^{k} S_i .$$

Простейшая формулировка условий справедливости такой замены — предположение о том, что к последовательности S_i можно применить закон больших чисел. А именно, пусть существует неотрицательно определенная матрица C такая, что при $k \to \infty$:

$$\frac{1}{k}(C_k^* - C_k) \to 0, \quad \frac{1}{k} C_k \to C . \tag{7}$$

В силу результатов приложения 1 из асимптотической нормальности ξ и соотношений (7) следует, что распределение статистики:

$$\eta = \frac{1}{\sqrt{k}}\xi$$

сходится к нормальному распределению $N(0; C)$. При этом, если некоторый случайный вектор τ имеет распределение $N(0; C)$, то распределение случайной величины $q(\eta)$ сходится к распределению $q(\tau)$ для произвольной интегрируемой по Риману по любому кубу функции q: $R^p \to R^1$. Для проверки нулевой гипотезы предлагается пользоваться статистикой $q(\eta)$ при подходящей функции q, а процентные точки брать соответственно распределению $q(\tau)$. В этом и состоит рассматриваемый метод проверки гипотез о люсианах в асимптотике растущей размерности. Для реальных расчетов целесообразно использовать линейные или квадратические функции q от координат вектора η.

Отклонения от нулевой гипотезы приводят, как правило, к нарушению равенств (3) и (4). Случайный вектор η при этом обычно остается асимптотически нормальным, но с другими параметрами, что может быть обычным образом использовано для построения оптимального решающего правила, соответствующего заданной альтернативе (например, согласно лемме Неймана — Пирсона). Поведение при альтернативах для некоторых гипотез изучено в [15, 16], здесь его не будем рассматривать, поскольку вычисление мощности не требует новых идей.

Несмещенные оценки параметров асимптотического распределения вектора попарных расстояний. Применим описанный выше метод для проверки гипотезы согласованности люсианов. Исходные данные — люсианы:

$$A_j = (X_{1j}, X_{2j}, \ldots, X_{kj}), j = 1, 2, \ldots, s.$$

В качестве i-й выборки возьмем совокупность испытаний Бернулли, стоящих на i-м месте в рассматриваемых люсианах:

$$X_{i1}, X_{i2}, \ldots, X_{is}. \qquad (8)$$

При справедливости нулевой гипотезы в (8) стоят независимые испытания Бернулли с одной и той же вероятностью успеха p_i; при нарушении ну-

левой гипотезы согласованности независимость испытаний Бернулли сохраняется, но вероятности успеха могут различаться.

В качестве вектора ξ, на основе которого строятся статистики для проверки согласованности, будем использовать вектор попарных расстояний между люсианами:

$$\xi = \{d(A_p, A_q),\ 1 \le p < q \le s\}, \tag{9}$$

в котором пары (p, q) упорядочены лексикографически,

$$d(A_p, A_q) = \sum_{i=1}^{k} \mu_i \,|\, X_{ip} - X_{iq} \,|, \quad \mu_i > 0\,. \tag{10}$$

В главе 1 это расстояние выведено из некоторой системы аксиом (напомним, что совокупность векторов из 0 и 1 размерности k находится во взаимно-однозначном соответствии с совокупностью подмножеств множества из k элементов; при этом 1 соответствует тому, что элемент входит в подмножество, а 0 — что не входит).

Из вида расстояния в формуле (10) следует, что введенный в (9) вектор ξ имеет вид (6) с:

$$\xi_i = \mu_i\{|X_{ip} - X_{iq}|,\ 1 \le p < q \le s\}. \tag{11}$$

Следовательно, для применения описанного выше метода проверки гипотез о люсианах в асимптотике растущей размерности достаточно построить на основе вектора ξ_i из (11) несмещенную оценку 0 и найти несмещенную оценку ковариационной матрицы этой оценки.

Чтобы применить общую схему, необходимо начать с построения статистики β такой, чтобы при всех p_i имело место равенство:

$$M(\,|\,X_{ip} - X_{iq}\,| - \beta) = 0,\ 1 \le p < q \le s.$$

Элементарный расчет дает:

$$M|X_{ip} - X_{iq}| = 2p_i(1 - p_i).$$

Как известно [22, с. 56–57], несмещенная оценка многочлена:

$$f(p) = \sum_{h=0}^{m} a_h p^h$$

по результатам m независимых испытаний Бернулли с вероятностью успеха p в каждом имеет вид:

$$f^*(p) = \sum_{h=0}^{m} a_h \frac{\gamma^{[h]}}{m^{[h]}}, \qquad (12)$$

где γ — общее число успехов в m испытаниях и использовано обозначение:

$$n^{[h]} = n(n-1) \ldots (n-h+1).$$

Ясно, что многочлены степени $m + 1$ и более высокой невозможно несмещенно оценить по результатам m испытаний.

В случае $f(p) = 2p(1 - p)$ в соответствии с (12) получаем несмещенную оценку:

$$\beta = \frac{2}{m-1}\left(\gamma - \frac{\gamma^2}{m}\right). \qquad (13)$$

Таким образом, можно применять общий метод проверки гипотез о люсианах в асимптотике растущей размерности с:

$$\xi_i = \mu_i\,(\{|X_{ip} - X_{iq}|,\ 1 \le p < q \le s\} - \beta_i e),$$

где коэффициенты β_i определяются с помощью формулы (13) по γ_i — общему числу единиц, стоящих на i-м месте в люсианах A_1, A_2, ..., A_s, а e — вектор размерности $s(s - 1) / 2$ с единичными координатами. Тогда несмещенная оценка 0, о которой идет речь в методе проверки гипотез по совокупности малых выборок, имеет вид:

$$\xi = \{d(A_p, A_q),\ 1 \le p < q \le s\} - \sum_{i=1}^{k} \mu_i \beta_i e\,.$$

Для использования статистики типа η, распределение которой приближается с помощью нормального распределения:

$$N\left(0;\ \frac{1}{k}\sum_{i=1}^{k} S_i\right),$$

необходимо уметь несмещенно оценивать ковариационные матрицы $Cov(\xi_i)$. Для этого достаточно найти математические ожидания элементов матрицы $M(\xi_i^T \xi_i)$ как функции (многочлены) от p_i, а затем использовать формулу (12) для получения несмещенных оценок.

Вычисление матрицы $M(\xi_i^T \xi_i)$ хотя и трудоемко, но не содержит каких-либо принципиальных трудностей. В [15] вычислены диагональные элементы рассматриваемой матрицы. Вычисление занимает около 2,5 книжных страниц (с. 299–301). Поэтому здесь приведен только окончательный итог.

Обозначим для краткости $p_i = p$. В [15] показано, что:

$$D = D(|X_{ip} - X_{iq}| - \beta_i) = \left(2 - \frac{4}{s}\right) p(1-p) - 4\frac{(s-2)(s-3)}{s(s-1)} p^2 (1-p)^2 .$$

Если двухэлементные множества $\{h, q\}$ и $\{r, t\}$ не имеют ни одного общего элемента, то:

$$C_1 = M(|X_{ih} - X_{iq}| - \beta_i)(|X_{ir} - X_{it}| - \beta_i) = \\ = -\frac{4}{s} p(1-p) + \frac{8(2s-3)}{s(s-1)} p^2 (1-p)^2 ,$$

а если имеют ровно один общий элемент, то:

$$C_2 = M(|X_{ih} - X_{iq}| - \beta_i)(|X_{ir} - X_{it}| - \beta_i) = \\ = \left(1 - \frac{4}{s}\right) p(1-p) - 4\frac{(s-2)(s-3)}{s(s-1)} p^2 (1-p)^2 .$$

С помощью формулы (12) получаем несмещенные оценки для D, C_1 и C_2 как многочленов от p:

$$D^* = \frac{2\gamma_i(s-\gamma_i)}{s^2(s-1)^2}\{(s-2)(s-1) - 2(\gamma_i - 1)(s-\gamma_i - 1)\},$$

$$C_1^* = \frac{4\gamma_i(s-\gamma_i)}{s^2(s-1)}\left\{\frac{2(2s-3)(\gamma_i - 1)(s-\gamma_i - 1)}{(s-1)(s-2)(s-3)} - 1\right\},$$

$$C_2^* = \frac{\gamma_i(s-\gamma_i)}{s^2(s-1)^2}\{(s-4)(s-1) - 4(\gamma_i - 1)(s-\gamma_i - 1)\}.$$

С помощью трех чисел D^*, C_1^*, C_2^* выписывается несмещенная оценка матрицы ковариаций вектора ξ_i/μ_i, которую обозначим B_i. Тогда асимптотически нормальный вектор ξ имеет нулевое математическое ожидание и ковариационную матрицу, несмещенно и состоятельно (в смысле соотношений (7)) оцениваемую с помощью:

$$Cov(\xi)^* = \sum_{i=1}^{k} \mu_i^2 B_i \,. \tag{14}$$

Асимптотическая нормальность доказывается, естественно, в схеме серий. Достаточным условием является существование положительной константы ε такой, что:

$$\mu_i \geq \varepsilon,\ \frac{1}{\mu_i} \geq \varepsilon,\ \ p_i \geq \varepsilon,\ 1 - p_i \geq \varepsilon \ . \tag{15}$$

при всех k и i, $1 \leq i \leq k$.

Поскольку D, C_1 и C_2 являются многочленами четвертой степени от p, то несмещенные оценки для них существуют при $s \geq 4$. Если же $s < 4$, то несмещенных оценок не существует. Поэтому указанным методом проверять согласованность можно лишь при числе люсианов $s \geq 4$.

Проверка согласованности люсианов. Пусть α — нормально распределенный случайный вектор размерности $s(s-1)/2$ с нулевым математическим ожиданием и ковариационной матрицей, определенной формулой (14). Согласно результатам приложения 1 для любой действительнозначной функции f, интегрируемой по Риману по любому гиперкубу, распределения случайных величин $f(\xi)$ и $f(\alpha)$ сближаются при $k \to \infty$. Это означает, что вместо распределения случайной величины $f(\xi)$ для построения критериев проверки гипотез можно использовать распределение случайной величины $f(\alpha)$. Более того, аналогичный результат верен при замене f на f_n (при слабых внутриматематических условиях регулярности, наложенных на последовательность функций f_n). Следовательно, для проверки гипотезы согласованности люсианов можно пользоваться любой статистикой $f_n(\xi)$, для которой могут быть вычислены на компьютере или заранее табулированы процентные точки распределения $f_n(\alpha)$, аппроксимирующего распределение $f_n(\xi)$.

В частности, можно использовать линейные статистики, представляющие собой скалярное произведение случайного вектора ξ и некоторого заданного детерминированного вектора коэффициентов a, т.е.

$$(\xi, a) = \sum_{i=1}^{k} \left(\mu_i \sum_{1 \le j < t \le s} a_{jt} (| X_{ij} - X_{it} | - \beta_i) \right). \qquad (16)$$

Линейные статистики имеют нулевое математическое ожидание и дисперсию, очевидным образом выражающуюся через матрицу коэффициентов $\|a_{ij}\|$ и числа D, C_1 и C_2, а потому несмещенно и состоятельно оцениваемую с помощью с помощью выписанных выше оценок для D, C_1 и C_2.

Отметим, что $(\xi, a) = 0$ при $a_{ij} \equiv 1$, $1 \le j < t \le s$. Это следует как из непосредственного вычисления дисперсии (ξ, a), так и из того, что (ξ, a) в рассматриваемом случае выражается через достаточную статистику $(\gamma_1, \gamma_2, ..., \gamma_k)$ и является несмещенной оценкой нуля, а семейство биномиальных распределений полно, т.е. существует только одна несмещенная оценка нуля — тождественный нуль. Таким образом, сумма координат вектора ξ, т.е. непосредственный аналог коэффициента ранговой конкордации Кендалла-Смита из теории ранговой корреляции, тождественно равна 0.

Распределение статистики (16) при альтернативах изучено в работе [16].

Рассмотрим два частных случая.

Первый частный случай. Проверка согласованности двух определенных люсианов (ответов двух экспертов), j-го и t-го, может осуществляться с помощью статистики (16), в которой отличен от 0 только член с $a_{jt} = 1$. Оценкой дисперсии является D^*.

Второй частный случай. Пусть необходимо проверить согласованность люсианов с одним из них, скажем, с j-м (например, люсианы отражают мнения экспертов, а j-й из них является наиболее компетентным — по априорной оценке, или «лицом, принимающим решения», или его мнение сильно отличается от мнений остальных). Это можно сделать с помощью статистики (16), в которой:

$$a_{jt} = 1,\ t = j+1, j+2, ..., s; \quad a_{tj} = 1,\ t = 1, 2, ..., j-1;$$
$$a_{qt} = 0,\ q \ne j,\ t \ne j,\ 1 \le q < t \le s.$$

Другими словами, она имеет вид:

$$W = \sum_{t=1}^{s} d(A_j, A_t) - (s-1) \sum_{i=1}^{k} \mu_i \beta_i, \qquad (17)$$

где расстояние d между люсианами определено в (10), а β_i — в (13) с заменой m на s и γ на γ_i. Используя полученные ранее несмещенные оценки элементов ковариационной матрицы, нетрудно показать, что несмещенная и состоятельная (в смысле формулы (7) выше) оценка дисперсии W имеет вид:

$$D^*(W) = \sum_{i=1}^{k} \mu_i^2 \frac{\gamma_i (s-\gamma_i)}{s^2} \{(s-2)^2 - 4(\gamma_i - 1)(s-\gamma_i - 1)\}.$$

Тогда при выполнении некоторых внутриматематических условий регулярности, например, условий (15), распределение статистики:

$$\frac{1}{\sqrt{D^*(W)}} W$$

сходится при $k \to \infty$, $s = \text{const}$ к стандартному нормальному распределению с математическим ожиданием 0 и дисперсией 1 (при справедливости гипотезы (1) согласованности люсианов).

Статистика (17) наряду со статистикой, предназначенной для проверки гипотезы однородности люсианов, включена в «Методические рекомендации» АМН СССР и УМС Минздрава СССР [19]. Последнюю статистику не расписываем здесь, поскольку для этого не требуются новые идеи.

Различные подходы к понятию согласованности. Обсудим условия, при выполнении которых люсианы естественно считать согласованными (а экспертов, чьи мнения отражают люсианы, имеющими единое мнение, искаженное случайными ошибками), т.е. обсудим различные методы проверки гипотезы (1).

Полное индивидуальное согласие имеет место, если никакие два эксперта не являются «несогласованными». Уровень значимости определяется описанным выше способом (первый частный случай). Однако наличие одной или нескольких пар экспертов, чьи мнения нельзя считать согласованными, не свидетельствует о необходимости отклонения гипотезы (1), поскольку парных проверок проводится много, а именно, $s(s-1) \geq 6$, а способы установления уровня значимости при множественных проверках, зависимых между собой, к настоящему времени плохо разработаны [5, раздел 11.1]. Проблема множественных проверок для количественных признаков обсуждается А. А. Любищевым [23, с. 36–39], выход дается дисперсионным анализом. Можно брать не все попарные проверки, а только для $[s/2]$ пар люсианов, при-

чем разбиение на пары проводить независимо от принятых люсианами значений, как это делает Т. Н. Дылько [24]. Тогда для проверки гипотезы (1) на уровне значимости α надо брать для проверки в каждой паре уровень значимости β, где β рассчитывается понятным образом, приближенно β = α / [*s*/2].

Полное согласие в целом означает, что для любого эксперта мнения всех остальных оказываются с ним согласованными при использовании статистики (17) (второй частный случай). Отсутствие подобного согласия для одного или нескольких экспертов не означает отклонения гипотезы согласованности люсианов (1) — по тем же причинам, что и в предыдущем случае.

Минимальное согласие имеют мнения экспертов, когда хотя бы для одного из них гипотеза согласованности не отвергается с помощью статистики (17). В этом случае групповое мнение целесообразно строить, выделяя «ядро», о чем подробнее сказано ниже.

Расстояние *d* между люсианами (см. формулу (10)) введено аксиоматически в главе 1 (напомним, что реализацию люсиана можно рассматривать как подмножество конечного множества). Там же из иной системы аксиом выведено другое расстояние — *D*-метрика. Рассмотрим проверку согласованности люсианов с использованием *D*-метрики. В этом случае расстояние между люсианами A_1 и A_2 имеет вид:

$$D(A_1, A_2) = \begin{cases} \dfrac{d(A_1, A_2)}{T(A_1, A_2)}, & T(A_1, A_2) \neq 0, \\ 0, & T(A_1, A_2) = 0, \end{cases}$$

где

$$T(A_1, A_2) = \sum_{i=1}^{k} \mu_i \max(X_{i1}, X_{i2}).$$

Ясно, что теория, основанная на *D*-метрике, из-за наличия знаменателя в только что приведенной формуле существенно сложнее теории, основанной на метрике *d*. Ясно, что описанный выше метод проверки гипотез о люсианах в асимптотике растущей размерности применить не удается. Чтобы продемонстрировать существенное усложнение ситуации, опишем лишь асимптотическое поведение расстояния $D(A_1, A_2)$ между двумя люсианами.

Теорема [25]. Пусть p_{1i} и p_{2i} отделены от 0 и 1, а μ_i отделены от 0 и $+\infty$. Тогда расстояние $D(A_1, A_2)$ между люсианами A_1 и A_2 асимптотически нормально при $k \to \infty$ с параметрами:

$$t_k = \frac{N_1}{N_2}, \quad q_k = \frac{N_1}{N_2}\sqrt{\frac{N_3}{N_1^2} + \frac{N_4}{N_2^2} - 2\frac{N_5}{N_1 N_2}},$$

т.е. для любого числа x справедливо предельное соотношение:

$$\lim_{k\to\infty} P\left\{\frac{D(A_1, A_2) - t_k}{q_k} \le x\right\} = \Phi(x),$$

где $\Phi(x)$ — функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1.

Величины N_j, j = 1, 2, 2, 4, 5, выражаются через μ_i и величины:

$$p_{3i} = p_{1i} + p_{2i} - 2p_{1i}p_{2i}, \; p_{4i} = p_{1i} + p_{2i} - p_{1i}p_{2i}$$

следующим образом:

$$N_1 = \sum_{i=1}^{k} \mu_i p_{3i}, \quad N_2 = \sum_{i=1}^{k} \mu_i p_{4i}, \quad N_3 = \sum_{i=1}^{k} \mu_i^2 p_{3i}(1 - p_{3i}),$$

$$N_4 = \sum_{i=1}^{k} \mu_i^2 p_{4i}(1 - p_{4i}), \quad N_5 = \sum_{i=1}^{k} \mu_i^2 p_{3i}(1 - p_{4i}).$$

Следствие 1. Пусть $p_{1i} = p_1$ и $p_{2i} = p_2$ при всех i, k, причем p_1 и p_2 лежат внутри отрезка (0; 1). Пусть μ_i отделены от 0 и $+\infty$. Тогда расстояние $D(A_1, A_2)$ между люсианами A_1 и A_2 асимптотически нормально при $k \to \infty$ с параметрами:

$$t_k = \frac{p_3}{p_4}, \quad q_k^2 = \frac{p_1 p_2 p_3}{p_4^3} \frac{\sum_{i=1}^{k} \mu_i^2}{\left(\sum_{i=1}^{k} \mu_i\right)^2},$$

где

$$p_3 = p_1 + p_2 - 2p_1p_2, \quad p_4 = p_1 + p_2 - p_1p_2.$$

Следствие 2. Пусть в предположениях следствия 1 $p_1 = p_2 = p$ и $\mu_i = 1$ при всех i, k. Тогда:

$$t_k = \frac{2(1-p)}{2-p}, \quad q_k = \frac{2(1-p)}{k(2-p)^3}.$$

Замечание. Пусть вследствие 2 $p = 1/2$. Тогда A_1 и A_2 — люсианы, равномерно распределенные на множестве всех последовательностей из 0 и 1 длины k. В частности, эти люсианы могут соответствовать независимым случайным множествам, равномерно распределенным на совокупности всех подмножеств конечного множества из k элементов, или независимым толерантностям, равномерно распределенным на множестве всех толерантностей, определенных на множества из m элементов, где $m(m-1)/2 = k$. По следствию 2 расстояние между люсианами $D(A_1, A_2)$ асимптотически нормально с математическим ожиданием 0,667 и дисперсией $0{,}296\ k^{-1}$. Напомним, что распределения коэффициентов ранговой корреляции Кендалла и Спирмена изучены (в основном) лишь при условии равномерности распределения случайных ранжировок на множестве всех возможных ранжировок фиксированного числа объектов. Для теории люсианов случай равномерности распределения — весьма частный, а для теории ранжировок — основной. Как уже говорилось, отказ от равномерности — привлекательная черта теории люсианов.

Классификация люсианов. Отсутствие согласованности в одном из перечисленных выше смыслов позволяет сделать заключение о целесообразности разбиения всех люсианов (например, если они выражают мнения экспертов) на группы близких между собой, т.е. о целесообразности классификации люсианов, точнее, их кластер-анализа. Поскольку введена мера близости между люсианами $d(A_1, A_2)$ или $D(A_1, A_2)$, то напрашивается следующий способ действий: провести разбиение на кластеры с помощью одного из алгоритмов, основанных на использовании меры близости, а затем проверить мнения в каждом классе на согласованность. Однако применение того или иного алгоритма кластер-анализа, вообще говоря, может нарушить предпосылки описанных выше способов описанных выше способов проверки согласованности (ср. обсуждение похожей проблемы, связанной с применением регрессионного анализа после кластер-анализа, в [5, гл. 11]). Поэтому опишем методы классификации, опирающиеся на результаты проверки согласованности.

Разбиение на кластеры, внутри каждого из которых имеет место «полное индивидуальное согласие», может быть проведено с помощью агломеративного иерархического алгоритма «дальнего соседа», дополненного ограни-

чением сверху на диаметр кластера. Это ограничение строится из статистических соображений, в отличие от методов, обычно используемых в кластер-анализе [5, гл. 5]. При этом в качестве меры близости между люсианами используют не расстояния d или D, а модуль статистики, применяемой для проверки согласованности двух люсианов, т.е. статистики (16), в которой только одно из чисел a_{ij} отлично от 0. Упомянутое ограничение таково: диаметр кластера не должен превосходить процентной точки предельного распределения, соответствующей используемому при анализе рассматриваемых данных уровню значимости (можно порекомендовать 5 %-й уровень значимости). В результате работы алгоритма получим кластеры, в которых имеется «полное индивидуальное согласие», причем объединение любых двух кластеров приведет к исчезновению этого свойства у объединения. Поскольку способ выделения итогового разбиения из иерархического дерева разбиений имеет вероятностно-статистическое обоснование, изложенное выше, то описанный метод классификации люсианов следует считать — в терминологии [26] — не методом анализа данных, а вероятностно-статистическим методом.

Кластеры «с полным согласием в целом» могут быть получены с помощью агломеративного иерархического алгоритма, в котором мерой близости двух кластеров является максимальное значение модуля статистики (17), когда j пробегает номера мнений (люсианов), вошедших в объединение рассматриваемых кластеров, а суммирование в (17) проводится по всем люсианам в этом объединении. Ограничение сверху на меру близости кластеров определяется процентной точкой предельного распределения статистики W, заданной формулой (17).

Кластеры «с минимальным согласием» можно получить, при фиксированном j выделяя совокупность люсианов, согласованных с A_j в смысле статистики W из (17).

На основе двух рассмотренных выше частных случаев линейной статистики (16) можно строить и другие способы классификации. Например, для каждого люсиана A_m можно выделить кластер «типа шара» (см. [5, гл. 5]) из люсианов, попарно согласованных с A_m. Все такие способы имеют вероятностно-статистическое обоснование, и потому к ним относится сказанное выше относительно выделения кластеров «с полным индивидуальным согласием».

Замечание. Проверка согласованности приведенными выше критериями может привести к отрицательному результату двумя способами — либо значение статистики окажется слишком большим, либо слишком малым.

Первое означает, что гипотеза согласованности люсианов (1) неверна, вторая — что неверна вероятностная модель реального явления или процесса, основанная на люсианах. С необходимостью учета второй возможности мы столкнулись при применении теории люсианов для анализа данных топокарт, полученных при проведении кинетокардиографии у больных инфарктом миокарда [17, 18].

Нахождение среднего. В результате классификации получаем согласованные (в одном из указанных выше смыслов) группы люсианов. Для каждой из них полезно рассмотреть среднее. В зависимости от конкретных приложений в прикладных исследованиях применяют либо среднее в виде последовательностей 0 и 1, т.е. в виде реализации люсиана, либо среднее в виде последовательности оценок вероятностей (p_1, p_2, ..., p_k). Кроме того, оно может находиться либо с помощью методов, подавляющих «засорения» («выбросы»), либо без учета возможности засорения. Рассмотрим все четыре возможности.

В соответствии с подходом главы 2 при отсутствии засорения эмпирическое среднее ищется как решение задачи:

$$\sum_{j=1}^{m} d(A_j, A) \to \min_{A \in X}, \qquad (18)$$

где $A_1, A_2, ..., A_m$ — люсианы, входящие в рассматриваемый кластер, X — множество, которому принадлежит среднее. Если X — совокупность последовательностей из 0 и 1, то правило (18) дает решение по правилу большинства.

Если X — пространство последовательностей вероятностей, то решением задачи (18) является та же последовательность 0 и 1, что и в первом случае. Поэтому в качестве среднего вместо решения задачи (18) целесообразно рассматривать просто последовательность частот.

Асимптотическое поведение средних при $m \to \infty$ вытекает из законов больших чисел, теорем, описывающих асимптотику решений экстремальных статистических задач (глава 2), и теоремы Муавра — Лапласа соответственно.

В работе [27] при анализе результатов эксперимента показано, что ответы реальных экспертов разбиваются на многочисленное «ядро», расположенное вокруг истинного мнения, и отдельных «диссидентов», разбросанных по периферии. Причем оценка истинного мнения по «ядру» является более точной, чем по всей совокупности, поскольку мнения «диссидентов» не отражают истинного мнения. Поэтому для построения группового мнения, в том числе среднего для совокупности люсианов, отражающих мнения экс-

пертов, естественно применять методы, подавляющие мнения «диссидентов», что соответствует методологии робастности.

«Ядро» может быть построено следующим образом. Решается задача (18) с конечным множеством X, состоящим из всех исходных люсианов: $X = \{A_1, A_2, ..., A_m\}$, т.е. из результатов наблюдений выбирается тот, что находится «в центре» совокупности результатов наблюдений. Пусть A_j является решением этой задачи. В качестве ядра предлагается рассматривать совокупность всех люсианов, которые попарно согласованы с A_j. Другой вариант: рассматривается кластер с «полным внутренним согласием», куда входит A_j. (При этом, очевидно, должно быть изменено (уменьшено) критическое значение критерия по сравнению с процедурой, приведшей к выделению группы, нахождением группового мнения которой мы занимаемся.) Затем групповое мнение ищется лишь для элементов «ядра». Описанная процедура особенно необходима в случае, когда не было предварительного разбиения совокупности люсианов на группы согласованных друг с другом. Новым по сравнению с [27] является придание вероятностного смысла порогу, выделяющему «ядро».

Обобщая идею выделения «ядра», приходим к «взвешенным итеративным методам оценивания среднего» (ВИМОС — оценкам среднего), введенным и изученным в работе [28]. Их применение для люсианов не требует специальных рассмотрений.

Таким образом, в настоящем разделе представлен ряд методов обработки специального вида объектов нечисловой природы — люсианов. При этом для решения одной и той же задачи, например, задачи классификации, предлагается ряд методов, точно так же, как для решения классической задачи проверки однородности двух независимых выборок имеется большое число методов [5, гл. 4].

3.5. МЕТОД ПАРНЫХ СРАВНЕНИЙ

Пример практического применения метода парных сравнений. Деятельность предприятия по реализации товаров и услуг всегда сопряжена с рядом проблем, от качества решения которых зависит его будущее. Руководителю службы маркетинга необходимо знать факторы, сдерживающие продажи, и оценить степень важности каждого из них. При кажущейся очевидности и простоте решения далеко не вся управленческая команда дает однозначный ответ: какая из проблем на текущий момент является наиболее важной. Необходим экспертный опрос на эту тему.

Целью исследования факторов, влияющих на объемы продаж, является их ранжирование по степени важности. Для этого среди 25 сотрудников отдела сбыта, а также 10 руководителей завода ГАРО (Великий Новгород) А. А. Пивнем был проведен опрос, в котором предлагалось сравнить попарно факторы, определив более важный среди двух. Итог определялся как среднее арифметическое сумм баллов, набранных каждым фактором у всех опрашиваемых.

Были проанализированы следующие 15 факторов:

- потребительские свойства изделий (качество, надежность, показатели назначения и т.д.);
- уровень цен;
- срок поставки продукции;
- информация о предлагаемых к продаже изделиях;
- уровень гарантийного и сервисного обслуживания;
- работа дилеров, представительств;
- рекламная деятельность;
- численность персонала;
- мотивация труда;
- инициативность персонала;
- маркетинговая деятельность;
- оснащенность техническими средствами;
- квалификация персонала;
- корпоративная культура;
- репутация Компании.

В результате анализа результатов парных сравнений построена структурная схема, показывающих степень влияния факторов на объемы продаж (рис. 1).

Наибольшую значимость на сегодняшний день имеет срок поставки продукции и квалификация персонала. Меняются подходы к продвижению товаров на рынке. Ранее успешно применяемые способы продаж (почтовая рассылка рекламы, участие в специализированных выставках, публикации в газетах и специализированных изданиях, конференции и т.д.) сегодня требуют иного качественного подхода. Срок поставки продукции, как правило, связан с производственно-технологическим циклом изготовления и настройки изделий. Мотивация труда, равно как и уровень гарантийного и сервисного обслуживания, имеют также большое значение. Разрабатывается и утверждается новая система оплаты труда, которая позволяет устранить возникающие противоречия. Отдел сервисного обслуживания гаражного

оборудования должен разработать концепцию развития сервисной сети с целью наиболее полного удовлетворения потребителя, а значит и завоевания преимуществ в конкурентной борьбе.

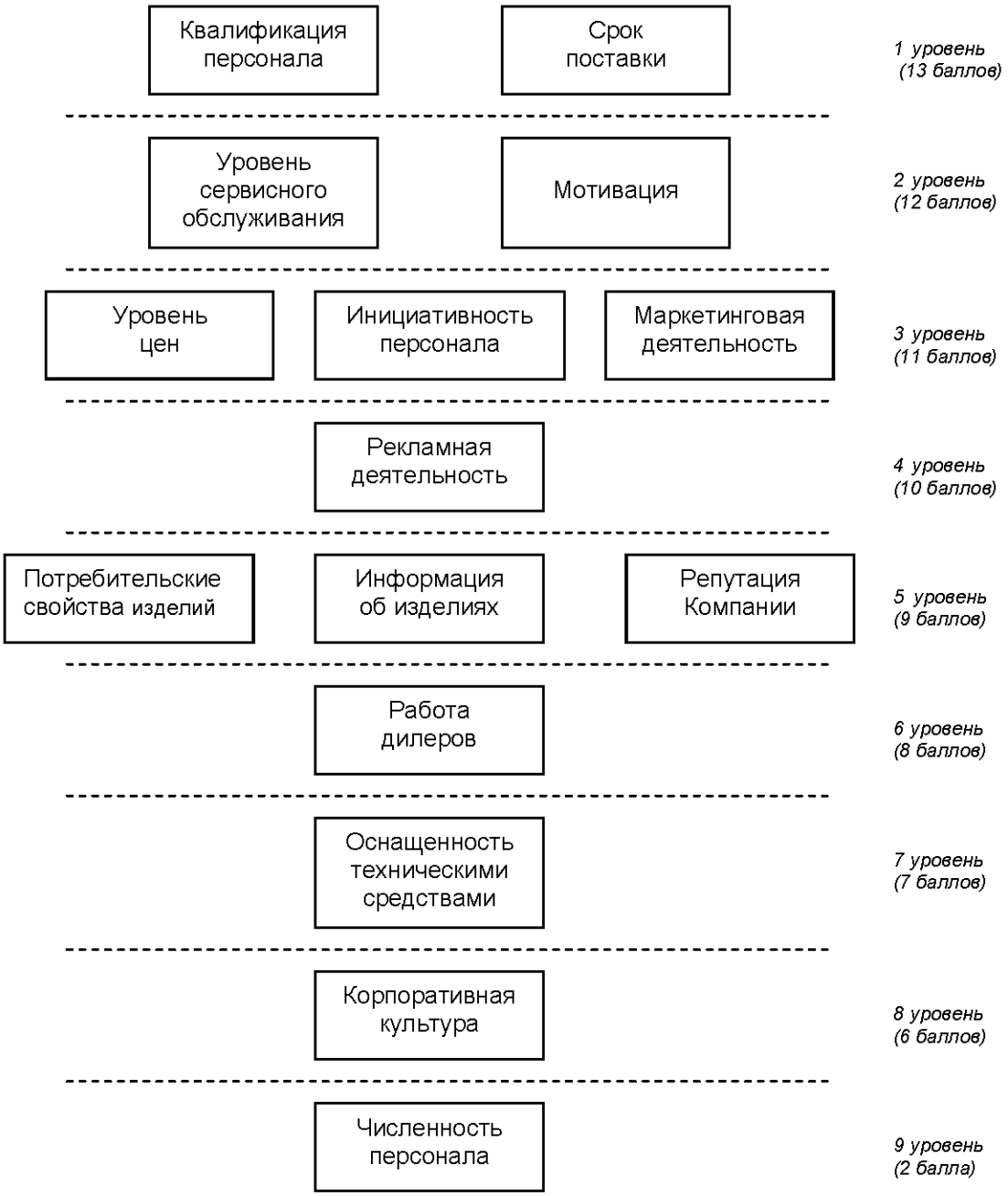


Рис. 1. Распределение факторов по их значимости

Среди проблем более низкого уровня значимости необходимо отметить место корпоративной культуры. Понимание и осознание себя, как части сплоченного коллектива — сложный процесс. Достижение синергетического эффекта возможно только в коллективе, в котором отдельный сотрудник по-

нимает и делает свою работу через понимание целей и задач всей Компании. Формированию корпоративной культуры следует уделить особое внимание.

Проведенный анализ дает возможность Компании сосредоточить свои усилия на наиболее важных на данный момент обозначенных проблемах. Выбор пути решения каждой из них определяется возможностями Компании и опытом руководителей.

Вероятностное моделирование парных сравнений. Напомним общую модель парных сравнений, введенную в главе 1.

Пусть t объектов $A_1, A_2, ..., A_t$ сравниваются попарно каждым из n экспертов. Следовательно, возможных пар для сравнения имеется $s = t(t-1)/2$. Эксперт с номером γ делает r_γ повторных сравнений для каждой из s возможностей. Пусть $X(i, j, \gamma, \delta)$, $i, j = 1, 2, ..., t$, $i \neq j$, $\gamma = 1, 2, ..., n$; $\delta = 1, 2, ..., r_\gamma$, — случайная величина, принимающая значения 1 или 0 в зависимости от того, предпочитает ли эксперт γ объект A_i или объект A_j в δ-м сравнении двух объектов. Обычно принимают, что все сравнения проводятся независимо друг от друга, так что случайные величины $X(i, j, \gamma, \delta)$ независимы в совокупности, если не считать того, что $X(i, j, \gamma, \delta) + X(j, i, \gamma, \delta) = 1$. Положим:

$$P(X(i, j, \gamma, \delta) = 1) = \pi(i, j, \gamma, \delta).$$

Ясно, что описанная модель парных сравнений представляет собой частный случай люсиана (в другой терминологии — бернуллиевского вектора). В этой модели число наблюдений равно числу неизвестных параметров, поэтому для получения статистических выводов необходимо наложить те или иные априорные условия на $\pi(i, j, \gamma, \delta)$, например:

- $\pi(i, j, \gamma, \delta) = \pi(i, j, \gamma)$ (нет эффекта от повторений);
- $\pi(i, j, \gamma, \delta) = \pi(i, j)$ (нет эффекта от повторений и от экспертов).

Теорию независимых парных сравнений целесообразно разделить на две части — непараметрическую, в которой статистические задачи ставятся непосредственно в терминах $\pi(i, j, \gamma, \delta)$, и параметрическую, в которой вероятности $\pi(i, j, \gamma, \delta)$ выражаются через меньшее число иных параметров. Ряд результатов непараметрической теории парных сравнений непосредственно вытекает из теории люсианов.

В параметрической теории парных сравнений наиболее популярна линейная модель, в которой предполагается, что каждому объекту A_i можно сопоставить некоторую «ценность» V_i так, что вероятность предпочтения $\pi(i, j)$

(т.е. предполагается дополнительно, что эффект от повторений и от экспертов отсутствует) выражается следующим образом:

$$\pi(i, j) = H(V_i - V_j), \tag{1}$$

где $H(x)$ — функция распределения, симметричная относительно 0, т.е.

$$H(-x) = 1 - H(x) \tag{2}$$

при всех x.

Широко применяются модели Терстоуна — Мостеллера и Брэдли — Терри, в которых $H(x)$ — соответственно функции нормального и логистического распределений. С прикладной точки зрения эти две модели практически совпадают. Действительно, поскольку функция $\Phi(x)$ стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1 и функция:

$$\Psi(x) = e^x (1 + e^x)^{-1}$$

стандартного логистического распределения удовлетворяют соотношению (см. главу 1):

$$\sup_{x \in R^1} | \Phi(x) - \Psi(1{,}7x) | < 0{,}01,$$

то для обоснованного выбора по статистическим данным между моделями Терстоуна — Мостеллера и Брэдли — Терри необходимо не менее тысячи наблюдений. Ясно, что при реальном проведении экспертного опроса число наблюдений по крайней мере на порядок меньше.

Соотношение (1) вытекает из следующей модели поведения эксперта: он измеряет «ценность» V_i и V_j объектов A_i и A_j, но с ошибками ε_i и ε_j соответственно, а затем сравнивает свои оценки ценности объектов $y_i = V_i + \varepsilon_i$ и $y_j = V_j + \varepsilon_j$. Если $y_i > y_j$, то он предпочитает A_i, в противном случае — A_j. Тогда:

$$\pi(i, j) = P(\varepsilon_j - \varepsilon_i < V_i - V_j) = H(V_i - V_j). \tag{3}$$

Обычно предполагают, что субъективные ошибки эксперта ε_i и ε_j независимы и имеют одно и то же непрерывное распределение. Тогда функция

распределения $H(x)$ из соотношения (3) непрерывна и удовлетворяет функциональному уравнению (2).

Пример. При опросе экспертов (август 2001 г.) попарно сравнивались четыре компании ТНК, Лукойл, Юкос, Татнефть, продающие автомобильное топливо. Сравнение проводилось по качеству бензина. При t = 4 пар для сравнения имеется $s = t(t-1)/2 = 6$. Результаты парных сравнений приведены в табл.1. По ним необходимо определить взаимное положение четырех компаний на оси «качество бензина», т.е. найти их «ценности» V_1, V_2, V_3, V_4.

Таблица 1

Сравнение компаний по качеству бензина

Пары	Частота выбора первого элемента пары	Частота выбора второго элемента пары
ТНК — Лукойл	π(1,2) = 0,508	π(2,1) = 0,492
ТНК — Юкос	π(1,3) = 0,331	π(3,1) = 0,669
ТНК — Татнефть	π(1,4) = 0,990	π(4,1) = 0,010
Лукойл — Юкос	π(2,3) = 0,338	π(3,2) = 0,662
Лукойл — Татнефть	π(2,4) = 0,990	π(4,2) = 0,010
Юкос — Татнефть	π(3,4) = 0,997	π(4,3) = 0,003

Применим модель Терстоуна — Мостеллера, согласно которой погрешности мнений экспертов ε_i являются независимыми нормально распределенными случайными величинами с нулевым математическим ожиданием и дисперсией σ^2.

Легко видеть, что «ценности» V_1, V_2, V_3, V_4 измерены в шкале интервалов. Начало координат можно выбрать произвольно, поскольку вероятности результатов сравнения зависят только от попарных разностей «ценностей» V_1, V_2, V_3, V_4. Например, можно положить V_4 = 0. Единицу измерения также можно выбрать произвольно. При изменении единицы измерения меняется σ^2, точнее, единица измерения однозначно связана с величиной σ. Дисперсия разности $\varepsilon_i - \varepsilon_j$ равна $2\sigma^2$. В соответствии с формулой (3) удобно выбрать единицу измерения так, чтобы $2\sigma^2 = 1$, т.е. $\sigma = 1/\sqrt{2}$. Тогда H в формуле (3) — это функция Ф стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1.

В соответствии с (3) имеем систему шести уравнений с тремя неизвестными:

$$\Phi(V_1 - V_2) = \pi(1, 2) = 0{,}508,$$
$$\Phi(V_1 - V_3) = \pi(1, 3) = 0{,}331,$$
$$\Phi(V_1) = \pi(1, 4) = 0{,}990,$$
$$\Phi(V_2 - V_3) = \pi(2, 3) = 0{,}338,$$
$$\Phi(V_2) = \pi(2, 4) = 0{,}990,$$
$$\Phi(V_3) = \pi(3, 4) = 0{,}997.$$

Применяя к каждому из этих уравнений преобразование Φ^{-1}, получаем систему шести линейных уравнений с тремя неизвестными:

$$V_1 - V_2 = a_1 = \Phi^{-1}(0{,}508) = 0{,}020054,$$
$$V_1 - V_3 = a_2 = \Phi^{-1}(0{,}331) = -0{,}437154,$$
$$V_1 = a_3 = \Phi^{-1}(0{,}990) = 2{,}326348,$$
$$V_2 - V_3 = a_4 = \Phi^{-1}(0{,}338) = -0{,}417928,$$
$$V_2 = a_5 = \Phi^{-1}(0{,}990) = 2{,}326348,$$
$$V_3 = a_6 = \Phi^{-1}(0{,}997) = 2{,}747781.$$

(Значения Φ^{-1} взяты из табл. 1.3 сборника [9].)

В полученной системе число уравнений больше числа неизвестных, т.е. система переопределена. Дальнейшие расчеты могут проводиться разными способами. Простейший из них состоит в том, чтобы выбрать три уравнения, а именно, третье, пятое и шестое, которые и дают искомые значения:

$$V_1 = V_2 = 2{,}326348,\ V_3 = 2{,}747781.$$

Таким образом, качество бензина лучше всего у Юкоса, оно несколько хуже у ТНК и Лукойла, одинаковых по этому показателю, а Татнефть значительно хуже тройки лидеров. Можно показать, что если модель Терстоуна — Мостеллера верна и число экспертов достаточно велико, то отбрасывание «лишних» уравнений является корректным способом обработки экспертных данных, поскольку дает состоятельные оценки «ценностей» $V_1, V_2, \ldots, V_n$.

Однако ясно, что при отбрасывании трех уравнений из шести часть информации теряется. Например, первое уравнение показывает, что по мнению экспертов, качество бензина у ТНК несколько лучше, у Лукойла. Поэтому

целесообразно применить метод наименьших квадратов для оценивания V_1, V_2, V_3, V_4. А именно, рассмотрим функцию трех переменных:

$$f(V_1, V_2, V_3) = (V_1 - V_2 - a_1)^2 + (V_1 - V_3 - a_2)^2 + (V_1 - a_3)^2 + \\ + (V_2 - V_3 - a_4)^2 + (V_2 - a_5)^2 + (V_3 - a_6)^2 .$$

Оценки по методу наименьших квадратов — это результат минимизации функции $f(V_1, V_2, V_3)$ по совокупности переменных V_1, V_2, V_3. Для минимизации этой функции достаточно приравнять 0 частные производные этой функции по V_1, V_2, V_3. Имеем:

$$\frac{\partial f}{\partial V_1} = 2(V_1 - V_2 - a_1) + 2(V_1 - V_3 - a_2) + 2(V_1 - a_3),$$

$$\frac{\partial f}{\partial V_2} = -2(V_1 - V_2 - a_1) + 2(V_2 - V_3 - a_4) + 2(V_2 - a_5),$$

$$\frac{\partial f}{\partial V_3} = -2(V_1 - V_3 - a_2) - 2(V_2 - V_3 - a_4) + 2(V_3 - a_6).$$

Приравнивая частные производные 0, деля на 2, раскрывая скобки и перенося свободные члены в правую часть, получаем систему трех линейных уравнений с тремя неизвестными:

$$3V_1 - V_2 - V_3 = a_1 + a_2 + a_3,$$
$$-V_1 + 3V_2 - V_3 = -a_1 + a_4 + a_5,$$
$$-V_1 - V_2 + 3V_3 = -a_2 - a_4 + a_6.$$

Решение этой системы не представляет трудностей.

Вообще говоря, не всегда сравниваемые объекты можно представить точками на прямой, т.е. не всегда их можно линейно упорядочить. Возможно, более соответствует данным опроса экспертов представление объектов точками на плоскости или в пространстве большей размерности. В статистике парных сравнений [29] разработаны методы проверки адекватности модели Терстоуна — Мостеллера и других параметрических моделей. Для этого обычно используются критерии типа хи-квадрат.

3.6. СТАТИСТИКА НЕЧЕТКИХ МНОЖЕСТВ

Нечеткие множества — частный вид объектов нечисловой природы. Поэтому при обработке выборки, элементами которой являются нечеткие

множества, могут быть использованы различные методы анализа статистических данных произвольной природы — расчет средних, непараметрических оценок плотности, построение диагностических правил и т.д.

Среднее значение нечеткого множества. Однако иногда используются методы, учитывающие специфику нечетких множеств. Например, пусть носителем нечеткого множества является конечная совокупность действительных чисел $\{x_1, x_2, ..., x_n\}$. Тогда под средним значением нечеткого множества иногда понимают число. А именно, среднее значение нечеткого множества определяют по формуле:

$$M(A) = \frac{\sum_{i=1}^{n} x_i \mu_A(x_i)}{\sum_{i=1}^{n} \mu_A(x_i)},$$

где $\mu_A(x_i)$ — функция принадлежности нечеткого множества A. Если знаменатель равен 1, то эта формула определяет математическое ожидание случайной величины, для которой вероятность попасть в точку x_i равна $\mu_A(x_i)$. Такое определение наиболее естественно, когда нечеткое множество A интерпретируется как нечеткое число.

Очевидно, наряду с $M(A)$ может оказаться полезным использование эмпирических средних, определяемых (согласно статистике в пространствах произвольной природы) путем решения соответствующих оптимизационных задач. Для конкретных расчетов необходимо ввести то или иное расстояние между нечеткими множествами.

Расстояния в пространствах нечетких множеств. Как известно, многие методы статистики нечисловых данных базируются на использовании расстояний (или показателей различия) в соответствующих пространствах нечисловой природы. Расстояние между нечеткими подмножествами A и B множества $X = \{x_1, x_2, \ldots, x_k\}$ можно определить как:

$$d(A,B) = \sum_{j=1}^{k} | \mu_A(x_j) - \mu_B(x_j) |,$$

где $\mu_A(x_j)$ — функция принадлежности нечеткого множества A, а $\mu_B(x_j)$ — функция принадлежности нечеткого множества B.

Может использоваться и другое расстояние:

$$D(A,B)=\frac{\sum_{j=1}^{k}|\mu_A(x_j)-\mu_B(x_j)|}{\sum_{j=1}^{k}(\mu_A(x_j)+\mu_B(x_j))}.$$

(Примем это расстояние равным 0, если функции принадлежности тождественно равны 0.)

В соответствии с аксиоматическим подходом к выбору расстояний (метрик) в пространствах нечисловой природы разработан обширный набор систем аксиом, из которых выводится тот или иной вид расстояний (метрик) в конкретных пространствах, в том числе в пространствах нечетких множеств (см. главу 1). При использовании вероятностных моделей расстояние между случайными нечеткими множествами (т.е. между случайными элементами со значениями в пространстве нечетких множеств) само является случайной величиной, имеющей в ряде постановок асимптотически нормальное распределение [25].

Проверка гипотез о нечетких множествах. Пусть ответ эксперта — нечеткое множество. Естественно считать, что его ответ, как показание любого средства измерения, содержит погрешности. Если есть несколько экспертов, то в качестве единой оценки (группового мнения) естественно взять эмпирическое среднее их ответов. Но возникает естественный вопрос: действительно ли все эксперты измеряют одно и то же? Может быть, глядя на реальный объект, они оценивают его с разных сторон? Например, на научную статью можно смотреть как с теоретической точки зрения, так и с прикладной, и соответствующие оценки будут, скорее всего, различны (если они совпадают, то работа либо никуда не годится, либо является выдающейся).

Итак, возник вопрос: как проверить согласованность мнений экспертов? Надо сначала определить понятие согласованности. Пусть A — нечеткий ответ эксперта. Будем считать, что соответствующая функция принадлежности есть сумма двух слагаемых:

$$\mu_A(u)=\mu_{N(A)}(u)+\xi_A(u),$$

где $N(A)$ — «истинное» нечеткое множество, а $\xi_A(u)$ — «погрешность» эксперта как прибора. Естественно рассмотреть две постановки.

Мнения экспертов $A(1)$, $A(2)$, … , $A(m)$ будем считать согласованными, если:

$$N(A(1)) = N(A(2)) = \ldots = N(A(m)).$$

Рассмотрим две группы экспертов. В первой у всех «истинное» мнение $N(A)$, а во второй у всех — $N(B)$. Две группы будем считать согласованными по мнениям, если:

$$N(A) = N(B).$$

Согласованность определена. Как же ее проверить? Если экспертов достаточно много, то эти гипотезы можно проверять отдельно для каждого элемента множества — общего носителя нечетких ответов. Проверка последней гипотезы переходит в проверку однородности двух независимых выборок [5, гл. 4]. Здесь ограничимся приведенными выше постановками основных гипотез (ср. с аналогичными гипотезами, рассмотренными выше для люсианов).

Восстановление зависимости между нечеткими переменными. Рассмотрим две нечеткие переменные A и B. Пусть каждый из n испытуемых выдает в ответ на вопрос два нечетких множества A_i и B_i, $i = 1, 2, \ldots, n$. Необходимо восстановить зависимость B от A, другими словами, наилучшим образом приблизить B с помощью A.

Для иллюстрации основной идеи ограничимся парной линейной регрессией нечетких множеств. Нечеткое множество C назовем линейной функцией от нечеткого множества A, если для любого x из носителя A функции принадлежности множеств A и C таковы, что $\mu_C(x) = \mu_A(y)$ при $x = \alpha y + \beta$. Другими словами,

$$\mu_C(x) = \mu_A((x - \beta)/\alpha)$$

для любого x из носителя A. В таком случае естественно писать:

$$C = \alpha A + \beta.$$

Однако нечеткие переменные, как и привычные статистикам числовые переменные, обычно несколько отклоняются от линейной связи. Наилучшее

линейное приближение нечеткой переменной B с помощью линейной функции от нечеткой переменной A естественно искать, решая задачу минимизации по α, β расстояния от B до C. Пусть:

$$\rho(B, \alpha_0 A + \beta_0) = min\, \rho(B, \alpha A + \beta),$$

где ρ — некоторое расстояние между нечеткими множествами, а минимизация проводится по всем возможным значениям α и β. Тогда наилучшей линейной аппроксимацией B является $\alpha_0 A + \beta_0$. Если рассматриваемый минимум равен 0, то имеет место точная линейная зависимость.

Для восстановления зависимости по выборочным парам нечетких переменных естественно воспользоваться подходом, развитым в статистике в пространствах произвольной природы для параметрической регрессии (аппроксимации). В соответствии с рассмотрениями главы 2 в качестве наилучших оценок параметров линейной зависимости следует рассматривать:

$$(\alpha^*, \beta^*) = Arg \min_{\alpha,\beta} \sum_{k=1}^{n} \rho(B_i, \alpha A_i + \beta).$$

Тогда наилучшим линейным приближением B является $C^* = \alpha^* A + \beta^*$.

Вероятностно-статистическая теория регрессионного анализа нечетких переменных [30] строится как частный случай аналогичной теории для переменных произвольной природы (глава 2). В частности, при обычных предположениях оценки α^*, β^* являются состоятельными, т.е. $\alpha^* \to \alpha_0$ и $\beta^* \to \beta_0$ при $n \to \infty$.

Кластер-анализ нечетких переменных. Строить группы сходных между собой нечетких переменных (кластеры) можно многими способами. Опишем два семейства алгоритмов.

Пусть на пространстве, в котором лежат результаты наблюдений, т.е. на пространстве нечетких множеств, заданы две меры близости ρ и τ (например, это могут быть введенные выше расстояния d и D). Берется один из результатов наблюдений (нечеткое множество) и вокруг него описывается шар радиуса R, определяемый мерой близости ρ. (Напомним, что шаром с центром в x относительно ρ называется множество всех элементов y рассматриваемого пространства таких, что $\rho(x, y) \le R$.) Берутся результаты наблюдений (элементы выборки), попавшие в этот шар, и находится их эмпирическое среднее относительно второй меры близости τ. Оно берется за новый центр,

вокруг которого снова описывается шар радиуса R относительно ρ, и процедура повторяется. (Чтобы алгоритм был полностью определен, необходимо сформулировать правило выбора элемента эмпирического среднего в качестве нового центра, если эмпирическое среднее состоит более чем из одного элемента.)

Когда центр шара зафиксируется (перестанет меняться), попавшие в этот шар элементы объявляются первым кластером и исключаются из дальнейшего рассмотрения. Алгоритм применяется к совокупности оставшихся результатов наблюдений, выделяет из нее второй кластер и т.д.

Всегда ли центр шара остановится? При реальных расчетах в течение многих лет так было всегда. Соответствующая теория была построена лишь в 1977 г. [31]. Было доказано, что описанный выше процесс всегда остановится через конечное число шагов. Причем число шагов до остановки оценивается через максимально возможное число результатов наблюдений в шаре радиуса R относительно ρ.

Обширное семейство образуют алгоритмы кластер-анализа типа «Дендрограмма», известные также под названием «агломеративные иерархические алгоритмы средней связи». На первом шаге алгоритма из этого семейства каждый результат наблюдения рассматривается как отдельный кластер. Далее на каждом шагу происходит объединение двух самых близких кластеров. Название «Дендрограмма» объясняется тем, что результат работы алгоритма обычно представляется в виде дерева. Каждая его ветвь соответствует кластеру, появляющемуся на каком-либо шагу работы алгоритма. Слияние ветвей соответствует объединению кластеров, а ствол — заключительному шагу, когда все наблюдения оказываются объединенными в один кластер.

Для работы алгоритмов кластер-анализа типа «Дендрограмма» необходимо определить расстояние между кластерами. Естественно использовать ассоциативные средние (которыми, как известно, являются обобщенные средние по Колмогорову) всевозможных попарных расстояний между элементами двух рассматриваемых кластеров. Итак, расстояние между кластерами K и L, состоящими из n_1 и n_2 элементов соответственно, определяется по формуле:

$$\tau(K,L) = F^{-1}\left(\frac{1}{n_1 n_2}\sum_{i\in K}\sum_{j\in L} F(\rho(X_i, X_j))\right),$$

где ρ — некоторое расстояние между нечеткими множествами, F — строго монотонная функция (строго возрастающая или строго убывающая).

Соображения теории измерений позволяют ограничить круг возможных алгоритмов типа «Дендрограмма». Естественно принять, что единица измерения расстояния выбрана произвольно. Тогда согласно результатам раздела 3.1 из всех обобщенных средних по Колмогорову годятся только степенные средние, т.е. $F(z) = z^{\lambda}$ при $\lambda \neq 0$ или $F(z) = \ln z$. Чтобы получить разбиение на кластеры, надо «разрезать» дерево на определенной высоте, т.е. объединять кластеры лишь до тех пор, пока расстояние между ними меньше заранее выбранной константы. При альтернативном подходе заранее фиксируется число кластеров. Рассматривают и двухкритериальную постановку, когда минимизируют сумму (или максимум) внутрикластерных разбросов и число кластеров. Для решения задачи двухкритериальной минимизации либо один из критериев заменяют на ограничение, либо два критерия «свертывают» в один, либо применяют иные подходы (последовательная оптимизация, построение поверхности Парето и др.).

При классификации нечетких множеств полезны многие подходы, рассмотренные в [5, гл. 5], а именно, все подходы, основанные только на использовании расстояний.

Сбор и описание нечетких данных. Разработано большое количество процедур описания нечеткости. Так, согласно Э. Борелю, понятие «Куча» описывается с помощью функции распределения: при каждом конкретном x значение функции принадлежности — это доля людей, считающих совокупность из x зерен кучей. Результат подобного опроса может дать и кривую иного вида, например, по поводу понятия «молодой» (слева будут отделены «дети», а справа — «люди зрелого и пожилого возраста»). Нечеткая толерантность может оцениваться с помощью случайных толерантностей (см. выше).

Целесообразно попытаться выделить наиболее практически полезные простые формы функций принадлежности. Видимо, наиболее простой является «ступенька» — внутри некоторого интервала функция принадлежности равна 1, а вне этого интервала равна 0. Это — простейший способ «размывания» числа путем замены его интервалом. Нечеткое множество описывается двумя числами — концами интервала. Оценки этих чисел можно получить с помощью экспертов. Статистическая теория подобных нечетких множеств рассмотрена в главе 4.

Тремя числами $a < b < c$ описывается функция принадлежности типа треугольника. При этом левее числа a и правее числа c функция принадлежности равна 0. В точке b функция принадлежности принимает значение 1. На отрезке $[a; b]$ функция принадлежности линейно растет от 0 до 1, а на отрезке $[b; c]$ — линейно убывает от 1 до 0. Оценки трех чисел $a < b < c$ получают при опросе экспертов.

Следующий по сложности вид функции принадлежности — типа трапеции — описывается четырьмя числами $a < b < c < d$. Левее a и правее d функция принадлежности равна 0. На отрезке $[a; b]$ она линейно возрастает от 0 до 1, на отрезке $[b; c]$ во всех точках равна 1, а на отрезке $[c; d]$ линейно убывает от 1 до 0. Для оценивания четверки чисел $a < b < c < d$ используют экспертов.

Ряд результатов статистики нечетких данных приведен в первой монографии российского автора по нечетким множествам [30] и во многих дальнейших публикациях.

3.7. СТАТИСТИКА НЕЧИСЛОВЫХ ДАННЫХ В ЭКСПЕРТНЫХ ОЦЕНКАХ

Развитие статистики нечисловых данных во многом стимулировалось запросами теории и практики экспертных оценок. Рассмотрим взаимоотношение этих двух научно-практических областей подробнее.

Современная теория измерений и экспертные оценки. Как проводить анализ собранных рабочей группой ответов экспертов? Для более углубленного рассмотрения проблем экспертных оценок понадобятся некоторые понятия *репрезентативной теории измерений*, служащей основой теории экспертных оценок, прежде всего той ее части, которая связана с анализом заключений экспертов, выраженных в качественном (а не в количественном) виде.

Как уже отмечалось, получаемые от экспертов мнения часто выражены в *порядковой шкале*. Другими словами, эксперт может сказать (и обосновать), что один тип продукции будет более привлекателен для потребителей, чем другой, один показатель качества продукции более важен, чем другой, первый технологический объект более опасен, чем второй, и т.д. Но эксперт не в состоянии обосновать, *во сколько раз* или *на сколько* более важен, соответственно, более опасен. Поэтому экспертов часто просят дать ранжировку (упорядочение) объектов экспертизы, т.е. расположить их в порядке возрас-

тания (или, точнее, неубывания) интенсивности интересующей организаторов экспертизы характеристики.

Рассмотрим в качестве примера применения результатов теории измерений, относящихся к средним величинам в порядковой шкале, один сюжет, связанный с ранжировками и рейтингами.

Методы средних баллов. В настоящее время распространены экспертные, маркетинговые, квалиметрические, социологические и иные опросы, в которых опрашиваемых просят выставить баллы объектам, изделиям, технологическим процессам, предприятиям, проектам, заявкам на выполнение научно-исследовательских работ, идеям, проблемам, программам, политикам и т.п. Затем рассчитывают средние баллы и рассматривают их *как интегральные* (*т.е. обобщенные, итоговые*) *оценки,* выставленные коллективом опрошенных экспертов. Какими формулами пользоваться для вычисления средних величин? Ведь существует очень много разных видов средних величин.

По традиции обычно применяют *среднее арифметическое*. Однако специалисты по теории измерений уже более 30 лет знают, что *такой способ некорректен*, поскольку баллы обычно измерены в *порядковой* шкале (см. раздел 3.1). Обоснованным является использование медиан в качестве средних баллов. Однако полностью игнорировать средние арифметические нецелесообразно из-за их привычности и распространенности. Поэтому *представляется рациональным использовать одновременно оба метода — и метод средних арифметических рангов (баллов), и методов медианных рангов.* Такая рекомендация находится в согласии с общенаучной *концепцией устойчивости* [1], рекомендующей применять различные методы для обработки одних и тех же данных с целью выделить выводы, получаемые одновременно при всех методах. Такие выводы, видимо, соответствуют реальной действительности, в то время как заключения, меняющиеся от метода к методу, зависят от субъективизма исследователя, выбирающего метод обработки исходных экспертных оценок.

Пример сравнения восьми проектов. Рассмотрим конкретный пример применения только что сформулированного подхода.

По заданию руководства фирмы анализировались восемь проектов, предлагаемых для включения в план стратегического развития фирмы. Они обозначены следующим образом: Д, Л, М-К, Б, Г-Б, Сол, Стеф, К (по фамилиям менеджеров, предложивших их для рассмотрения). Все проекты были направлены 12 экспертам, включенным в экспертную комиссию, организо-

ванную по решению Правления фирмы. В табл. 1 приведены ранги восьми проектов, присвоенные им каждым из 12 экспертов в соответствии с представлением экспертов о целесообразности включения проекта в стратегический план фирмы. При этом эксперт присваивает ранг 1 самому лучшему проекту, который обязательно надо реализовать. Ранг 2 получает от эксперта второй по привлекательности проект и т.д. Наконец, ранг 8 — наиболее сомнительный проект, который реализовывать стоит лишь в последнюю очередь.

Таблица 1

Ранги 8 проектов по степени привлекательности

№ эксперта	Д	Л	М-К	Б	Г-Б	Сол	Стеф	К
1	5	3	1	2	8	4	6	7
2	5	4	3	1	8	2	6	7
3	1	7	5	4	8	2	3	6
4	6	4	2,5	2,5	8	1	7	5
5	8	2	4	6	3	5	1	7
6	5	6	4	3	2	1	7	8
7	6	1	2	3	5	4	8	7
8	5	1	3	2	7	4	6	8
9	6	1	3	2	5	4	7	8
10	5	3	2	1	8	4	6	7
11	7	1	3	2	6	4	5	8
12	1	6	5	3	8	4	2	7

Примечание. Эксперт № 4 считает, что проекты М-К и Б равноценны, но уступают лишь одному проекту — проекту Сол. Поэтому проекты М-К и Б должны были бы стоять на втором и третьем местах и получить ранги 2 и 3. Поскольку они равноценны, то получают средний (связанный) ранг (2 + 3) / 2 = 5 / 2 = 2,5.

Анализируя результаты работы экспертов (т.е. упомянутую таблицу), члены аналитической подразделения Рабочей группы, обрабатывавшие ответы экспертов по заданию Правления фирмы, были вынуждены констатировать, что полного согласия между экспертами нет, а потому данные, приведенные в табл. 1, следует подвергнуть тщательному математическому анализу.

Метод средних арифметических рангов. Сначала для получения группового мнения экспертов был применен метод средних арифметических

рангов. Прежде всего была подсчитана сумма рангов, присвоенных проектам (см. табл. 1). Затем эта сумма была разделена на число экспертов, в результате рассчитан средний арифметический ранг (именно эта операция дала название методу). По средним рангам строится итоговая ранжировка (в другой терминологии — упорядочение), исходя из принципа — чем меньше средний ранг, чем лучше проект. Наименьший средний ранг, равный 2,625, у проекта Б, — следовательно, в итоговой ранжировке он получает ранг 1. Следующая по величине сумма, равная 3,125, у проекта М-К, — и он получает итоговый ранг 2. Проекты Л и Сол имеют одинаковые средние (равные 3,25), значит, с точки зрения экспертов они равноценны (при рассматриваемом способе сведения вместе мнений экспертов), а потому они должны бы стоять на 3 и 4 местах и получают средний ранг (3 + 4) / 2 = 3,5. Дальнейшие результаты приведены в табл. 2.

Итак, ранжировка по суммам рангов (или, что в данном случае то же самое, по средним арифметическим рангам) имеет вид:

$$Б < М\text{-}К < \{Л, Сол\} < Д < Стеф < Г\text{-}Б < К . \qquad (1)$$

Здесь запись типа «А<Б» означает, что проект А предшествует проекту Б (т.е. проект А лучше проекта Б). Поскольку проекты Л и Сол получили одинаковую сумму рангов, то по рассматриваемому методу они эквивалентны, а потому объединены в группу (в фигурных скобках). В терминологии математической статистики ранжировка (1) имеет одну связь.

Метод медиан рангов. Значит, наука сказала свое слово, итог расчетов — ранжировка (1), и на ее основе предстоит принимать решение? Так был поставлен вопрос при обсуждении полученных результатов на заседании Правления фирмы. Но тут наиболее знакомый с современной эконометрикой член Правления вспомнил, что ответы экспертов измерены в порядковой шкале, а потому для них неправомерно проводить усреднение методом средних арифметических. Надо использовать метод медиан.

Что это значит? Надо взять ответы экспертов, соответствующие одному из проектов, например, проекту Д. Это ранги 5, 5, 1, 6, 8, 5, 6, 5, 6, 5, 7, 1. Затем их надо расположить в порядке неубывания (проще было бы сказать — «в порядке возрастания», но поскольку некоторые ответы совпадают, то приходится использовать непривычный термин «неубывание»). Получим последовательность: 1, 1, 5, 5, 5, 5, 5, 6, 6, 6, 7, 8. На центральных местах — шестом и седьмом — стоят 5 и 5. Следовательно, медиана равна 5.

Таблица 2

Результаты расчетов по методу средних арифметических и методу медиан для данных, приведенных в таблице 1

Показатели	Д	Л	М-К	Б	Г-Б	Сол	Стеф	К
Сумма рангов	60	39	37,5	31.5	76	39	64	85
Среднее арифметическое рангов	5	3,25	3,125	2,625	6,333	3,25	5,333	7,083
Итоговый ранг по среднему арифметическому	5	3,5	2	1	7	3,5	6	8
Медианы рангов	5	3	3	2,25	7,5	4	6	7
Итоговый ранг по медианам	5	2,5	2,5	1	8	4	6	7

Медианы совокупностей из 12 рангов, соответствующих определенным проектам, приведены в предпоследней строке табл. 2. (При этом медианы вычислены по обычным правилам статистики — как среднее арифметическое центральных членов вариационного ряда.) Итоговое упорядочение комиссии экспертов по методу медиан приведено в последней строке таблицы. Ранжировка (т.е. упорядочение — итоговое мнение комиссии экспертов) по медианам имеет вид:

$$Б < \{М\text{-}К, Л\} < Сол < Д < Стеф < К < Г\text{-}Б \ . \qquad (2)$$

Поскольку проекты Л и М-К имеют одинаковые медианы баллов, то по рассматриваемому методу ранжирования они эквивалентны, а потому объединены в группу (кластер), т.е. с точки зрения математической статистики ранжировка (4) имеет одну связь.

Сравнение ранжировок по методу средних арифметических и методу медиан. Сравнение ранжировок (1) и (2) показывает их близость (похожесть). Можно принять, что проекты М-К, Л, Сол упорядочены как М-К < Л < Сол, но из-за погрешностей экспертных оценок в одном методе признаны равноценными проекты Л и Сол (ранжировка (1)), а в другом — проекты М-К и Л (ранжировка (2)). Существенным является только расхождение, касающееся упорядочения проектов К и Г-Б: в ранжировке (3) Г-Б < К, а в ранжировке (4), наоборот, К < Г-Б. Однако эти проекты — наименее привлекательные из восьми рассматриваемых, и при выборе наиболее привлекательных проектов

для дальнейшего обсуждения и использования на указанное расхождение можно не обращать внимания.

Рассмотренный пример демонстрирует сходство и различие ранжировок, полученных по методу средних арифметических рангов и по методу медиан, а также пользу от совместного применения этих методов.

Метод согласования кластеризованных ранжировок. Проблема состоит в выделении общего нестрогого порядка из набора кластеризованных ранжировок (в другой терминологии — ранжировок со связями). Этот набор может отражать мнения нескольких экспертов или быть получен при обработке мнений экспертов различными методами. Рассмотрим *метод согласования кластеризованных ранжировок, позволяющий «загнать» противоречия внутрь специальным образом построенных кластеров (групп), в то время как упорядочение кластеров соответствует одновременно всем исходным упорядочениям.*

В различных прикладных областях возникает необходимость анализа нескольких кластеризованных ранжировок объектов. К таким областям относятся, прежде всего, инженерный бизнес, менеджмент, экономика, социология, экология, прогнозирование, научные и технические исследования и т.д. Особенно те их разделы, что связаны с экспертными оценками (см., например, [5, 32]). В качестве объектов могут выступать образцы продукции, технологии, математические модели, проекты, кандидаты на должность и др. Кластеризованные ранжировки могут быть получены как с помощью экспертов, так и объективным путем, например, при сопоставлении математических моделей с экспериментальными данными с помощью того или иного критерия качества. Описанный ниже метод был разработан в связи с проблемами химической безопасности биосферы и экологического страхования [32].

В настоящем пункте рассматривается метод построения кластеризованной ранжировки, согласованной (в раскрытом ниже смысле) со всеми рассматриваемыми кластеризованными ранжировками. При этом противоречия между отдельными исходными ранжировками оказываются заключенными внутри кластеров согласованной ранжировки. В результате упорядоченность кластеров отражает общее мнение экспертов, точнее, то общее, что содержится в исходных ранжировках.

В кластеры заключены объекты, по поводу которых некоторые из исходных ранжировок *противоречат* друг другу. Для их упорядочения необходимо провести новые исследования. Эти исследования могут быть как формально-математическими (например, вычисление медианы Кемени, упо-

рядочения по средним рангам или по медианам и т.п.), так и требовать привлечения новой информации из соответствующей прикладной области, возможно, проведения дополнительных научных или прикладных работ.

Введем необходимые понятия, затем сформулируем алгоритм согласования кластеризованных ранжировок в общем виде и рассмотрим его свойства.

Пусть имеется конечное число объектов, которые мы для простоты изложения будем изображать натуральными числами 1, 2, 3, ..., k и называть их совокупность «носителем». *Под кластеризованной ранжировкой, определенной на заданном носителе, понимаем следующую математическую конструкцию*. Пусть объекты разбиты на группы, которые будем называть кластерами. В кластере может быть и один элемент. Входящие в один кластер объекты будем заключать в фигурные скобки. Например, объекты 1, 2, 3, ..., 10 могут быть разбиты на 7 кластеров: {1}, {2,3}, {4}, {5,6,7}, {8}, {9}, {10}. В этом разбиении один кластер {5, 6, 7} содержит три элемента, другой — {2, 3} — два, остальные пять — по одному элементу. Кластеры не имеют общих элементов, а объединение их (как множеств) есть все рассматриваемое множество объектов (весь носитель).

Вторая составляющая кластеризованной ранжировки — это строгий линейный порядок между кластерами. Задано, какой из них первый, какой второй, и т.д. Будем изображать упорядоченность с помощью знака «<». При этом кластеры, состоящие из одного элемента, будем для простоты изображать без фигурных скобок. Тогда кластеризованную ранжировку на основе введенных выше кластеров можно изобразить так:

$$A = [\ 1 < \{2,3\} < 4 < \{5, 6, 7\} < 8 < 9 < 10\ ] .$$

Конкретные кластеризованные ранжировки будем заключать в квадратные скобки. Если для простоты речи термин «кластер» применять только к кластеру не менее чем из двух элементов, то можно сказать, что в кластеризованную ранжировку A входят два кластера {2, 3} и {5, 6, 7} и 5 отдельных элементов.

Введенная описанным образом кластеризованная ранжировка является бинарным отношением на носителе — множестве {1, 2, 3, ..., 10}. Его структура такова. Задано отношение эквивалентности с 7-ю классами эквивалентности, а именно, {2, 3}, {5, 6, 7}, а остальные состоят 5 классов из оставшихся 5 отдельных элементов. Затем введен строгий линейный порядок между классами эквивалентности.

Введенный математический объект известен в литературе как «ранжировка со связями» (М. Холлендер, Д. Вулф), «упорядочение» (Дж. Кемени, Дж. Снелл [33]), «квазисерия» (Б. Г. Миркин), «совершенный квазипорядок» (Ю. А. Шрейдер [34, с. 127, 130]). Учитывая разнобой в терминологии, было признано полезным ввести специальный термин *«кластеризованная ранжировка»,* поскольку в нем явным образом названы основные элементы изучаемого математического объекта — кластеры, рассматриваемые на этапе согласования ранжировок как классы эквивалентности, и ранжировка — строгий совершенный порядок между ними (в терминологии Ю. А. Шрейдера [34, гл. IV]).

Следующее важное понятие — *противоречивость*. Оно определяется для четверки — две кластеризованные ранжировки на одном и том же носителе и два различных объекта — элементы того же носителя. При этом два элемента из одного кластера будем связывать символом равенства «=», как эквивалентные.

Пусть A и B — две кластеризованные ранжировки. *Пару объектов* (a, b) *назовем «противоречивой» относительно кластеризованных ранжировок* A *и* B*, если эти два элемента по-разному упорядочены в* A *и* B*, т.е.* $a < b$ в A *и* $a > b$ *в* B *(первый вариант противоречивости) либо* $a > b$ *в* A *и* $a < b$ *в* B *(второй вариант противоречивости).* Отметим, что в соответствии с этим определением пара объектов (a, b), эквивалентная хотя бы в одной кластеризованной ранжировке, не может быть противоречивой: эквивалентность $a = b$ не образует «противоречия» ни с $a < b$, ни с $a > b$. Это свойство оказывается полезным при выделении противоречивых пар.

В качестве примера рассмотрим, кроме A, еще две кластеризованные ранжировки:

$$B = [\{1, 2\} < \{3, 4, 5\} < 6 < 7 < 9 < \{8, 10\}],$$
$$C = [3 < \{1, 4\} < 2 < 6 < \{5, 7, 8\} < \{9, 10\}].$$

Совокупность противоречивых пар объектов для двух кластеризованных ранжировок A *и* B *назовем «ядром противоречий» и обозначим* $S(A, B)$*.* Для рассмотренных выше в качестве примеров трех кластеризованных ранжировок A, B и C, определенных на одном и том же носителе {1, 2, 3, ..., 10}, имеем:

$$S(A,B) = [(8, 9)], S(A,C) = [(1, 3), (2,4)],$$
$$S(B,C) = [(1, 3), (2, 3), (2, 4), (5, 6), (8,9)].$$

Как при ручном, так и при программном нахождении ядра можно в поисках противоречивых пар просматривать пары (1, 2), (1, 3), (1, 4), ..., (1, k), затем (2, 3), (2, 4), ..., (2, k), потом (3, 4), ..., (3, k), и т.д., вплоть до последней пары (k-1, k).

Пользуясь понятиями дискретной математики, «ядро противоречий» можно изобразить *графом* с вершинами в точках носителя. При этом *противоречивые пары задают ребра этого графа.* Граф для $S(A, B)$ имеет только одно ребро (одна связная компонента более чем из одной точки). Граф для $S(A, C)$ — 2 ребра (две связные компоненты более чем из одной точки). Граф для $S(B, C)$ — 5 ребер (три связные компоненты более чем из одной точки, а именно, {1, 2, 3, 4}, {5, 6} и {8, 9}).

Каждую кластеризованную ранжировку, как и любое бинарное отношение, можно задать матрицей $\|x(a, b)\|$ из 0 и 1 порядка k x k. При этом $x(a, b) = 1$ тогда и только тогда, когда $a < b$ либо $a = b$. В первом случае $x(b, a) = 0$, а во втором $x(b, a) = 1$. При этом хотя бы одно из чисел $x(a, b)$ и $x(b, a)$ равно 1. Из определения противоречивости пары (a, b) вытекает, что для нахождения всех таких пар достаточно поэлементно перемножить две матрицы $\|x(a, b)\|$ и $\|y(a, b)\|$, соответствующие двум кластеризованным ранжировкам, и отобрать те и только те пары, для которых $x(a, b)y(a, b) = x(b, a)y(b, a) = 0$.

Алгоритм согласования некоторого числа (двух или более) кластеризованных ранжировок состоят из трех этапов. На первом *выделяются противоречивые пары* объектов во всех парах кластеризованных ранжировок. На втором формируются кластеры итоговой кластеризованной ранжировки (т.е. классы эквивалентности — *связные компоненты графов*, соответствующих объединению попарных ядер противоречий). На третьем этапе эти *кластеры (классы эквивалентности) упорядочиваются*. Для установления порядка между кластерами произвольно выбирается один объект из первого кластера и второй — из второго, порядок между кластерами устанавливается такой же, какой имеет быть между выбранными объектами в любой из рассматриваемых кластеризованных ранжировок. (Если в одной из исходных кластеризованных ранжировок имеет быть равенство, а в другой — неравенство, то при построении итоговой кластеризованной ранжировки используется неравенство.)

Корректность подобного упорядочивания, т.е. его независимость от выбора той или иной пары объектов, вытекает из соответствующих теорем, доказанных в работе [32].

Два объекта из разных кластеров согласующей кластеризованной ранжировки могут оказаться эквивалентными в одной из исходных кластеризованных ранжировок (т.е. находиться в одном кластере). В таком случае надо рассмотреть упорядоченность этих объектов в какой-либо другой из исходных кластеризованных ранжировок. Если же во всех исходных кластеризованных ранжировках два рассматриваемых объекта находились в одном кластере, то естественно считать (и это является уточнением к этапу 3 алгоритма), что они находятся в одном кластере и в согласующей кластеризованной ранжировке.

Результат согласования кластеризованных ранжировок A, B, C, ... обозначим $f(A, B, C, ...)$. Тогда:

$$f(A, B) = [1 < 2 < 3 < 4 < 5 < 6 < 7 < \{8, 9\} < 10],$$
$$f(A, C) = [\{1, 3\} < \{2, 4\} < 6 < \{5, 7\} < 8 < 9 < 10],$$
$$f(B, C) = [\{1, 2, 3, 4\} < \{5, 6\} < 7 < \{8, 9\} < 10],$$
$$f(A, B, C) = f(B, C) = [\{1, 2, 3, 4\} < \{5, 6\} < 7 < \{8, 9\} < 10].$$

Итак, в случае $f(A, B)$ дополнительного изучения с целью упорядочения требуют только объекты 8 и 9. В случае $f(A, C)$ кластер $\{5, 7\}$ появился не потому, что относительно объектов 5 и 7 имеется противоречие, а потому, что в обеих исходных ранжировках эти объекты не различаются. В случае $f(B, C)$ четыре объекта с номерами 1, 2, 3, 4 объединились в один кластер, т.е. кластеризованные ранжировки оказались настолько противоречивыми, что процедура согласования не позволила провести достаточно полную декомпозицию задачи нахождения итогового мнения экспертов.

Обсудим некоторые свойства алгоритмов согласования.

1. Пусть $D = f(A, B, C, ...)$. Если $a < b$ в согласующей кластеризованной ранжировке D, то $a<b$ или $a=b$ в каждой из исходных ранжировок A, B, C, ..., причем хотя бы в одной из них справедливо строгое неравенство.

2. Построение согласующих кластеризованных ранжировок может осуществляться поэтапно. В частности,

$$f(A, B, C) = f(f(A, B), f(A, C), f(B, C)).$$

Ясно, что ядро противоречий для набора кластеризованных ранжировок является объединением таких ядер для всех пар рассматриваемых ранжировок.

3. Построение согласующих кластеризованных ранжировок нацелено на выделение общего упорядочения в исходных кластеризованных ранжировках. Однако при этом некоторые общие свойства исходных кластеризованных ранжировок могут теряться. Так, при согласовании ранжировок B и C, рассмотренных выше, противоречия в упорядочении элементов 1 и 2 не было — в ранжировке B эти объекты входили в один кластер, т.е. «1» = «2», в то время как «1» < «2» в кластеризованной ранжировке C. Значит, при их отдельном рассмотрении можно принять упорядочение «1» < «2». Однако в $f(B, C)$ они попали в один кластер, т.е. возможность их упорядочения исчезла. Это связано с поведением объекта 3, который «перескочил» в C на первое место и «увлек с собой в противоречие» пару (1, 2), образовав противоречивые пары и с 1, и с 2. Другими словами, связная компонента графа, соответствующего ядру противоречий, сама по себе не всегда является полным графом. Недостающие ребра при этом соответствуют парам типа (1, 2), которые сами по себе не являются противоречивыми, но «увлекаются в противоречие» другими парами.

4. Необходимость согласования кластеризованных ранжировок возникает, в частности, при разработке методики применения экспертных оценок в задачах экологического страхования и химической безопасности биосферы. Как уже говорилось, популярным является метод упорядочения по средним рангам, в котором итоговая ранжировка строится на основе средних арифметических рангов, выставленных отдельными экспертами [5, 35]. Однако из теории измерений известно (см. раздел 3.1), что более обоснованным является использование не средних арифметических, а медиан. Вместе с тем метод средних арифметических рангов весьма известен и широко применяется, так что просто отбросить его нецелесообразно. Поэтому было принято решение об одновременном применении обеих методов. Реализация этого решения потребовала разработки методики согласования двух указанных кластеризованных ранжировок.

5. Область применения рассматриваемого метода не ограничивается экспертными оценками. Он может быть использован, например, для сравнения качества математических моделей, используемых для описания процесса испарения жидкости. Имелись данные экспериментов и результаты расчетов по 8 математическим моделям. Сравнивать модели можно по различным критериям качества. Например, по сумме модулей относительных отклонений расчетных и экспериментальных значений. Можно действовать и по-другому. В каждой экспериментальной точке упорядочить модели по каче-

ству, а потом получить единые оценки методами средних рангов и медиан. Использовались и иные методы. Затем применялись методы согласования кластеризованных ранжировок, полученных различными способами. В результате оказалось возможным упорядочить модели по качеству и использовать это упорядочение при разработке банка математических моделей, используемого в задачах химической безопасности биосферы.

6. Рассматриваемый метод согласования кластеризованных ранжировок построен в соответствии с *методологией теории устойчивости* [1], согласно которой результат обработки данных, инвариантный относительно метода обработки, соответствует реальности, а результат расчетов, зависящий от метода обработки, отражает субъективизм исследователя, а не объективные соотношения.

Основные математические задачи анализа экспертных оценок. Ясно, что при анализе мнений экспертов можно применять самые разнообразные статистические методы, описывать их — значит описывать практически всю прикладную статистику. Тем не менее, можно выделить основные широко используемые в настоящее время методы математической обработки экспертных оценок — это проверка согласованности мнений экспертов (или классификация экспертов, если нет согласованности) и усреднение мнений экспертов внутри согласованной группы с целью формирования итогового мнения экспертной комиссии.

Поскольку ответы экспертов во многих процедурах экспертного опроса — не числа, а такие объекты нечисловой природы, как градации качественных признаков, ранжировки, разбиения, результаты парных сравнений, нечеткие предпочтения и т.д., то для их анализа оказываются полезными методы статистики нечисловых данных.

Почему ответы экспертов часто носят нечисловой характер? Наиболее общий ответ состоит в том, что люди не мыслят числами. В мышлении человека используются образы, слова, но не числа. Поэтому требовать от эксперта ответ в форме чисел — значит насиловать его разум. Даже в экономике менеджеры и предприниматели, принимая решения, лишь частично опираются на численные расчеты. Это видно из условного (т.е. определяемого произвольно принятыми соглашениями, обычно оформленными в виде нормативных актов и инструкций) характера балансовой прибыли, амортизационных отчислений и других экономических показателей. Поэтому фраза типа «фирма стремится к максимизации прибыли» не может иметь строго определенного смысла. Достаточно спросить: «Максимизация прибыли — за

какой период?» И сразу станет ясно, что степень оптимальности принимаемых решений зависит от горизонта планирования (на экономико-математическом уровне этот сюжет рассмотрен в монографии [1]).

Эксперт может сравнить два объекта, сказать, какой из двух лучше (метод парных сравнений), дать им оценки типа «хороший», «приемлемый», «плохой», упорядочить несколько объектов по привлекательности, но обычно не может ответить, во сколько раз или на сколько один объект лучше другого. Другими словами, ответы эксперта обычно измерены в порядковой шкале, или являются ранжировками, результатами парных сравнений и другими объектами нечисловой природы, но не числами. *Распространенное заблуждение состоит в том, что ответы экспертов стараются рассматривать как числа, занимаются «оцифровкой» их мнений, приписывая этим мнениям численные значения — баллы, которые потом обрабатывают с помощью различных методов прикладной статистики как результаты обычных физико-технических измерений.* В случае произвольности «оцифровки» выводы, полученные в результате подобной обработки данных, могут не иметь отношения к реальности. В связи с «оцифровкой» уместно вспомнить классическую притчу о человеке, который ищет потерянные ключи под фонарем, хотя потерял их в кустах. На вопрос, почему он так делает, отвечает: «Под фонарем светлее». Это, конечно, верно. Но, к сожалению, весьма малы шансы найти потерянные ключи под фонарем. Так и с «оцифровкой» нечисловых данных. Она дает возможность имитации научной деятельности и получения «научно обоснованных» результатов, но не возможность найти истину.

Проверка согласованности мнений экспертов и классификация экспертных мнений. Ясно, что мнения разных экспертов различаются. Важно понять, насколько велико это различие. Если мало — усреднение мнений экспертов позволит выделить то общее, что есть у всех экспертов, отбросив случайные отклонения в ту или иную сторону. Если велико — усреднение является чисто формальной процедурой. Так, если представить себе, что ответы экспертов равномерно покрывают поверхность бублика, то формальное усреднение укажет на центр дырки от бублика, а такого мнения не придерживается ни один эксперт. Из сказанного ясна важность проблемы проверки согласованности мнений экспертов.

Разработан ряд методов такой проверки. Статистические методы проверки согласованности зависят от математической природы ответов экспертов. Так, соответствующие статистические теории весьма трудны, если эти

ответы — ранжировки или разбиения, и достаточно просты, если ответы — результаты независимых парных сравнений. Отсюда вытекает рекомендация по организации экспертного опроса: не старайтесь сразу получить от эксперта ранжировку или разбиение, ему трудно это сделать, да и имеющиеся математические методы не позволяют далеко продвинуться в анализе подобных данных.

Например, рекомендуют проверять согласованность ранжировок с помощью коэффициента ранговой конкордации Кендалла — Смита. Но давайте вспомним, какая статистическая модель при этом используется. Проверяется нулевая гипотеза, согласно которой ранжировки независимы и равномерно распределены на множестве всех ранжировок. Если эта гипотеза принимается, то конечно, ни о какой согласованности мнений экспертов говорить нельзя. А если отклоняется? Тоже нельзя. Например, может быть два (или больше) центра, около которых группируются ответы экспертов. Нулевая гипотеза отклоняется. Но разве можно говорить о согласованности?

Эксперту гораздо легче на каждом шагу сравнивать только два объекта. Пусть он занимается парными сравнениями. *Непараметрическая теория парных сравнений* (*теория люсианов*) *позволяет решать более сложные задачи, чем статистика ранжировок или разбиений.* В частности, вместо гипотезы равномерного распределения можно рассматривать гипотезу однородности, т.е. вместо совпадения всех распределений с одним фиксированным (равномерным) можно проверять лишь совпадение распределений мнений экспертов между собой, что естественно трактовать как согласованность их мнений. Таким образом, удается избавиться от неестественного предположения равномерности.

При отсутствии согласованности экспертов естественно разбить их на группы сходных по мнению. Это можно сделать различными методами статистики объектов нечисловой природы, относящимися к кластер-анализу, предварительно введя метрику в пространство мнений экспертов. Как известно, идея американского математика Джона Кемени об аксиоматическом введении метрик нашла многочисленных продолжателей. Однако методы кластер-анализа обычно являются эвристическими. В частности, обычно невозможно с позиций статистической теории строго обосновать «законность» объединения двух кластеров в один. Имеется важное исключение — *для независимых парных сравнений* (*люсианов*) *разработаны методы, позволяющие проверять возможность объединения кластеров как статистическую гипотезу* (см. разделы 3.4 и 3.5 выше). Это — еще один аргумент за то, чтобы

рассматривать теорию люсианов как ядро математических методов экспертных оценок.

Нахождение итогового мнения комиссии экспертов. Пусть мнения комиссии экспертов или какой-то ее части признаны согласованными. Каково же итоговое (среднее, общее) мнение комиссии? Согласно идее Джона Кемени следует найти среднее мнение как решение *оптимизационной задачи*. А именно, надо минимизировать суммарное расстояние от кандидата в средние до мнений экспертов. Найденное таким способом среднее мнение называют «медианой Кемени».

Математическая сложность состоит в том, что мнения экспертов лежат в некотором пространстве объектов нечисловой природы. Общая теория подобного усреднения рассмотрена выше (глава 2). В частности, показано, что в силу закона больших чисел (в пространствах произвольной природы) среднее мнение при увеличении числа экспертов (чьи мнения независимы и одинаково распределены) приближается к некоторому пределу, который, как известно, является *математическим ожиданием* (случайного элемента, имеющего то же распределение, что и ответы экспертов).

В конкретных пространствах нечисловых мнений экспертов вычисление медианы Кемени может быть достаточно сложным делом. Кроме свойств самого пространства, велика роль конкретных метрик. Так, в пространстве ранжировок при использовании метрики, связанной с коэффициентом ранговой корреляции Кендалла, необходимо проводить достаточно сложные расчеты, в то время как применение показателя различия на основе коэффициента ранговой корреляции Спирмена приводит к упорядочению по средним арифметическим рангам [1].

Бинарные отношения и расстояние Кемени. Как известно, бинарное отношение A на конечном множестве $Q = \{q_1, q_2, ..., q_k\}$ — это подмножество *декартова квадрата* $Q^2 = \{(q_m, q_n), m,n = 1, 2, \dots, k\}$. При этом пара (q_m, q_n) входит в соответствующее бинарному отношению A подмножество тогда и только тогда, когда между q_m и q_n имеется рассматриваемое отношение A.

Напомним, что каждую кластеризованную ранжировку, как и любое бинарное отношение, можно задать квадратной матрицей $\|x(a, b)\|$ из 0 и 1 порядка k Ч k. При этом $x(a, b) = 1$ тогда и только тогда, когда $a < b$ либо $a = b$. В первом случае $x(b, a) = 0$, а во втором $x(b, a) = 1$. При этом хотя бы одно из чисел $x(a, b)$ и $x(b, a)$ равно 1.

В экспертных методах используют, в частности, такие бинарные отношения, как ранжировки (упорядочения, т.е. разбиения на группы, между ко-

торыми имеется строгий порядок), отношения эквивалентности, толерантности (отношения сходства). Как следует из сказанного выше, каждое бинарное отношение A можно описать матрицей $\| a(i, j) \|$ из 0 и 1, причем $a(i, j) = 1$ тогда и только тогда, когда q_i и q_j находятся в отношении A, и $a(i, j) = 0$ в противном случае.

Определение. *Расстоянием Кемени между бинарными отношениями A и B, описываемыми матрицами* $\|a(i, j)\|$ и $\|b(i, j)\|$ *соответственно, называется число:*

$$d(A, B) = \sum \left| a(i, j) - b(i, j) \right|,$$

где суммирование производится по всем i, j от 1 *до k, т.е. расстояние Кемени между бинарными отношениями равно сумме модулей разностей элементов, стоящих на одних и тех же местах в соответствующих им матрицах.*

Легко видеть, что расстояние Кемени — это число несовпадающих элементов в матрицах $\|a(i, j)\|$ и $\|b(i, j)\|$.

Расстояние Кемени основано на некоторой системе аксиом. Эта система аксиом и вывод из нее формулы для расстояния Кемени между упорядочениями содержится в книге [33]. Она сыграла большую роль в развитии в нашей стране такого научного направления, как анализ нечисловой информации (см. историю вопроса в монографиях [1, 5], а также в главе 1). В дальнейшем под влиянием работ Дж. Кемени были предложены различные системы аксиом для получения расстояний в тех или иных нужных для социально-экономических, технических, медицинских и иных исследований пространствах (см. раздел 1.8).

Медиана Кемени и законы больших чисел. С помощью расстояния Кемени находят итоговое мнение комиссии экспертов. Пусть A_1, A_2, A_3, ... , A_p — ответы p экспертов, представленные в виде бинарных отношений. Для их усреднения используют *медиану Кемени:*

$$\text{Arg min} \sum d(A_i, A),$$

где *Arg min* — то или те значения A, при которых достигает минимума указанная сумма расстояний Кемени от ответов экспертов до текущей переменной A, по которой и проводится минимизация. Таким образом,

$$\sum d(A_i, A) = d(A_1, A) + d(A_2, A) + d(A_3, A) + \ ... \ + d(A_p, A).$$

Кроме медианы Кемени, используют *среднее по Кемени*, в котором вместо $d(A_i, A)$ стоит $d^2(A_i, A)$. Среднее по Кемени рассматривается в книге [33].

Медиана Кемени — частный случай определения эмпирического среднего в пространствах нечисловой природы. Для нее справедлив закон больших чисел, т.е. эмпирическое среднее приближается при росте числа составляющих (т.е. p — числа слагаемых в сумме), к теоретическому среднему:

$$Arg\ min \sum d(A_i, A) \rightarrow Arg\ min\ M\ d(A_1, A)\ .$$

Здесь M — символ математического ожидания. Предполагается, что есть основания рассматривать ответы p экспертов A_1 , A_2, $A_3, \ldots, A_p$ как независимые одинаково распределенные случайные элементы (т.е. как случайную выборку) в соответствующем пространстве нечисловой природы, например, в пространстве упорядочений или отношений эквивалентности. Систематически эмпирические и теоретические средние и соответствующие различные варианты законов больших чисел рассмотрены выше в главе 2.

Законы больших чисел показывают, во-первых, что медиана Кемени обладает *устойчивостью* по отношению к незначительному изменению состава экспертной комиссии; во-вторых, при увеличении числа экспертов она *приближается к некоторому пределу*. Его естественно рассматривать как *истинное мнение* экспертов, от которого каждый из них несколько отклонялся по случайным причинам.

Вычисление медианы Кемени — задача целочисленного программирования. Для ее нахождения используется различные алгоритмы дискретной математики, в частности, основанные на методе ветвей и границ. Применяют также алгоритмы, основанные на идее случайного поиска, поскольку для каждого бинарного отношения нетрудно найти множество его соседей.

Рассмотрим упрощенный пример вычисления медианы Кемени. Пусть дана квадратная матрица (порядка 9) попарных расстояний для множества бинарных отношений из 9 элементов A_1 , A_2 , A_3 , ..., A_9 (см. табл. 3). Пусть требуется найти в этом множестве *медиану* для множества из 5 элементов $\{A_2, A_4, A_5, A_8, A_9\}$.

Таблица 3

Матрица попарных расстояний

Элементы	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9
A_1	0	2	13	1	7	4	10	3	11
A_2	2	0	5	6	1	3	2	5	1

Элементы	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9
A_3	13	5	0	2	2	7	6	5	7
A_4	1	6	2	0	5	4	3	8	8
A_5	7	1	2	5	0	10	1	3	7
A_6	4	3	7	4	10	0	2	1	5
A_7	10	2	6	3	1	2	0	6	3
A_8	3	5	5	8	3	1	6	0	9
A_9	11	1	7	8	7	5	3	9	0

В соответствии с определением медианы Кемени следует ввести в рассмотрение функцию:

$$C(A) = \sum d(A_i, A) = d(A_2, A)+d(A_4, A)+d(A_5, A)+d(A_8, A)+d(A_9, A),$$

рассчитать ее значения для всех A_1, A_2, A_3, ..., A_9 и выбрать наименьшее. Проведем расчеты:

$$C(A_1) = d(A_2, A_1)+d(A_4, A_1)+d(A_5, A_1)+d(A_8, A_1)+d(A_9, A_1) = \\ = 2 + 1 +7 +3 +11 = 24,$$

$$C(A_2) = d(A_2, A_2)+d(A_4, A_2)+d(A_5, A_2)+d(A_8, A_2)+d(A_9, A_2) = \\ = 0 + 6 + 1 + 5 + 1 = 13,$$

$$C(A_3) = d(A_2, A_3)+d(A_4, A_3)+d(A_5, A_3)+d(A_8, A_3)+d(A_9, A_3) = \\ = 5 + 2 + 2 + 5 +7 = 21,$$

$$C(A_4) = d(A_2, A_4)+d(A_4, A_4)+d(A_5, A_4)+d(A_8, A_4)+d(A_9, A_4) = \\ = 6 + 0 + 5 + 8 + 8 = 27,$$

$$C(A_5) = d(A_2, A_5)+d(A_4, A_5)+d(A_5, A_5)+d(A_8, A_5)+d(A_9, A_5) = \\ = 1 + 5 + 0 +3 + 7 = 16,$$

$$C(A_6) = d(A_2, A_6)+d(A_4, A_6)+d(A_5, A_6)+d(A_8, A_6)+d(A_9, A_6) = \\ = 3 + 4 + 10 + 1 + 5 = 23,$$

$$C(A_7) = d(A_2, A_7)+d(A_4, A_7)+d(A_5, A_7)+d(A_8, A_7)+d(A_9, A_7) = \\ = 2 + 3 +1 + 6 + 3 = 15,$$

$$C(A_8) = d(A_2, A_8)+d(A_4, A_8)+d(A_5, A_8)+d(A_8, A_8)+d(A_9, A_8) = \\ = 5 + 8 + 3 + 0 +9 = 25,$$

$$C(A_9) = d(A_2, A_9)+d(A_4, A_9)+d(A_5, A_9)+d(A_8, A_9)+d(A_9, A_9) = \\ = 1 + 8 + 7 + 9 + 0 = 25.$$

Из всех вычисленных сумм наименьшая равна 13, и достигается она при $A=A_2$, следовательно, медиана Кемени — это множество $\{A_2\}$, состоящее из одного элемента A_2.

Дадим обзор недавних публикаций по тематике настоящей главы. Подробное доказательство теорем о характеризация средних величин шкалами измерения содержится в статье [36]. Современный этап развития теории люсианов отражен в [37]. Статистика нечетких данных развита в [38].

Методу проверки гипотез по совокупности малых выборок и его применению в теории статистического контроля посвящена статья [39]. Обзор развития теории экспертных оценок в нашей стране дан в [40]. Новая парадигма анализа статистических и экспертных данных в задачах экономики и управления представлена в работе [41]. О развитии теории принятия решений как научной области и экспертных оценок как ее части рассказано в [42].

Конкретным методам сбора и анализа экспертных оценок посвящено достаточно много статей. Укажем несколько. Экспертные технологии применяются [43] при оценивании вероятностей редких событий (в связи с разработкой автоматизированной системы прогнозирования и предотвращения авиационных происшествий). Методика проведения анализа экспертных упорядочений с целью построения итогового мнения комиссии экспертов развита в [44]. Исследование итогового ранжирования мнений группы экспертов с помощью медианы Кемени с целью оценки кредитного риска проведено в [45, 46]. Весь спектр экспертных технологий применяется для определения приоритетности реализации НИОКР на предприятиях ракетно-космической отрасли [47].

Темы докладов, рефератов, исследовательских работ

1. Показатели разброса, связи, показатели различия (в том числе метрики) в порядковой шкале.

2. Ранговые методы математической статистики как инвариантные методы анализа порядковых данных.

3. Показатели разброса, связи, показатели различия (в том числе метрики) в шкале интервалов.

4. Показатели разброса, связи, показатели различия (в том числе метрики) в шкале отношений.

5. Теорема В. В. Подиновского: любое изменение коэффициентов весомости единичных показателей качества продукции приводит к изменению упорядочения изделий по средневзвешенному показателю.

6. Рассчитайте мощность статистик W и N, рассматриваемых в теории равномерно распределенных случайных толерантностей.

7. Изучите распределение при альтернативах статистики T, используемой для проверки однородности двух групп люсианов (при безграничном росте объемов групп).

8. Несмещенные оценки в прикладной статистике.

9. Применение метода проверки гипотез по совокупности малых выборок в задачах обнаружения эффекта и проверки однородности [5, гл.4].

10. По данным примера в разделе 3.5 найдите методом наименьших квадратов взаимное положение четырех нефтяных компаний на оси «качество бензина», т.е. найдите их «ценности» V_1, V_2, V_3, V_4.

11. Методы оценивания функции принадлежности нечеткого множества.

12. Описание данных для выборок, элементы которых — нечеткие множества.

13. Регрессионный анализ нечетких переменных.

14. Непараметрические оценки плотности распределения вероятностей в пространстве нечетких множеств.

15. Классификация мнений экспертов и проверка согласованности.

16. Использование люсианов в теории и практике экспертных оценок.

17. Математические методы формирования итогового мнения комиссии экспертов.

Контрольные вопросы и задачи

1. Какие средние величины целесообразно использовать при расчете средней заработной платы (или среднего дохода)?

2. Постройте пример, показывающий некорректность использования среднего арифметического $f(X_1, X_2) = (X_1 + X_2) / 2$ в порядковой шкале, используя допустимое преобразование $g(x) = x^2$ (при положительных усредняемых величинах x).

3. Постройте пример, показывающий некорректность использования среднего геометрического в порядковой шкале. Другими словами, приведите пример чисел x_1, x_2, y_1, y_2 и строго возрастающего преобразования f: $R^1 \to R^1$ таких, что

$$(x_1 x_2)^{1/2} < (y_1 y_2)^{1/2}, \quad [f(x_1)f(x_2)]^{1/2} > [f(y_1)f(y_2)]^{1/2}.$$

4. Приведите пример чисел x_1, x_2, y_1, y_2 и строго возрастающего преобразования f: $R^1 \to R^1$ таких, что

$$[(x_1)^2 + (x_2)^2]^{1/2} < [(y_1)^2 + (y_2)^2]^{1/2},$$
$$[(f(x_1))^2 + (f(x_2))^2]^{1/2} > [(f(y_1))^2 + (f(y_2))^2]^{1/2}.$$

5. Как случайные толерантности используются в теории нечетких толерантностей?

6. В теории люсианов (раздел 3.4) выведите из общего вида несмещенной оценки многочлена от p по результатам m независимых испытаний Бернулли с вероятностью успеха p в каждом (формула (12)) несмещенную оценку в случае $f(p) = 2p(1 - p)$ (формула (13)).

7. Выпишите несмещенную оценку для функции $f(p) = p^3 - 3p^2 + 2p$, где p — параметр биномиального распределения.

8. Как можно проводить кластерный анализ совокупности нечетких множеств?

9. Чем метод средних арифметических рангов отличает от метода медиан рангов?

10. Почему необходимо согласование кластеризованных ранжировок и как оно проводится?

11. В чем состоит проблема согласованности ответов экспертов?

12. Как бинарные отношения используются в экспертизах?

13. Как бинарные отношения описываются матрицами из 0 и 1?

14. Что такое расстояние Кемени и медиана Кемени?

15. Чем закон больших чисел для медианы Кемени отличается от «классического» закона больших чисел, известного в статистике?

16. В таблице приведены упорядочения 7 инвестиционных проектов, представленные 7 экспертами.

Таблица к задаче 16

Упорядочения проектов экспертами

Эксперты	Упорядочения
1	1 < {2, 3} < 4 < 5 < {6, 7}
2	{1, 3} < 4 < 2< 5< 7 < 6
3	1 < 4 < 2 < 3 < 6 < 5 < 7
4	1 < {2, 4} < 3 < 5 < 7 <6
5	2 < 3 < 4 < 5 <1 <6 <7
6	1 < 3 < 2 < 5 < 6 < 7 < 4
7	1 < 5 < 3 < 4 < 2 < 6 < 7

Найдите:

а) итоговое упорядочение по средним арифметическим рангам;

б) итоговое упорядочение по медианам рангов;

в) кластеризованную ранжировку, согласующую эти два упорядочения.

17. Выпишите матрицу из 0 и 1, соответствующую бинарному отношению (кластеризованной ранжировке) $5 < \{1, 3\} < 4 < 2 < \{6, 7\}$.

18. Найдите расстояние Кемени между бинарными отношениями — упорядочениями $A = [3 < 2 < 1 < \{4, 5\}]$ и $B = [1 < \{2, 3\} < 4 < 5]$.

19. Дана квадратная матрица (порядка 9) попарных расстояний (мер различия) для множества бинарных отношений из 9 элементов $A_1, A_2, A_3, \ldots, A_9$. Найдите в этом множестве медиану для множества из 5 элементов $\{A_2, A_3, A_5, A_6, A_9\}$.

Таблица к задаче 19

Попарные расстояния между бинарными отношениями

0	5	3	6	7	4	10	3	11
5	0	5	6	10	3	2	5	7
3	5	0	8	2	7	6	5	7
6	6	8	0	5	4	3	8	8
7	10	2	5	0	10	8	3	7
4	3	7	4	10	0	2	3	5
10	2	6	3	8	2	0	6	3
3	5	5	8	3	3	6	0	9
11	7	7	8	7	5	3	9	0

Литература

1. *Орлов, А. И.* Устойчивость в социально-экономических моделях / А. И. Орлов. — Москва : Наука, 1979. — 296 с.

2. *Колмогоров, А. Н.* Об определении среднего / А. Н. Колмогоров // Избранные труды. Математика и механика. — Москва : Наука, 1985. — С. 136–138.

3. *Орлов, А. И.* Связь между средними величинами и допустимыми преобразованиями шкалы / А. И. Орлов // Математические заметки. — 1981. — Т. 30. — № 4. — С. 561–568.

4. *Пфанцагль, И.* Теория измерений / И. Пфанцагль. — Москва : Мир, 1976. — 165 с.

5. *Орлов, А. И.* Эконометрика : учебник для вузов / А. И. Орлов. —3-е изд., испр. и доп. — Москва : Экзамен, 2004. — 576 с.

6. *Толстова, Ю. Н.* Измерение в социологии / Ю. Н. Толстова. — Москва : Инфра-М, 1998. — 352 с.

7. *Кендэл, М.* Ранговые корреляции / М. Кендэл. — Москва : Статистика, 1975. — 216 с.

8. *Маамяги, А. В.* Некоторые задачи статистического анализа классификаций / А. В. Маамяги. — Таллинн : АН ЭССР, 1982. — 24 с.

9. *Большев, Л. Н.* Таблицы математической статистики / Л. Н. Большев, Н. В. Смирнов. — 3-е изд. — Москва : Наука, 1983. — 474 с.

10. *Крамер, Г.* Математические методы статистики / Г. Крамер. — Москва : Мир, 1975. — 648 с.

11. *Уилкс, С.* Математическая статистика / С. Уилкс. — Москва : Наука, 1967. — 632 с.

12. *Кендалл, М. Дж.* Статистические выводы и связи / М. Дж. Кендалл, А. Стьюарт. — Москва : Наука, 1973. — 900 с.

13. *Орлов, А. И.* Парные сравнения в асимптотике Колмогорова / А. И. Орлов // Экспертные оценки в задачах управления. — Москва : Изд-во Института проблем управления АН СССР, 1982. — С. 58–66.

14. *Орлов, А. И.* Статистика объектов нечисловой природы и экспертные оценки / А. И. Орлов // Экспертные оценки. Вопросы кибернетики. — Вып. 58. — Москва : Научный Совет АН СССР по комплексной проблеме «Кибернетика», 1979. — С. 17–33.

15. *Орлов, А. И.* Случайные множества с независимыми элементами (люсианы) и их применения / А. И. Орлов // Алгоритмическое и программное обеспечение прикладного статистического анализа : сборник статей. — Т. 36. — Москва : Наука, 1980. — С. 287–308.

16. *Рыданова, Г. В.* Некоторые вопросы статистического анализа случайных бинарных векторов : диссертация на соискание ученой степени кандидата физико-математических наук / Г. В. Рыданова. — Москва : МГУ им. М. В. Ломоносова, 1987. — 139 с.

17. Кинетотопография в диагностике инфаркта миокарда / Г. А. Аксенова, Е. С. Кузьмина, А. И. Орлов, Н. К. Розова // Актуальные вопросы клинической и экспериментальной медицины. — Москва : 4-е Главное Управление при Минздраве СССР, 1979. — С. 24–26.

18. Кинетокардиография в определении зон асинергии у больных инфарктом миокарда / В. Г. Попов, Г. А. Аксенова, А. И. Орлов [и др.] // Клиническая медицина. — 1982. — Т. LX. — № 3. — С. 25–30.

19. Методические рекомендации по проведению экспертной оценки планируемых и законченных научных работ в области медицины (по проблемам союзного значения) / составители Г. В. Раушенбах, О. В. Филиппов. — Москва : АМН СССР — Ученый медицинский совет Минздрава СССР, 1982. — 36 с.

20. *Леман, Э.* Проверка статистических гипотез / Э. Леман. — Москва : Наука, 1979. — 408 с.

21. *Боровков, А. А.* Математическая статистика : учебное пособие для вузов / А. А. Боровков. — Москва : Наука, 1984. — 472 с.

22. *Лумельский, Я. П.* Статистические оценки результатов контроля качества / Я. П. Лумельский. — Москва : Изд-во стандартов, 1979. — 200 с.

23. *Любищев, А. А.* Дисперсионный анализ в биологии / А. А. Любищев. — Москва : Изд-во МГУ им. М. В. Ломоносова, 1986. — 200 с.

24. *Дылько, Т. Н.* Проверка гипотез в экспертном оценивании / Т. Н. Дылько // Вестник Белорусского государственного университета. — Сер. 1. Физика, математика и механика. — 1988. — № 2. — С. 36–40.

25. *Орлов, А. И.* Метрика подобия: аксиоматическое введение, асимптотическая нормальность / А. И. Орлов, Г. В. Раушенбах // Статистические методы оценивания и проверки гипотез : межвузовский сборник научных трудов. — Пермь : Изд-во ПГНИУ, 1986. — С. 148–157.

26. Рекомендации. Прикладная статистика. Методы обработки данных. Основные требования и характеристики / А. И. Орлов, Н. Г. Миронова, В. Н. Фомин, А. Н. Черчинцев. — Москва : ВНИИ стандартизации Госстандарта СССР, 1987. — 62 с.

27. *Тюрин, Ю. Н.* К проблеме обработки рядов ранжировок / Ю. Н. Тюрин, А. П. Василевич // Статистические методы анализа экспертных оценок. — Т. 29. — Москва : Наука, 1977. — С. 96–111.

28. *Орлов, А. И.* Некоторые вероятностные вопросы теории классификации / А. И. Орлов // Прикладная статистика. — Т. 45. — Москва : Наука, 1983. — С. 166–179.

29. *Дэвид, Г.* Метод парных сравнений / Г. Дэвид. — Москва : Статистика, 1978. — 144 с.

30. *Орлов, А. И.* Задачи оптимизации и нечеткие переменные / А. И. Орлов. — Москва : Знание, 1980. — 64 с.

31. *Орлов, А. И.* Сходимость эталонных алгоритмов / А. И. Орлов // Прикладной многомерный статистический анализ. — Т. 33. Ученые записки по статистике. — Москва : Наука, 1978. — С. 361–364.

32. *Горский, В. Г.* Метод согласования кластеризованных ранжировок / В. Г. Горский, А. А. Гриценко, А. И. Орлов // Автоматика и телемеханика. — 2000. — № 3. — С. 159–167.

33. *Кемени, Дж.* Кибернетическое моделирование: Некоторые приложения / Дж. Кемени, Дж. Снелл. — Москва : Советское радио, 1972. — 192 с.

34. *Шрейдер, Ю. А.* Равенство, сходство, порядок / Ю. А. Шрейдер. — Москва : Наука, 1971. — 256 с.

35. Менеджмент / под редакцией Ж. В. Прокофьевой. — Москва : Знание, 2000. — 288 с.

36. *Орлов, А. И.* Характеризация средних величин шкалами измерения / А. И. Орлов // Научный журнал КубГАУ. — 2017. — № 134. — С. 877–907.

37. *Орлов, А. И.* Теория люсианов / А. И. Орлов // Научный журнал КубГАУ. — 2014. — № 101. — С. 275–304.

38. *Орлов, А. И.* Статистика нечетких данных / А. И. Орлов // Научный журнал КубГАУ. — 2016. — № 119. — С. 75–91.

39. *Орлов, А. И.* Метод проверки гипотез по совокупности малых выборок и его применение в теории статистического контроля / А. И. Орлов // Научный журнал КубГАУ. — 2014. — № 104. — С. 38–52.

40. *Орлов, А. И.* Теория экспертных оценок в нашей стране / А. И. Орлов // Научный журнал КубГАУ. — 2013. — № 93. — С. 1–11.

41. *Орлов, А. И.* Новая парадигма анализа статистических и экспертных данных в задачах экономики и управления / А. И. Орлов // Научный журнал КубГАУ. — 2014. — № 98. — С. 1254–1260.

42. *Орлов, А. И.* О развитии теории принятия решений и экспертных оценок / А. И. Орлов // Научный журнал КубГАУ. — 2021. — № 167. — С. 177–198.

43. *Орлов, А. И.* Экспертные технологии и их применение при оценивании вероятностей редких событий / А. И. Орлов, Ю. Г. Савинов, А. Ю. Богданов // Заводская лаборатория. Диагностика материалов. — 2014. — Т. 80. — № 3. — С. 63–69.

44. *Орлов, А. И.* Анализ экспертных упорядочений / А. И. Орлов // Научный журнал КубГАУ. — 2015. — № 112. — С. 21–51.

45. *Жуков, М. С.* Задача исследования итогового ранжирования мнений группы экспертов с помощью медианы Кемени / М. С. Жуков, А. И. Орлов // Научный журнал КубГАУ. — 2016. — № 122. — С. 785–806.

46. *Жуков, М. С.* Экспертные оценки в рисках / М. С. Жуков, А. И. Орлов, С. Г. Фалько // Контроллинг. — 2017. — № 4 (66). — С. 24–27.

47. *Орлов, А. И.* Определение приоритетности реализации НИОКР на предприятиях ракетно-космической отрасли / А. И. Орлов, А. Д. Цисарский // Контроллинг. — 2020. — № 2 (76). — С. 58–65.

ГЛАВА 4. СТАТИСТИКА ИНТЕРВАЛЬНЫХ ДАННЫХ

В статистике интервальных данных элементы выборки — не числа, а интервалы. Это приводит к алгоритмам и выводам, принципиально отличающимся от классических. Настоящая глава посвящена основным идеям и подходам асимптотической статистики интервальных данных. Приведены результаты, связанные с основополагающими в рассматриваемой области прикладной математической статистики понятиями нотны и рационального объема выборки. Рассмотрен ряд задач оценивания характеристик и параметров распределения, проверки гипотез, регрессионного, кластерного и дискриминантного анализов.

4.1. ОСНОВНЫЕ ИДЕИ СТАТИСТИКИ ИНТЕРВАЛЬНЫХ ДАННЫХ

Перспективная и быстро развивающаяся область статистических исследований последних лет — математическая статистика интервальных данных. Речь идет о развитии методов прикладной математической статистики в ситуации, когда статистические данные — не числа, а интервалы, в частности, порожденные наложением ошибок измерения на значения случайных величин. Полученные результаты отражены, в частности, в выступлениях на проведенной в «Заводской лаборатории» дискуссии [1] и в докладах международной конференции ИНТЕРВАЛ-92 [2]. Приведем основные идеи весьма перспективного для вероятностно-статистических методов и моделей принятия решений асимптотического направления в статистике интервальных данных.

В настоящее время признается необходимым изучение устойчивости (робастности) оценок параметров к малым отклонениям исходных данных и предпосылок модели. Однако популярная среди теоретиков модель засорения (Тьюки — Хьюбера) представляется не вполне адекватной. Эта модель нацелена на изучение влияния больших «выбросов». Поскольку любые реальные измерения лежат в некотором фиксированном диапазоне, а именно, заданном в техническом паспорте средства измерения, то зачастую выбросы не могут быть слишком большими. Поэтому представляются полезными иные, более общие схемы устойчивости, введенные в монографии [3], в которых, например, учитываются отклонения распределений результатов наблюдений от предположений модели.

В одной из таких схем изучается влияние интервальности исходных данных на статистические выводы. Необходимость такого изучения стала оче-

видной следующим образом. В государственных стандартах СССР по прикладной статистике в обязательном порядке давалось справочное приложение «Примеры применения правил стандарта». При разработке ГОСТ 11.011-83 (в настоящее время отменен как нормативный документ, но может использоваться как научная публикация) [4] были переданы для анализа реальные данные о наработке резцов до предельного состояния (в часах). Оказалось, что все эти данные представляли собой либо целые числа, либо полуцелые (т.е. после умножения на 2 становящиеся целыми). Ясно, что исходная длительность наработок искажена. Необходимо учесть в статистических процедурах наличие такого искажения исходных данных. Как это сделать?

Первое, что приходит в голову — модель группировки данных, согласно которой для истинного значения X проводится замена на ближайшее число из множества $\{0{,}5n,\ n = 1, 2, 3, ...\}$. Однако эту модель целесообразно подвергнуть сомнению, а также рассмотреть иные модели. Так, возможно, что X надо приводить к ближайшему сверху элементу указанного множества — если проверка качества поставленных на испытание резцов проводилась раз в полчаса. Другой вариант: если расстояния от X до двух ближайших элементов множества $\{0{,}5n,\ n = 1, 2, 3, ...\}$ примерно равны, то естественно ввести рандомизацию при выборе заменяющего числа, и т.д.

Целесообразно построить новую математико-статистическую модель, согласно которой **результаты наблюдений — не числа, а интервалы**. Например, если в таблице приведено значение 53,5, то это значит, что реальное значение — какое-то число от 53,0 до 54,0, т.е. какое-то число в интервале [53,5 – 0,5; 53,5 + 0,5], где 0,5 — максимально возможная погрешность. Принимая эту модель, мы попадаем в новую научную область — статистику интервальных данных [5,6]. Статистика интервальных данных идейно связана с интервальной математикой, в которой в роли чисел выступают интервалы (см., например, монографию [7]). Это направление математики является дальнейшим развитием всем известных правил приближенных вычислений, посвященных выражению погрешностей суммы, разности, произведения, частного через погрешности тех чисел, над которыми осуществляются перечисленные операции.

В интервальной математике сумма двух интервальных чисел $[a, b]$ и $[c, d]$ имеет вид $[a, b] + [c, d] = [a + c, b + d]$, а разность определяется по формуле $[a, b] - [c, d] = [a - d, b - c]$. Для положительных a, b, c, d произведение определяется формулой $[a, b] \times [c, d] = [ac, bd]$, а частное имеет вид $[a, b] / [c, d] = [a / d, b / c]$. Эти формулы получены при решении соответствующих

оптимизационных задач. Пусть x лежит в отрезке $[a, b]$, а y — в отрезке $[c, d]$. Каково минимальное и максимальное значение для $x + y$? Очевидно, $a + c$ и $b + d$ соответственно. Минимальные и максимальные значения для $x - y$, xy, x / y указывают нижние и верхние границы для интервальных чисел, задающих результаты арифметических операций. А от арифметических операций можно перейти ко всем остальным математическим алгоритмам. Так строится интервальная математика.

Как видно из сборника трудов Международной конференции [2], исследователям удалось решить, в частности, ряд задач теории интервальных дифференциальных уравнений, в которых коэффициенты, начальные условия и решения описываются с помощью интервалов. По мнению ряда специалистов, статистика интервальных данных является частью интервальной математики [7]. Впрочем, есть точка зрения, согласно которой такое включение нецелесообразно, поскольку статистика интервальных данных использует несколько иные подходы к алгоритмам анализа реальных данных, чем сложившиеся в интервальной математике (подробнее см. ниже).

В настоящей главе развиваем асимптотические методы статистического анализа интервальных данных при больших объемах выборок и малых погрешностях измерений. В отличие от классической математической статистики, сначала устремляется к бесконечности объем выборки и только потом — уменьшаются до нуля погрешности. В частности, еще в начале 1980-х гг. с помощью такой асимптотики были сформулированы правила выбора метода оценивания в ГОСТ 11.011-83 (в настоящее время отменен как нормативный документ, но может использоваться как научная публикация) [4].

Разработана [8] общая схема исследования, включающая расчет нотны (максимально возможного отклонения статистики, вызванного интервальностью исходных данных) и рационального объема выборки (превышение которого не дает существенного повышения точности оценивания). Она применена к оцениванию математического ожидания и дисперсии [1], медианы и коэффициента вариации [9], параметров гамма-распределения [4, 10] и характеристик аддитивных статистик [8], при проверке гипотез о параметрах нормального распределения, в т.ч. с помощью критерия Стьюдента, а также гипотезы однородности с помощью критерия Смирнова [9]. Изучено асимптотическое поведение оценок метода моментов и оценок максимального правдоподобия (а также более общих — оценок минимального контраста), проведено асимптотическое сравнение этих методов в случае интервальных данных, найдены общие условия, при которых, в отличие от классической

математической статистики, метод моментов дает более точные оценки, чем метод максимального правдоподобия [11].

Разработаны подходы к рассмотрению интервальных данных в основных постановках регрессионного, дискриминантного и кластерного анализов [12]. В частности, изучено влияние погрешностей измерений и наблюдений на свойства алгоритмов регрессионного анализа, разработаны способы расчета нотн и рациональных объемов выборок, введены и исследованы новые понятия многомерных и асимптотических нотн, доказаны соответствующие предельные теоремы [12, 13]. Начата разработка интервального дискриминантного анализа, в частности, рассмотрено влияние интервальности данных на показатель качества классификации [12, 14]. Основные идеи и результаты рассматриваемого направления в статистике интервальных данных приведены в публикациях обзорного характера [5, 6].

Как показала, в частности, международная конференция ИНТЕРВАЛ-92, в области асимптотической математической статистики интервальных данных мы имеем мировой приоритет. По нашему мнению, со временем во все виды статистического программного обеспечения должны быть включены алгоритмы интервальной статистики, «параллельные» обычно используемым алгоритмам прикладной математической статистики. Это позволит в явном виде учесть наличие погрешностей у результатов наблюдений, сблизить позиции метрологов и статистиков.

Многие из утверждений статистики интервальных данных весьма отличаются от аналогов из классической математической статистики. В частности, не существует состоятельных оценок; средний квадрат ошибки оценки, как правило, асимптотически равен сумме дисперсии оценки, рассчитанной согласно классической теории, и некоторого положительного числа (равного квадрату так называемые *нотны* — максимально возможного отклонения значения статистики из-за погрешностей исходных данных) — в результате метод моментов оказывается иногда точнее метода максимального правдоподобия [11]; нецелесообразно увеличивать объем выборки сверх некоторого предела (называемого рациональным объемом выборки) — вопреки классической теории, согласно которой чем больше объем выборки, тем точнее выводы.

В стандарт [4] был включен раздел 5, посвященный выбору метода оценивания при неизвестных параметрах формы и масштаба и известном параметре сдвига и основанный на концепциях статистики интервальных дан-

ных. Теоретическое обоснование этого раздела стандарта опубликовано лишь через 5 лет в статье [10].

Следует отметить, что хотя в 1982 г. при разработке стандарта [4] были сформулированы основные идеи статистики интервальных данных, однако из-за недостатка времени они не были полностью реализованы в ГОСТ 11.011-83 (в настоящее время отменен как нормативный документ, но может использоваться как научная публикация), и этот стандарт написан в основном в классической манере. Развитие идей статистики интервальных данных продолжается уже в течение 25 лет, и еще многое необходимо сделать! Большое значение статистики интервальных данных для современной прикладной статистики обосновано в [15, 16].

Ведущая научная школа в области статистики интервальных данных — это школа проф. А. П. Вощинина, активно работающая с конца 1970-х гг. Полученные результаты отражены в ряде монографий (см. прежде всего [17, 18, 19]), статей [1, 20, 21], докладов, в частности, в трудах [2] Международной конференции ИНТЕРВАЛ-92, диссертациях [22, 23]. В частности, изучены проблемы регрессионного анализа, планирования эксперимента, сравнения альтернатив и принятия решений в условиях интервальной неопределенности.

Рассматриваемое ниже наше научное направление отличается нацеленностью на асимптотические результаты, полученные при больших объемах выборок и малых погрешностях измерений, поэтому оно и названо **асимптотической статистикой интервальных данных.**

Сформулируем сначала основные идеи асимптотической математической статистики интервальных данных, а затем рассмотрим реализацию этих идей на перечисленных выше примерах. Следует сразу подчеркнуть, что основные идеи достаточно просты, в то время как их проработка в конкретных ситуациях зачастую оказывается достаточно трудоемкой.

Пусть существо реального явления описывается выборкой x_1 , x_2 , ..., x_n . В вероятностной теории математической статистики, из которой мы исходим (см. терминологическую статью [24]), выборка — это набор независимых в совокупности одинаково распределенных случайных величин. Однако беспристрастный и тщательный анализ подавляющего большинства реальных задач показывает, что статистику известна отнюдь не выборка x_1 , x_2 , ..., x_n, а величины:

$$y_j = x_j + \varepsilon_j, j = 1, 2, ..., n ,$$

где $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ — некоторые погрешности измерений, наблюдений, анализов, опытов, исследований (например, инструментальные ошибки).

Одна из причин появления погрешностей — запись результатов наблюдений с конечным числом значащих цифр. Дело в том, что для случайных величин с непрерывными функциями распределения событие, состоящее в попадании хотя бы одного элемента выборки в множество рациональных чисел, согласно правилам теории вероятностей имеет вероятность 0, а такими событиями в теории вероятностей принято пренебрегать. Поэтому при рассуждениях о выборках из нормального, логарифмически нормального, экспоненциального, равномерного, гамма-распределений, распределения Вейбулла-Гнеденко и др. приходится принимать, что эти распределения имеют элементы исходной выборки x_1 , x_2 , ..., x_n, в то время как статистической обработке доступны лишь искаженные значения $y_j = x_j + \varepsilon_j$.

Введем обозначения:

$$x = (x_1, x_2, ..., x_n),\ y = (y_1, y_2, ..., y_n),\ \varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n).$$

Пусть статистические выводы основываются на статистике $f : R^n \to R^1$, используемой для оценивания параметров и характеристик распределения, проверки гипотез и решения иных статистических задач. Принципиально важная для статистики интервальных данных идея такова: СТАТИСТИК ЗНАЕТ ТОЛЬКО $f(y)$, НО НЕ $f(x)$.

Очевидно, в статистических выводах необходимо отразить различие между $f(y)$ и $f(x)$. Одним из двух основных понятий статистики интервальных данных является понятие нотны.

Определение. Величину максимально возможного (по абсолютной величине) отклонения, вызванного погрешностями наблюдений ε, известного статистику значения $f(y)$ от истинного значения $f(x)$, т.е.

$$N_{ff(x)} = \sup | f(y) - f(x) | ,$$

где супремум берется по множеству возможных значений вектора погрешностей ε (см. ниже), будем называть НОТНОЙ.

Если функция f имеет частные производные второго порядка, а ограничения на погрешности имеют вид:

$$| \varepsilon_i | \le \Delta,\ i = 1,2,..., n, \qquad (1)$$

причем Δ мало, то приращение функции f с точностью до бесконечно малых более высокого порядка описывается главным линейным членом, т.е.

$$f(y)-f(x)=\sum_{1\le i\le n}\frac{\partial f(x)}{\partial x_i}\varepsilon_i+O(\Delta^2).$$

Чтобы получить асимптотическое (при $\Delta\to 0$) выражение для нотны, достаточно найти максимум и минимум линейной функции (главного линейного члена) на кубе, заданном неравенствами (1). Легко видеть, что максимум достигается, если положить:

$$\varepsilon_i=\begin{cases}\Delta, \dfrac{\partial f(x)}{\partial x_i}\ge 0,\\ -\Delta, \dfrac{\partial f(x)}{\partial x_i}<0,\end{cases}$$

а минимум, отличающийся от максимума только знаком, достигается при $\varepsilon_i^{'}=-\varepsilon_i$. Следовательно, *нотна* с точностью до бесконечно малых более высокого порядка имеет вид:

$$N_f(x)=(\sum_{1\le i\le n}|\frac{\partial f(x)}{\partial x_i}|)\Delta.$$

Это выражение назовем *асимптотической нотной*.

Условие (1) означает, что исходные данные представляются статистику в виде интервалов $[y_i-\Delta; y_i+\Delta], i=1,2,...,n$ (отсюда и название этого научного направления). Ограничения на погрешности могут задаваться разными способами — кроме абсолютных ошибок используются относительные или иные показатели различия между x и y.

Если задана не предельная абсолютная погрешность Δ, а предельная относительная погрешность δ, т.е. ограничения на погрешности вошедших в выборку результатов измерений имеют вид:

$$|\varepsilon_i|\le\delta|x_i|, i=1,2,...,n,$$

то аналогичным образом получаем, что нотна с точностью до бесконечно малых более высокого порядка, т.е. асимптотическая нотна, имеет вид:

$$N_f(x)=(\sum_{1\le i\le n}|x_i\frac{\partial f(x)}{\partial x_i}|)\delta.$$

При практическом использовании рассматриваемой концепции необходимо провести тотальную замену символов x на символы y. В каждом конкретном случае удается показать, что в силу малости погрешностей разность $N_f(y) - N_f(x)$ является бесконечно малой более высокого порядка сравнительно с $N_f(x)$ или $N_f(y)$.

Основные результаты в вероятностной модели. В классической вероятностной модели элементы исходной выборки x_1, x_2, ..., x_n рассматриваются как независимые одинаково распределенные случайные величины. Как правило, существует некоторая константа $C > 0$ такая, что в смысле сходимости по вероятности:

$$\lim_{n\to\infty} N_f(x) = C\Delta. . \qquad (2)$$

Соотношение (2) доказывается отдельно для каждой конкретной задачи.

При использовании классических статистических методов в большинстве случаев используемая статистика $f(x)$ является асимптотически нормальной. Это означает, что существуют константы a и σ^2 такие, что

$$\lim_{n\to\infty} P\left(\sqrt{n}\,\frac{f(x)-a}{\sigma} < x\right) = \Phi(x),$$

где $\Phi(x)$ — функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1. При этом обычно оказывается, что

$$\lim_{n\to\infty} \sqrt{n}(Mf(x)-a) = 0$$

и

$$\lim_{n\to\infty} nDf(x) = \sigma^2,$$

а потому в классической математической статистике средний квадрат ошибки статистической оценки равен:

$$M(f(x)-a)^2 = (Mf(x)-a)^2 + Df(x) = \frac{\sigma^2}{n}$$

с точностью до членов более высокого порядка.

В статистике интервальных данных ситуация совсем иная — обычно можно доказать, что средний квадрат ошибки равен:

$$\max_{\{\varepsilon\}} M(f(y)-a)^2 = \frac{\sigma^2}{n} + N_f^2(y) + o(\Delta^2 + \frac{1}{n}). \qquad (3)$$

Из соотношения (3) вытекает ряд важных следствий. Прежде всего отметим, что правая часть этого равенства, в отличие от правой части соответствующего классического равенства, не стремится к 0 при безграничном возрастании объема выборки. Она остается больше некоторого положительного числа, а именно, квадрата нотны. Следовательно, статистика $f(x)$ не является состоятельной оценкой параметра a. Более того, состоятельных оценок вообще не существует.

Пусть доверительным интервалом для параметра a, соответствующим заданной доверительной вероятности γ, в классической математической статистике является интервал $(c_n(\gamma); d_n(\gamma))$. В статистике интервальных данных аналогичный доверительный интервал является более широким. Он имеет вид $(c_n(\gamma) - N_f(y); d_n(\gamma) + N_f(y))$. Таким образом, его длина увеличивается на две нотны. Следовательно, при увеличении объема выборки длина доверительного интервала не может стать меньше, чем $2C\Delta$ (см. формулу (2)).

В статистике интервальных данных методы оценивания параметров имеют другие свойства по сравнению с классической математической статистикой. Так, при больших объемах выборок метод моментов может быть заметно лучше, чем метод максимального правдоподобия (т.е. иметь меньший средний квадрат ошибки — см. формулу (3)), в то время как в классической математической статистике второй из названных методов всегда не хуже первого.

Рациональный объем выборки. Анализ формулы (3) показывает, что в отличие от классической математической статистики нецелесообразно безгранично увеличивать объем выборки, поскольку средний квадрат ошибки остается всегда большим квадрата нотны. Поэтому представляется полезным ввести понятие «рационального объема выборки» n_{rat}, при достижении которого продолжать наблюдения нецелесообразно.

Как установить «рациональный объем выборки»? Можно воспользоваться идеей «принципа уравнивания погрешностей», выдвинутой в монографии [3]. Речь идет о том, что вклад погрешностей различной природы в общую погрешность должен быть примерно одинаков. Этот принцип дает

возможность выбирать необходимую точность оценивания тех или иных характеристик в тех случаях, когда это зависит от исследователя. В статистике интервальных данных в соответствии с «принципом уравнивания погрешностей» предлагается определять рациональный объем выборки n_{rat} из условия равенства двух величин — метрологической составляющей, связанной с нотной, и статистической составляющей — в среднем квадрате ошибки (3), т.е. из условия:

$$\frac{\sigma^2}{n_{rat}} = N_f^2(y), \quad n_{rat} = \frac{\sigma^2}{N_f^2(y)}.$$

Для практического использования выражения для рационального объема выборки неизвестные теоретические характеристики необходимо заменить их оценками. Это делается в каждой конкретной задаче по-своему.

Исследовательскую программу в области статистики интервальных данных можно «в двух словах» сформулировать так: для любого алгоритма анализа данных (алгоритма прикладной статистики) необходимо вычислить нотну и рациональный объем выборки. Или иные величины из того же понятийного ряда, возникающие в многомерном случае, при наличии нескольких выборок и при иных обобщениях описываемой здесь простейшей схемы. Затем проследить влияние погрешностей исходных данных на точность оценивания, доверительные интервалы, значения статистик критериев при проверке гипотез, уровни значимости и другие характеристики статистических выводов. Очевидно, классическая математическая статистика является частью статистики интервальных данных, выделяемой условием $\Delta = 0$.

4.2. ИНТЕРВАЛЬНЫЕ ДАННЫЕ В ЗАДАЧАХ ОЦЕНИВАНИЯ

Поясним теоретические концепции статистики интервальных данных на простых примерах.

Пример 1. Оценивание математического ожидания. Пусть необходимо оценить математическое ожидание случайной величины с помощью обычной оценки — среднего арифметического результатов наблюдений, т.е.

$$f(x) = \frac{x_1 + x_2 + ... + x_n}{n}.$$

Тогда при справедливости ограничений (1) на абсолютные погрешности имеем $N_f(x) = \Delta$. Таким образом, нотна полностью известна и не зависит от многомерной точки, в которой берется. Вполне естественно: если каждый результат наблюдения известен с точностью до Δ, то и среднее арифметическое известно с той же точностью. Ведь возможна систематическая ошибка — если к каждому результату наблюдению добавить Δ, то и среднее арифметическое увеличится на Δ.

Поскольку

$$D(\overline{x}) = \frac{D(x_1)}{n},$$

то в обозначениях предыдущего пункта:

$$\sigma^2 = D(x_1).$$

Следовательно, рациональный объем выборки равен:

$$n_{rat} = \frac{D(x_1)}{\Delta^2}.$$

Для практического использования полученной формулы надо оценить дисперсию результатов наблюдений. Можно доказать, что, поскольку Δ мало, это можно сделать обычным способом, например, с помощью несмещенной выборочной оценки дисперсии

$$s^2(y) = \frac{1}{n-1}\sum_{1 \le i \le n}(y_i - \overline{y})^2.$$

Здесь и далее рассуждения часто идут на двух уровнях. Первый — это уровень «истинных» случайных величин, обозначаемых *«x»*, описывающих реальность, но неизвестных специалисту по анализу данных. Второй — уровень известных этому специалисту величин *«y»*, отличающихся погрешностями от истинных. Погрешности малы, поэтому функции от *x* отличаются от функций от *y* на некоторые бесконечно малые величины. Эти соображения и позволяют использовать $s^2(y)$ как оценку $D(x_1)$.

Итак, выборочной оценкой рационального объема выборки является

$$n_{sample-rat} = \frac{s^2(y)}{\Delta^2}.$$

Уже на этом первом рассматриваемом примере видим, что рациональный объем выборки находится не где-то вдали, а непосредственно рядом с теми объемами, с которыми имеет дело любой практически работающий статистик. Например, если статистик знает, что $\Delta = \frac{\sigma}{6}$, то $n_{rat} = 36$. А именно такова погрешность контрольных шаблонов во многих технологических процессах! Поэтому, занимаясь управлением качеством, необходимо обращать внимание на действующую на предприятии систему измерений.

По сравнению с классической математической статистикой [27, п. 4.3; 51, п. 8.1] доверительный интервал для математического ожидания (для заданной доверительной вероятности γ) имеет другой вид:

$$(\bar{y} - \Delta - u(\gamma)\frac{s}{\sqrt{n}}; \bar{y} + \Delta + u(\gamma)\frac{s}{\sqrt{n}}), \qquad (4)$$

где $u(\gamma)$ — квантиль порядка $(1 + \gamma) / 2$ стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1.

По поводу формулы (4) была довольно жаркая дискуссия среди специалистов. Отмечалось, что она получена на основе Центральной Предельной Теоремы теории вероятностей и может быть использована при любом распределении результатов наблюдений (с конечной дисперсией). Если же имеется дополнительная информация, то, по мнению отдельных специалистов, формула (4) может быть уточнена. Например, если известно, что распределение x_i является нормальным, в качестве $u(\gamma)$ целесообразно использовать квантиль распределения Стьюдента. К этому надо добавить, что по небольшому числу наблюдений нельзя надежно установить нормальность, а при росте объема выборки квантили распределения Стьюдента приближаются к квантилям нормального распределения. Вопрос о том, часто ли результаты наблюдений имеют нормальное распределение, подробно обсуждался среди специалистов. Выяснилось, что распределения встречающихся в практических задачах результатов измерений почти всегда отличны от нормальных [25]. А также и от распределений из иных параметрических семейств, описываемых в учебниках.

Применительно к оцениванию математического ожидания (но не к оцениванию других характеристик или параметров распределения) факт существования границы возможной точности, определяемой точностью исходных данных, не0днократно отмечался в литературе ([26, с. 230–234], [31, с. 121] и др.).

Пример 2. Оценивание дисперсии. Для статистики $f(y) = s^2(y)$, где $s^2(y)$ — выборочная дисперсия (несмещенная оценка теоретической дисперсии), при справедливости ограничений (1) на абсолютные погрешности имеем:

$$N_f(y) = \frac{2\Delta}{n-1}\sum_{i=1}^{n} | y_i - \overline{y} | + O(\Delta^2).$$

Можно показать, что нотна $N_f(y)$ сходится к

$$2\Delta M | x_1 - M(x_1) |$$

по вероятности с точностью до $o(\Delta)$, когда n стремится к бесконечности. Это же предельное соотношение верно и для нотны $N_f(x)$, вычисленной для исходных данных. Таким образом, в данном случае справедлива формула (2) с

$$C = 2M | x_1 - M(x_1) | .$$

Известно [27, п. 4.3; 51, п. 8.1], что случайная величина:

$$\frac{s^2 - \sigma^2}{\sqrt{n}}$$

является асимптотически нормальной с математическим ожиданием 0 и дисперсией $D(x_1^2)$.

Из сказанного вытекает, что в статистике интервальных данных асимптотический доверительный интервал для дисперсии σ^2 (соответствующий доверительной вероятности γ) имеет вид:

$$(s^2(y) - A; \quad s^2 + A),$$

где

$$A = \frac{u(\gamma)}{\sqrt{n(n-1)}} \sqrt{\sum_{i=1}^{n} (y_i^2 - \frac{1}{n}\sum_{j=1}^{n} y_j^2)^2} + \frac{2\Delta}{n-1}\sum_{i=1}^{n} | y_i - \overline{y} |,$$

где $u(\gamma)$ обозначает тот же самый квантиль стандартного нормального распределения, что и выше в случае оценивания математического ожидания.

Рациональный объем выборки при оценивании дисперсии равен:

$$n_{rat} = \frac{D(x_1^2)}{4\Delta^2 (M \mid x_1 - M(x_1) \mid)^2},$$

а выборочную оценку рационального объема выборки $n_{sample-rat}$ можно вычислить, заменяя теоретические моменты на соответствующие выборочные и используя доступные статистику результаты наблюдений, содержащие погрешности.

Что можно сказать о численной величине рационального объема выборки? Как и в случае оценивания математического ожидания, она отнюдь не выходит за пределы обычно используемых объемов выборок. Так, если распределение результатов наблюдений x_i является нормальным с математическим ожиданием 0 и дисперсией σ^2, то в результате вычисления моментов случайных величин в предыдущей формуле получаем, что

$$n_{rat} = \frac{\sigma^2}{\pi\Delta^2},$$

где π — отношение длины окружности к диаметру, $\pi = 3{,}141592...$ Например, если $\Delta = \sigma / 6,$ то $n_{rat} = 11.$ Это меньше, чем при оценивании математического ожидания в предыдущем примере.

Пример 3. Аддитивные статистики. Пусть $g : R^1 \to R^1$ — некоторая непрерывная функция. Аддитивные статистики имеют вид:

$$f(x) = \frac{1}{n} \sum_{1 \le i \le n} g(x_i).$$

Тогда

$$\sum_{1 \le i \le n} \mid \frac{\partial f(x)}{\partial x_i} \mid = \frac{1}{n} \sum_{1 \le i \le n} \mid \frac{dg(x_i)}{dx_i} \mid \to M \mid \frac{dg(x_1)}{dx_1} \mid,$$

$$\sum_{1 \le i \le n} \mid x_i \frac{\partial f(x)}{\partial x_i} \mid = \frac{1}{n} \sum_{1 \le i \le n} \mid x_i \frac{dg(x_i)}{dx_i} \mid \to M \mid x_1 \frac{dg(x_1)}{dx_1} \mid$$

по вероятности при $n \to \infty,$ если математические ожидания в правых частях двух последних соотношений существуют. Применяя рассмотренные выше

общие соображения, получаем, что при малых фиксированных Δ и δ и достаточно больших n значения $f(y)$ могут принимать любые величины из разрешенных (например, записываемых заданным числом значащих цифр) в замкнутом интервале:

$$[f(x)-\Delta M\,|\,\frac{dg(x_1)}{dx_1}\,|;\,f(x)+\Delta M\,|\,\frac{dg(x_1)}{dx_1}\,|] \qquad (5)$$

при ограничениях (1) на абсолютные ошибки и в замкнутом интервале:

$$[f(x)-\delta M\,|\,x_1\frac{dg(x_1)}{dx_1}\,|;\,f(x)+\delta M\,|\,x_1\frac{dg(x_1)}{dx_1}\,|] \qquad (6)$$

при ограничениях на относительные погрешности результатов наблюдений. Обратим внимание, что длины этих интервалов независимы от объема выборки, в частности, не стремятся к 0 при его росте.

К каким последствиям это приводит в задачах статистического оценивания? Поскольку для статистик аддитивного типа:

$$f(x)=\frac{1}{n}\sum_{1\le i\le n} g(x_i)\to Mg(x_1) \qquad (7)$$

по вероятности при $n\to\infty$, если математическое ожидание в правой части формулы (7) существует, то аддитивную статистику $f(x)$ естественно рассматривать как непараметрическую оценку этого математического ожидания. Термин «непараметрическая» означает, что не делается предположений о принадлежности функции распределения выборки к тому или иному параметрическому семейству распределения. Распределение статистики $f(x)$ зависит от распределения результатов наблюдений. Однако для любого распределения результатов наблюдений с конечной дисперсией статистика $f(x)$ является состоятельной и асимптотически нормальной оценкой для математического ожидания, указанного в правой части формулы (7).

Как известно, в рамках классической математической статистики в предположении существования ненулевой дисперсии $Dg(x_1)$ в силу асимптотической нормальности аддитивной статистики $f(x)$ асимптотический доверительный интервал, соответствующий доверительной вероятности γ, имеет вид:

$$[f(x)-u(\frac{1+\gamma}{2})\frac{s(g(x))}{\sqrt{n}};\,f(x)+u(\frac{1+\gamma}{2})\frac{s(g(x))}{\sqrt{n}}],$$

где $s(g(x))$ — выборочное среднее квадратическое отклонение, построенное по $g(x_1), g(x_2), \ldots, g(x_n)$, а $u(\frac{1+\gamma}{2})$ — квантиль стандартного нормального распределения порядка $\frac{1+\gamma}{2}$.

В рассматриваемой модели порождения интервальных данных вместо $f(x)$ необходимо использовать $f(y)$, а вместо $g(x_i)$ — соответственно $g(y_i)$, $i = 1, 2, \ldots, n$. При этом доверительный интервал необходимо расширить с учетом формул (5) и (6).

В соответствии с проведенными рассуждениями для аддитивных статистик асимптотическая нотна имеет вид:

$$N_f(x) = \Delta M \mid \frac{dg(x_1)}{dx_1} \mid$$

при ограничениях (1) на абсолютную погрешность и

$$N_f(x) = \delta M \mid x_1 \frac{dg(x_1)}{dx_1} \mid$$

при ограничениях на относительную погрешность. В первом случае нотна является обобщением понятия предельной абсолютной систематической ошибки, во втором — предельной относительной систематической ошибки. Отметим, что, как и в примерах 1 и 2, асимптотическая нотна не зависит от точки, в которой вычисляется. Таким образом, она является константой для конкретного метода статистического анализа данных.

Поскольку n велико, а Δ и δ малы, то можно пренебречь отличием выборочного среднего квадратического отклонения $s(g(y))$, вычисленного по выборке преобразованных значений $g(y_1), g(y_2), \ldots, g(y_n)$, от выборочного среднего квадратического отклонения $s(g(x))$, построенного по выборке $g(x_1), g(x_2), \ldots, g(x_n)$. Разность этих двух величин является бесконечно малой, они приближаются к одной и той же положительной константе.

В статистике интервальных данных выборочный доверительный интервал для $Mg(x_1)$ имеет вид:

$$[f(y) - N_f(y) - u(\frac{1+\gamma}{2})\frac{s(g(y))}{\sqrt{n}}; f(y) + N_f(y) + u(\frac{1+\gamma}{2})\frac{s(g(y))}{\sqrt{n}}].$$

В асимптотике его длина такова:

$$2N_f(x) + 2u(\frac{1+\delta}{2})\frac{\sigma}{\sqrt{n}}, \tag{8}$$

где σ^2 — дисперсия $g(x_1)$, в то время как в классической теории математической статистики имеется только второе слагаемое. Соотношение (8) — аналог суммарной ошибки у метрологов [26]. Поскольку первое слагаемое положительно, то оценивание $Mg(x_1)$ с помощью $f(y)$ не является состоятельным.

Для аддитивных статистик при больших n максимум (по возможным погрешностям) среднего квадрата отклонения оценки имеет вид:

$$\max_{\varepsilon} M[f(y) - Mg(x_1)]^2 = N_f^2(x) + \frac{Dg(x_1)}{n} \tag{9}$$

с точностью до членов более высокого порядка. Исходя из принципа уравнивания погрешностей в общей схеме устойчивости [3], нецелесообразно второе слагаемое в (9) делать меньше первого за счет увеличения объема выборки n. Рациональный объем выборки, т.е. тот объем, при котором равны погрешности оценивания (или проверки гипотез), вызванные погрешностями исходных данных, и статистические погрешности, рассчитанные по обычным правилам математической статистики (при $\varepsilon_i \equiv 0$), для аддитивных статистик согласно (9) имеет вид:

$$n_{rat} = \frac{Dg(x_1)}{N_f^2(x)}. \tag{10}$$

В качестве примера рассмотрим экспоненциально распределенные результаты наблюдений x_i с $M(x_1) = D(x_1) = 1$. Оцениваем математическое ожидание с помощью выборочного среднего арифметического при ограничениях на относительную погрешность. Тогда согласно формуле (10):

$$N_f(x) = \delta, \quad n_{rat} = \frac{1}{\delta^2}.$$

В частности, если относительная погрешность измерений δ = 10 %, то рациональный объем выборки равен 100. Формуле (10) соответствует также рассмотренный выше пример 1.

Пример 4. Оценивание медианы распределения с помощью выборочной медианы. Хотя нельзя выделить главный линейный член из-за недифференцируемости функции *f(x)*, выражающей выборочную медиану через элементы выборки, непосредственно из определения нотны следует, что при ограничениях на абсолютные погрешности:

$$N_f(x) = \Delta,$$

а при ограничениях на относительные погрешности:

$$N_f(x) = \delta x_{med}$$

с точностью до бесконечно малых более высокого порядка, где x_{med} — теоретическая медиана. Доверительный интервал для медианы имеет вид:

$$[a_1(x) - N_f(x); a_2(x) + N_f(x)],$$

где $[a_1(x); a_2(x)]$ — доверительный интервал для медианы, вычисленный по классическим правилам непараметрической статистики [27]. Для нахождения рационального объема выборки можно использовать асимптотическую дисперсию выборочной медианы. Она, как известно (см., например, [28, с. 178]), равна:

$$\sigma^2(M) = \frac{1}{4np^2(x_{med})},$$

где $p(x_{med})$ — плотность распределения результатов измерений в точке x_{med}. Следовательно, рациональный объем выборки имеет вид:

$$n_{rat} = \frac{1}{4p^2(x_{med})\Delta^2}, \quad n_{rat} = \frac{1}{4p^2(x_{med})x_{med}^2\delta^2}$$

при ограничениях на абсолютные и относительные погрешности результатов измерений соответственно. Для практического использования этих формул следует оценить плотность распределения результатов измерений в одной точке — теоретической медиане. Это можно сделать с помощью тех или иных непараметрических оценок плотности [27].

Если результаты наблюдений имеют стандартное нормальное распределение с математическим ожиданием 0 и дисперсией 1, то

$$n_{rat} = \frac{\pi}{2\Delta^2} \approx \frac{1{,}57}{\Delta^2}.$$

В этом случае рациональный объем выборки в $\pi/2$ раз больше, чем для оценивания математического ожидания (пример 1 выше). Однако для других распределений рассматриваемое соотношение объемов может быть иным, в частности, меньше 1. Как вытекает из статьи А. Н. Колмогорова 1931 г. [29], рассматриваемое соотношение объемов может принимать любое значение между 0 и 3.

Пример 5. Оценивание коэффициента вариации. Рассмотрим выборочный коэффициент вариации:

$$v = f(y_1, y_2, ..., y_n) = \frac{\left\{\frac{1}{n-1}\sum_{1\le i\le n}(y_i - \bar{y})^2\right\}^{1/2}}{\frac{1}{n}\sum_{1\le i\le n} y_i} = \frac{s(y)}{\bar{y}}.$$

Как нетрудно подсчитать,

$$\frac{\partial f}{\partial x_i} = \frac{n\bar{x}(x_i - \bar{x}) - (n-1)s^2(x)}{n(n-1)(\bar{x})^2 s(x)}.$$

В случае ограничений на относительную погрешность:

$$\lim_{n\to\infty} N_f(x) = \frac{\delta}{(M(x_1))^2\sigma} M\,|\,x_1\{[x_1 - M(x_1)]M(x_1) - \sigma^2\}\,|\,.$$

На основе этого предельного соотношения и формулы для асимптотической дисперсии выборочного коэффициента вариации, приведенной в [27], могут быть найдены по описанной выше схеме доверительные границы для теоретического коэффициента вариации и рациональный объем выборки.

Замечание. Отметим, что формулы для рационального объема выборки получены на основе асимптотической теории, а применяются для получения конечных объемов — 36 и 100 в примерах 1–3. Как всегда при использовании асимптотических результатов математической статистики, необходимы дополнительные исследования для изучения точности асимптотических формул при конечных объемах выборок.

Рассмотрим классическую в прикладной математической статистике параметрическую задачу оценивания. Исходные данные — выборка x_1, x_2, ..., x_n, состоящая из n действительных чисел. В вероятностной модели простой случайной выборки ее элементы x_1, x_2, ..., x_n считаются набором реализаций n независимых одинаково распределенных случайных величин. Будем считать,

что эти величины имеют плотность *f(x)*. В параметрической статистической теории предполагается, что плотность *f(x)* известна с точностью до конечномерного параметра, т.е., $f(x) = f(x,\theta_0)$ при некотором $\theta_0 \in \Theta \subseteq R^k$. Это, конечно, весьма сильное предположение, которое требует обоснования и проверки; однако в настоящее время параметрическая теория оценивания широко используется в различных прикладных областях.

Все результаты наблюдений определяются с некоторой точностью, в частности, записываются с помощью конечного числа значащих цифр (обычно 2–5). Следовательно, все реальные распределения результатов наблюдений дискретны. Обычно считают, что эти дискретные распределения достаточно хорошо приближаются непрерывными. Уточняя это утверждение, приходим к уже рассматривавшейся модели, согласно которой статистику доступны лишь величины:

$$yj = xj + \varepsilon j\,,\ j = 1, 2, ..., n\,,$$

где x_i — «истинные» значения, $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ — погрешности наблюдений (включая погрешности дискретизации). В вероятностной модели принимаем, что *n* пар:

$$(x_1, \varepsilon_1), (x_2, \varepsilon_2), ..., (x_n, \varepsilon_n)$$

образуют простую случайную выборку из некоторого двумерного распределения, причем x_1, x_2, ..., x_n — выборка из распределения с плотностью $f(x) = f(x,\theta_0)$. Необходимо учитывать, что x_i и ε_i — реализации зависимых случайных величин (если считать их независимыми, то распределение y_i будет непрерывным, а не дискретным). Поскольку систематическую ошибку, как правило, нельзя полностью исключить [26, с. 141], то необходимо рассматривать случай $M\varepsilon_i \neq 0$. Нет оснований априори принимать и нормальность распределения погрешностей (согласно сводкам экспериментальных данных о разнообразии форм распределения погрешностей измерений, приведенным в [26, с. 148] и [27, с. 71–77], в подавляющем большинстве случаев гипотеза о нормальном распределении погрешностей оказалась неприемлемой для средств измерений различных типов). Таким образом, все три распространенных представления о свойствах погрешностей не адекватны реальности. Влияние погрешностей наблюдений на свойства статистических моделей необходимо изучать на основе иных моделей, а именно, моделей интервальной статистики.

Пусть ε — характеристика величины погрешности, например, средняя квадратическая ошибка $\varepsilon = \sqrt{M(\varepsilon_i^2)}$. В классической математической статистике ε считается пренебрежимо малой ($\varepsilon \to 0$) при фиксированном объеме выборки n. Общие результаты доказываются в асимптотике $n \to \infty$. Таким образом, в классической математической статистике сначала делается предельный переход $\varepsilon \to 0$, а затем предельный переход $n \to \infty$. В статистике интервальных данных принимаем, что объем выборки достаточно велик ($n \to \infty$), но всем измерениям соответствует одна и та же характеристика погрешности $\varepsilon \neq 0$. Полезные для анализа реальных данных предельные теоремы получаем при $\varepsilon \to 0$. В статистике интервальных данных сначала делается предельный переход $n \to \infty$, а затем предельный переход $\varepsilon \to 0$. Итак, в обеих теориях используются одни и те же два предельных перехода: $n \to \infty$ и $\varepsilon \to 0$, но в разном порядке. Утверждения обеих теорий принципиально различны.

В дальнейшем изложение идет на примере оценивания параметров гамма-распределения, хотя аналогичные результаты можно получить и для других параметрических семейств, а также для задач проверки гипотез (см. ниже) и т.д. Наша цель — продемонстрировать основные черты подхода статистики интервальных данных. Его разработка была стимулирована подготовкой ГОСТ 11.011-83 (в настоящее время отменен, но может использоваться как научная публикация) [4].

Отметим, что постановки статистики объектов нечисловой природы соответствуют подходу, принятому в общей теории устойчивости [3, 27]. В соответствии с этим подходом выборке $x = (x_1, x_2, ..., x_n)$ ставится в соответствие множество допустимых отклонений $G(x)$, т.е. множество возможных значений вектора результатов наблюдений $y = (y_1, y_2, ..., y_n)$. Если известно, что абсолютная погрешность результатов измерений не превосходит Δ, то множество допустимых отклонений имеет вид:

$$G(x, \Delta) = \{y : | y_i - x_i | \le \Delta, i = 1,2,...,n\}.$$

Если известно, что относительная погрешность не превосходит δ, то множество допустимых отклонений имеет вид:

$$G(x, \delta) = \{y : | \frac{y_i}{x_i} - 1 | \le \delta, i = 1,2,...,n\}.$$

Теория устойчивости позволяет учесть «наихудшие» отклонения, т.е. приводит к выводам типа минимаксных, в то время как конкретные модели погрешностей позволяют делать заключения о поведении статистик «в среднем».

Оценки параметров гамма-распределения. Как известно, случайная величина X имеет гамма-распределение, если ее плотность такова [4]:

$$f(x;a,b)=\begin{cases}\dfrac{1}{\Gamma(a)}x^{a-1}b^{-a}\exp\{-\dfrac{x}{b}\}, x>0,\\ 0, x\le 0,\end{cases}$$

где a — параметр формы, b — параметр масштаба, $\Gamma(a)$ — гамма-функция. Отметим, что есть и иные способы параметризации семейства гамма-распределений [30].

Поскольку $M(X) = ab$, $D(X) = ab^2$, то оценки метода имеют вид:

$$\hat{a}=\frac{(\bar{x})^2}{s^2},\ \hat{b}=\frac{\bar{x}}{\hat{a}}=\frac{s^2}{\bar{x}},$$

где $\bar{x}$ — выборочное среднее арифметическое, а s^2 — выборочная дисперсия. Можно показать, что при больших n:

$$M(\hat{a}-a)^2=\frac{2a(a+1)}{n},\ M(\hat{b}-b)^2=\frac{b^2}{n}\left(2+\frac{3}{a}\right) \qquad (11)$$

с точностью до бесконечно малых более высокого порядка.

Оценка максимального правдоподобия a^* имеет вид [4]:

$$a^*=H(\frac{1}{n}\sum_{1\le i\le n}\ln(\frac{\bar{x}}{x_i})), \qquad (12)$$

где $H(\bullet)$ — функция, обратная к функции:

$$Q(a)=\ln a-\frac{d\Gamma(a)}{da}\Big/\Gamma(a).$$

При больших n с точностью до бесконечно малых более высокого порядка:

$$M(a^*-a)^2=\frac{a}{n(a\psi'(a)-1)},\quad \psi(a)=\frac{d\Gamma(a)}{da}\Big/\Gamma(a).$$

Как и для оценок метода моментов, оценка максимального правдоподобия b^* параметра масштаба имеет вид:

$$b^*=\bar{x}/a^*.$$

При больших n с точностью до бесконечно малых более высокого порядка:

$$M(b^* - b)^2 = \frac{b^2 \psi'(a)}{n(a\psi'(a) - 1)}.$$

Используя свойства гамма-функции, можно показать [4], что при больших a:

$$M(a^* - a)^2 = \frac{a(2a-1)}{n}, \quad M(b^* - b)^2 = \frac{2b^2}{n}.$$

с точностью до бесконечно малых более высокого порядка. Сравнивая с формулами (11), убеждаемся в том, что средние квадраты ошибок для оценок метода моментов больше соответствующих средних квадратов ошибок для оценок максимального правдоподобия. Таким образом, с точки зрения классической математической статистики оценки максимального правдоподобия имеют преимущество по сравнению с оценками метода моментов.

Необходимость учета погрешностей измерений. Положим:

$$v = f(x_1, x_2, ..., x_n) = \frac{1}{n} \sum_{1 \le i \le n} \ln\left(\frac{\bar{x}}{x_i}\right).$$

Из свойств функции $H(\bullet)$ следует [4, с. 14], что при малых:

$$a^* \sim 1/(2v). \qquad (13)$$

В силу состоятельности оценки максимального правдоподобия a^* из формулы (13) следует, что $v \to 0$ по вероятности при $a \to \infty$.

Согласно модели статистики интервальных данных результатами наблюдений являются не x_i, а y_i, вместо v по реальным данным рассчитывают:

$$w = f(y_1, y_2, ..., y_n) = \frac{1}{n} \sum_{1 \le i \le n} \ln\left(\frac{\bar{y}}{y_i}\right).$$

Имеем:

$$w - v = \ln\left(\frac{\bar{y}}{\bar{x}}\right) - \frac{1}{n} \sum_{1 \le i \le n} \ln\left(1 + \frac{\varepsilon_i}{x_i}\right). \qquad (14)$$

В силу закона больших чисел при достаточно малой погрешности ε, обеспечивающей возможность приближения $\ln(1+\alpha) \sim \alpha$ для слагаемых в формуле (14), или, что эквивалентно, при достаточно малых предельной абсолютной погрешности Δ в формуле (1) или достаточно малой предельной относительной погрешности δ имеем при $n \to \infty$:

$$w - v \to \frac{M(\varepsilon_i)}{M(x_i)} - M\left(\frac{\varepsilon_i}{x_i}\right) = c$$

по вероятности (в предположении, что все погрешности одинаково распределены). Таким образом, наличие погрешностей вносит сдвиг, вообще говоря, не исчезающий при росте объема выборки. Следовательно, если $c \neq 0$, то оценка максимального правдоподобия не является состоятельной. Имеем:

$$a^*(y) - a^* \approx -\frac{c}{2v^2},$$

где величина $a^*(y)$ определена по формуле (12) с заменой x_i на y_i, $i = 1, 2, \ldots, n$. Из формулы (13) следует [4], что

$$a^*(y) - a \approx -2(a^*)^2 c, \tag{15}$$

т.е. влияние погрешностей измерений увеличивается по мере роста a.

Из формул для v и w следует, что с точностью до бесконечно малых более высокого порядка:

$$w - v \approx \sum_{1 \le i \le n} \frac{\partial f}{\partial x_i} \varepsilon_i = \frac{1}{n} \sum_{1 \le i \le n} \left(\frac{1}{\bar{x}} - \frac{1}{x_i}\right) \varepsilon_i . \tag{16}$$

С целью нахождения асимптотического распределения w выделим, используя формулу (16) и формулу для v, главные члены в соответствующих слагаемых:

$$w = \ln M(x_1) + \frac{1}{n} \sum_{1 \le i \le n} \left\{ \frac{x_i - M(x_1)}{M(x_i)} - \ln x_i + \left(\frac{1}{M(x_1)} - \frac{1}{x_i}\right) \varepsilon_i \right\} + O_p\left(\frac{1}{n}\right). \tag{17}$$

Таким образом, величина w представлена в виде суммы независимых одинаково распределенных случайных величин (с точностью до зависящего от случая остаточного члена порядка $1 / n$). В каждом слагаемом выделяются две части — одна, соответствующая v, и вторая, в которую входят ε_i. На основе представления (17) можно показать, что при $n \to \infty, \varepsilon \to 0$ распределения случайных величин v и w асимптотически нормальны, причем

$$M(w) \approx M(v) + c, \quad D(w) \approx D(v).$$

Из асимптотического совпадения дисперсий v и w, вида параметров асимптотического распределения (при $a \to \infty$) оценки максимального правдоподобия a^* и формулы (15) вытекает одно из основных соотношений статистики интервальных данных:

$$M(a^*(y) - a)^2 \approx 4a^4c^2 + \frac{a(2a-1)}{n}. \qquad (18)$$

Соотношение (18) уточняет утверждение о несостоятельности a^*. Из него следует также, что не имеет смысла безгранично увеличивать объем выборки n с целью повышения точности оценивания параметра a, поскольку при этом уменьшается только второе слагаемое в (18), а первое остается постоянным.

В соответствии с общим подходом статистики интервальных данных в стандарте [4] предлагается определять рациональный объем выборки n_{rat} из условия «уравнивания погрешностей» (это условие было впервые предложено в монографии [3]) различных видов в формуле (18), т.е. из условия:

$$4a^4c^2 = \frac{a(2a-1)}{n_{rat}}.$$

Упрощая это уравнение в предположении $a \to \infty$, получаем, что

$$n_{rat} = \frac{1}{2a^2c^2}.$$

Согласно сказанному выше, целесообразно использовать лишь выборки с объемами $n \le n_{rat}$. Превышение рационального объема выборки n_{rat} не дает существенного повышения точности оценивания.

Применение методов теории устойчивости. Найдем асимптотическую нотну. Как следует из вида главного линейного члена в формуле (17), решение оптимизационной задачи:

$$w - v \to \max, \quad |\varepsilon_i| \le \Delta,$$

соответствующей ограничениям на абсолютные погрешности, имеет вид:

$$\varepsilon_i = \begin{cases} \Delta, & \dfrac{1}{\bar{x}} - \dfrac{1}{x_i} \ge 0, \\ -\Delta, & \dfrac{1}{\bar{x}} - \dfrac{1}{x_i} < 0 \end{cases}.$$

Однако при этом пары (x_i, ε_i) не образуют простую случайную выборку, т.к. в выражения для ε_i входит $\bar{x}$. Однако при $n \to \infty$ можно заменить $\bar{x}$ на $M(x_1)$. Тогда получаем, что

$$w - v \approx A\Delta$$

при $a > 1$, где

$$A = M\left|\frac{1}{M(x_1)} - \frac{1}{x_1}\right| = \int_0^\infty \left|\frac{1}{ab} - \frac{1}{x}\right| f(x; a, b)dx.$$

Таким образом, с точностью до бесконечно малых более высокого порядка нотна имеет вид:

$$N_{a^*}(y) = 2(a^*)^2 c, \quad c = A\Delta.$$

Применим полученные результаты к построению доверительных интервалов. В постановке классической математической статистики (т.е. при $\varepsilon = 0$) доверительный интервал для параметра формы a, соответствующий доверительной вероятности γ, имеет вид [4]:

$$\left[a^* - u\left(\frac{1+\gamma}{2}\right)\sigma^*(a^*); \quad a^* + u\left(\frac{1+\gamma}{2}\right)\sigma^*(a^*)\right],$$

где $u\left(\frac{1+\gamma}{2}\right)$ — квантиль порядка $\frac{1+\gamma}{2}$ стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1,

$$[\sigma^*(a^*)]^2 = \frac{a^*}{n(a^*\psi'(a^*)-1)}, \quad \psi(a) = \frac{d\Gamma(a)}{da} \Big/ \Gamma(a).$$

В постановке статистики интервальных данных (т.е. при $\varepsilon \neq 0$) следует рассматривать доверительный интервал:

$$[a^* - 2(a^*)^2 \mid c \mid - u\left(\frac{1+\gamma}{2}\right)\sigma^*(a^*); \quad a^* + 2(a^*)^2 \mid c \mid + u\left(\frac{1+\gamma}{2}\right)\sigma^*(a^*)],$$

где

$$c = \frac{M(\varepsilon_i)}{M(x_i)} - M\left(\frac{\varepsilon_i}{x_i}\right)$$

в вероятностной постановке (пары (x_i, ε_i) образуют простую случайную выборку) и $c = A\Delta$ в оптимизационной постановке. Как в вероятностной, так и в оптимизационной постановках длина доверительного интервала не стремится к 0 при $n \to \infty$.

Если ограничения наложены на предельную относительную погрешность, задана величина δ, то значение c можно найти с помощью следующих правил приближенных вычислений [32, с. 142].

(I) Относительная погрешность суммы заключена между наибольшей и наименьшей из относительных погрешностей слагаемых.

(II) Относительная погрешность произведения и частного равна сумме относительных погрешностей сомножителей или, соответственно, делимого и делителя.

Можно показать, что в рамках статистики интервальных данных с ограничениями на относительную погрешность правила (I) и (II) являются строгими утверждениями при $\delta \to 0$.

Обозначим относительную погрешность некоторой величины t через ОП(t), абсолютную погрешность — через АП(t).

Из правила (I) следует, что ОП($\bar{x}$) = δ, а из правила (II) — что

$$ОП\left(\frac{\bar{x}}{x_i}\right) = 2\delta.$$

Поскольку рассмотрения ведутся при $a \to \infty$, то в силу неравенства Чебышева:

$$\frac{\bar{x}}{x_i} \to 1 \qquad (19)$$

по вероятности при $a \to \infty$, поскольку и числитель, и знаменатель в (19) с близкой к 1 вероятностью лежат в промежутке $[ab - db\sqrt{a};\quad ab + db\sqrt{a}]$, где константа d может быть определена с помощью упомянутого неравенства Чебышева.

Поскольку при справедливости (19) с точностью до бесконечно малых более высокого порядка:

$$\ln\left(\frac{\bar{x}}{x_i}\right) \approx \frac{\bar{x}}{x_i} - 1,$$

то с помощью трех последних соотношений имеем:

$$ОП\left(\frac{\bar{x}}{x_i}\right) = АП\left(\frac{\bar{x}}{x_i}\right) = АП\left(\ln\left(\frac{\bar{x}}{x_i}\right)\right) = 2\delta. \qquad (20)$$

Применим еще одно правило приближенных вычислений [32, с. 142].

(III). Предельная абсолютная погрешность суммы равна сумме предельных абсолютных погрешностей слагаемых.

Из (20) и правила (III) следует, что

$$АП(v) = 2\delta.$$

Из этого соотношения и (15) вытекает [4, с. 44, ф-ла (18)], что

$$АП(a^*) = 4a^2\delta,$$

откуда в соответствии с ранее полученной формулой для рационального объема выборки с заменой $c = 2\delta$ получаем, что

$$n_{rat} = \frac{1}{8a^2\delta^2}.$$

В частности, при a = 5,00, δ = 0,01 получаем $n_{rat} = 50$, т.е. в ситуации, в которой были получены данные о наработке резцов до предельного состоя-

ния (см. табл. 1, составленную согласно [4, с. 29]), проводить более 50 наблюдений нерационально.

Таблица 1

Наработка резцов до предельного состояния, ч

№ п/п	Наработка, ч	№ п/п	Наработка, ч	№ п/п	Наработка, ч
1	9	18	47,5	35	63
2	17,5	19	48	36	64,5
3	21	20	50	37	65
4	26,5	21	51	38	67,5
5	27,5	22	53,5	39	68,5
6	31	23	55	40	70
7	32,5	24	56	41	72,5
8	34	25	56	42	77,5
9	36	26	56,5	43	81
10	36,5	27	57,5	44	82,5
11	39	28	58	45	90
12	40	29	59	46	96
13	41	30	59	47	101,5
14	42,5	31	60	48	117,5
15	43	32	61	49	127,5
16	45	33	61,5	50	130
17	46	34	62		

В соответствии с ранее проведенными рассмотрениями асимптотический доверительный интервал для a, соответствующий доверительной вероятности $\gamma = 0{,}95$, имеет вид:

$$[a^* - 4(a^*)^2\delta - 1{,}96\sqrt{\frac{a^*(2a^*-1)}{n}};\quad a^* + 4(a^*)^2\delta + 1{,}96\sqrt{\frac{a^*(2a^*-1)}{n}}].$$

В частности, при $a^* = 5{,}00,\ \delta = 0{,}01,\ n = 50$ имеем асимптотический доверительный интервал [2,12; 7,86] вместо [3,14; 6,86] при $\delta = 0$.

При больших a в силу соображений, приведенных при выводе формулы (19), можно связать между собой относительную и абсолютную погрешности результатов наблюдений x_i:

$$\delta = \frac{\Delta}{M(x_1)} = \frac{\Delta}{ab}. \qquad (21)$$

Следовательно, при больших a имеем:

$$c = 2\delta = A\Delta, \quad A = \frac{2\delta}{\Delta} = \frac{2}{ab}.$$

Таким образом, проведенные рассуждения дали возможность вычислить асимптотику интеграла, задающего величину A.

Сравнение методов оценивания. Изучим влияние погрешностей измерений (с ограничениями на абсолютную погрешность) на оценку $\hat{a}$ метода моментов. Имеем:

$$АП(\bar{x}) = \Delta, \quad АП((\bar{x})^2) \approx 2\bar{x}\Delta \approx 2ab\Delta.$$

Погрешность s^2 зависит от способа вычисления s^2. Если используется формула:

$$s^2 = \frac{1}{n-1}\sum_{1\le i\le n}(x_i - \bar{x})^2, \tag{22}$$

то необходимо использовать соотношения:

$$АП\ (x_i - \bar{x}) = 2\Delta, \quad АП\ [(x_i - \bar{x})^2] \approx 2\,|\,x_i - \bar{x}\,|\,\Delta.$$

По сравнению с анализом влияния погрешностей на оценку a^* здесь возникает новый момент — необходимость учета погрешностей в случайной составляющей отклонения оценки $\hat{a}$ от оцениваемого параметра, в то время как при рассмотрении оценки максимального правдоподобия погрешности давали лишь смещение. Примем в соответствии с неравенством Чебышева:

$$|\,x_i - \bar{x}\,| \sim \sqrt{D(x_1)}, \tag{23}$$

тогда

$$АП[(x_i - \bar{x})^2] \sim 2b\sqrt{a}\Delta, \quad АП(s^2) \sim 2b\sqrt{a}\Delta.$$

Замечание. Если вычислять s^2 по формуле:

$$s^2 = \frac{1}{n-1}\sum_{1\le i\le n} x_i^2 - \frac{n}{n-1}(\bar{x})^2, \tag{24}$$

то аналогичные вычисления дают, что

$$АП(s^2) \sim 4ab\Delta,$$

т.е. погрешность при больших a существенно больше. Хотя правые части формул (22) и (24) тождественно равны, но погрешности вычислений по этим формулам весьма отличаются. Связано это с тем, что в формуле (24) последняя операция — нахождение разности двух больших чисел, примерно равных по величине (для выборки из гамма-распределения при большом значении параметра формы).

Из полученных результатов следует, что

$$ÀÏ\ (\hat{a}) = ÀÏ\left(\frac{(\overline{x})^2}{s^2}\right) \approx \frac{2\Delta}{b}(1+\sqrt{a}).$$

При выводе этой формулы использована линеаризация влияния погрешностей (выделение главного линейного члена). Используя связь (21) между абсолютной и относительной погрешностями, можно записать:

$$ÀÏ\ (\hat{a}) \approx 2a(1+\sqrt{a})\delta.$$

Эта формула отличается от приведенной в [4, с. 44 (19)]:

$$ÀÏ\ (\hat{a}) \approx 2a(1+3\sqrt{a})\delta,$$

поскольку в [4] вместо (23) использовалась оценка:

$$|x_i - \overline{x}| < 3\sqrt{D(x_1)}.$$

Используя соотношение (23), мы характеризуем влияние погрешностей «в среднем».

Доверительный интервал, соответствующий доверительной вероятности 0,95, имеет вид:

$$\left[\hat{a} - 2\hat{a}(1+\sqrt{\hat{a}})\delta - 1{,}96\sqrt{\frac{2\hat{a}(\hat{a}+1)}{n}};\ \hat{a} + 2\hat{a}(1+\sqrt{\hat{a}})\delta + 1{,}96\sqrt{\frac{2\hat{a}(\hat{a}+1)}{n}}\right].$$

Если $\hat{a} = 5{,}00;\ \delta = 0{,}01;\ n = 50$, то получаем доверительный интервал [2,54; 7,46] вместо [2,86; 7,14] при $\delta = 0$. Хотя при $\delta = 0$ доверительный интервал для a при использовании оценки метода моментов $\hat{a}$ шире, чем при использовании оценки максимального правдоподобия a^*, при $\delta = 0{,}01$ результат сравнения длин интервалов противоположен.

Необходимо выбрать способ сравнения двух методов оценивания параметра a, поскольку в длины доверительных интервалов входят две составляющие — зависящая от доверительной вероятности и не зависящая от нее. Выберем $\delta = 0{,}68$, т.е. $u\left(\frac{1+\gamma}{2}\right) = 1{,}00$. Тогда оценке максимального правдоподобия a^* соответствует полудлина доверительного интервала:

$$\nu(a^*) = 4a^2\delta + \sqrt{\frac{a(2a-1)}{n}}, \qquad (25)$$

а оценке $\hat{a}$ метода моментов соответствует полудлина доверительного интервала:

$$\nu(\hat{a}) = 2a(1+\sqrt{a})\delta + \sqrt{\frac{2a(a+1)}{n}}. \qquad (26)$$

Ясно, что больших a или больших n справедливо неравенство $\nu(a^*) > \nu(\hat{a})$, т.е. метод моментов лучше метода максимального правдоподобия, вопреки классическим результатам Р. Фишера при $\delta = 0$ [33, с. 99].

Из (25) и (26) элементарными преобразованиями получаем следующее правило принятия решений. Если

$$\delta\sqrt{n} \geq \frac{\sqrt{2a(a+1)} - \sqrt{a(2a-1)}}{4a^2 - 2a(1+\sqrt{a})} = B(a),$$

то $\nu(a^*) \geq \nu(\hat{a})$ и следует использовать $\hat{a}$; а если $\delta\sqrt{n} < B(a)$, то $\nu(a^*) < \nu(\hat{a})$ и надо применять a^*. Для выбора метода оценивания при обработке реальных данных целесообразно использовать $B(\hat{a})$ (см. раздел 5 в ГОСТ 11.011-83 (в настоящее время отменен, но может использоваться как научная публикация) [4, с. 10–11]).

Пример анализа реальных данных опубликован в [4].

На основе рассмотрения проблем оценивания параметров гамма-распределения можно сделать некоторые общие выводы. Если в классической теории математической статистики:

А. Существуют состоятельные оценки a_n параметра a,

$$\lim_{n\to\infty} M(a_n - a)^2 = 0.$$

Б. Для повышения точности оценивания объем выборки целесообразно безгранично увеличивать.

В. Оценки максимального правдоподобия лучше оценок метода моментов, то в статистике интервальных данных, учитывающей погрешности измерений, соответственно:

а) не существует состоятельных оценок: для любой оценки a_n существует константа c такая, что

$$\lim_{n\to\infty} M(a_n - a)^2 \geq c > 0;$$

б) не имеет смысла рассматривать объемы выборок, большие «рационального объема выборки» n_{rat};

в) оценки метода моментов в обширной области параметров (a, n, δ) лучше оценок максимального правдоподобия, в частности, при $a \to \infty$ и при $n \to \infty$.

Ясно, что приведенные выше результаты справедливы не только для рассмотренной задачи оценивания параметров гамма-распределения, но и для многих других постановок прикладной математической статистики.

Метрологические, методические, статистические и вычислительные погрешности. Целесообразно выделить ряд видов погрешностей статистических данных. Погрешности, вызванные неточностью измерения исходных данных, называем *метрологическими*. Их максимальное значение можно оценить с помощью нотны. Впрочем, выше на примере оценивания параметров гамма-распределения показано, что переход от максимального отклонения к реально имеющемуся в вероятностно-статистической модели не меняет выводы (с точностью до умножения предельных значений погрешностей Δ или δ на константы). Как правило, метрологические погрешности не убывают с ростом объема выборки.

Методические погрешности вызваны неадекватностью вероятностно-статистической модели, отклонением реальности от ее предпосылок. Неадекватность обычно не исчезает при росте объема выборки. Методические погрешности целесообразно изучать с помощью «общей схемы устойчивости» [3, 27], обобщающей популярную в теории робастных статистических процедур модель засорения большими выбросами. В настоящей главе методические погрешности не рассматриваются.

Статистическая погрешность — это та погрешность, которая традиционно рассматривается в математической статистике. Ее характеристики — дисперсия оценки, дополнение до 1 мощности критерия при фиксированной альтернативе и т.д. Как правило, статистическая погрешность стремится к 0 при росте объема выборки.

Вычислительная погрешность определяется алгоритмами расчета, в частности, правилами округления. На уровне чистой математики справедливо тождество правых частей формул (22) и (24), задающих выборочную дисперсию s^2, а на уровне вычислительной математики формула (22) дает при определенных условиях существенно больше верных значащих цифр, чем вторая [34, с. 51–52].

Выше на примере задачи оценивания параметров гамма-распределения рассмотрено совместное действие метрологических и вычислительных погрешностей, причем погрешности вычислений оценивались по классическим правилам для ручного счета [32]. Оказалось, что при таком подходе оценки метода моментов имеют преимущество перед оценками максимального правдоподобия в обширной области изменения параметров. Однако, если учитывать только метрологические погрешности, как это делалось выше в примерах 1–5, то с помощью аналогичных выкладок можно показать, что оценки этих двух типов имеют (при достаточно больших n) одинаковую погрешность.

Вычислительную погрешность здесь подробно не рассматриваем. Ряд интересных результатов о ее роли в статистике получили Н.Н. Ляшенко и М. С. Никулин [35].

Проведем сравнение методов оценивания параметров в более общей постановке.

В теории оценивания параметров классической математической статистики установлено, что метод максимального правдоподобия, как правило, лучше (в смысле асимптотической дисперсии и асимптотического среднего квадрата ошибки), чем метод моментов. Однако в интервальной статистике это, вообще говоря, не так, что продемонстрировано выше на примере оце-

нивания параметров гамма-распределения. Сравним эти два метода оценивания в случае интервальных данных в общей постановке. Поскольку метод максимального правдоподобия – частный случай метода минимального контраста, начнем с разбора этого несколько более общего метода.

Оценки минимального контраста. Пусть X — пространство, в котором лежат независимые одинаково распределенные случайные элементы $x_1, x_2, ..., x_n, ...$ Будем оценивать элемент пространства параметров Θ с помощью функции контраста $f : X \times \Theta \to R^1$. Оценкой минимального контраста называется:

$$\theta_n = Arg \min \left\{ \sum_{1 \le i \le n} f(x_i, \theta), \theta \in \Theta \right\}.$$

Если множество θ_n состоит из более чем одного элемента, то оценкой минимального контраста называют также любой элемент θ_n.

Оценками минимального контраста являются, в частности, многие робастные статистики [3, 36]. Эти оценки широко используются в статистике объектов нечисловой природы [3, 27], поскольку при $X = \Theta$ переходят в переходят в эмпирические средние, а если $X = \Theta$ — пространство бинарных отношений — в медиану Кемени.

Пусть в X имеется мера μ (заданная на той же σ-алгебре, что участвует в определении случайных элементов x_i), и $p(x;\theta)$ — плотность распределения x_i по мере μ. Если

$$f(x;\theta) = -\ln p(x;\theta),$$

то оценка минимального контраста переходит в оценку максимального правдоподобия.

Асимптотическое поведение оценок минимального контраста в случае пространств X и Θ общего вида хорошо изучено [37], в частности, известны условия состоятельности оценок. Здесь ограничимся случаем $X = R^1$, но при этом введя погрешности измерений ε_i. Примем также, что $\Theta = (\theta_{\min}, \theta_{\max}) \subseteq R^1$.

В рассматриваемой математической модели предполагается, что статистику известны лишь искаженные значения $y_i = x_i + \varepsilon_i$, $i = 1,2,...,n$. Поэтому вместо θ_n он вычисляет:

$$\theta_n^* = Arg \min \left\{ \sum_{1 \le i \le n} f(y_i, \theta), \theta \in \Theta \right\}.$$

Будем изучать величину $\theta_n^* - \theta_n$ в предположении, что погрешности измерений ε_i малы. Цель этого изучения — продемонстрировать идеи статистики интервальных данных при достаточно простых предположениях. Поэтому естественно следовать условиям и ходу рассуждений, которые обычно принимаются при изучении оценок максимального правдоподобия [38, п. 33.3].

Пусть θ_0 — истинное значение параметра, функция $f(x;\theta)$ трижды дифференцируема по θ, причем

$$\left|\frac{\partial^3 f(x;\theta)}{\partial\theta^3}\right| < H(x)$$

при всех x, θ. Тогда

$$\frac{\partial f(x;\theta)}{\partial\theta} = \frac{\partial f(x;\theta_0)}{\partial\theta} + \frac{\partial^2 f(x;\theta_0)}{\partial\theta^2}(\theta-\theta_0) + \frac{1}{2}\alpha(x)H(x)(\theta-\theta_0)^2, \qquad (27)$$

где $|\alpha(x)|<1$.

Используя обозначения векторов $x = (x_1, x_2, ..., x_n)$, $y = (y_1, y_2, ..., y_n)$, введем суммы:

$$B_0(x) = \frac{1}{n}\sum_{1\le i\le n}\frac{\partial f(x_i;\theta_0)}{\partial\theta}, \quad B_1(x) = \frac{1}{n}\sum_{1\le i\le n}\frac{\partial^2 f(x_i;\theta_0)}{\partial\theta^2}, \quad R(x) = \frac{1}{n}\sum_{1\le i\le n}H(x_i).$$

Аналогичным образом введем функции $B_0(y)$, $B_1(y)$, $R(y)$, в которых вместо x_i стоят y_i, $i = 1, 2, \ldots, n$.

Поскольку в соответствии с теоремой Ферма оценка минимального контраста θ_n удовлетворяет уравнению:

$$\sum_{1\le i\le n}\frac{\partial f(x_i;\theta_n)}{\partial\theta} = 0, \qquad (28)$$

то, подставляя в (27) x_i вместо x и суммируя по $i = 1, 2, \ldots, n$, получаем, что

$$0 = B_0(x) + B_1(x)(\theta_n-\theta_0) + \frac{\beta R(x)}{2}(\theta_n-\theta_0)^2, \quad |\beta|<1, \qquad (29)$$

откуда

$$\theta_n - \theta_0 = \frac{-B_0(x)}{B_1(x) + \dfrac{\beta R(x)}{2}(\theta_n - \theta_0)}. \qquad (30)$$

Решения уравнения (28) будем также называть оценками минимального контраста. Хотя уравнение (28) — лишь необходимое условие минимума, такое словоупотребление не будет вызывать трудностей.

Теорема 1. Пусть для любого x выполнено соотношение (27). Пусть для случайной величины x_1 с распределением, соответствующим значению параметра $\theta = \theta_0$, существуют математические ожидания:

$$M\frac{\partial f(x_1,\theta_0)}{\partial \theta_0} = 0, \quad M\frac{\partial^2 f(x_1,\theta_0)}{\partial \theta_0^2} = A \neq 0, \quad MH(x_1) = M < +\infty. \qquad (31)$$

Тогда существуют оценки минимального контраста θ_n такие, что $\theta_n \to \theta_0$ при $n \to \infty$ (в смысле сходимости по вероятности).

Доказательство. Возьмем $\varepsilon > 0$ и $\delta > 0$. В силу закона больших чисел (теорема Хинчина) существует $n(\varepsilon, \delta)$ такое, что для любого $n > n(\varepsilon, \delta)$ справедливы неравенства:

$$P\{|B_0| \geq \delta^2\} < \varepsilon/3, \quad P\{|B_1| < |A|/2\} < \varepsilon/3, \quad P\{R(x) > 2M\} < \varepsilon/3.$$

Тогда с вероятностью не менее $1-\varepsilon$ одновременно выполняются соотношения:

$$|B_0| \leq \delta^2, \quad |B_1| \geq |A|/2, \quad R(x) \leq 2M. \qquad (32)$$

При $\theta \in [\theta_0 - \delta; \theta_0 + \delta]$ рассмотрим многочлен второй степени:

$$y(\theta) = B_0(x) + B_1(x)(\theta - \theta_0) + \frac{\beta R(x)}{2}(\theta - \theta_0)^2$$

(см. формулу (29)). С вероятностью не менее $1-\varepsilon$ выполнены соотношения:

$$|B_0 + \frac{\beta R(x)}{2}(\theta - \theta_0)^2| \leq |B_0| + \frac{R(x)\delta^2}{2} \leq \delta^2(M+1), \quad |B_1\delta| \geq \frac{|A|\,\delta}{2}.$$

Если $0 < 2(M+1)\delta < |A|$, то знак $y(\theta)$ в точках $\theta_1 = \theta_0 - \delta$ и $\theta_2 = \theta_0 + \delta$ определяется знаком линейного члена $B_1(\theta_i - \theta_0)$, $i = 1,2$, следовательно, знаки $y(\theta_1)$ и $y(\theta_2)$ различны, а потому существует $\theta_n \in [\theta_0 - \delta; \theta_0 + \delta]$ такое, что $y(\theta_n) = 0$, что и требовалось доказать.

Теорема 2. Пусть выполнены условия теоремы 1 и, кроме того, для случайной величины x_1, распределение которой соответствует значению параметра $\theta = \theta_0$, существует математическое ожидание:

$$M\left(\frac{\partial f(x_1;\theta_0)}{\partial \theta_0}\right) = \sigma^2.$$

Тогда оценка минимального контраста имеет асимптотически нормальное распределение:

$$\lim_{n\to\infty} P\left\{\sqrt{n}\frac{|A|}{\sigma}(\theta_n - \theta_0) < x\right\} = \Phi(x) \qquad (33)$$

для любого x, где $\Phi(x)$ — функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1.

Доказательство. Из центральной предельной теоремы вытекает, что числитель в правой части формулы (30) асимптотически нормален с математическим ожиданием 0 и дисперсией σ^2. Первое слагаемое в знаменателе формулы (30) в силу условий (31) и закона больших чисел сходится по вероятности к $A \neq 0$, а второе слагаемое по тем же основанием и с учетом теоремы 1 — к 0. Итак, знаменатель сходится по вероятности к $A \neq 0$. Доказательство теоремы 2 завершает ссылка на теорему о наследовании сходимости (Приложение 1).

Нотна оценки минимального контраста. Аналогично (30) нетрудно получить, что

$$\theta_n^* - \theta_0 = \frac{-B_0(y)}{B_1(y) + \dfrac{\beta(y)R(y)}{2}(\theta_n^* - \theta_0)}, \quad |\beta(y)| < 1. \qquad (34)$$

Следовательно, $\theta_n^* - \theta_n$ есть разность правых частей формул (30) и (34). Найдем максимально возможное значение (т.е. нотну) величины $|\theta_n^* - \theta_n|$ при ограничениях (1) на абсолютные погрешности результатов измерений.

Покажем, что при $\Delta \to 0$ для некоторого $C>0$ нотна имеет вид:

$$N_{\theta_n}(x) = \sup_{\{\varepsilon\}} |\theta_n^* - \theta_n| = C\Delta(1 + o(1)). \qquad (35)$$

Поскольку $\theta_n^* - \theta_n = (\theta_n^* - \theta_0) + (\theta_0 - \theta_n)$, то из (33) и (35) следует, что

$$\sup_{\{\varepsilon\}} M(\theta_n^* - \theta_n)^2 = \left(C^2\Delta^2 + \frac{\sigma^2}{A^2 n} \right)(1 + o(1)). \tag{36}$$

Можно сказать, что наличие погрешностей ε_i приводит к появлению систематической ошибки (смещения) у оценки метода максимального правдоподобия, и нотна является максимально возможным значением этой систематической ошибки.

В правой части (36) первое слагаемое — квадрат асимптотической нотны, второе соответствует статистической ошибке. Приравнивая их, получаем рациональный объем выборки:

$$n_{rat} = \left(\frac{\sigma}{CA\Delta} \right)^2.$$

Остается доказать соотношение (35) и вычислить C. Укажем сначала условия, при которых $\theta_n^* \to \theta_0$ (по вероятности) при $n \to \infty$ одновременно с $\Delta \to 0$.

Теорема 3. Пусть существуют константа Δ_0 и функции $g_1(x)$, $g_2(x)$, $g_3(x)$ такие, что при $0 \le \Delta \le \Delta_0$ и $-1 \le \gamma \le 1$ выполнены неравенства (ср. формулу (27)):

$$\begin{aligned} &\left| \frac{\partial f(x;\theta_0)}{\partial\theta} - \frac{\partial f(x+\gamma\Delta;\theta_0)}{\partial\theta} \right| \le g_1(x)\Delta, \\ &\left| \frac{\partial^2 f(x;\theta_0)}{\partial\theta^2} - \frac{\partial^2 f(x+\gamma\Delta;\theta_0)}{\partial\theta^2} \right| \le g_2(x)\Delta, \\ &| H(x) - H(x+\gamma\Delta) | \le g_3(x)\Delta \end{aligned} \tag{37}$$

при всех x. Пусть для случайной величины x_1, распределение которой соответствует $\theta = \theta_0$, существуют $m_1 = Mg_1(x_1)$, $m_2 = Mg_2(x_1)$ и $m_3 = Mg_3(x_1)$. Пусть выполнены условия теоремы 1. Тогда $\theta_n^* \to \theta_0$ (по вероятности) при $\Delta \to 0$, $n \to \infty$.

Доказательство проведем по схеме доказательства теоремы 1. Из неравенств (37) вытекает, что

$$\begin{aligned} &| B_0(y) - B_0(x) | \le \Delta \left(\frac{1}{n} \sum_{1 \le i \le n} g_1(x_i) \right), \\ &| B_1(y) - B_1(x) | \le \Delta \left(\frac{1}{n} \sum_{1 \le i \le n} g_2(x_i) \right), \\ &| R(y) - R(x) | \le \Delta \left(\frac{1}{n} \sum_{1 \le i \le n} g_3(x_i) \right). \end{aligned} \tag{38}$$

Возьмем $\varepsilon > 0$ и $\delta > 0$. В силу закона больших чисел (теорема Хинчина) существует $n(\varepsilon, \delta)$ такое, что для любого $n > n(\varepsilon, \delta)$ справедливы неравенства:

$$P\left\{|B_0| \geq \frac{\delta^2}{2}\right\} < \frac{\varepsilon}{6}, \quad P\left\{|B_1| < \frac{3|A|}{4}\right\} < \frac{\varepsilon}{6}, \quad P\left\{R(x) > \frac{3M}{2}\right\} < \frac{\varepsilon}{6},$$

$$P\left\{\frac{1}{n}\sum_{1 \leq i \leq n} g_j(x_i) > 2m_j\right\} < \frac{\varepsilon}{6}, \quad j = 1,2,3.$$

Тогда с вероятностью не менее $1-\varepsilon$ одновременно выполняются соотношения:

$$|B_0| < \frac{1}{2}\delta^2, \quad |B_1| \geq \frac{3|A|}{4}, \quad R(x) \leq \frac{3M}{2}, \quad \frac{1}{n}\sum_{1 \leq i \leq n} g_j(x_i) \leq 2m_j, \quad j = 1,2,3.$$

В силу (38) при этом:

$$|B_0(y)| < \frac{1}{2}\delta^2 + 2\Delta m_1, \quad |B_1(y)| \geq \frac{3|A|}{4} - 2\Delta m_2, \quad R(y) \leq \frac{3M}{2} + 2\Delta m_3.$$

Пусть

$$0 \leq \Delta \leq \min\left\{\frac{1}{4}\frac{\delta^2}{m_1};\ \frac{1}{8}\frac{|A|}{m_2};\ \frac{1}{4}\frac{M}{m_3}\right\}.$$

Тогда с вероятностью не менее $1-\varepsilon$ одновременно выполняются соотношения (ср. (32)):

$$|B_0(y)| \leq \delta^2, \quad |B_1(y)| \geq |A|/2, \quad R(y) \leq 2M.$$

Завершается доказательство дословным повторением такового в теореме 1, с единственным отличием — заменой в обозначениях x на y.

Теорема 4. Пусть выполнены условия теоремы 3 и, кроме того, существуют математические ожидания (при $\theta = \theta_0$):

$$M\left|\frac{\partial^2 f(x_1, \theta_0)}{\partial x \partial \theta}\right|, \quad M\left|\frac{\partial^3 f(x_1, \theta_0)}{\partial x \partial \theta^2}\right|. \tag{39}$$

Тогда выполнено соотношение (35) с:

$$C = \frac{1}{|A|} M\left|\frac{\partial^2 f(x_1, \theta_0)}{\partial x \partial \theta}\right|. \tag{40}$$

Доказательство. Воспользуемся следующим элементарным соотношением. Пусть a и b — бесконечно малые по сравнению с Z и B соответственно. Тогда с точностью до бесконечно малых более высокого порядка:

$$\frac{Z+a}{B+b}-\frac{Z}{B}=\frac{aB-bZ}{B^2}.$$

Чтобы применить это соотношение к анализу $\theta_n^*-\theta_n$ в соответствии с (30), (34) и теоремой 2, положим:

$$Z=B_0(x),\quad a=B_0(y)-B_0(x),\quad B=B_1(x),\quad b=(B_1(y)-B_1(x))+\frac{\beta(y)R(y)}{2}(\theta_n^*-\theta_0).$$

В силу условий теоремы 4 при малых ε_i с точностью до членов более высокого порядка:

$$B_0(y)-B_0(x)=\frac{1}{n}\sum_{1\le i\le n}\frac{\partial^2 f(x_i;\theta_0)}{\partial x_i\partial\theta_0}\varepsilon_i,\quad B_1(y)-B_1(x)=\frac{1}{n}\sum_{1\le i\le n}\frac{\partial^3 f(x_i;\theta_0)}{\partial x_i\partial\theta_0^2}\varepsilon_i.$$

При $\Delta\to 0$ эти величины бесконечно малы, а потому с учетом сходимости $B_1(x)$ к A и теоремы 3:

$$\theta_n^*-\theta_n=\frac{1}{A^2}\{(B_0(y)-B_0(x))A-(B_1(y)-B_1(x))B_0(x)\}=\frac{1}{A^2 n}\sum_{1\le i\le n}\gamma_i\varepsilon_i$$

с точностью до бесконечно малых более высокого порядка, где

$$\gamma_i=\frac{\partial^2 f(x_i;\theta_0)}{\partial x_i\partial\theta_0}A-\frac{\partial^3 f(x_i;\theta_0)}{\partial x_i\partial\theta_0^2}B_0(x).$$

Ясно, что задача оптимизации:

$$\begin{cases}\displaystyle\sum_{1\le i\le n}\gamma_i\varepsilon_i\to\max\\ |\varepsilon_i|\le\Delta,\quad i=1,2,\ldots,n,\end{cases}\tag{41}$$

имеет решение:

$$\varepsilon_i=\begin{cases}\Delta, & \gamma_i\ge 0,\\ -\Delta, & \gamma_i<0,\end{cases}$$

при этом максимальное значение линейной формы есть $\Delta\sum_{1\le i\le n}|\gamma_i|$. Поэтому

$$\sup_{\{\varepsilon\}}|\theta_n^*-\theta_n|=\frac{\Delta}{A^2n}\sum_{1\le i\le n}|\gamma_i|. \tag{42}$$

С целью упрощения правой части (42) воспользуемся тем, что

$$\frac{1}{n}\sum_{1\le i\le n}|\gamma_i|=\frac{|A|}{n}\sum_{1\le i\le n}\left|\frac{\partial^2 f(x_i;\theta_0)}{\partial x\partial\theta_0}\right|+\alpha\frac{|B_0(x)|}{n}\sum_{1\le i\le n}\left|\frac{\partial^3 f(x_i;\theta_0)}{\partial x\partial\theta_0^2}\right|, \tag{43}$$

где $|\alpha|\le 1$. Поскольку при $n\to\infty$:

$$\frac{1}{n}\sum_{1\le i\le n}\left|\frac{\partial^3 f(x_i;\theta_0)}{\partial x\partial\theta_0^2}\right|\to M\left|\frac{\partial^3 f(x_1;\theta_0)}{\partial x\partial\theta_0^2}\right|<+\infty,\quad B_0(x)\to 0$$

по вероятности, то второе слагаемое в (43) сходится к 0, а первое в силу закона больших чисел с учетом (39) сходится к CA^2, где C определено в (40). Теорема 4 доказана.

Оценки метода моментов. Пусть $g:R^k\to R^1,\quad h_j:R^1\to R^1, j=1,2,...,k,$ — некоторые функции. Рассмотрим аналоги выборочных моментов:

$$m_j=\frac{1}{n}\sum_{1\le i\le n}h_j(x_i), j=1,2,...,k.$$

Оценки метода моментов имеют вид:

$$\hat\theta_n(x)=g(m_1,m_2,...,m_k)$$

(функции g и h_j должны удовлетворять некоторым дополнительным условиям [39, с. 80], которые здесь не приводим). Очевидно, что

$$\hat\theta_n(y)-\hat\theta_n(x)=\sum_{1\le j\le k}\frac{\partial g}{\partial m_j}(m_j(y)-m_j(y)),$$
$$m_j(y)-m_j(x)=\frac{1}{n}\sum_{1\le i\le n}\frac{dh_j(x_i)}{dx_i}\varepsilon_i,\quad j=1,2,...,k, \tag{44}$$

с точностью до бесконечно малых более высокого порядка, а потому с той же точностью:

$$\hat\theta_n(y)-\hat\theta_n(x)=\frac{1}{n}\sum_{1\le i\le n}\left(\sum_{1\le j\le k}\frac{\partial g}{\partial m_j}\frac{dh_j(x_i)}{dx_i}\right)\varepsilon_i. \tag{45}$$

Теорема 5. Пусть при $\theta = \theta_0$ существуют математические ожидания:

$$M_j = Mm_j = Mh_j(x_1), \quad M\left(\frac{dh_j(x_1)}{dx_1}\right), \quad j = 1,2,...,n,$$

функция g дважды непрерывно дифференцируема в некоторой окрестности точки $(M_1, M_2, ..., M_k)$. Пусть существует функция $t: R^1 \to R^1$ такая, что

$$\sup_{|x-y|\le\Delta} \left| h_j(y) - h_j(x) - \frac{dh_j(x)}{dx}(y-x) \right| \le t(x)\Delta^2, \quad j = 1,2,...,k, \tag{46}$$

причем $Mt(x_1)$ существует.

Тогда

$$\sup_{\{\varepsilon\}} | \hat{\theta}_n(y) - \hat{\theta}_n(x) | = C_1\Delta$$

с точностью до бесконечно малых более высокого порядка, причем

$$C_1 = M\left| \sum_{1\le j\le k} \frac{\partial g(M_1, M_2, ..., M_k)}{\partial m_j} \frac{dh_j(x_1)}{dx_1} \right|.$$

Доказательство теоремы 5 сводится к обоснованию проведенных ранее рассуждений, позволивших получить формулу (45). В условиях теоремы 5 собраны предположения, достаточные для такого обоснования. Так, условие (46) дает возможность обосновать соотношения (44); существование $M\left(\frac{dh_j(x_1)}{dx_1}\right)$ обеспечивает существование C_1, и т.д. Завершает доказательство ссылка на решение задачи оптимизации (41) и применение закона больших чисел.

Полученные в теоремах 4 и 5 нотны оценок минимального контраста и метода моментов, асимптотические дисперсии этих оценок (см. теорему 2 и [40] соответственно) позволяют находить рациональные объемы выборок, строить доверительные интервалы с учетом погрешностей измерений, а также сравнивать оценки по среднему квадрату ошибки (36). Подобное сравнение было проведено для оценок максимального правдоподобия и метода моментов параметров гамма-распределения. Установлено, что классический

вывод о преимуществе оценок максимального правдоподобия [33, с. 99–100] неверен в случае $\Delta > 0$.

4.3. ИНТЕРВАЛЬНЫЕ ДАННЫЕ В ЗАДАЧАХ ПРОВЕРКИ ГИПОТЕЗ

С позиций статистики интервальных данных целесообразно изучить все практически используемые процедуры прикладной математической статистики, установить соответствующие нотны и рациональные объемы выборок. Это позволит устранить разрыв между математическими схемами прикладной статистики и реальностью влияния погрешностей наблюдений на свойства статистических процедур. Статистика интервальных данных — часть теории устойчивых статистических процедур, развитой в монографии [3]. Часть, более адекватная реальной статистической практике, чем некоторые другие постановки, например, с засорением нормального распределения большими выбросами.

Рассмотрим подходы статистики интервальных данных в задачах проверки статистических гипотез. Пусть принятие решения основано на сравнении рассчитанного по выборке значения статистики критерия $f = f(y_1, y_2, ..., y_n)$ с граничным значением C: если $f>C$, то гипотеза отвергается, если же $f \leq C$, то принимается. С учетом погрешностей измерений выборочное значение статистики критерия может принимать любое значение в интервале $[f(y) - N_f(y); f(y) + N_f(y)]$. Это означает, что «истинное» значение порога, соответствующее реально используемому критерию, находится между $C\text{-}N_f(y)$ и $C+N_f(y)$, а потому уровень значимости описанного правила (критерия) лежит между $1 - P(C + N_f(y))$ и $1 - P(C - N_f(y))$, где $P(Z)=P(f<Z)$.

Пример 1. Пусть $x_1, x_2, ..., x_n$ — выборка из нормального распределения с математическим ожиданием a и единичной дисперсией. Необходимо проверить гипотезу H_0: $a = 0$ при альтернативе $H_1 : a \neq 0$.

Как известно из любого учебного курса математической статистики, следует использовать статистику $f = \sqrt{n}\,|\,\bar{y}\,|$ и порог $C = \Phi(1 - \alpha/2)$, где α — уровень значимости, Ф (•) — функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1. В частности, C = 1,96 при $\alpha = 0{,}05$.

При ограничениях (1) на абсолютную погрешность $N_f(y) = \sqrt{n}\Delta$. Например, если Δ = 0,1, а n = 100, то $N_f(y)$ = 1,0. Это означает, что истинное значение порога лежит между 0,96 и 2,96, а истинный уровень значимости —

между 0,003 и 0,34. Можно сделать и другой вывод: нулевую гипотезу H_0 допустимо отклонить на уровне значимости 0,05 лишь тогда, когда $f > 2{,}96$.

Если же $n = 400$ при $\Delta = 0{,}1$, то $N_f(y) = 2{,}0$ и $C - N_f(y) = -0{,}04$, в то время как $C + N_f(y) = 3{,}96$. Таким образом, даже в случае $x = 0$ гипотеза H_0 может быть отвергнута только из-за погрешностей измерений результатов наблюдений.

Вернемся к общему случаю проверки гипотез. С учетом погрешностей измерений граничное значение C_α в статистике интервальных данных целесообразно заменить на $C_\alpha + N_f(y)$. Такая замена дает гарантию, что вероятность отклонения нулевой гипотезы H_0, когда она верна, не более α. При проверке гипотез аналогом статистической погрешности, рассмотренной выше в задачах оценивания, является C_α. Суммарная погрешность имеет вид $C_\alpha + N_f(y)$. Исходя из принципа уравнивания погрешностей [3], целесообразно определять рациональный объем выборки из условия:

$$C_\alpha = N_f(y).$$

Если $f = |f_1|$, где f_1 при справедливости H_0 имеет асимптотически нормальное распределение с математическим ожиданием 0 и дисперсией σ^2 / n, то

$$C_\alpha = u\left(1 - \frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}} \qquad (47)$$

при больших n, где $u(1-\alpha/2)$ — квантиль порядка $1-\alpha/2$ стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1. Из (47) вытекает, что в рассматриваемом случае:

$$n_{rat} = \left[\frac{u(1-\alpha/2)\sigma}{N_f(y)}\right]^2.$$

В условиях примера 1 $f_1 = \bar{y}$ и

$$n_{rat} = \frac{3{,}84}{\Delta^2} = 384.$$

Пример 2. Рассмотрим статистику одновыборочного критерия Стьюдента:

$$t = \sqrt{n}\frac{\bar{y}}{s(y)} = \frac{\sqrt{n}}{v},$$

где v — выборочный коэффициент вариации. Тогда с точностью до бесконечно малых более высокого порядка нотна для t имеет вид:

$$N_t(y) = \frac{\sqrt{n}}{v^2} N_v(y),$$

где $N_v(y)$ — рассмотренная ранее нотна для выборочного коэффициента вариации.

Поскольку распределение статистики Стьюдента t сходится к стандартному нормальному, то небольшое изменение предыдущих рассуждений дает:

$$n_{rat} = \frac{v^4 u^2 (1-\alpha/2)}{N_v^2(y)}.$$

Пример 3. Рассмотрим двухвыборочный критерий Смирнова, предназначенный для проверки однородности (совпадения) функций распределения двух независимых выборок [41]. Статистика этого критерия имеет вид:

$$D_{mn} = \sup_x \left| F_m(x) - G_n(x) \right|,$$

где $F_m(x)$ — эмпирическая функция распределения, построенная по первой выборке объема m, извлеченной из генеральной совокупности с функцией распределения $F(x)$, а $G_n(x)$ — эмпирическая функция распределения, построенная по второй выборке объема n, извлеченной из генеральной совокупности с функцией распределения G(x). Нулевая гипотеза имеет вид $H_0 : F(x) \equiv G(x)$, альтернативная состоит в ее отрицании: $H_1 : F(x) \neq G(x)$ при некотором x. Значение статистики сравнивают с порогом $D(\alpha, m, n)$, зависящим от уровня значимости α и объемов выборок m и n. Если значение статистики не превосходит порога, то принимают нулевую гипотезу, если больше порога — альтернативную. Пороговые значения $D(\alpha, m, n)$ берут из таблиц [42]. Описанный критерий иногда неправильно называют критерием Колмогорова — Смирнова. История вопроса описана в [43].

При ограничениях (1) на абсолютные погрешности и справедливости нулевой гипотезы $H_0 : F(x) \equiv G(x)$ нотна имеет вид (при больших объемах выборок):

$$N_D = \sup_x | F(x+\Delta) - F(x-\Delta) | .$$

Если $F(x) = G(x) = x$ при $0 \le x \le 1$, то $N_D = 2\Delta$. С помощью условия $C_\alpha = N_f(y)$ при уровне значимости $\alpha = 0{,}05$ и достаточно больших объемах выборок (т.е. используя асимптотическое выражение для порога согласно [42]) получаем, что выборки имеет смысл увеличивать, если

$$\frac{mn}{m+n} \le \frac{0{,}46}{\Delta^2}.$$

Правая часть этой формулы при $\Delta = 0{,}1$ равна 46. Если $m = n$, то последнее неравенство переходит в $n \le 92$.

Теоретические результаты в области статистических методов входят в практику через алгоритмы расчетов, воплощенные в программные средства (пакеты программ, диалоговые системы). Ввод данных в современном статистической программной системе должен содержать запросы о погрешностях результатов измерений. На основе ответов на эти запросы вычисляются нотны рассматриваемых статистик, а затем — доверительные интервалы при оценивании, разброс уровней значимости при проверке гипотез, рациональные объемы выборок. Необходимо использовать систему алгоритмов и программ статистики интервальных данных, «параллельную» подобным системам для классической математической статистики.

4.4. ЛИНЕЙНЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ ИНТЕРВАЛЬНЫХ ДАННЫХ

Перейдем к многомерному статистическому анализу. Сначала с позиций асимптотической математической статистики интервальных данных рассмотрим оценки метода наименьших квадратов (МНК).

Статистическое исследование зависимостей — одна из наиболее важных задач, которые возникают в различных областях науки и техники. Под словами «исследование зависимостей» имеется в виду выявление и описание существующей связи между исследуемыми переменными на основании результатов статистических наблюдений. К методам исследования зависимостей относятся регрессионный анализ, многомерное шкалирование, идентификация параметров динамических объектов, факторный анализ, дисперсионный анализ, корреляционный анализ и др. Однако многие реальные ситуации характеризуются наличием данных интервального типа, причем из-

вестны допустимые границы погрешностей (например, из технических паспортов средств измерения).

Если какая-либо группа объектов характеризуется переменными $X_1, X_2, ..., X_m$ и проведен эксперимент, состоящий из n опытов, где в каждом опыте эти переменные измеряются один раз, то экспериментатор получает набор чисел: $X_{1j}, X_{2j}, ..., X_{mj}$ $(j = 1, ..., n)$.

Однако процесс измерения, какой бы физической природы он ни был, обычно не дает однозначный результат. Реально результатом измерения какой-либо величины X являются два числа: X_H — нижняя граница и X_B — верхняя граница. Причем $X_{ИСТ} \in [X_H, X_B]$, где $X_{ИСТ}$ — истинное значение измеряемой величины. Результат измерения можно записать как X: $[X_H, X_B]$. Интервальное число X может быть представлено другим способом, а именно, X: $[X_m, \Delta_x]$, где $X_H = X_m - \Delta_x$, $X_H = X_m + \Delta_x$. Здесь X_m — центр интервала (как правило, не совпадающий с $X_{ИСТ}$), а Δ_x — максимально возможная погрешность измерения.

Метод наименьших квадратов для интервальных данных. Пусть математическая модель задана следующим образом:

$$y = Q(x, b) + \varepsilon,$$

где $x = (x_1, x_2, ..., x_m)$ — вектор влияющих переменных (факторов), поддающихся измерению; $b = (b_1, b_2, ..., b_r)$ — вектор оцениваемых параметров модели; y — отклик модели (скаляр); $Q(x, b)$ — скалярная функция векторов x и b; наконец, ε — случайная ошибка (невязка, погрешность).

Пусть проведено n опытов, причем в каждом опыте измерены (один раз) значения отклика (y) и вектора факторов (x). Результаты измерений могут быть представлены в следующем виде:

$$X = \{x_{ij}; i = 1, ..., n; j = 1, ..., m\},\ Y = (y_1, y_2, ..., y_n),\ E = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n),$$

где X — матрица значений измеренного вектора (x) в n опытах; Y — вектор значений измеренного отклика в n опытах; E — вектор случайных ошибок. Тогда выполняется матричное соотношение:

$$Y = Q(X, b) + E,$$

где $Q(X, b) = (Q(x_1, b), Q(x_2, b), ..., Q(x_n, b))^T$, причем $x_1, x_2, ..., x_n$ — m-мерные вектора, которые составляют матрицу $X = (x_1, x_2, ..., x_n)^T$.

Введем меру близости $d(Y, Q)$ между векторами Y и Q. В МНК в качестве $d(Y, Q)$ берется квадратичная форма взвешенных квадратов ε_i^2 невязок $\varepsilon_i = y_i - Q(x_i, b)$, т.е.

$$d(Y, Q) = [Y - Q(X, b)]^T W[Y - Q(X, b)],$$

где $W = \{w_{ij}, i, j = 1, \ldots, n\}$ — матрица весов, не зависящая от b. Тогда в качестве оценки b можно выбрать такое b^*, при котором мера близости $d(Y, Q)$ принимает минимальное значение, т.е.

$$b^* = \{b : d(Y, Q) \to \min_{\{b\}}\}.$$

В общем случае решение этой экстремальной задачи может быть не единственным. Поэтому в дальнейшем будем иметь в виду одно из этих решений. Оно может быть выражено в виде некоторой вектор-функции $b^* = f(X, Y)$, где $f(X, Y) = (f_1(X, Y), f_2(X, Y), \ldots, f_r(X, Y))^T$, причем действительнозначные функции $f_i(X, Y)$ непрерывны и дифференцируемы по $(X, Y) \in Z$, где Z — область определения функции $f(X, Y)$. Эти свойства функции $f(X, Y)$ дают возможность использовать подходы статистики интервальных данных.

Преимущество метода наименьших квадратов заключается в сравнительной простоте и универсальности вычислительных процедур. Однако не всегда оценка МНК является состоятельной (при функции $Q(X, b)$, не являющейся линейной по векторному параметру b), что ограничивает его применение на практике.

Важным частным случаем является линейный МНК, когда $Q(x, b)$ есть линейная функция от b:

$$y = b_o x_o + b_1 x_1 + \ldots + b_m x_m + \varepsilon = b x^T + \varepsilon,$$

где, возможно, $x_o = 1$, а b_o — свободный член линейной комбинации. Как известно, в этом случае МНК-оценка имеет вид:

$$b^* = (X^T W X)^{-1} X^T W Y.$$

Если матрица $X^T W X$ не вырождена, то эта оценка является единственной. Если матрица весов W единичная, то

$$b^* = (X^T X)^{-1} X^T Y.$$

Пусть выполняются следующие предположения относительно распределения ошибок ε_i:

- ошибки ε_i имеют нулевые математические ожидания $M\{\varepsilon_i\} = 0$;
- результаты наблюдений имеют одинаковую дисперсию $D\{\varepsilon_i\} = \sigma^2$;
- ошибки наблюдений некоррелированы, т.е. $cov\{\varepsilon_i, \varepsilon_j\} = 0$.

Тогда, как известно, оценки МНК являются наилучшими линейными оценками, т.е. состоятельными и несмещенными оценками, которые представляют собой линейные функции результатов наблюдений и обладают минимальными дисперсиями среди множества всех линейных несмещенных оценок. Далее именно этот наиболее практически важный частный случай рассмотрим более подробно.

Как и в других постановках асимптотической математической статистики интервальных данных, при использовании МНК измеренные величины отличаются от истинных значений из-за наличия погрешностей измерения. Запишем истинные данные в следующей форме:

$$X_R = \{x_{ij}^R; i = \overline{1,n}; j = \overline{1,m}\}, Y_R = (y_1^R, y_2^R, ..., y_n^R),$$

где R — индекс, указывающий на то, что значение истинное. Истинные и измеренные данные связаны следующим образом:

$$X = X_R + \Delta X, \;\; Y = Y_R + \Delta Y,$$

где $\Delta X = \{\Delta x_{ij}; i = \overline{1,n}; j = \overline{1,m}\}, \Delta Y = (\Delta y_1, \Delta y_2, ..., \Delta y_n)$. Предположим, что погрешности измерения отвечают граничным условиям:

$$|\Delta x_{ij}| \le \Delta_j^x \;\; (i = 1,2,...,n, j = 1,2,...,m), \;\; ,|\Delta y_i| \le \Delta^y \;\; (i = 1,2,...,n), \qquad (48)$$

аналогичным ограничениям (1).

Пусть множество W возможных значений (X_R, Y_R) входит в Z — область определения функции $f(X, Y)$. Рассмотрим b^{*R} — оценку МНК, рассчитанную по истинным значениям факторов и отклика, и b^* — оценку МНК, найденную по искаженным погрешностями данным. Тогда

$$\Delta b^* = b^{*R} - b^* = f(X_R, Y_R) - f(X, Y).$$

Ввести понятие *нотны* придется несколько иначе, чем это было сделано выше, поскольку оценивается не одномерный параметр, а вектор. Положим:

$$n(1) = (\sup \Delta b_1^*, \sup \Delta b_2^*, ..., \sup \Delta b_r^*)^T, \quad n(2) = -(\inf \Delta b_1^*, \inf \Delta b_2^*, ..., \inf \Delta b_r^*)^T.$$

Будем называть $n(1)$ нижней *нотной*, а $n(2)$ верхней *нотной*. Предположим, что при безграничном возрастании числа измерений n, т.е. при $n \to \infty$, вектора $n(1)$, $n(2)$ стремятся к постоянным значениям $N(1)$, $N(2)$ соответственно. Тогда $N(1)$ будем называть нижней асимптотической *нотной*, а $N(2)$ — верхней асимптотической *нотной*.

Рассмотрим доверительное множество $B_\alpha = B_\alpha(n, b^{*R})$ для вектора параметров b, т.е. замкнутое связное множество точек в r-мерном евклидовом пространстве такое, что $P(b \in B_\alpha) = \alpha$, где α — доверительная вероятность, соответствующая B_α (α ≈ 1). Другими словами, $B_\alpha\ (n,\ b^{*R})$ есть область рассеивания (аналог эллипсоида рассеивания) случайного вектора b^{*R} с доверительной вероятностью α и числом опытов n.

Из определения верхней и нижней *нотн* следует, что всегда $b^{*R} \in [b^* - n(1); b^* + n(2)]$. (т.е. по каждой координате выполнено соответствующее неравенство). В соответствии с определением нижней асимптотической нотны и верхней асимптотической нотны можно считать, что $b^{*R} \in [b^* - N(1); b^* + N(2)]$ при достаточно большом числе наблюдений n. Этот многомерный интервал описывает r-мерный гиперпараллелепипед P.

Каким-либо образом разобьем P на L гиперпараллелепипедов. Пусть b_k — внутренняя точка k-го гиперпараллелепипеда. Учитывая свойства доверительного множества и устремляя L к бесконечности, можно утверждать, что $P(b \in C) \geq \alpha$, где

$$C = \lim_{L \to \infty} \bigcup_{1 \leq k \leq L} B_\alpha(n, b_k).$$

Таким образом, множество C характеризует неопределенность при оценивании вектора параметров b. Его можно назвать доверительным множеством в статистике интервальных данных.

Введем некоторую меру $M(X)$, характеризующую «величину» множества $X \subseteq R^r$. По определению меры она удовлетворяет условию: если

$X = Z \cup Y$ и $Z \cap Y = \varnothing$, то $M(X)=M(Z)+M(Y)$. Примерами такой меры являются площадь для $r = 2$ и объем для $r = 3$. Тогда:

$$M(C) = M(P) + M(F), \tag{49}$$

где $F = C \setminus P$. Здесь $M(F)$ характеризует меру статистической неопределенности, в большинстве случаев она убывает при увеличении числа опытов n. В то же время $M(P)$ характеризует меру интервальной (метрологической) неопределенности, и, как правило, $M(P)$ стремится к некоторой постоянной величине при увеличении числа опытов n. Пусть теперь требуется найти то число опытов, при котором статистическая неопределенность составляет δ-ю часть общей неопределенности, т.е.

$$M(F) = \delta M(C), \tag{50}$$

где $\delta < 1$. Тогда, подставив соотношение (50) в равенство (49) и решив уравнение относительно n, получим искомое число опытов. В асимптотической математической статистике интервальных данных оно называется «рациональным объемом выборки». При этом δ есть «степень малости» статистической неопределенности $M(P)$ относительно всей неопределенности. Она выбирается из практических соображений. При использовании «принципа уравнивания погрешностей» согласно [3] имеем $\delta = 1 / 2$.

Метод наименьших квадратов для линейной модели. Рассмотрим наиболее важный для практики частный случай МНК, когда модель описывается линейным уравнением (см. выше).

Для простоты описания преобразований пронормируем переменные x_{ij}, y_i, следующим образом:

$$x_{ij}^0 = (x_{ij} - \bar{x}_j) / s(x_j), \quad y_i^0 = (y_i - \bar{y}) / s(y),$$

где

$$\bar{x}_j = \frac{1}{n}\sum_{1\le i\le n} x_{ij}, \quad s^2(x_j) = \frac{1}{n}\sum_{1\le i\le n}(x_{ij} - \bar{x}_j)^2, \quad \bar{y} = \frac{1}{n}\sum_{1\le i\le n} y_i, \quad s^2(y) = \frac{1}{n}\sum_{1\le i\le n}(y_i - \bar{y})^2.$$

Тогда

$$\bar{x}_j^0 = 0 \quad s^2(x_j^0) = \frac{1}{n}\sum_{1\le i\le n}(x_{ij}^0 - \bar{x}_j^0)^2 = 1, \quad \bar{y}^0 = 0 \quad s^2(y^0) = \frac{1}{n}\sum_{1\le i\le n}(y_i^0 - \bar{y}^0)^2 = 1, \quad j = 1,2,\ldots, m..$$

В дальнейшем изложении будем считать, что рассматриваемые переменные пронормированы описанным образом, и верхние индексы ^{0} опустим. Для облегчения демонстрации основных идей примем достаточно естественные предположения.

1. Для рассматриваемых переменных существуют следующие пределы:

$$\lim_{n\to\infty}\frac{1}{n}\sum_{1\le i\le n}x_{ij}x_{ik}=0,\quad j,k=1,2,...,m.$$

2. Количество опытов n таково, что можно пользоваться асимптотическими результатами, полученными при $n\to\infty$.

3. Погрешности измерения удовлетворяют одному из следующих типов ограничений:

Тип 1. Абсолютные погрешности измерения ограничены согласно (48).

Тип 2. Относительные погрешности измерения ограничены:

$$|\Delta x_{ij}|\le\delta_j^x|x_{ij}|\quad(i=1,2,...,n,j=1,2,...,m),\quad,|\Delta y_i|\le\delta^y|y_i|\quad(i=1,2,...,n).$$

Тип 3. Ограничения наложены на сумму погрешностей:

$$\sum_{j=1}^{m}|\Delta x_{ij}|\le\alpha_x\quad(i=1,2,...,n,j=1,2,...,m),\ |\Delta y_i|\le\alpha_y\quad(i=1,2,...,n)).$$

(Поскольку все переменные отнормированы, т.е. представляют собой относительные величины, то различие в размерностях исходных переменных не влияет на возможность сложения погрешностей.)

Перейдем к вычислению нотны оценки МНК. Справедливо равенство:

$$\Delta b^*=b^{*R}-b^*=(X_R^TX_R)^{-1}X_R^TY_R-(X^TX)^{-1}X^TY=(X_R^TX_R)^{-1}X_R^TY_R-$$
$$-((X_R+\Delta X)^T(X_R+\Delta X))^{-1}(X_R+\Delta X)(Y_R+\Delta Y).$$

Воспользуемся следующей теоремой из теории матриц [44].

Теорема. Если функция $f(\lambda)$ разлагается в степенной ряд в круге сходимости $|\lambda-\lambda_0|<r$, т.е.

$$f(\lambda)=\sum_{k=0}^{\infty}\alpha_k(\lambda-\lambda_0)^k,$$

то это разложение сохраняет силу, если скалярный аргумент заменить любой матрицей A, характеристические числа которой λ_k, $k=1,\ldots,n$, лежат внутри круга сходимости.

Из этой теоремы вытекает, что

$$(E - A)^{-1} = \sum_{P=0}^{\infty} A^P ,$$

если

$$| \lambda_k | < 1, \ k = 1,..., n .$$

Легко убедиться, что

$$((X_R + \Delta X)^T (X_R + \Delta X))^{-1} = -Z(E - \Delta \cdot Z)^{-1},$$

где

$$Z = -(X_R^T X_R)^{-1}, \ \Delta = X_R^T \Delta X + \Delta X^T X_R + \Delta X^T \Delta X.$$

Это вытекает из последовательности равенств:

$$((X_R + \Delta X)^T (X_R + \Delta X))^{-1} = (X_R^T X_R + X_R^T \Delta X + \Delta X^T X_R + \Delta X^T \Delta X \ )^{-1} = (X_R^T X_R + \Delta)^{-1} =$$
$$= ((E + \Delta (X_R^T X_R)^{-1} X_R^T X_R)^{-1} = (X_R^T X_R)^{-1} (E + \Delta (X_R^T X_R)^{-1})^{-1} = -Z(E - \Delta \cdot Z)^{-1}.$$

Применим приведенную выше теорему из теории матриц, полагая $A = \Delta Z$ и принимая, что собственные числа этой матрицы удовлетворяют неравенству $|\lambda_k| < 1$. Тогда получим:

$$((X_R + \Delta X)^T (X_R + \Delta X))^{-1} = -Z \sum_{P=0}^{\infty} (\Delta \cdot Z)^P = (X^T{}_R X_R)^{-1} \sum_{P=0}^{\infty} (-\Delta \cdot (X^T{}_R X_R)^{-1})^P .$$

Подставив последнее соотношение в заключение упомянутой теоремы, получим:

$$\Delta b^* = (X_R^T X_R)^{-1} X_R^T Y_R - ((X_R^T X_R)^{-1} \sum_{P}^{\infty} (-\Delta \cdot (X_R^T X_R)^{-1})^P)(X_R + \Delta X)^T (Y_R + \Delta Y) =$$

$$= (X_R^T X_R)^{-1} X_R^T Y_R - ((X_R^T X_R)^{-1} \sum_{P}^{\infty} (-\Delta \cdot (X_R^T X_R)^{-1})^P)(X_R^T Y_R + \Delta X^T Y_R + X_R^T \Delta Y + \Delta X^T \Delta Y).$$

Для дальнейшего анализа понадобится вспомогательное утверждение. Исходя из предположений 1–3, докажем, что:

$$(X_R^T X_R)^{-1} \approx \frac{1}{n} E.$$

Доказательство. Справедливо равенство:

$$X_R^T X_R = n \begin{pmatrix} D^*(x_1) & \cdots & \operatorname{cov}^*(x_1, x_m) \\ \cdots & \cdots & \cdots \\ \operatorname{cov}^*(x_1, x_m) & \cdots & D^*(x_m) \end{pmatrix} = n^*(x),$$

где $D^*(x_i)\operatorname{cov}^*(x_i, x_j)$ — состоятельные и несмещенные оценки дисперсий и коэффициентов ковариации, т.е.

$$D^*(x_i) = D(x_i)+o(1/n),\ \operatorname{cov}^*(x_i, x_j)-\operatorname{cov}^*(x_i, x_j)+o(1/n),$$

тогда

$$X_R^T X_R = n\operatorname{cov}^*(x) = n(\| \operatorname{cov}(x_i, x_j) \| + O(1/n)),$$

где

$$O(1/n) = \| a_{ij} = O(1/n),\ i = \overline{1,n}, j = \overline{1,m} \|.$$

Другими словами, каждый элемент матрицы, обозначенной как $O(1/n)$, есть бесконечно малая величина порядка $1/n$. Для рассматриваемого случая $cov(x) = E$, поэтому

$$X_R^T X_R = n \operatorname{cov}^*(x) = n(E + O(1/n)) \cdot$$

Предположим, что n достаточно велико и можно считать, что собственные числа матрицы $O(1/n)$ меньше единицы по модулю, тогда

$$\left(X_R^T X_R\right)^{-1} = \frac{1}{n}(E + O(1/n))^{-1} \approx \frac{1}{n}(E + O(1/n)) = \frac{1}{n} E + O(1/n^2)) \approx \frac{1}{n} E,$$

что и требовалось доказать.

Подставим доказанное асимптотическое соотношение в формулу для приращения b^*, получим:

$$\Delta b^* = b^{*R} - \frac{1}{n}\sum_{P=0}^{\infty}(-\Delta\cdot\frac{1}{n})^P(nb^{*R} + \Delta X^T Y_R + X_R^T\Delta Y + \Delta X^T\Delta Y) =$$

$$= b^{*R} - \frac{1}{n}\sum_{P=0}^{\infty}(-(X_R^T\Delta X + \Delta X^T X_R + \Delta X^T\Delta X)\cdot(\frac{1}{n})^P(nb^{*R} + \Delta X^T Y_R + X_R^T\Delta Y + \Delta X^T\Delta Y) =$$

$$= b^{*R} - \frac{1}{n}(E - (X_R^T\Delta X + \Delta X^T X_R + \Delta X^T\Delta X)\frac{1}{n} + (X_R^T\Delta X + \Delta X^T X_R + \Delta X^T\Delta X)^2(\frac{1}{n})^2)\cdot$$

$$\cdot(nb^{*R} + \Delta X^T Y_R + X_R^T\Delta Y + \Delta X^T\Delta Y).$$

Выразим Δb^* относительно приращений ΔX, ΔY до 2-го порядка:

$$\Delta b^* = b^{*R} - \frac{1}{n}(E - (X_R^T\Delta X + \Delta X^T X_R + \Delta X^T\Delta X)\frac{1}{n} + (X_R^T\Delta X X_R^T\Delta X + \Delta X^T X_R\Delta X^T X_R +$$

$$+ \Delta X^T X_R X_R^T\Delta X + X_R^T\Delta X\Delta X^T X_R)(\frac{1}{n})^2)\cdot(nb^{*R} + \Delta X^T Y_R + X_R^T\Delta Y + \Delta X^T\Delta Y);$$

$$\Delta b^* = b^{*R} - \frac{1}{n}(E - (X_R^T\Delta X + \Delta X^T X_R + \Delta X^T\Delta X)\frac{1}{n})\cdot(nb^{*R} + \Delta X^T Y_R + X_R^T\Delta Y + \Delta X^T\Delta Y);$$

$$\Delta b^* = \frac{1}{n}(X_R^T\Delta X + \Delta X^T X_R + \Delta X^T\Delta X)b^{*R} - \frac{1}{n}(\Delta X^T Y_R + X_R^T\Delta Y + \Delta X^T\Delta Y) =$$

$$= \frac{1}{n}[(X_R^T\Delta X + \Delta X^T X_R + \Delta X^T\Delta X)b^{*R} - (\Delta X^T Y_R + X_R^T\Delta Y + \Delta X^T\Delta Y)].$$

Перейдем от матричной к скалярной форме, опуская индекс (R):

$$\Delta b_k^* = \frac{1}{n}\{\sum_j^m\sum_i^n(x_{ik}\Delta x_{ij} + \Delta x_{ik}x_{ij})b_j^* - \sum_i^n(\Delta x_{ik}y_i + x_{ik}\Delta y_i)\};$$

$$\Delta b_k^* = \frac{1}{n}\{2\sum_i^n x_{ik}\Delta x_{ik}b_k^* + \sum_{j\neq k}^m\sum_i^n(x_{ik}\Delta x_{ij} + \Delta x_{ik}x_{ij})b_j^* - \sum_i^n(\Delta x_{ik}y_i + x_{ik}\Delta y_i)\} =$$

$$= \frac{1}{n}\{2\sum_i^n x_{ik}\Delta x_{ik}b_k^* + \sum_{j\neq k}^m\sum_i^n[(x_{ik}\Delta x_{ij} + \Delta x_{ik}x_{ij})b_j^* - \frac{1}{m-1}\Delta x_{ik}y_i] - \sum_i^n x_{ik}\Delta y_i\} =$$

$$= \frac{1}{n}\{\sum_{j\neq k}^m\sum_i^n[\frac{2}{m-1}x_{ik}\Delta x_{ik}b_k^* + (x_{ik}\Delta x_{ij} + \Delta x_{ik}x_{ij})b_j^* - \frac{1}{m-1}\Delta x_{ik}y_i] - \sum_i^n x_{ik}\Delta y_i\} =$$

$$= \frac{1}{n}\{\sum_{j\neq k}^m\sum_i^n[(\frac{2}{m-1}x_{ik}b_k^* + x_{ij}b_j^* - \frac{1}{m-1}y_i)\Delta x_{ik} - x_{ik}b_j^*\Delta x_{ij}] - \sum_i^n x_{ik}\Delta y_i\}$$

Будем искать $\max(|\Delta b_k^*|)$ по Δx_{ij} и Δy_i ($i = 1, \ldots, n; j = 1, \ldots, m$). Для этого рассмотрим все три ранее введенных типа ограничений на ошибки измерения.

Тип 1 (абсолютные погрешности измерения ограничены). Тогда:

$$\max_{\Delta x,\Delta y}(|\Delta b_k^*|)=\frac{1}{n}\{\sum_{j\neq k}^{m}\sum_{i}^{n}[|(\frac{2}{m-1}x_{ik}b_k^*+x_{ij}b_j^*-\frac{1}{m-1}y_i)|\Delta_k^x+|x_{ik}b_j^*|\Delta_j^x]-\sum_{i}^{n}|x_{ik}|\Delta y\}.$$

Тип 2 (относительные погрешности измерения ограничены). Аналогично получим:

$$\max_{\Delta x,\Delta y}(|\Delta b_k^*|)=\frac{1}{n}\left\{\sum_{j\neq k}^{m}\sum_{i}^{n}\left[\left(\frac{2}{m-1}x_{ik}b_k^*+x_{ij}b_j^*-\frac{1}{m-1}y_i\right)x_{ik}\right|\delta_k^x+|x_{ik}x_{ij}b_j^*|\delta_j^x\Big]-\sum_{i}^{n}|x_{ik}y_i|\delta^y\right\}$$

Тип 3 (ограничения наложены на сумму погрешностей). Предположим, что $|\Delta b_k^*|$ достигает максимального значения при таких значениях погрешностей Δx_{ij} и Δy_i, которые мы обозначим как:

$$\{\Delta x_{ij}^*,\qquad i=\overline{1,2,\ldots,n},\ j=1,2,\ldots,m\},\ \{\Delta y_i^*,\qquad i=1,2,\ldots,n\}.$$

тогда:

$$\max_{\Delta x,\Delta y}(|\Delta b_k^*|)=\frac{1}{n}\{\sum_{j\neq k}^{m}\sum_{i}^{n}[(\frac{2}{m-1}x_{ik}b_k^*+x_{ij}b_j^*-\frac{1}{m-1}y_i)\Delta x_{ik}^*+x_{ik}b_j^*\Delta x_{ij}^*]-\sum_{i}^{n}x_{ik}\Delta y_i^*\}.$$

Ввиду линейности последнего выражения и выполнения ограничения типа 3:

$$\max_{\Delta x,\Delta y}(|\Delta b_k^*|)=\frac{1}{n}\{\sum_{j\neq k}^{m}\sum_{i}^{n}[|\frac{2}{m-1}x_{ik}b_k^*+x_{ij}b_j^*-\frac{1}{m-1}y_i|\cdot|\Delta x_{ik}^*|+|x_{ik}b_j^*|\cdot|\Delta x_{ij}^*|]-\sum_{i}^{n}|x_{ik}|\cdot|\Delta y_i^*|\},$$

$$\sum_{j}^{m}|\Delta x_{ij}^*|=\alpha_x\qquad(j=1,2,\ldots,m),\qquad|\Delta y_i^*|=\alpha_y.$$

Для простоты записей выкладок сделаем следующие замены:

$$|\Delta x_{ij}|=\alpha_{ij}\geq 0,\qquad C_k=n\sum_{i}^{n}|x_{ik}|\cdot|\Delta y_i^*|\geq 0,$$

$$K_i^k=\sum_{j\neq k}^{m}|\frac{2}{m-1}x_{ik}b_k^*+x_{ij}b_j^*-\frac{1}{m-1}y_i|\geq 0,$$

$$|x_{ik}b_j^*|=R_{ij}^k\geq 0.$$

Теперь для достижения поставленной цели можно сформулировать следующую задачу, которая разделяется на *m* типовых задач оптимизации:

$$f_k(\{\alpha_{ij}\}) \to \max_{\alpha_{ij}} \quad (i = 1,2,...,n;\ j = 1,2,...,m;\ k = 1,2,...,m),$$

где

$$f_k(\{\alpha_{ij}\}) = \frac{1}{n}\{\sum_{i}^{n} K_i^k \alpha_{ik} + \sum_{j \neq m}^{m}\sum_{i}^{n} R_{ij}^k \alpha_{ij}\} + C_k,$$

при ограничениях:

$$\sum_{j}^{m} \alpha_{ij} = \alpha_x \qquad (j = 1,2,...,m).$$

Перепишем минимизируемые функции в следующем виде:

$$f_k = \frac{1}{n}\sum_{i}^{n}(K_i^k \alpha_{ik} + \sum_{j \neq m}^{m} R_{ij}^k \alpha_{ij}) + C_k = \frac{1}{n}\sum_{i}^{n} f_i^k + C_k.$$

Очевидно, что $f_i^k > 0$.

Легко видеть, что

$$n \cdot \max_{\alpha_{ij}}(f_k) = \max_{\alpha_{i1}}(f_1^k) + \max_{\alpha_{i2}}(f_2^k) + ... + \max_{\alpha_{in}}(f_n^k) + C_k = \sum_{i}^{n} \max_{\alpha_{ii}}(f_i^k) + C_k,$$

где

$$i = 1,2,...,n;\ j = 1,2...,m.$$

Следовательно, необходимо решить *nm* задач:

$$\{f_i^k\} \to \max_{\alpha_{ij}} \quad (i = 1.2.,,,,n;\ \ j = 1.2,...,m;\ \ k = 1,2,...,m)$$

при ограничениях «типа равенства»:

$$\sum_{j}^{m} \alpha_{ij} = \alpha_x \qquad (i = 1,2,...,n),$$

где

$$f_i^k = K_i^k \alpha_{ik} + \sum_{j \neq m}^{m} R_{ij}^k \alpha_{ij} = \sum_{j}^{m} S_{ij}^k \alpha_{ij},$$

причем

$$S_{ij}^k = \begin{cases} K_i^k, если & j = k, \\ R_{ij}^k, если & j \neq k. \end{cases}$$

Сформулирована типовая задача поиска экстремума функции. Она легко решается. Поскольку

$$\max_{a_{ij}} (f_i^k) = \max_{j} (S_{ij}^k) \cdot \alpha_x,$$

то максимальное отклонение МНК-оценки k-го параметра равно:

$$\max_{\Delta X, \Delta Y} (|\Delta b_k^*|) = \max_{\alpha_{ij}} (f_k) = \frac{1}{n} \alpha_x \sum_{i=1}^{n} \max_{j} (S_{ij}^k) + \frac{1}{n} C_k, \ (i = 1,2,...,n; j = 1,2,...,m).$$

Кроме рассмотренных выше трех видов ограничений на погрешности могут представлять интерес и другие, но для демонстрации типовых результатов ограничимся только этими тремя видами.

Оценивание линейной регрессионной связи. В качестве примера рассмотрим оценивание линейной регрессионной связи случайных величин y и x_1, x_2, ..., x_m с нулевыми математическими ожиданиями. Пусть эта связь описывается соотношением:

$$y = \sum_{j=1}^{m} b_j x_j + e,$$

где b_1, b_2, ..., b_m — постоянные, а случайная величина e некоррелирована с x_1, x_2, ..., x_m. Допустим, необходимо оценить неизвестные параметры b_1, b_2, ..., b_m по серии независимых испытаний:

$$y_i = \sum_{j=1}^{m} b_j x_{ij} + e_i, \quad (i = 1,2,...,n).$$

Здесь при каждом $i = 1, 2, \ldots, n$ имеем новую независимую реализацию рассматриваемых случайных величин. В этой частной схеме оценки наименьших квадратов b_1^{*R}, b_2^{*R}, ..., b_m^{*R} параметров $b_1, b_2, \ldots, b_m$ являются, как известно, состоятельными [45].

Пусть величины $x_1, x_2, \ldots, x_m$ в дополнение к попарной независимости имеют единичные дисперсии. Тогда из закона больших чисел [45] следует существование следующих пределов (ср. предположение 1 выше):

$$\lim_{n\to\infty}\{\frac{1}{n}\sum_{i}^{n} x_{ij}^R\} = M\{x_j\} = 0 \qquad (j = \overline{1,m}),$$

$$\lim_{n\to\infty}\{\frac{1}{n}\sum_{i}^{n} (x_{ij}^R - M\{x_j\})^2\} = D\{x_j\} = 1 \qquad (j = \overline{1,m}),$$

$$\lim_{n\to\infty}\{\frac{1}{n}\sum_{i}^{n} (x_{ij}^R - M\{x_j\})(x_{ik}^R - M\{x_k\})\} = 0 \qquad (j,k = \overline{1,m}),$$

$$\lim_{n\to\infty}\{\frac{1}{n}\sum_{i}^{n} y_i^R\} = M\{y\} = b_1 M\{x_1\} + \ldots + b_m M\{x_m\} + M\{e\} = 0,$$

$$\lim_{n\to\infty}\{\frac{1}{n}\sum_{i}^{n} (y_i^R - M\{y\})^2\} = D\{y\} = b_1^2 + \ldots + b_m^2 + \sigma^2,$$

где σ — среднее квадратическое отклонение случайной величины e.

Пусть измерения производятся с погрешностями, удовлетворяющими ограничениям типа 1, тогда максимальное приращение величины $|\Delta b^*_k|$, как показано выше, равно:

$$\max_{\Delta x,\Delta y}(|\Delta b^*_k|) = \frac{1}{n}\{\sum_{j\neq k}^{m}\sum_{i}^{n}[|\frac{2}{m-1}x_{ik}^R b^*_k + x_{ij}^R b^*_j - \frac{1}{m-1}y_i^r|\cdot\Delta_k^x + |x_{ik}^R b^*_j|\cdot\Delta_j^x] + \sum_{i}^{n}|x_{ik}^R|\cdot\Delta y\}.$$

Перейдем к предельному случаю и выпишем выражение для нотны:

$$N_k = \lim_{n\to\infty}\{\max_{\Delta x,\Delta y}(|\Delta b^*_k|)\} =$$

$$= \sum_{j\neq k}^{m}[M\{|\frac{2}{m-1}x_k b_k + x_j b_j - \frac{1}{m-1}y|\}\cdot\Delta_k^x + M\{|x_k b_j|\}\cdot\Delta_j^x + M\{|x_k|\}\cdot\Delta y.$$

В качестве примера рассмотрим случай $m = 2$. Тогда

$$N_1 = M\{|2x_1 b_1 + x_2 b_2 - y|\}\Delta_1^x + M\{b_2 x_1\}\Delta_2^x + M\{|x_1|\}\Delta y,$$
$$N_2 = M\{|2x_2 b_2 + x_1 b_1 - y|\}\Delta_2^x + M\{b_1 x_2\}\Delta_1^x + M\{|x_2|\}\Delta y.$$

Приведенное выше выражение для максимального приращения метрологической погрешности не может быть использовано в случае $m = 1$. Для $m = 1$ выведем выражение для нотны, исходя из соотношения:

$$\Delta b_k^* = \frac{1}{n}\left\{\sum_j^m \sum_i^n (x_{ik}\Delta x_{ij} + \Delta x_{ik}x_{ij})b_j^* - \sum_i^n (\Delta x_{ik}y_i + x_{ik}\Delta y_i)\right\}.$$

Подставив $m = 1$, получим:

$$\Delta b^* = \frac{1}{n}\{\sum_i^n (2x_i\Delta x_i)b^* - \sum_i^n (\Delta x_i y_i + x_i\Delta y_i)\} = \frac{1}{n}\{\sum_i^n ((2x_i b^* - y_i)\Delta x_i + x_i\Delta y_i)\}.$$

Следовательно, *нотна* выглядит так:

$$N_f = M\{|2xb^* - y|\}\,\Delta^x + M\{|x|\}\,\Delta^y.$$

Для нахождения рационального объема выборки необходимо сделать следующее.

Этап 1. Выразить зависимость размеров и меры области рассеивания $B_\alpha(n, b)$ от числа опытов n (см. выше).

Этап 2. Ввести меру неопределенности и записать соотношение между статистической и интервальной неопределенностями.

Этап 3. По результатам этапов 1 и 2 получить выражение для рационального объема выборки.

Для выполнения этапа 1 определим область рассеивания следующим образом. Пусть доверительным множеством $B_\alpha(n, b)$ является m-мерный куб со сторонами длиною $2K$, для которого

$$P(b \in B_\alpha(n, b^{*R})) = \alpha.$$

Исследуем случайный вектор b^* и

$$b^{*R} = (X_R^T X_R)^{-1} X_R^T Y_R = (X_R^T X_R)^{-1} X_R^T (X_R b + e) =$$
$$= (X_R^T X_R)^{-1} X_R^T X_R b + (X_R^T X_R)^{-1} X_R^T e = b + (X_R^T X_R)^{-1} X_R^T e.$$

Как известно, если элементы матрицы $A = \{a_{ij}\}$ — случайные, т.е. A — случайная матрица, то ее математическим ожиданием является матрица, составленная из математических ожиданий ее элементов, т.е. $M\{A\} = \{M\{a_{ij}\}\}$.

Утверждение 1. Пусть $A = \{a_{ij}\}$ и $B = \{b_{ij}\}$ — случайные матрицы порядка $(m \times n)$ и $(n \times r)$ соответственно, причем любая пара их элементов (a_{ij}, b_{kl}) состоит из независимых случайных величин. Тогда математическое ожидание произведения матриц равно произведению математических ожиданий сомножителей, т.е. $M\{AB\} = M\{A\}\, M\{B\}$.

Доказательство. На основании определения математического ожидания матрицы заключаем, что

$$A \cdot B = \{\sum_{k}^{n} a_{ik} \cdot b_{kj}\} \to M\{A \cdot B\} = \{M\{\sum_{k}^{n} a_{ik} \cdot b_{kj}\}\} = \{\sum_{k}^{n} M\{a_{ik} \cdot b_{kj}\}\},$$

но так как случайные величины a_{ik}, b_{kj} независимы, то

$$M\{A \cdot B\} = \{\sum_{k}^{n} M\{a_{ik}\} \cdot M\{b_{kj}\}\} = M\{A\} \cdot M\{B\},$$

что и требовалось доказать.

Утверждение 2. Пусть $A = \{a_{ij}\}$ и $B = \{b_{ij}\}$ — случайные матрицы порядка $(m \times n)$ и $(n \times r)$ соответственно. Тогда математическое ожидание суммы матриц равно сумме математических ожиданий слагаемых:, т.е. $M\{A+B\} = M\{A\} + M\{B\}$.

Доказательство. На основании определения математического ожидания матрицы заключаем, что

$$M\{A+B\} = \{M\{a_{ij}+b_{ij}\}\} = \{M\{a_{ij}\} + M\{b_{ij}\}\} = M\{A\} + M\{B\},$$

что и требовалось доказать.

Найдем математическое ожидание и ковариационную матрицу вектора b^* с помощью утверждений 1, 2 и выражения для b^{*R}, приведенного выше. Имеем

$$M\{b^{*R}\} = b + M\{(X_R^T X_R)^{-1} X_R^T e\} = b + M\{(X_R^T X_R)^{-1} X_R^T\} \cdot M\{e\}.$$

Но так как M{e} = 0, то $M\{b^{*R}\} = b$. Это означает что оценка МНК является несмещенной.

Найдем ковариационную матрицу:

$$D\{b^{*R}\} = M\{(b^{*R} - b)(b^{*R} - b)^T\} = M\{(X_R^T X_R)^{-1} X_R^T \cdot e \cdot e^T \cdot X_R (X_R^T X_R)^{-1}\}.$$

Можно доказать, что

$$D\{b^{*R}\} = M\{(X_R^T X_R)^{-1} X_R^T \cdot M\{e \cdot e^T\} \cdot X_R (X_R^T X_R)^{-1}\},$$

но

$$M\{e \cdot e^T\} = D\{e\} = \sigma^2 E,$$

поэтому

$$D\{b^{*R}\} = M\{(X_T^R X_R)^{-1} X_R^T (\sigma^2 E) X_R (X_R^T X_R)^{-1}\} = \sigma^2 M\{(X_R^T X_R)^{-1}\}.$$

Как выяснено ранее, для достаточно большого количества опытов n выполняется приближенное равенство:

$$(X_R^T X_R)^{-1} \approx \frac{1}{n} E, \qquad (51)$$

тогда при больших n:

$$D\{b^{*R}\} = \frac{\sigma^2}{n} E.$$

Осталось определить вид распределения вектора b^{*R} . Из выражения для b^{*R}, приведенного выше, и асимптотического соотношения (51) следует, что

$$b^{*R} = b + \frac{1}{n} X_R^T e \ .$$

Можно показать, что вектор b^{*R} имеет асимптотически нормальное распределение, т.е.

$$b^{*R} \in N(b, \frac{\sigma^2}{n} E).$$

Тогда совместная функция плотности распределения вероятностей случайных величин b^{*R_1}, $b^{*R_2}, \ldots, b^{*R_m}$ будет иметь в асимптотике вид:

$$f(b^{*R}) = \frac{1}{(2\pi)^{m/2} \cdot (\det C)^{1/2}} \cdot \exp[-\frac{1}{2}(b^{*R} - b)^T \cdot C^{-1} \cdot (b^{*R} - b)], \qquad (52)$$

где

$$C = D(b^{*R}) = \frac{\sigma^2}{n} E.$$

Тогда справедливы соотношения:

$$C^{-1} = \frac{n}{\sigma^2} E \ , \quad \det C = \det(\frac{n}{\sigma^2} E) = (\frac{\sigma^2}{n})^m.$$

Подставим в формулу (52), получим:

$$f(b^{*R}) = \frac{1}{(2\pi)^{m/2}(\sigma^2/n)^{m/2}} \exp\left[-\frac{1}{2}(b^{*R} - b)^T C^{-1}(b^{*R} - b)\right] =$$
$$= \frac{1}{\left(\sigma\sqrt{2\pi/n}\right)^m} \exp\left[-\frac{n}{2\sigma^2}(b^{*R} - b)^T E(b^{*R} - b)\right] =$$
$$= \frac{1}{\left(\sigma\sqrt{2\pi/n}\right)^m} \exp\left[-\frac{n}{2\sigma^2}(\beta_1^2 + \beta_2^2 + \ldots + \beta_m^2)\right]$$

где

$$\beta_i = b^{*R}_i - b_i, \quad i = 1,2,\ldots m \quad .$$

Вычислим асимптотическую вероятность попадания описывающего реальность вектора параметров b в m-мерный куб с длиной стороны, равной $2k$, и с центром b^{*R}.

$$P(-k<\beta_1<k,-k<\beta_2<k,\ldots,-k<\beta_m<k)=$$

$$= \frac{1}{(\sigma\sqrt{2\pi/n})^m}\{\int_{-k}^{k}\ldots\int_{-k}^{k}\exp[-\frac{n}{2\sigma^2}(\beta_1^2+\beta_2^2+\ldots+\beta_m^2)]\cdot d\beta_1\ldots d\beta_m\}=$$

$$=\frac{1}{\left(\sigma\sqrt{2\pi/n}\right)^m}\left\{\int_{-k}^{r}\exp\left[-\frac{n}{2\sigma^2}\beta_1^2\right]d\beta_1\ldots\int_{-k}^{r}\exp\left[-\frac{n}{2\sigma^2}\beta_m^2\right]d\beta_m\right\}.$$

Сделаем замену:

$$t_i=\sqrt{n/2}\cdot\frac{1}{\sigma}\beta_i \quad , \quad i=1,2,\ldots,m.$$

Тогда

$$P=P(-k<\beta_1<k,-k<\beta_2<k,\ldots,-k<\beta_m<k,)=$$

$$=\frac{(\sigma\sqrt{2/n})^m}{(\sigma\sqrt{2\pi/n})^m}[\int_{-T}^{T}e^{-t^2}dt]^m=[(1/\sqrt{\pi})\int_{-T}^{T}e^{-t^2}dt]]^m=[\Phi_0(T)]^m,$$

где $T=(n/2)^{1/2}(k/\sigma)$, а $\Phi_0(T)$ — интеграл Лапласа:

$$\Phi_0(T)=\Phi(\sqrt{2}T)-\Phi(-\sqrt{2}T),$$

где $\Phi(t)$ — функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1. Из последнего соотношения получаем:

$$T=\Phi_0^{-1}(P^{1/m}),$$

где $\Phi^{-1}(P)$ — обратная функция Лапласа. Отсюда следует, что

$$k=\sigma\,(2/n)^{1/2}\,\Phi_0^{-1}(P^{1/m}). \qquad (53)$$

Напомним, что доверительная область $B_\alpha\,(n,\ b)$ — это m-мерный куб, длина стороны которого равна K, т.е.

$$P(b\in B_\alpha(n,b))=P(-K<\beta_1<K,\ -K<\beta_2<K,\ \ldots,\ -K<\beta_m<K)=\alpha.$$

Подставляя $P=\alpha$ в формулу (53), получим:

$$K=k=\sigma\,(2/n)^{1/2}\,\Phi_0^{-1}(\alpha^{1/m}). \qquad (54)$$

Соотношение (54) выражает зависимость размеров доверительной области (т.е. длины ребра куба K) от числа опытов n, среднего квадратического отклонения σ ошибки e и доверительной вероятности α. Это соотношение понадобится для определения рационального объема выборки.

Переходим к этапу 2. Необходимо ввести меру разброса (неопределенности) и установить соотношение между статистической и интервальной (метрологической) неопределенностями с соответствии с ранее сформулированным общим подходом.

Пусть A — некоторое измеримое множество точек в m-мерном евклидовом пространстве, характеризующее неопределенность задания вектора $a \in A$. Тогда необходимо ввести некую меру $M(A)$, измеряющую степень неопределенности. Такой мерой может служить m-мерный объем $V(A)$ множества A (т.е. его мера Лебега или Жордана), $M(A) = V(A)$.

Пусть P — m-мерный параллелепипед, характеризующий интервальную неопределенность. Длины его сторон равны значениям *нотн* $2N_1$, $2N_2$, ..., $2N_m$, а центр a (точка пересечений диагоналей параллелепипеда) находится в точке b^{*R}. Пусть C — измеримое множество точек, характеризующее общую неопределенность. В рассматриваемом случае это m-мерный параллелепипед, длины сторон которого равны $2(N_1 + K)$, $2(N_2 + K)$, …, $2(N_m + K)$, а центр находится в точке b^{*R}.

Тогда

$$M(P) = V(P) = 2^m N_1 N_2 \ldots N_m, \tag{55}$$

$$M(C) = V(C) = 2^m (N_1 + K)(N_2 + K) \ldots (N_m + K). \tag{56}$$

Справедливо соотношение (49), согласно которому $M(C) = M(P) + M(F)$, где множество $F = C \setminus P$ характеризует статистическую неопределенность.

На этапе 3 получаем по результатам этапов 1 и 2 выражение для рационального объема выборки. Найдем то число опытов, при котором статистическая неопределенность составит δ 100% от общей неопределенности, т.е. согласно правилу (50):

$$M(F) = M(C) - M(P) = \delta M(C) \tag{57}$$

где $0 < \delta < 1$. Подставив (55) и (56) в (57), получим:

$$2^m \prod_{i=1}^{m} (N_i + K) - 2^m \prod_{i=1}^{m} (N_i) = 2^m \delta \prod_{i=1}^{m} (N_i + K).$$

Следовательно,

$$(1-\delta)\prod_{i=1}^{m}(N_i+K)/\prod_{i=1}^{m}(N_i)=1.$$

Преобразуем эту формулу:

$$(1-\delta)\prod_{i=1}^{m}(1+K/N_i)=1,$$

откуда

$$\prod_{i}^{m}(1+K/N_i)=1/(1-\delta).$$

Если статистическая погрешность мала относительно метрологической, т.е. величины K/N_i малы, то

$$\prod_{i}^{m}(1+K/N_i)\approx 1+\sum_{i}^{m}(K/N_i).$$

При m = 1 эта формула является точной. Из нее следует, что для дальнейших расчетов можно использовать соотношение:

$$1+\sum_{i}^{m}(K/N_i)=1/(1-\delta).$$

Отсюда нетрудно найти K:

$$K=\frac{\delta}{1-\delta}(1/\sum_{i=1}^{m}(1/N_i)). \qquad (58)$$

Подставив в формулу (58) зависимость $K = K(n)$, полученную в формуле (54), находим приближенное (асимптотическое) выражение для рационального объема выборки:

$$n_{рац}=2\left(\frac{1-\delta}{\delta}\sigma\sum_{i=1}^{m}(1/N_i)\cdot\Phi^{-1}(\alpha^{1/m})\right)^2.$$

При $m = 1$ эта формула также справедлива, более того, является точной.

Переход от произведения к сумме является обоснованным при достаточно малом δ, т.е. при достаточно малой статистической неопределенности по сравнению с метрологической. В общем случае можно находить K и затем рациональный объем выборки тем или иным численным методом.

Пример 1. Представляет интерес определение $n_{рац}$ для случая, когда $m = 2$, поскольку простейшая линейная регрессия с $m = 2$ широко применяется. В этом случае базовое соотношение имеет вид:

$$(1 + K / N_1)(1 + K / N_2) = 1 / (1 - \delta).$$

Решая это уравнение относительно K, получаем:

$$K = 0{,}5\{-(N_1 + N_2) + [(N_1 + N_2)^2 + 4 N_1 N_2 (\delta / (1 - \delta)]^{1/2}\}.$$

Далее, подставив в формулу (54), получим уравнение для рационального объема выборки в случае $m = 2$:

$$\sigma (2/n)^{1/2} \Phi^{-1}(\alpha^{1/2}) = 0{,}5\{-(N_1 + N_2) + [(N_1 + N_2)^2 + 4 N_1 N_2 (\delta / (1 - \delta)]^{1/2}\}.$$

Следовательно,

$$n_{rat} = \frac{8\left\{\Phi^{-1}(\sqrt{\alpha})\right\}^2}{\left\{-\frac{N_1}{\sigma} - \frac{N_2}{\sigma} + \sqrt{\left(\frac{N_1}{\sigma} + \frac{N_2}{\sigma}\right)^2 + 4\frac{N_1 N_2 \delta}{\sigma^2 (1-\delta)}}\right\}^2}.$$

При использовании «принципа уравнивания погрешностей» согласно [3] $\delta = 1/2$. При доверительной вероятности $\alpha = 0{,}95$ имеем $\sqrt{\alpha} = 0{,}9747$ и согласно [42] $\Phi^{-1}(\sqrt{\alpha}) = 1{,}954$. Для этих численных значений:

$$n_{rat} = \frac{30{,}545}{\left\{-\frac{N_1}{\sigma} - \frac{N_2}{\sigma} + \sqrt{\left(\frac{N_1}{\sigma} + \frac{N_2}{\sigma}\right)^2 + 4\frac{N_1 N_2}{\sigma^2}}\right\}^2}.$$

Если $\frac{N_1}{\sigma}=\frac{N_2}{\sigma}=0{,}1,$ то $n_{rat}=4455.$ Если же $\frac{N_1}{\sigma}=\frac{N_2}{\sigma}=0{,}5,$ то $n_{rat}=178.$ Если первое из этих чисел превышает обычно используемые объемы выборок, то второе находится в «рабочей зоне» регрессионного анализа.

Парная регрессия. Наиболее простой и одновременно наиболее широко применяемый частный случай парной регрессии рассмотрим подробнее. Модель имеет вид:

$$y_i = ax_i + b + \varepsilon_i, \quad i = 1,2,\ldots,n.$$

Здесь x_i — значения фактора (независимой переменной), y_i — значения отклика (зависимой переменной), ε_i — статистические погрешности, a, b — неизвестные параметры, оцениваемые методом наименьших квадратов. Она переходит в модель (используем альтернативную запись линейной модели):

$$y = X\beta + \varepsilon,$$

если положить:

$$y=\begin{pmatrix} y_1 \\ \ldots \\ y_n \end{pmatrix}, \quad X=\begin{pmatrix} x_1 & 1 \\ \ldots & \ldots \\ x_n & 1 \end{pmatrix}, \quad \varepsilon=\begin{pmatrix} \varepsilon_1 \\ \ldots \\ \varepsilon_n \end{pmatrix}, \quad \beta=\begin{pmatrix} a \\ b \end{pmatrix}.$$

Естественно принять, что погрешности факторов описываются матрицей:

$$\Delta X = \alpha = \begin{pmatrix} \Delta x_1 & 0 \\ \ldots & \ldots \\ \Delta x_n & 0 \end{pmatrix} = \begin{pmatrix} \alpha_1 & 0 \\ \ldots & \ldots \\ \alpha_n & 0 \end{pmatrix}.$$

В рассматриваемой модели интервального метода наименьших квадратов:

$$X = X_R + \alpha, \quad y = y_R + \begin{pmatrix} \gamma_1 \\ \ldots \\ \gamma_n \end{pmatrix},$$

где X, y — наблюдаемые (т.е. известные статистику) значения фактора и отклика, X_R, y_R — истинные значения переменных, α, γ — погрешности измере-

ний переменных. Пусть $\beta*$ — оценка метода наименьших квадратов, вычисленная по наблюдаемым значениям переменных, β_R^* — аналогичная оценка, найденная по истинным значениям. В соответствии с ранее проведенными рассуждениями:

$$\beta*-\beta=[-(X^TX)^{-1}\Delta(X^TX)^{-1}X^T+(X^TX)^{-1}\alpha^T]y+(X^TX)^{-1}X^T\gamma \tag{59}$$

с точностью до бесконечно малых более высокого порядка по $|\alpha|$ и $|\gamma|$. В формуле (59) использовано обозначение $\Delta=X^T\alpha+\alpha^TX$. Вычислим правую часть в (59), выделим главный линейный член и найдем нотну.

Легко видеть, что

$$X^TX=\begin{pmatrix}\sum x_i^2 & \sum x_i\\ \sum x_i & n\end{pmatrix}, \tag{60}$$

где суммирование проводится от 1 до n. Для упрощения обозначений в дальнейшем до конца настоящего пункта не будем указывать эти пределы суммирования. Из (60) вытекает, что

$$(X^TX)^{-1}=\begin{pmatrix}n & -\sum x_i\\ -\sum x_i & \sum x_i^2\end{pmatrix}\Big/[n\sum x_i^2-(\sum x_i)^2]. \tag{61}$$

Легко подсчитать, что

$$X^T\alpha+\alpha^TX=\begin{pmatrix}2\sum x_i\alpha_i & \sum\alpha_i\\ \sum\alpha_i & n\end{pmatrix}. \tag{62}$$

Положим:

$$S_0^2=\frac{1}{n}\sum(x_i-\bar{x})^2.$$

Тогда знаменатель в (61) равен $n^2S_0^2$. Из (61) и (62) следует, что

$$(X^TX)^{-1}(X^T\alpha+\alpha^TX)=\frac{1}{n^2S_0^2}\begin{pmatrix}2n\sum x\alpha-\sum x\sum\alpha & n\sum\alpha\\ -2\sum x\sum x\alpha+\sum x^2\sum\alpha & -\sum x\sum\alpha\end{pmatrix}. \tag{63}$$

Здесь и далее опустим индекс i, по которому проводится суммирование. Это не может привести к недоразумению, поскольку всюду суммирование проводится по индексу i в интервале от 1 до n. Из (61) и (63) следует, что

$$(X^TX)^{-1}(X^T\alpha+\alpha^TX)(X^TX)^{-1}=\frac{1}{n^4S_0^4}\begin{pmatrix}A & B\\ C & D\end{pmatrix}, \tag{64}$$

где

$$A=2n^2\sum x\alpha-2n\sum x\sum\alpha,$$
$$B=C=-2n\sum x\sum x\alpha+\left(\sum x\right)^2\sum\alpha+n\sum\alpha\sum x^2,$$
$$D=2\left(\sum x\right)^2\sum x\alpha-2\sum\alpha\sum x\sum x^2.$$

Наконец, вычисляем основной множитель в (59):

$$(X^TX)^{-1}(X^T\alpha+\alpha^TX)(X^TX)^{-1}X^T=\frac{1}{n^4S_0^4}\begin{pmatrix}z_{11} & z_{12} & \dots & z_{1i} & \dots & z_{1n}\\ z_{21} & z_{22} & \dots & z_{2i} & \dots & z_{2n}\end{pmatrix}, \tag{65}$$

где

$$z_{1i}=Ax_i+B,\ \ z_{2i}=Cx_i+D,\ \ i=1,2,\dots,n.$$

Перейдем к вычислению второго члена с α в (59). Имеем:

$$(X^TX)^{-1}\alpha^T=\frac{1}{n^2S_0^2}\begin{pmatrix}w_{11} & \dots & w_{1i} & \dots & w_{1n}\\ w_{21} & \dots & w_{2i} & \dots & w_{2n}\end{pmatrix}, \tag{67}$$

где

$$w_{1i}=n\alpha_i,\ \ w_{2i}=-\alpha_i\sum x,\ \ i=1,2,\dots,n.$$

Складывая правые части (65) и (67) и умножая на y, получим окончательный вид члена с α в (59):

$$\{(X^TX)^{-1}(X^T\alpha+\alpha^TX)(X^TX)^{-1}X^T+(X^TX)^{-1}\alpha^T\}y=\begin{pmatrix}F\\ G\end{pmatrix}, \tag{68}$$

где

$$F = \left(\sum xy\right)\left(2n^2\sum x\alpha - 2n\sum x\sum\alpha\right)/n^4S_0^4 + \left(\sum y\alpha\right)/nS_0^2 + \\ + \left(\sum y\right)\left(n\sum\alpha\sum x^2 + \sum\alpha\left(\sum x\right)^2 - 2n\sum x\sum x\alpha\right)/n^4S_0^4,$$
$$G = \left(\sum xy\right)\left(-2n\sum x\sum x\alpha + n\sum\alpha\sum x^2 + \sum\alpha\left(\sum x\right)^2\right)/n^4S_0^4 - \\ - \left(\sum y\alpha\right)\left(\sum x\right)/n^2S_0^2 + \left(\sum y\right)\left(2\sum x\alpha\left(\sum x\right)^2 - 2\sum\alpha\sum x\sum x^2\right)/n^4S_0^4. \qquad (69)$$

Для вычисления нотны выделим главный линейный член. Сначала найдем частные производные. Имеем:

$$\frac{\partial F}{\partial\alpha_j} = \left(\sum xy\right)(2n^2x_j - 2n\sum x)/n^4S_0^4 + y_j/nS_0^2 + \\ + \left(\sum y\right)\left(n\sum x^2 + \left(\sum x\right)^2 - 2n\left(\sum x\right)x_j\right)/n^4S_0^4;$$
$$\frac{\partial G}{\partial\alpha_j} = \left(\sum xy\right)\left(-2n(\sum x)x_j + n\sum x^2 + \left(\sum x\right)^2\right)/n^4S_0^4 - \\ - y_j\left(\sum x\right)/n^2S_0^2 + \left(\sum y\right)\left(2x_j\left(\sum x\right)^2 - 2\sum x\sum x^2\right)/n^4S_0^4. \qquad (70)$$

Если ограничения имеют вид:

$$|\alpha_j| \le \Delta, \quad j = 1,2,\ldots,n,$$

то максимально возможное отклонение оценки a^* параметра a из-за погрешностей α_j таково:

$$N_a(x) = \sum_{1\le j\le n}\left|\frac{\partial F}{\partial\alpha_j}\right|\Delta + O(\Delta^2),$$

где производные заданы формулой (70).

Пример 2. Пусть вектор (x, y) имеет двумерное нормальное распределение с нулевыми математическими ожиданиями, единичными дисперсиями и коэффициентом корреляции ρ. Тогда

$$\lim_{\Delta\to 0}\lim_{n\to\infty}\frac{N_a(x)}{\Delta} = \lim_{n\to\infty}\sum_{1\le j\le n}\left|\frac{\partial F}{\partial\alpha_j}\right| = M\,|\,2\rho x + y\,| = \sqrt{\frac{2(1+8\rho^2)}{\pi}} \qquad (71)$$

При этом

$$\lim_{n\to\infty}\frac{\partial G}{\partial \alpha_j}=\rho,$$

следовательно, максимально возможному изменению параметра b^* соответствует сдвиг всех x_i в одну сторону, т.е. наличие систематической ошибки при определении x-ов. В то же время согласно (71) значения α_j в асимптотике выбираются по правилу:

$$\alpha_j=\begin{cases}\Delta, & 2\rho x_j+y_j>0,\\ -\Delta, & 2\rho x_j+y_j\le 0.\end{cases}$$

Таким образом, максимальному изменению a^* соответствуют не те α_j, что максимальному изменению b^*. В этом — новое по сравнению с одномерным случаем. В зависимости от вида ограничений на возможные отклонения, в частности, от вида метрики в пространстве параметров, будут «согласовываться» отклонения по отдельным параметрам. Ситуация аналогична той, что возникает в классической математической статистике в связи с оптимальным оцениванием параметров. Если параметр одномерен, то ситуация с оцениванием достаточно прозрачна — есть понятие эффективных оценок, показателем качества оценки является средний квадрат ошибки, а при ее несмещенности — дисперсия. В случае нескольких параметров возникает необходимость соизмерить точность оценивания по разным параметрам. Есть много критериев оптимальности (см., например, [46]), но нет признанных правил выбора среди них.

Вернемся к формуле (59). Интересно, что отклонения вектора параметров, вызванные отклонениями значений факторов α и отклика γ, входят в (59) аддитивно. Хотя

$$\sup_{\alpha,\gamma}\|\beta^*-\beta\|\ne\sup_{\alpha}|\{-(X^TX)^{-1}\Delta(X^TX)^{-1}X^T+(X^TX)^{-1}\alpha^T\}y|$$
$$+\sup_{\gamma}|(X^TX)^{-1}X^T\gamma|,$$

но для отдельных компонент (не векторов!) имеет место равенство.

В случае парной регрессии

$$(X^TX)^{-1}X^T\gamma=\frac{1}{n^2S_0^2}\left(\sum\gamma_i(nx_i-\sum x);\ \sum\gamma_i(-x_i\sum x+\sum x^2)\right)^T. \qquad (72)$$

Из формул (68), (69) и (72) следует, что

$$\beta^* - \beta = \begin{pmatrix} a^*(X,y) - a^*(X_R, y_R) \\ b^*(X,y) - b^*(X_R, y_R) \end{pmatrix} = \begin{pmatrix} F + F_1 \\ G + G_1 \end{pmatrix},$$

где F и G определены в (69), а

$$F_1 = \frac{1}{n^2 S_0^2}\left(n\sum yx - \sum x \sum y\right), \quad G_1 = \frac{1}{n^2 S_0^2}\left(\sum y \sum x^2 - \sum yx \sum x\right).$$

Итак, продемонстрирована возможность применения основных подходов статистики интервальных данных в регрессионном анализе.

Пример использования интервального регрессионного анализа. Методы статистики интервальных данных наряду с классическими методами оказываются полезными не только в традиционных статистических задачах, но и во многих других областях, в частности, в экономике и управлении промышленными предприятиями [27, 47]. Пример использования статистики интервальных данных в инвестиционном менеджменте подробно описан в [27] (см. также раздел 4.7 ниже). Перспективы применения статистики интервальных данных в контроллинге рассмотрены в [48]. Компьютерный анализ данных и использование статистических методов в информационных системах управления предприятием при решении задач контроллинга рассмотрены в [49]. Рассмотрим практический пример применения интервального регрессионного анализа при анализе и прогнозировании затрат предприятия[7] [50].

Выпуск продукции y зависит от величины суммарных переменных затрат x. Условные исходные данные для предприятия «Омега» приведены в табл. 1. Необходимо построить уравнение регрессии и найти *нотну*. В данном случае $n = 12$, $k = 2$. Зависимость ищется в виде $y = ax + b$.

Таблица 1

Исходные данные для предприятия «Омега», тыс. руб.

№ п/п	x	y	№ п/п	x	y
1	15,1	89,0	7	44,3	145,9
2	16,8	110,8	8	46,0	151,8
3	25,0	104,4	9	46,8	153,7
4	30,7	116,1	10	53,4	161,8
5	33,2	127,8	11	56,5	175,8
6	44,2	143,3	12	65,4	193,4

[7] Пример рассмотрен и расчеты проведены Е. А. Гуськовой.

Пусть как для x, так и для y максимально возможная погрешность $\lambda = 10$. Можно доказать [12], что указанное значение λ допустимо считать малым, поскольку под «малостью» следует понимать малость относительно типовых значений x и y. Построим уравнение регрессии согласно методу наименьших квадратов:

$$y = 63{,}32 + 1{,}914x.$$

Оценим максимально возможное изменение (приращение) вектора (a^*, b^*) оценок параметров линейной зависимости методом наименьших квадратов при изменении исходных данных, когда α и γ малы (см. формулу (59) выше). Для этого найдем нотны — максимально возможные изменения координат этого вектора в предположении $|\alpha| \le \lambda$ и $|\gamma| \le \lambda$:

$$N_{a^*}(x, y) = 0{,}87;\ N_{b^*}(x, y) = 32{,}98.$$

Найдем доверительные интервалы для параметров a и b согласно [27, п. 5.1] при доверительной вероятности 0,95. Для параметра a (т.е. для переменных затрат на единицу выпуска) нижняя доверительная граница $a_H(0{,}95) = 1{,}595$, а верхняя — $a_B(0{,}95) = 2{,}233$. Доверительный интервал для параметра a с учетом нотны равен [1,595 – 0,87; 2,233 + 0,87] или [0,73; 3,1]. Ширина «классического» доверительного интервала $d_1 = a_{\text{В}}(0{,}95) — a_{\text{Н}}(0{,}95)$ равна 0,63, что несколько меньше, чем нотна 0,87.

Для параметра b (т.е. для постоянных затрат) нижняя доверительная граница $b_H(0{,}95) = 58{,}51$, а верхняя — $b_B(0{,}95) = 68{,}13$. Ширина «классического» доверительного интервала для параметра b^* равна 9,63, т.е. почти в 3 раза меньше, чем нотна 32,98. Доверительный интервал для параметра b с учетом нотны равен [58,51 – 32,98; 68,13 + 32,98] или [25,53; 101,12].

Итак, восстановленная зависимость с учетом метрологических и статистических погрешностей имеет вид:

$$y = (1{,}914 \pm 1{,}187)x + (63{,}32 \pm 37{,}79).$$

Исходя из погрешностей коэффициентов линейной зависимости, можно указать нижнюю и верхнюю доверительные границы для функции:

$$y_{\min} = 0{,}727x + 25{,}53,\ y_{\max} = 3{,}101x + 101{,}11.$$

Более точно доверительные границы для значения функции в определенной точке можно указать, если найти нотну и статистическую погрешность не для коэффициентов, а непосредственно для значения функции [27, п. 5.1].

Полученные результаты дают возможность оценивать точность прогнозирования с помощью восстановленной зависимости, рассчитывая нижние и верхние границы для значения зависимой переменной. Например, при x=100 нижняя и верхняя границы интервала равны:

$$y_{\text{н}}(100) = (1{,}914 - 1{,}187) \times 100 + 63{,}32 - 37{,}79 = 98{,}23;$$
$$y_{\text{в}}(100) = (1{,}914 + 1{,}187) \times 100 + 63{,}32 + 37{,}79 = 411{,}21.$$

Некоторые замечания. На основе использования вероятностных моделей регрессионного анализа [27, гл. 5.1] удается построить доверительные границы для восстановленной зависимости. Однако при практическом применении вероятностных моделей не всегда легко обосновать предположения, наложенные на вектор невязок ε (независимость и одинаковую распределенность его координат). Кроме того, при моделировании экономических явлений и процессов обычно нет оснований использовать нормально распределенные случайные величины [27, гл. 4.1], следовательно, нельзя применять методы регрессионного анализа, основанные на нормальном распределении погрешностей. При этом объем данных обычно таков, что применение асимптотических формул непараметрического регрессионного анализа [27, гл. 5] не вполне оправдано. Поэтому описанный выше подход интервального регрессионного анализа представляется не менее оправданным, чем подход на основе вероятностных моделей. В этом мы согласны с А. П. Вощининым [21]. Представляется необходимым использование интервального регрессионного анализа в различных областях научных и прикладных исследований, прежде всего, в технических, экономических, управленческих разработках.

4.5. ИНТЕРВАЛЬНЫЙ ДИСКРИМИНАНТНЫЙ АНАЛИЗ

Перейдем к задачам классификации в статистике интервальных данных. Как известно [27], важная их часть — задачи дискриминации (диагностики, распознавания образов с учителем). В этих задачах заданы классы (полностью или частично, с помощью обучающих выборок), и необходимо

принять решение — к какому этих классов отнести вновь поступающий объект.

В линейном дискриминантном анализе правило принятия решений основано на линейной функции *f(x)* от распознаваемого вектора $x \in R^k$. Рассмотрим для простоты случай двух классов. Правило принятия решений определяется константой C — при $f(x) > C$ распознаваемый объект относится к первому классу, при $f(x) \le C$ — ко второму.

В первоначальной вероятностной модели Р.Фишера предполагается, что классы заданы обучающими выборками объемов N_1 и N_2 соответственно из многомерных нормальных распределений с разными математическими ожиданиями, но одинаковыми ковариационными матрицами. В соответствии с леммой Неймана — Пирсона, дающей правило принятия решений при поверке статистических гипотез, дискриминантная функция является линейной. Для ее практического использования теоретические характеристики распределения необходимо заменить на выборочные. Тогда дискриминантная функция приобретает следующий вид:

$$f(x) = \left(x - \frac{1}{2}(\bar{x}_1 + \bar{x}_2) \right)^T S^{-1}(\bar{x}_1 - \bar{x}_2) .$$

Здесь $\bar{x}_1$ — выборочное среднее арифметическое по первой выборке $x_\alpha^{(1)}$, $\alpha = 1,2,...,N_1$, а $\bar{x}_2$ — выборочное среднее арифметическое по второй выборке $x_\beta^{(2)}$, $\beta = 1,2,...,N_2$. В роли S может выступать любая состоятельная оценка общей для выборок ковариационной матрицы. Обычно используют следующую оценку, естественным образом сконструированную на основе выборочных ковариационных матриц:

$$S = \frac{\sum_{\alpha=1}^{N_1}(x_\alpha^{(1)} - \bar{x}_1)(x_\alpha^{(1)} - \bar{x}_1)^T + \sum_{\beta=1}^{N_2}(x_\beta^{(2)} - \bar{x}_2)(x_\beta^{(2)} - \bar{x}_2)^T}{N_1 + N_2 - 2}.$$

В соответствии с подходом статистики интервальных данных считаем, что специалисту по анализу данных известны лишь значения с погрешностями:

$$y_\alpha^{(1)} = x_\alpha^{(1)} + \varepsilon_\alpha^{(1)}, \ \alpha = 1,2,...,N_1, \quad y_\beta^{(2)} = x_\beta^{(2)} + \varepsilon_\beta^{(2)}, \ \beta = 1,2,...,N_2.$$

Таким образом, вместо *f(x)* статистик делает выводы на основе искаженной линейной дискриминантной функции $f_1(x)$, в которой коэффициенты рассчитаны не по исходным данным $x_\alpha^{(1)}, x_\beta^{(2)}$, а по искаженным погрешностями значениям $y_\alpha^{(1)}, y_\beta^{(2)}$.

Это — модель с искаженными параметрами дискриминантной функции. Следующая модель — такая, в которой распознаваемый вектор x также известен с ошибкой. Далее, константа C может появляться в модели различными способами. Она может задаваться априори абсолютно точно. Может задаваться с какой-то ошибкой, не связанной с ошибками, вызванными конечностью обучающих выборок. Может рассчитываться по обучающим выборкам, например, с целью уравнять ошибки классификации, т.е. провести плоскость дискриминации через середину отрезка, соединяющего центры классов. Итак — целый спектр моделей ошибок.

На какие статистические процедуры влияют ошибки в исходных данных? Здесь тоже много постановок. Можно изучать влияние погрешностей измерений на значения дискриминантной функции f, например, в той точке, куда попадает вновь поступающий объект x. Очевидно, случайная величина *f(x)* имеет некоторое распределение, определяемое распределениями обучающих выборок. Выше описана модель Р. Фишера с нормально распределенными совокупностями. Однако реальные данные, как правило, не подчиняются нормальному распределению [27]. Тем не менее линейный статистический анализ имеет смысл и для распределений, не являющихся нормальными (при этом вместо свойств многомерного нормального распределения приходится опираться на многомерную центральную предельную теорему и теорему о наследовании сходимости [3]). В частности, приравняв метрологическую ошибку, вызванную погрешностями исходных данных, и статистическую ошибку, получим условие, определяющее рациональность объемов выборок. Здесь два объема выборок, а не один, как в большинстве рассмотренных постановок статистики интервальных данных. С подобным мы сталкивались ранее при рассмотрении двухвыборочного критерия Смирнова.

Естественно изучать влияние погрешностей исходных данных не при конкретном x, а для правила принятия решений в целом. Может представлять интерес изучение характеристик этого правила по всем x или по какому-либо отрезку. Более интересно рассмотреть показатель качества классификации, связанный с пересчетом на модель линейного дискриминантного анализа [27] (см. также раздел 2.8 выше).

Математический аппарат изучения перечисленных моделей развит выше в предыдущих пунктах настоящей главы. Некоторые результаты приведе-

ны в [14]. Из-за большого объема выкладок ограничимся приведенными здесь замечаниями.

4.6. ИНТЕРВАЛЬНЫЙ КЛАСТЕР-АНАЛИЗ

Кластер-анализ, как известно [27], имеет целью разбиение совокупности объектов на группы сходных между собой. Многие методы кластер-анализа основаны на использовании расстояний между объектами. (Степень близости между объектами может измеряться также с помощью мер близости и показателей различия, для которых неравенство треугольника выполнено не всегда.) Рассмотрим влияние погрешностей измерения на расстояния между объектами и на результаты работы алгоритмов кластер-анализа.

С ростом размерности p евклидова пространства диагональ единичного куба растет как $\sqrt{p}$. А какова погрешность определения евклидова расстояния? Пусть двум рассматриваемым объектам соответствуют $X_0 = (x_1, x_2, ..., x_p)$ и $Y_0 = (y_1, y_2, ..., y_p)$ — вектора размерности p. Они известны с погрешностями $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_p)$ и $\delta = (\delta_1, \delta_2, ..., \delta_p)$, т.е. статистику доступны лишь вектора $X = X_0 + \varepsilon,\ Y = Y_0 + \delta$. Легко видеть, что

$$\rho^2(X,Y) = \rho^2(X_0,Y_0) + 2\sum_{1\le i\le p}(x_i - y_i)(\varepsilon_i - \delta_i) + \sum_{1\le i\le p}(\varepsilon_i - \delta_i)^2. \qquad (73)$$

Пусть ограничения на абсолютные погрешности имеют вид:

$$|\ \varepsilon_i\ | \le \Delta,\ |\ \delta_i\ | \le \Delta,\ \ i = 1,2,...,n.$$

Такая запись ограничений предполагает, что все переменные имеют примерно одинаковый разброс. Трудно ожидать этого, если переменные имеют различные размерности. Однако рассматриваемые ограничения на погрешности естественны, если переменные предварительно стандартизованы, т.е. центрированы и отнормированы (т.е. из каждого значения вычтено среднее арифметическое, а разность поделена на выборочное среднее квадратическое отклонение).

Пусть $p\Delta^2 \to 0$. Тогда последнее слагаемое в (73) не превосходит $4p\Delta^2$, поэтому им можно пренебречь. Тогда из (73) следует, что нотна евклидова расстояния имеет вид:

$$N_{\rho^2}(X_0,Y_0) = 4\sum_{1\le i\le p}|\ x_i - y_i\ |\ \Delta$$

с точностью до бесконечно малых более высокого порядка. Если случайные величины $|x_i - y_i|$ имеют одинаковые математические ожидания и для них справедлив закон больших чисел (эти предположения естественны, если переменные перед применением кластер-анализа стандартизованы), то существует константа C такая, что

$$N_{\rho^2}(X_0, Y_0) = Cp\Delta$$

с точностью до бесконечно малых более высокого порядка при малых Δ, больших p и $p\Delta^2 \to 0$.

Из рассмотрений настоящего пункта вытекает, что

$$\rho(X,Y) = \rho(X_0,Y_0) + \theta \frac{Cp\Delta}{2\rho(X_0,Y_0)} \qquad (74)$$

при некотором θ таком, что $|\theta| < 1$.

Какое минимальное расстояние является различимым? По аналогии с определением рационального объема выборки при проверке гипотез предлагается уравнять слагаемые в (74), т.е. определять минимально различимое расстояние $\rho_{\min}$ из условия:

$$\rho_{\min} = \frac{Cp\Delta}{2\rho_{\min}}, \quad \rho_{\min} = \sqrt{\frac{Cp\Delta}{2}}. \qquad (75)$$

Естественно принять, что расстояния, меньшие $\rho_{\min}$, не отличаются от 0, т.е. точки, лежащие на расстоянии $\rho \le \rho_{\min}$, не различаются между собой.

Каков порядок величины C? Если x_i и y_i независимы и имеют стандартное нормальное распределение с математическим ожиданием 0 и дисперсией 1, то, как легко подсчитать, $M|x_i - y_i| = 2/\sqrt{\pi} = 1{,}13$ и соответственно $C = 4{,}51$. Следовательно, в этой модели:

$$\rho_{\min} = 1{,}5\sqrt{p\Delta}.$$

Формула (75) показывает, что хотя с ростом размерности пространства p растет диаметр (длина диагонали) единичного куба — естественной области расположения значений переменных, с той же скоростью растет и есте-

ственное квантование расстояния с помощью порога неразличимости $\rho_{\min}$, т.е. увеличение размерности (вовлечение новых переменных), вообще говоря, не улучшает возможности кластер-анализа.

Можно сделать выводы и для конкретных алгоритмов. В дендрограммах (например, результатах работы иерархических агломеративных алгоритмах ближнего соседа, дальнего соседа, средней связи) можно порекомендовать склеивать (т.е. объединять) уровни, отличающиеся менее чем на $\rho_{\min}$. Если все уровни склеятся, то можно сделать вывод, что у данных нет кластерной структуры, они однородны. В алгоритмах типа «Форель» центр тяжести текущего кластера определяется с точностью $\pm\Delta$ по каждой координате, а порог для включения точки в кластер (радиус шара R) из-за погрешностей исходных данных может измениться согласно (74) на

$$\pm\frac{2{,}25}{R}p\Delta.$$

Поэтому кроме расчетов с R рекомендуется провести также расчеты с радиусами R_1 и R_2, где

$$R_1 = R\left(1-\frac{2{,}25}{R^2}p\Delta\right), \quad R_2 = R\left(1+\frac{2{,}25}{R^2}p\Delta\right),$$

и сравнить полученные разбиения. Быть адекватными реальности могут только выводы, общие для всех трех расчетов. Эти рекомендации развивают общую идею [3] о целесообразности проведения расчетов при различных значениях параметров алгоритмов с целью выделения выводов, инвариантных по отношению к выбору конкретного алгоритма.

4.7. ИНТЕРВАЛЬНЫЕ ДАННЫЕ В ИНВЕСТИЦИОННОМ МЕНЕДЖМЕНТЕ

Методы статистики интервальных данных оказываются полезными не только в традиционных технических и эконометрических задачах, но и во многих других областях, например, в инвестиционном менеджменте.

Основная идея формулируется так. Все знают, что любое инженерное измерение проводится с некоторой погрешностью. Эту погрешность обычно приводят в документации и учитывают при принятии решений. Ясно, что и

любое экономическое измерение также проводится с погрешностью. А вот какова она? Необходимо уметь ее оценивать, поскольку ошибки при принятии экономических решений обходятся дорого.

Например, как принимать решение о выгодности или невыгодности инвестиционного проекта? Как сравнивать инвестиционные проекты между собой? Как известно, для решения этих задач используют такие экономические характеристики, как *NPV* (*Net Present Value*) — чистая текущая стоимость (этот термин переводится с английского также как чистый дисконтированный доход, чистое приведенное значение и др.), внутренняя норма доходности, срок окупаемости, показатели рентабельности и др.

С экономической точки зрения инвестиционные проекты описываются финансовыми потоками, т.е. функциями от времени, значениями которых являются платежи (и тогда значения этих функций отрицательны) и поступления (значения функций положительны). Сравнение инвестиционных проектов — это сравнение функций от времени с учетом внешней среды, проявляющейся в виде дисконт-функции (как результата воздействия социальных, технологических, экологических, экономических и политических факторов), и представлений законодателя или инвестора — обычно ограничений на финансовые потоки платежей и на горизонт планирования. Основная проблема при сравнении инвестиционных проектов такова: *что лучше — меньше, но сейчас, или больше, но потом*? Как правило, чем больше вкладываем сейчас, тем больше получаем в более или менее отдаленном будущем. Вопрос в том, достаточны ли будущие поступления, чтобы покрыть нынешние платежи и дать приемлемую для инвестора прибыль?

В настоящее время широко используются различные теоретические подходы к сравнению инвестиционных проектов и облегчающие расчеты компьютерные системы, в частности, ТЭО-ИНВЕСТ, *Project Expert, COMFAR, PROPSIN*, Альт-Инвест. Однако ряд важных моментов в них не учтен.

Введем основные понятия.

Дисконт-функция как функция от времени показывает, сколько стоит для фирмы 1 руб. в заданный момент времени, если его привести к начальному моменту. Если дисконт-функция — константа для разных отраслей, товаров и проектов, то эта константа называется дисконт-фактором, или просто дисконтом. Дисконт-функция определяется совместным действием различных факторов, в частности, реальной процентной ставки и индекса инфляции. Реальная процентная ставка описывает «нормальный» рост экономики

(т.е. без инфляции). В стабильной ситуации доходность от вложения средств в различные отрасли, в частности, в банковские депозиты, примерно одинакова. Сейчас она, по оценке ряда экспертов, около 12 %. Итак, нынешний 1 руб. превращается в 1,12 руб. через год, а потому 1 руб. через год соответствует 1/1,12 = 0,89 руб. сейчас — это и есть максимум дисконта.

Обозначим дисконт буквой C. Если q — банковский процент (плата за депозит), т.е. вложив в начале года в банк 1 руб., в конце года получим $(1 + q)$ руб., то дисконт определяется по формуле $C=1/(1+q)$. При таком подходе полагают, что банковские проценты одинаковы во всех банках. Более правильно было бы считать q, а потому и C, нечисловыми величинами, а именно, интервалами $[q_1; q_2]$ и $[C_1; C_2]$. Следовательно, экономические выводы должны быть исследованы на *устойчивость* (применяют и термин *«чувствительность»*) по отношению к возможным отклонениям.

Как функцию времени t дисконт-функцию обозначим $C(t)$. При постоянстве дисконт-фактора имеем $C(t) = C^t$. Если $q = 0{,}12$, $C = 0{,}89$, то 1 руб. за 2 года превращается в $1{,}12^2 = 1{,}2544$, через 3 — в 1,4049. Итак, 1 руб., получаемый через 2 года, соответствует 1/1,2544 = 0,7972 руб., т.е. 79,72 коп. сейчас, а 1 руб., обещанный через 3 года, соответствует 0,71 руб. сейчас. Другими словами, $C(2) = 0{,}80$, а $C(3) = 0{,}71$. Если дисконт-фактор зависит от времени, в первый год равен C_1, во второй — C_2, в третий — C_3, ..., в t-й год — C_t, то $C(t) = C_1 C_2 C_3 \ldots C_t$.

Рассмотрим характеристики потоков платежей. Срок окупаемости — тот срок, за который доходы покроют расходы. Обычно предполагается, что после этого проект приносит только прибыль. Это верно не всегда. Простейший вариант, для которого не возникает никаких парадоксов, состоит в том, что все инвестиции (капиталовложения) делаются сразу, в начале, а затем инвестор получает только доход. Сложности возникают, если проект состоит из нескольких очередей, вложения распределены во времени.

Примитивный способ расчета срока окупаемости состоит в делении объема вложений A на ожидаемый ежегодный доход B. Тогда срок окупаемости равен A/B. Этот способ некорректен. Если дисконт-фактор равен C, то максимально возможный суммарный доход равен:

$$BC + BC^2 + BC^3 + BC^4 + BC^5 + \ldots =$$
$$= BC(1 + C + C^2 + C^3 + C^4 + \ldots) = BC / (1 - C).$$

Если A/B меньше $C/(1 - C)$, то можно рассчитать срок окупаемости проекта, но он будет больше, чем A/B. Если же A/B больше или равно $C/(1 - C)$, то проект не окупится никогда. Поскольку максимум C равен 0,89, то проект не окупится никогда, если A/B не меньше 8,09.

Пусть вложения равны 1 млн. руб., ежегодная прибыль составляет 500 тыс., т.е. $A/B = 2$, дисконт-фактор $C = 0.8$. При примитивном подходе (при $C = 1$) срок окупаемости равен 2 годам. А на самом деле? За k лет будет возвращено:

$$BC(1 + C + C^2 + C^3 + C^4 + ... + C^k) = BC(1 - C^{k+1}) / (1 - C).$$

Срок окупаемости k получаем из уравнения $1 = 0{,}5 \times 0{,}8(1 - 0{,}8^{k+1})/(1 - 0{,}8)$, откуда $k = 2{,}11$. Он оказался равным 2,11 лет, т.е. увеличился примерно на 6 недель. Это немного. Однако если $B = 0{,}2$, то имеем уравнение $1 = 0{,}2 \times 0{,}8\ (1 - 0{,}8^{k+1}) / (1 - 0{,}8)$. У этого уравнения нет корней, поскольку $A/B = 5 > C / (1 - C) = 0.8/(1 - 0{,}8) = 4$. Проект не окупится никогда. Прибыль можно ожидать лишь при $A/B < 4$. Рассмотрим промежуточный случай, $B = 0{,}33$, с «примитивным» сроком окупаемости 3 года. Тогда имеем уравнение $1 = 0{,}33 \times 0{,}8(1 - 0{,}8^{k+1})/ (1 - 0{,}8)$, откуда $k = 5{,}40$.

Рассмотрим финансовый поток $a(0)$, $a(1)$, $a(2)$, $a(3)$, ..., $a(t)$, ... (для простоты примем, что платежи или поступления происходят раз в год). Выше рассмотрен поток с одним платежом $a(0) = (-A)$ и дальнейшими поступлениями $a(1) = a(2) = a(3) = ... = a(t) = ... = B$. Чистая текущая стоимость (*Net Present Value*, сокращенно *NPV*) рассчитывается для финансового потока путем приведения затрат и поступлений к начальному моменту времени:

$$NPV = a(0) + a(1)C(1) + a(2)C(2) + a(3)C(3) + ... + a(t)C(t) + ...,$$

где $C(t)$ — дисконт-функция. В простейшем случае, когда дисконт-фактор не меняется год от года и имеет вид $C=1/(1+q)$, формула для *NPV* конкретизируется:

$$NPV = NPV(q) = a(0) + a(1)/(1 + q) +$$
$$+a(2)/(1 + q)^2 + a(3)/(1 + q)^3 + ... + a(t)/(1 + q)^t + ...$$

Пусть, например, $a(0)= -10$, $a(1)=3$, $a(2)=4$, $a(3)=5$. Пусть $q=0{,}12$, тогда

$$NPV(0{,}12) = -10 + 3 \times 0{,}89 + 4 \times 0.80 + 5 \times 0{,}71 =$$
$$= -10 + 2{,}67 + 3{,}20 + 3{,}55 = -0{,}58.$$

Итак, проект невыгоден для вложения капитала, поскольку $NPV(0{,}12)$ отрицательна. При отсутствии дисконтирования (при $C = 1$, $q = 0$) вывод иной:

$$NPV(0) = -10 + 3 + 4 + 5 = 2,$$

проект выгоден.

Срок окупаемости и сам вывод о прибыльности проекта зависят от неизвестного дисконт-фактора C или даже от неизвестной дисконт-функции — ибо какие у нас основания считать будущую дисконт-функцию постоянной? Экономическая история России последних лет показывает, что банки часто меняют проценты платы за депозит. Часто предлагают использовать норму дисконта, равную *приемлемой для инвестора норме дохода на капитал.* Это значит, что экономисты явным образом обращаются к инвестору как к эксперту, который должен назвать им некоторое число исходя из своего опыта и интуиции (т.е. экономисты перекладывают свою работу на инвестора). Кроме того, при этом игнорируется изменение указанной нормы во времени,

Приведем пример исследования NPV на устойчивость (чувствительность) к малым отклонениям значений дисконт-функции. Для этого надо найти максимально возможное отклонение NPV при допустимых отклонениях значений дисконт-функции (или, если угодно, значений банковских процентов). В качестве примера рассмотрим:

$$NPV = NPV\,(a(0), a(1), C(1), a(2), C(2), a(3), C(3)) = \\ = a(0) + a(1)C(1) + a(2)C(2) + a(3)C(3).$$

Предположим, что изучается устойчивость (чувствительность) для ранее рассмотренных значений:

$$a(0) = -10,\ a(1) = 3,\ a(2) = 4,\ a(3) = 5,\ C(1) = 0{,}89,\ C(2) = 0{,}80,\ C(3) = 0{,}71.$$

Пусть максимально возможные отклонения $C(1)$, $C(2)$, $C(3)$ равны $\pm 0{,}05$. Тогда максимум значений NPV равен:

$$NPVmax = -10 + 3 \times 0{,}94 + 4 \times 0.85 + 5 \times 0{,}76 = -10 + 2{,}82 + 3{,}40 + 3{,}80 = 0{,}02,$$

в то время как минимум значений NPV есть

$$NPVmin = -10 + 3 \times 0{,}84 + 4 \times 0.75 + 5 \times 0{,}66 = \\ = -10 + 2{,}52 + 3{,}00 + 3{,}30 = -1{,}18.$$

Для *NPV* получаем интервал от (–1,18) до (+0,02). В нем есть и положительные, и отрицательные значения. Следовательно, нет однозначного заключения — проект убыточен или выгоден. Для принятия решения не обойтись без экспертов.

Для иных характеристик, например, внутренней нормы доходности, выводы аналогичны. Дополнительные проблемы вносит неопределенность горизонта планирования, а также будущая инфляция. Если считать, что финансовый поток должен учитывать инфляцию, то это означает, что до принятия решений об инвестициях необходимо на годы вперед спрогнозировать рост цен, а это до сих пор еще не удавалось ни одной государственной или частной исследовательской структуре. Если же рост цен не учитывать, то отдаленные во времени доходы могут «растаять» в огне инфляции. На практике риски учитывают, увеличивая q на десяток-другой процентов.

4.8. СТАТИСТИКА ИНТЕРВАЛЬНЫХ ДАННЫХ В ПРИКЛАДНОЙ СТАТИСТИКЕ

Кратко рассмотрим положение статистики интервальных данных (СИД) среди других методов описания неопределенностей и анализа данных.

Нечеткость и СИД. С формальной точки зрения описание нечеткости интервалом — это частный случай описания ее нечетким множеством. В СИД функция принадлежности нечеткого множества имеет специфический вид — она равна 1 в некотором интервале и 0 вне его. Такая функция принадлежности описывается всего двумя параметрами (границами интервала). Эта простота описания делает математический аппарат СИД гораздо более прозрачным, чем аппарат теории нечеткости в общем случае. Это, в свою очередь, позволяет исследователю продвинуться дальше, чем при использовании функций принадлежности произвольного вида.

Интервальная математика и СИД. Можно было бы сказать, что СИД — часть интервальной математики, что СИД так соотносится с прикладной математической статистикой, как интервальная математика — с математикой в целом. Однако исторически сложилось так, что интервальная математика занимается прежде всего вычислительным погрешностями. С точки зрения интервальной математики две известные формулы для выборочной дисперсии, рассмотренные выше, имеют разные погрешности. А с точки зрения СИД эти две формулы задают одну и ту же функцию, и поэтому им соответствуют совпадающие нотны и рациональные объемы выборок. Интервальная математика прослеживает процесс вычислений, СИД этим не

занимается. Необходимо отметить, что типовые постановки СИД могут быть перенесены в другие области математики, и, наоборот, вычислительные алгоритмы прикладной математической статистики и СИД заслуживают изучения. Однако и то, и другое — скорее дело будущего. Из уже сделанного отметим применение методов СИД при анализе такой характеристики финансовых потоков, как NPV — чистая текущая стоимость (см. выше).

Математическая статистика и СИД. Как уже отмечалось, математическая статистика и СИД отличаются тем, в каком порядке делаются предельные переходы $n \to \infty$ и $\Delta \to 0$. При этом СИД переходит в математическую статистику при $\Delta = 0$. Правда, тогда исчезают основные особенности СИД: нотна становится равной 0, а рациональный объем выборки — бесконечности. Рассмотренные выше методы СИД разработаны в предположении, что погрешности малы (но не исчезают), а объем выборки велик. СИД расширяет классическую математическую статистику тем, что в исходных статистических данных каждое число заменяет интервалом. С другой стороны, можно считать СИД новым этапом развития математической статистики.

Статистика объектов нечисловой природы и СИД. Статистика объектов нечисловой природы (СОНП) расширяет область применения классической математической статистики путем включения в нее новых видов статистических данных. Естественно, при этом появляются новые виды алгоритмов анализа статистических данных и новый математический аппарат (в частности, происходит переход от методов суммирования к методам оптимизации). С точки зрения СОНП частному виду новых статистических данных — интервальным данным — соответствует СИД. Напомним, что одно из двух основных понятий СИД — нотна — определяется как решение оптимизационной задачи. Однако СИД, изучая классические методы прикладной статистики применительно к интервальным данным, по математическому аппарату ближе к классике, чем другие части СОНП, например, статистика бинарных отношений.

Робастные методы статистики и СИД. Если понимать робастность согласно [3] как теорию устойчивости статистических методов по отношению к допустимым отклонениям исходных данных и предпосылок модели, то в СИД рассматривается одна из естественных постановок робастности. Однако в массовом сознании специалистов термин «робастность» закрепился за моделью засорения выборки большими выбросами (модель Тьюки — Хубера), хотя эта модель не имеет большого практического значения [27]. К этой модели СИД не имеет отношения.

Теория устойчивости и СИД. Общей схеме устойчивости [3] математических моделей социально-экономических явлений и процессов по отношению к допустимым отклонениям исходных данных и предпосылок моделей СИД полностью соответствует. Она посвящена математико-статистическим моделям, используемым при анализе статистических данных, а допустимые отклонения — это интервалы, заданные ограничениями на погрешности. СИД можно рассматривать как пример теории, в которой учет устойчивости позволил сделать нетривиальные выводы. Отметим, что с точки зрения общей схемы устойчивости [3] устойчивость по Ляпунову в теории дифференциальных уравнений — весьма частный случай, в котором из-за его конкретности удалось весьма далеко продвинуться.

Минимаксные методы, типовые отклонения и СИД. Постановки СИД относятся к минимаксным. За основу берется максимально возможное отклонение. Это — «подход пессимиста», применяемый, например, в теории антагонистических игр. Использование минимаксного подхода позволяет подозревать СИД в завышении роли погрешностей измерения. Однако примеры изучения вероятностно-статистических моделей погрешностей, проведенные, в частности, при разработке методов оценивания параметров гамма-распределения [4], показали, что это подозрение не подтверждается. Влияние погрешностей измерений по порядку такое же, только вместо максимально возможного отклонения (нотны) приходится рассматривать математическое ожидание соответствующего отклонения (см. выше). Подчеркнем, что применение в СИД вероятностно-статистических моделей погрешностей не менее перспективно, чем минимаксных.

Подход научной школы А. П. Вощинина и СИД. Если в математической статистике неопределенность только статистическая, то в научной школе А. П. Вощинина — только интервальная. Можно сказать, что СИД лежит между классической прикладной математической статистикой и областью исследований научной школы А. П. Вощинина. Другое отличие состоит в том, что в этой школе разрабатывают новые методы анализа интервальных данных, а в СИД в настоящее время изучается устойчивость классических статистических методов по отношению к малым погрешностям. Подход СИД оправдывается распространенностью этих методов, однако в дальнейшем следует переходить к разработке новых методов, специально предназначенных для анализа интервальных данных.

Анализ чувствительности и СИД. При анализе чувствительности, как и в СИД, рассчитывают производные по используемым переменным, или непосредственно находят изменения при отклонении переменной на $\pm 10\,\%$

от базового значения. Однако этот анализ делают по каждой переменной отдельно. В СИД все переменные рассматриваются совместно, и находится максимально возможное отклонение (нотна). При малых погрешностях удается на основе главного члена разложения функции в многомерный ряд Тейлора получить удобную формулу для нотны. Можно сказать, что СИД — это многомерный анализ чувствительности.

По нашему мнению, во все виды статистического программного обеспечения должны быть включены алгоритмы интервальной статистики, «параллельные» обычно используемым в настоящее время алгоритмам прикладной математической статистики. Это позволит в явном виде учесть наличие погрешностей у результатов наблюдений (измерений, испытаний, анализов, опытов).

Основным идеям статистики интервальных данных посвящены работы [52, 53]. Оценку погрешностей характеристик финансовых потоков инвестиционных проектов в ракетно-космической промышленности предлагается проводить на основе применения методов статистики интервальных данных [54].

Темы докладов, рефератов, исследовательских работ

1. Классическая математическая статистика как предельный случай статистики интервальных данных.

2. Концепция рационального объема выборки.

3. Сравнение методов оценивания параметров и характеристик распределений в статистике интервальных данных и в классической математической статистике.

4. Подход к проверке гипотез в статистике интервальных данных.

5. Метод наименьших квадратов для интервальных данных.

6. Различные способы учета погрешностей исходных данных в статистических процедурах.

7. Статистика интервальных данных как часть теории устойчивости (с использованием монографии [3]).

Контрольные вопросы и задачи

1. Покажите на примерах, что в задачах принятия решений исходные данные часто имеют интервальный характер.

2. В чем особенности подхода статистики интервальных данных в задачах оценивания параметров?

3. В чем особенности подхода статистики интервальных данных в задачах проверки гипотез?

4. Какие новые нюансы проявляются в статистике интервальных данных при переходе к многомерным задачам?

5. Выполните операции над интервальными числами:

Вариант 1:

а) [1, 2] + [3, 4];

б) [4, 5] – [2, 3];

в) [3, 4] × [5, 7];

г) [10, 20] : [4, 5].

Вариант 2:

а) [0, 2] + [3, 5];

б) [3, 5] – [2, 4];

в) [2, 4] × [5, 8];

г) [15, 25] : [1, 5].

6. Выпишите формулу для асимптотической нотны (ошибки определения переменных по абсолютной величине не превосходят константы t, предполагающейся малой) для функции:

$$f(x_1, x_2) = 5\,(x_1)^2 + 10\,(x_2)^2 + 7\,x_1x_2.$$

Вычислите асимптотическую нотну в точке $(x_1, x_2) = (1, 2)$ при $t = 0{,}1$.

7. Выпишите формулу для асимптотической нотны (ошибки определения переменных по абсолютной величине не превосходят константы t, предполагающейся малой) для функции:

$$f(x_1, x_2) = 4\,(x_1)^2 + 12\,(x_2)^2 - 3\,x_1x_2.$$

Вычислите асимптотическую нотну в точке $(x_1, x_2) = (2, 1)$ при $t = 0{,}05$.

Литература

1. Дискуссия по анализу интервальных данных // Заводская лаборатория. — 1990. — Т. 56. — № 7. — С. 75–95.

2. Сборник трудов Международной конференции по интервальным и стохастическим методам в науке и технике (ИНТЕРВАЛ-92) : в 2 томах. — Москва : МЭИ, 1992. — 216 с.; 152 с.

3. *Орлов, А. И.* Устойчивость в социально-экономических моделях / А. И. Орлов. — Москва : Наука, 1979. — 296 с.

4. ГОСТ 11.011-83. Прикладная статистика. Правила определения оценок и доверительных границ для параметров гамма-распределения = Applied statistics. Regulations for determinations of estimates and confidence limits for parameters of gamma distribution : государственный стандарт Союза ССР : издание официальное : утвержден Постановлением Государственного комитета СССР по стандартам от 27 июня 1983 г. № 2684 : введен впервые : дата введения 1 января 1985 г. — Москва : Изд-во стандартов, 1984. — 53 с. (В настоящее время отменен как нормативный документ, но может использоваться как научная публикация.)

5. *Orlov, A. I.* Interval statistics // Interval Computations. — 1992. — № 1(3). — P. 44–52.

6. *Орлов, А. И.* Основные идеи интервальной математической статистики / А. И. Орлов // Наука и технология в России. — 1994. — № 4 (6). — С. 8–9.

7. *Шокин, Ю. И.* Интервальный анализ / Ю. И. Шокин. — Новосибирск : Наука, 1981. — 112 с.

8. *Орлов, А. И.* О развитии реалистической статистики / А. И. Орлов // Статистические методы оценивания и проверки гипотез : межвузовский сборник научных трудов. — Пермь : Изд-во ПГНИУ, 1990. — С. 89–99.

9. *Орлов, А. И.* Некоторые алгоритмы реалистической статистики / А. И. Орлов // Статистические методы оценивания и проверки гипотез : межвузовский сборник научных трудов. — Пермь : Изд-во ПГНИУ, 1991. — С. 77–86.

10. *Орлов, А. И.* О влиянии погрешностей наблюдений на свойства статистических процедур (на примере гамма-распределения) / А. И. Орлов // Статистические методы оценивания и проверки гипотез : межвузовский сборник научных трудов. — Пермь : Изд-во ПГНИУ, 1988 — С. 45–55.

11. *Орлов, А. И.* Интервальная статистика: метод максимального правдоподобия и метод моментов / А. И. Орлов // Статистические методы оценивания и проверки гипотез : межвузовский сборник научных трудов. — Пермь : Изд-во ПГНИУ, 1995. — С. 114–124.

12. *Орлов, А. И.* Интервальный статистический анализ / А. И. Орлов // Статистические методы оценивания и проверки гипотез : межвузовский сборник научных трудов. — Пермь : Изд-во ПГНИУ, 1993. — С. 149–158.

13. *Биттар, А. Б.* Метод наименьших квадратов для интервальных данных : дипломная работа / А. Б. Биттар. — Москва : МЭИ, 1994. — 38 с.

14. *Пузикова, Д. А.* Об интервальных методах статистической классификации / Д. А. Пузикова // Наука и технология в России. — 1995. — № 2 (8). — С. 12–13.

15. *Орлов, А. И.* Пути развития статистических методов: непараметрика, робастность, бутстреп и реалистическая статистика / А. И. Орлов // Надежность и контроль качества. — 1991. — № 8. — С. 3–8.

16. *Орлов, А. И.* Современная прикладная статистика / А. И. Орлов // Заводская лаборатория. — 1998. — Т. 64. — № 3. — С. 52–60.

17. *Вощинин, А. П.* Метод оптимизации объектов по интервальным моделям целевой функции / А. П. Вощинин. — Москва : МЭИ, 1987. — 109 с.

18. *Вощинин, А. П.* Оптимизация в условиях неопределенности / А. П. Вощинин, Г. Р. Сотиров. — Москва : МЭИ ; София : Техника, 1989. — 224 с.

19. *Вощинин, А. П.* Оптимизация по регрессионным моделям и планирование эксперимента / А. П. Вощинин, Р. А. Акматбеков. — Бишкек : Илим, 1991. — 164 с.

20. *Вощинин, А. П.* Метод анализа данных с интервальными ошибками в задачах проверки гипотез и оценивания параметров неявных и линейно параметризованных функций / А. П. Вощинин // Заводская лаборатория. — 2000. — Т. 66. — № 3. — С. 51–65.

21. *Вощинин, А. П.* Интервальный анализ данных: развитие и перспективы / А. П. Вощинин // Заводская лаборатория. — 2002. — Т. 68. — № 1. — С. 118–126.

22. *Дывак, Н. П.* Разработка методов оптимального планирования эксперимента и анализа интервальных данных : специальность 05.13.01 «Системный анализ, управление и обработка информации» : автореферат диссертации на соискание ученой степени кандидата технических наук / Дывак Николай Петрович ; Московский энергетический институт. — Москва : МЭИ, 1992. — 20 с.

23. *Симов, С. Ж.* Разработка и исследование интервальных моделей при анализе данных и проектировании экспертных систем : специальность 05.13.01 «Системный анализ, управление и обработка информации» : автореферат диссертации на соискание ученой степени кандидата технических наук / Симов Симеон Живков ; Московский энергетический институт. — Москва : МЭИ, 1992. — 20 с.

24. *Орлов, А. И.* Термины и определения в области вероятностно-статистических методов / А. И. Орлов // Заводская лаборатория. — 1999. — Т. 65. — № 7. — С. 46–54.

25. *Орлов, А. И.* Часто ли распределение результатов наблюдений является нормальным? / А. И. Орлов // Заводская лаборатория. — 1991. — Т. 57. — № 7. — С. 64–66.

26. *Новицкий, П. В.* Оценка погрешностей результатов измерений / П. В. Новицкий, И. А. Зограф. — Ленинград : Энергоатомиздат, 1985. — 248 с.

27. *Орлов, А. И.* Эконометрика / А. И. Орлов. — 3-е изд., испр. и доп. — Москва : Экзамен, 2004. — 576 с.

28. *Дейвид, Г.* Порядковые статистики / Г. Дейвид. — Москва : Наука, 1979. — 340 с.

29. *Колмогоров, А. Н.* Метод медианы в теории ошибок / А. Н. Колмогоров // Теория вероятностей и математическая статистика : сборник статей. — Москва : Наука, 1986. — С. 111–114.

30. *Орлов, А. И.* Об оценивании параметров гамма-распределения / А. И. Орлов // Обозрение прикладной и промышленной математики. — 1997. — Т. 4. — Вып. 3. — С. 471–482.

31. *Гнеденко, Б. В.* Элементарное введение в теорию вероятностей / Б. В. Гнеденко, А. Я. Хинчин. — Москва : Наука, 1970. — 168 с.

32. *Бронштейн, И. Н.* Справочник по математике для инженеров и учащихся втузов / И. Н. Бронштейн, К. А. Семендяев. — Москва : Ленинград : ГИТТЛ, 1945. — 608 с.

33. *Кендалл, М.* Статистические выводы и связи / М. Кендалл, А. Стьюарт. — Москва : Наука, 1973. — 900 с.

34. Рекомендации. Прикладная статистика. Методы обработки данных. Основные требования и характеристики / А. И. Орлов, В. Н. Фомин [и др.]. — Москва : ВНИИСтандартизации, 1987. — 62 с.

35. *Ляшенко, Н. Н.* Машинное умножение и деление независимых случайных величин / Н. Н. Ляшенко, М. С. Никулин // Записки научных семинаров Ленингр. Отделения Математического ин-та АН СССР. Т. 153. — Ленинград : Наука, 1986. — С. 97–104.

36. *Хьюбер, П.* Робастность в статистике / П. Хьюбер. — Москва : Мир, 1984. — 303 с.

37. *Орлов, А. И.* Асимптотика решений экстремальных статистических задач / А. И. Орлов // Анализ нечисловых данных в системных исследованиях : сборник трудов. Вып. 10. — Москва : ВНИИ системных исследований АН СССР, 1982. — С. 4–12.

38. *Крамер, Г.* Математические методы статистики / Г. Крамер. — Москва : Мир, 1975. — 648 с.

39. *Боровков, А. А.* Математическая статистика / А. А. Боровков. — Москва : Наука, 1984. — 472 с.

40. *Кендалл, М.* Теория распределений / М. Кендалл, А. Стьюарт. — Москва : Наука, 1966. — 566 с.

41. *Смирнов, Н. В.* Оценка расхождения между эмпирическими кривыми распределения в двух независимых выборках / Н. В. Смирнов // Бюллетень МГУ. Сер. А. — 1939. — Т. 2. — № 2. — С. 3–14.

42. *Большев, Л. Н.* Таблицы математической статистики / Л. Н. Большев, Н. В. Смирнов. — Москва : Наука, 1983. — 474 с.

43. *Орлов, А. И.* О критериях Колмогорова и Смирнова / А. И. Орлов // Заводская лаборатория. — 1995. — Т. 61. — № 7. — С. 59–61.

44. *Гантмахер, Ф. Р.* Теория матриц / Ф. Р. Гантмахер. — Москва : Наука, 1966. — 576 с.

45. *Розанов, Ю. А.* Теория вероятностей, случайные процессы и математическая статистика / Ю. А. Розанов. — Москва : Наука, 1989. — 320 с.

46. *Налимов, В. В.* Логические основания планирования эксперимента / В. В. Налимов, Т. И. Голикова. — Москва : Металлургия, 1976. — 128 с.

47. *Орлов, А. И.* Менеджмент в техносфере / А. И. Орлов, В. Н. Федосеев. — Москва : Академия, 2003. — 384 с.

48. *Орлов, А. И.* Эконометрическая поддержка контроллинга / А. И. Орлов // Контроллинг. — 2002. — № 1. — С. 42–53.

49. *Орлов, А. И.* Информационные системы управления предприятием в решении задач контроллинга / А. И. Орлов, Е. А. Гуськова // Контроллинг. — 2003. — № 1(5). — С. 52–59.

50. *Гуськова, Е. А.* Интервальная линейная парная регрессия (обобщающая статья) / Е. А. Гуськова, А. И. Орлов // Заводская лаборатория. — 2005. — Т. 71. — № 3. — С. 57–63.

51. *Орлов, А. И.* Прикладная статистика / А. И. Орлов. — Москва : Экзамен, 2006. — 671 с.

52. *Орлов, А. И.* Основные идеи статистики интервальных данных / А. И. Орлов // Научный журнал КубГАУ. — 2013. — № 94. — С. 55–70.

53. *Орлов, А. И.* Статистика интервальных данных (обобщающая статья) / А. И. Орлов // Заводская лаборатория. Диагностика материалов. — 2015. — Т. 81. — № 3. — С. 61–69.

54. *Орлов, А. И.* Оценка погрешностей характеристик финансовых потоков инвестиционных проектов в ракетно-космической промышленности / А. И. Орлов // Научный журнал КубГАУ. — 2015. — № 109. — С. 238–264.

ПРИЛОЖЕНИЯ

Приложение 1

Теоретическая база нечисловой статистики

В настоящем приложении собраны основные математико-статистические утверждения, постоянно используемые при математическом обосновании методов нечисловой статистики. Эти утверждения отнюдь не всегда легко найти в литературе по теории вероятностей и математической статистике. Например, такие рассматриваемые далее теоремы и методы, как многомерная центральная предельная теорема, теоремы о наследовании сходимости и метод линеаризации, даже не включены в энциклопедию «Вероятность и математическая статистика» [1] — наиболее полный свод знаний по этой тематике. Последний факт наглядно демонстрирует разрыв между математической дисциплиной «теория вероятностей и математическая статистика» и потребностями прикладной статистики.

1. Законы больших чисел

Законы больших чисел позволяют описать поведение сумм случайных величин. Простейшим примером является следующая теорема Чебышева для пространства элементарных событий из конечного числа элементов.

Теорема Чебышева. Пусть случайные величины X_1, X_2, …, X_k попарно независимы и существует число C такое, что дисперсии всех этих случайных величин не превосходят C, т.е. $D(X_i) \le C$ при всех $i = 1, 2, \ldots, k$. Тогда для любого положительного ε выполнено неравенство:

$$P\left\{\left|\frac{X_1+X_2+\ldots+X_k}{k}-\frac{M(X_1)+M(X_2)+\ldots+M(X_k)}{k}\right|\ge\varepsilon\right\}\le\frac{C}{k\varepsilon^2}. \qquad (1)$$

Частным случаем теоремы Чебышева является теорема Бернулли — первый в истории вариант закона больших чисел.

Теорема Бернулли. Пусть m — число наступлений события A в k независимых (попарно) испытаниях, и p есть вероятность наступления события A в каждом из испытаний. Тогда при любом $\varepsilon > 0$ справедливо неравенство:

$$P\left\{\left|\frac{m}{k}-p\right|\ge\varepsilon\right\}\le\frac{p(1-p)}{k\varepsilon^2}. \qquad (2)$$

Ясно, что при росте k выражения в правых частях формул (1) и (2) стремятся к 0. Таким образом, среднее арифметическое попарно независимых случайных величин сближается со средним арифметическим их математических ожиданий.

Напомним, что до сих пор речь шла лишь о пространствах элементарных событий из конечного числа элементов. Однако приведенные теоремы верны и в общем случае, для произвольных пространств элементарных событий. Однако в условие закона больших чисел необходимо добавить требование существования дисперсий. Легко видеть, что если существуют дисперсии, то существуют и математические ожидания. Закон больших чисел в форме Чебышёва приобретает следующий вид.

Теорема Чебышева [2, с. 147]. Если $X_1, X_2, \ldots, X_k, \ldots$ — последовательность попарно независимых случайных величин, имеющих конечные дисперсии, ограниченные одной и той же постоянной:

$$D(X_1) \leq C,\ D(X_2) \leq C, \ldots\ D(X_k) \leq C, \ldots\ ,$$

то, каково бы ни было постоянное $\varepsilon > 0$,

$$\lim_{k\to\infty} P\left\{\left|\frac{1}{k}\sum_{j=1}^{k} X_j - \frac{1}{k}\sum_{j=1}^{k} MX_j\right| < \varepsilon\right\} = 1. \qquad (3)$$

С точки зрения прикладной статистики ограниченность дисперсий вполне естественна. Она вытекает, например, из ограниченности диапазона изменения практически всех величин, используемых при реальных расчетах.

В 1923 г. А. Я. Хинчин показал, что если случайные величины не только независимы, но и одинаково распределены, то существование у них математического ожидания является необходимым и достаточным условием для применимости закона больших чисел [2, с. 150].

Теорема [2, с. 150–151]. Для того чтобы для последовательности $X_1, X_2, \ldots, X_k, \ldots,$ (как угодно зависимых) случайных величин при любом положительном ε выполнялось соотношение (3), необходимо и достаточно, чтобы при $k \to \infty$:

$$M\frac{\left(\sum_{j=1}^{k}(X_j - MX_j)\right)^2}{k^2 + \left(\sum_{j=1}^{k}(X_j - MX_j)\right)^2} \to 0.$$

Законы больших чисел для случайных величин служат основой для аналогичных утверждений для случайных элементов в пространствах более сложной природы. В частности, в пространствах произвольной природы (см. главу 2). Однако здесь мы ограничимся классическими формулировками, служащими основой для современной прикладной статистики.

Смысл классических законов больших чисел состоит в том, что выборочное среднее арифметическое независимых одинаково распределенных случайных величин приближается (сходится) к математическому ожиданию этих величин. Другими словами, *выборочные средние сходятся к теоретическому среднему.*

Это утверждение справедливо и для других видов средних. Например, выборочная медиана сходится к теоретической медиане. Это утверждение — тоже закон больших чисел, но не классический.

Существенным продвижением в теории вероятностей во второй половине XX в. явилось введение средних величин в пространствах произвольной природы и получение для них законов больших чисел, т.е. утверждений, состоящих в том, что эмпирические (т.е. выборочные) средние сходятся к теоретическим средним. Эти результаты рассмотрены в главе 2.

2. Центральные предельные теоремы

Простейший вариант Центральной предельной теоремы (ЦПТ) теории вероятностей таков.

Центральная предельная теорема (для одинаково распределенных слагаемых). Пусть $X_1, X_2, \ldots, X_n, \ldots$ — независимые одинаково распределенные случайные величины с математическими ожиданиями $M(X_i) = m$ и дисперсиями $D(X_i) = \sigma^2$, $i = 1, 2, \ldots, n, \ldots$ Тогда для любого действительного числа x существует предел:

$$\lim_{n\to\infty} P\left(\frac{X_1 + X_2 + \ldots + X_n - nm}{\sigma\sqrt{n}} < x\right) = \Phi(x),$$

где $\Phi(x)$ — функция стандартного нормального распределения.

Эту теорему иногда называют теоремой Линдеберга — Леви [3, с. 122].

В ряде прикладных задач не выполнено условие одинаковой распределенности. В таких случаях центральная предельная теорема обычно остается справедливой, однако на последовательность случайных величин приходится

накладывать те или иные условия. Суть этих условий состоит в том, что ни одно слагаемое не должно быть доминирующим, вклад каждого слагаемого в среднее арифметическое должен быть пренебрежимо мал по сравнению с итоговой суммой. Наиболее часто используется теорема Ляпунова.

Центральная предельная теорема (для разнораспределенных слагаемых) — *теорема Ляпунова*. Пусть $X_1, X_2, \ldots, X_n, \ldots$ — независимые случайные величины с математическими ожиданиями $M(X_i) = m_i$ и дисперсиями $D(X_i) = \sigma_i^2 \neq 0$, $i = 1, 2, \ldots, n, \ldots$ Пусть при некотором $\delta > 0$ у всех рассматриваемых случайных величин существуют центральные моменты порядка $2 + \delta$ и безгранично убывает «дробь Ляпунова»:

$$\lim_{n\to\infty} \frac{1}{B_n^{2+\delta}} \sum_{k=1}^{n} M \mid X_k - m_k \mid^{2+\delta} = 0,$$

где

$$B_k^2 = \sum_{i=1}^{k} \sigma_i^2 = D\left(\sum_{i=1}^{k} X_i \right).$$

Тогда для любого действительного числа x существует предел:

$$\lim_{n\to\infty} P\left(\frac{X_1 + X_2 + \ldots + X_n - m_1 - m_2 - \ldots - m_n}{B_n} < x \right) = \Phi(x), \qquad (1)$$

где $\Phi(x)$ — функция стандартного нормального распределения.

В случае одинаково распределенных случайных слагаемых:

$$m_1 = m_2 = \ldots = m_n = m, \quad B_n = [D(X_1 + X_2 + \ldots + X_n)]^{1/2} = \sigma\sqrt{n},$$

и теорема Ляпунова переходит в теорему Линдеберга — Леви.

История получения центральных предельных теорем для случайных величин растянулась на два века — от первых работ Муавра в 30-х гг. XVIII в. для необходимых и достаточных условий, полученных Линдебергом и Феллером в 30-х гг. XX в.

Теорема Линдеберга — Феллера. Пусть $X_1, X_2, \ldots, X_n, \ldots$ — независимые случайные величины с математическими ожиданиями $M(X_i) = m_i$ и дисперси-

ями $D(X_i) = \sigma_i^2 \neq 0$, $i = 1, 2, \ldots, n, \ldots$ Предельное соотношение (1), т.е. центральная предельная теорема, выполнено тогда и только тогда, когда при любом $\tau > 0$:

$$\lim_{n\to\infty} \frac{1}{B_n^2} \sum_{k=1}^{n} \int_{|x-m_k|>\tau B_n} (x-m_k)^2 \, dF_k(x) = 0,$$

где $F_k(x)$ обозначает функцию распределения случайной величины X_k.

Доказательства перечисленных вариантов центральной предельной теоремы для случайных величин можно найти в классическом курсе теории вероятностей [2].

Для прикладной статистики и, в частности, для нечисловой статистики большое значение имеет многомерная центральная предельная теорема. В ней речь идет не о сумме случайных величин, а о сумме случайных векторов.

Необходимое и достаточное условие многомерной сходимости [3, с. 124]. Пусть F_n обозначает совместную функцию распределения k-мерного случайного вектора $(X_n^{(1)},\ldots,X_n^{(k)})$, $n = 1, 2, \ldots$, и $F_{\lambda n}$ — функция распределения линейной комбинации $\lambda_1 X_n^{(1)} + \lambda_2 X_n^{(2)} \ldots + \lambda_k X_n^{(k)}$. Необходимое и достаточное условие для сходимости F_n к некоторой k-мерной функции распределения F состоит в том, что $F_{\lambda n}$ имеет предел для любого вектора λ.

Приведенная теорема ценна тем, что сходимость векторов сводит к сходимости линейных комбинаций их координат, т.е. к сходимости обычных случайных величин, рассмотренных ранее. Однако она не дает возможности непосредственно указать предельное распределение. Это можно сделать с помощью следующей теоремы.

Теорема о многомерной сходимости. Пусть F_n и $F_{\lambda n}$ — те же, что в предыдущей теореме. Пусть F — совместная функция распределения k-мерного случайного вектора $(X_1,\ldots,X_k)$. Если функция распределения $F_{\lambda n}$ сходится при росте объема выборки к функции распределения F_λ для любого вектора λ, где F_λ — функция распределения линейной комбинации $\lambda_1 X_1 + \ldots + \lambda_k X_k$, то F_n сходится к F.

Здесь сходимость F_n к F означает, что для любого k-мерного вектора $(x_1,\ldots,x_k)$ такого, что функция распределения F непрерывна в $(x_1,\ldots,x_k)$, числовая последовательность $F_n(x_1,\ldots,x_k)$ сходится при росте n к числу $F(x_1,\ldots,x_k)$. Другими словами, сходимость функций распределения понимается ровно

также, как при обсуждении предельных теорем для случайных величин выше. Приведем многомерный аналог этих теорем.

Многомерная центральная предельная теорема [3]. Рассмотрим независимые одинаково распределенные k-мерные случайные вектора:

$$U_n' = (U_{1n},...,U_{kn}), \quad n = 1,2,...,$$

где штрих обозначает операцию транспонирования вектора. Предположим, что случайные вектора U_n имеют моменты первого и второго порядка, т.е.

$$M(U_n) = \mu, D(U_n) = \Sigma,$$

где μ — вектор математических ожиданий координат случайного вектора, Σ — его ковариационная матрица. Введем последовательность средних арифметических случайных векторов:

$$\overline{U}_n = (\overline{U}_{1n},...,\overline{U}_{kn}), \quad n = 1,2,..., \quad \overline{U}_{in} = \frac{1}{n}\sum_{j=1}^{n} U_{ij}.$$

Тогда случайный вектор $\sqrt{n}(\overline{U}_n - \mu)$ имеет асимптотическое k-мерное нормальное распределение $N_k(0,\Sigma)$, т.е. он асимптотически распределен так же, как k-мерная нормальная величина с нулевым математическим ожиданием, ковариационной матрицей Σ и плотностью:

$$N_k(u \mid 0,\Sigma) = (2\pi)^{-k/2} \mid \Sigma \mid^{-1/2} \exp\{-\frac{1}{2}u'\Sigma^{-1}u\}.$$

Здесь $|\Sigma|$ — определитель матрицы Σ. Другими словами, распределение случайного вектора $\sqrt{n}(\overline{U}_n - \mu)$ сходится к k-мерному нормальному распределению с нулевым математическим ожиданием и ковариационной матрицей Σ.

Напомним, что многомерным нормальным распределением с математическим ожиданием μ и ковариационной матрицей Σ называется распределение, имеющее плотность:

$$N_k(u \mid \mu,\Sigma) = (2\pi)^{-k/2} \mid \Sigma \mid^{-1/2} \exp\{-\frac{1}{2}[(u-\mu)^T\Sigma^{-1}(u-\mu)]\}.$$

Многомерная центральная предельная теорема показывает, что распределения сумм независимых одинаково распределенных случайных векторов при большом числе слагаемых хорошо приближаются с помощью нормальных распределений, имеющих такие же первые два момента (вектор математических ожиданий координат случайного вектора и его корреляционную матрицу), как и исходные вектора. От одинаковой распределенности можно отказаться, но это потребует некоторого усложнения символики. В целом из теоремы о многомерной сходимости вытекает, что многомерный случай ничем принципиально не отличается от одномерного.

Пример. Пусть $X_1, \ldots X_n, \ldots$ — независимые одинаково распределенные случайные величины. Рассмотрим k-мерные независимые одинаково распределенные случайные вектора:

$$U_n^{'} = (X_n, X_n^2, X_n^3, ..., X_n^k), \quad n = 1,2,...$$

Их математическое ожидание — вектор теоретических начальных моментов, а ковариационная матрица составлена из соответствующих центральных моментов. Тогда $\overline{U}_n$ — вектор выборочных центральных моментов. Многомерная центральная предельная теорема утверждает, что $\overline{U}_n$ имеет асимптотически нормальное распределение. Как вытекает из теорем о наследовании сходимости и о линеаризации (см. ниже), из распределения $\overline{U}_n$ можно вывести распределения различных функций от выборочных начальных моментов. А поскольку центральные моменты выражаются через начальные моменты, то аналогичное утверждение верно и для них.

3. Теоремы о наследовании сходимости

Суть проблемы наследования сходимости. Пусть распределения случайных величин X_n при $n \to \infty$ стремятся к распределению случайной величины X. При каких функциях f можно утверждать, что распределения случайных величин $f(X_n)$ сходятся к распределению $f(X)$, т.е. наследуется сходимость?

Хорошо известно, что для непрерывных функций f сходимость наследуется [3]. Однако в прикладной статистике и, в частности, в нечисловой статистике используются различные обобщения этого утверждения. Необходимость обобщений связана с тремя обстоятельствами.

1. Статистические данные могут моделироваться не только случайными величинами, но и случайными векторами, случайными множествами, случайными элементами произвольной природы (т.е. функциями на вероятностном пространстве со значениями в произвольном множестве).

2. Переход к пределу должен рассматриваться не только для случая безграничного возрастания объема выборки, но и в более общих случаях. Например, если в постановке статистической задачи участвуют несколько выборок объемов $n(1)$, $n(2)$, … , $n(k)$, то вполне обычным является предположение о безграничном росте всех этих объемов (что можно описать и как $min\ \{n(1), n(2), ..., n(k)\} \to \infty$).

3. Функция f не обязательно является непрерывной. Она может иметь разрывы. Кроме того, она может зависеть от параметров, по которым происходит переход к пределу. Например, может зависеть от объемов выборок. Если в постановке статистической задачи участвуют несколько выборок объемов $n(1)$, $n(2)$, ..., $n(k)$, то, как правило, необходимо рассматривать функции вида $f = f(n(1), n(2), ..., n(k))$.

Расстояние Прохорова и сходимость по направленному множеству. Введем необходимые для дальнейшего изложения понятия.

Для определения расстояния (метрики) Прохорова нужны предварительные определения. Пусть C — некоторое пространство, A — его подмножество, d — метрика в C. Введем понятие ε-окрестности множества A в метрике d:

$$S(A, \varepsilon) = \{x \in C : d(A, x) < \varepsilon\}.$$

Таким образом, ε-окрестность множества A — это совокупность всех точек пространства C, отстоящих от A не более чем на положительное число ε. При этом расстояние от точки x до множества A — это точная нижняя грань расстояний от x до точек множества A, т.е.

$$d(A, x) = \inf\{d(x, y) : y \in A\}.$$

Пусть P_1 и P_2 — две вероятностные меры на C (т.е. распределения двух случайных элементов со значениями в C). Пусть D_{12} — множество чисел $\varepsilon > 0$ таких, что

$$P_1(A) \le P_2(S(A, \varepsilon) + \varepsilon$$

для любого замкнутого подмножества A пространства C. Пусть D_{21} — множество чисел $\varepsilon > 0$ таких, что

$$P_2(A) \leq P_1(S(A, \varepsilon) + \varepsilon$$

для любого замкнутого подмножества A пространства C.

Расстояние Прохорова $L(P_1,P_2)$ между вероятностными мерами (его можно рассматривать и как расстояние между случайными элементами с распределениями P_1 и P_2 соответственно) вводится формулой:

$$L(P_1, P_2) = \max (\inf D_{12}, \inf D_{21}).$$

С помощью метрики Прохорова формализуется понятие сходимости распределений случайных элементов в произвольном пространстве.

Расстояние $L(P_1, P_2)$ введено академиком РАН Юрием Васильевичем Прохоровым в середине XX в. и широко используется в современной теории вероятностей.

Сходимость по направленному множеству [4, с. 95–96]. Бинарное отношение $\geq$ (упорядочение), заданное на множестве B, называется направлением на нем, если B не пусто и

(а) если m, n и p — такие элементы множества B, что $m \geq n$ и $n \geq p$, то $m \geq p$;

(б) $m \geq m$ для любого m из B;

(в) если m и n принадлежат B, то найдется элемент p из B такой, что $p \geq m$ и $p \geq n$.

Направленное множество — это пара $(B, \geq)$, где $\geq$ — направление на множестве B. Направленностью (или «последовательностью по направленному множеству») называется пара $(f, \geq)$, где f — функция, $\geq$ — направление на ее области определения. Пусть f: $B \to Y$, где Y — топологическое пространство. Направленность $(f, \geq)$ сходится в топологическом пространстве Y к точке y_0, если для любой окрестности U точки y_0 найдется p из B такое, что $f(q) \in U$ при любом $q \geq p$. В таком случае говорят также о сходимости по направленному множеству.

Пусть $B = \{(n(1), n(2), ..., n(k))\}$ — совокупность векторов, каждый из которых составлен из объемов k выборок. Пусть

$$(n(1), n(2), \dots , n(k)) \geq (n_1(1), n_1(2), \dots , n_1(k))$$

тогда и только тогда, когда $n(i) \geq n_1(i)$ при всех $i = 1, 2, \ldots, k$. Тогда $(B, \geq)$ — направленное множество, сходимость по которому эквивалентна сходимости при $\min \{n(1), n(2), ..., n(k)\} \to \infty$.

Чтобы охватить различные частные случаи, целесообразно предельные теоремы формулировать в терминах сходимости по направленному множеству. Будем писать $B = \{\alpha\}$. Пусть запись $\alpha \to \infty$ обозначает переход к пределу по направленному множеству.

Формулировка проблемы наследования сходимости. Пусть случайные элементы X_α со значениями в пространстве C сходятся при $\alpha \to \infty$ к случайному элементу X, где через $\alpha \to \infty$ обозначен переход к пределу по направленному множеству. Сходимость случайных элементов означает, что $L(X_\alpha, X) \to 0$ при $\alpha \to \infty$, где L — метрика Прохорова в пространстве C.

Пусть f_α: $C \to Y$ — некоторые функции. Какие условия надо на них наложить, чтобы из $L(X_\alpha, X) \to 0$ вытекало, что $L_1(f_\alpha(X_\alpha), f_\alpha(X)) \to 0$ при $\alpha \to \infty$, где L_1 — метрика Прохорова в пространстве Y? Другими словами, какие условия на функции f_α: $C \to Y$ гарантируют наследование сходимости?

В работах [5, 6] найдены необходимые и достаточные условия на функции f_α: $C \to Y$, гарантирующие наследование сходимости. Описанию этих условий посвящена оставшаяся часть настоящего раздела П-3.

Приведем для полноты изложения строгие формулировки математических предположений.

Математические предположения. Пусть C и Y — полные сепарабельные метрические пространства, Пусть выполнены обычные предположения измеримости: X_α и X — случайные элементы C, $f_\alpha(X_\alpha)$ и $f_\alpha(X)$ — случайные элементы в Y, рассматриваемые ниже подмножества пространств C и Y лежат в соответствующих σ–алгебрах измеримых подмножеств, и т.д.

Понадобятся некоторые *определения*. Разбиение $T_n = \{C_{1n}, C_{2n}, ..., C_{nn}\}$ пространства C — это такой набор подмножеств C_j, $j = 1, 2, ..., n$, этого пространства, что пересечение любых двух из них пусто, а объединение совпадает с C. Диаметром $diam(A)$ подмножества A множества C называется точная верхняя грань расстояний между элементами A, т.е.

$$diam(A) = \sup \{d(x, y), x \in A, y \in A\},$$

где $d(x, y)$ — метрика в пространстве C. Обозначим ∂A границу множества A, т.е. совокупность точек x таких, что любая их окрестность $U(x)$ имеет непу-

стое пересечение как с A, так и с $C\backslash A$. Колебанием $\delta(f, B)$ функции f на множестве B называется $\delta(f, B) = \sup \{|f(x) - f(y)|, x \in B, y \in B\}$.

Достаточное условие для наследования сходимости. Пусть $L(X_\alpha, X) \to 0$ при $\alpha \to \infty$. Пусть существует последовательность T_n разбиений пространства C такая, что $P(X \in \partial A) = 0$ для любого A из T_n и, основное условие, для любого $\varepsilon > 0$:

$$m_\varepsilon(\alpha, n) = \sum P(X \in A) \to 0 \qquad (1)$$

при $n \to \infty$ и $\alpha \to \infty$, где сумма берется по всем тем A из T_n, для которых колебание функции f_α на A больше ε, т.е. $\delta(f_\alpha, A) > \varepsilon$. Тогда $L_1(f_\alpha(X_\alpha), f_\alpha(X)) \to 0$ при $\alpha \to \infty$.

Необходимое условие для наследования сходимости. Пусть Y — конечномерное линейное пространство, $Y = R^k$. Пусть случайные элементы $f_\alpha(X)$ асимптотически ограничены по вероятности при $\alpha \to \infty$, т.е. для любого $\varepsilon > 0$ существуют число $S(\varepsilon)$ и элемент направленного множества $\alpha(\varepsilon)$ такие, что $P(\|f_\alpha(X)\| > S(\varepsilon)) < \varepsilon$ при $\alpha \geq \alpha(\varepsilon)$, где $\|f_\alpha(X)\|$ — норма (длина) вектора $f_\alpha(X)$. Пусть существует последовательность T_n разбиений пространства C такая, что

$$\lim_{n \to \infty} \max\{ diam(C_{jn}),\ C_{jn} \in T_n \} = 0 ,$$

т.е. последовательность T_n является безгранично измельчающейся. Самое существенное — пусть условие (1) не выполнено для последовательности T_n. Тогда существует последовательность случайных элементов X_α такая, что $L(X_\alpha, X) \to 0$ при $\alpha \to \infty$, но $L_1(f_\alpha(X_\alpha), f_\alpha(X))$ не сходится к 0 при $\alpha \to \infty$.

Несколько огрубляя, можно сказать, что *условие (1) является необходимым и достаточным для наследования сходимости*.

Пример 1. Пусть C и Y — конечномерные линейные пространства, функции f_α не зависят от α, т.е. $f_\alpha \equiv f$, причем функция f ограничена. Тогда условие (1) эквивалентно требованию интегрируемости по Риману — Стилтьесу функции f по мере $G(A) = P(X \in A)$. В частности, условие (1) выполнено для непрерывной функции f.

В конечномерных пространствах C вместо сходимости $L(X_\alpha, X) \to 0$ при $\alpha \to \infty$ можно говорить о слабой сходимости функций распределения случайных векторов X_α к функции распределения случайного вектора X. Речь идет о «сходимости по распределению», т.е. о сходимости во всех точках непре-

рывности функции распределения случайного вектора X. В этом случае разбиения могут состоять из многомерных параллелепипедов [5, гл. 2].

Пример 2. Полученные выше результаты дают обоснование для рассуждений типа следующего. Пусть по двум независимым выборкам объемов m и n соответственно построены статистики X_m и Y_n. Пусть известно, что распределения этих статистик сходятся при безграничном росте объемов выборок к стандартному нормальному распределению с математическим ожиданием 0 и дисперсией 1. Пусть $a(m, n)$ и $b(m, n)$ — некоторые коэффициенты. Тогда согласно результатам примера 1 распределение случайной величины $Z(m, n) = a(m, n)X_m + b(m, n)Y_n$ сближается с распределением нормально распределенной случайной величины с математическим ожиданием 0 и дисперсией $a^2(m, n) + b^2(m, n)$. Если же $a^2(m, n) + b^2(m, n) = 1$, например,

$$a(m,n)=\sqrt{\frac{m}{m+n}}. \quad b(m,n)=\sqrt{\frac{n}{m+n}},$$

то распределение $Z(m, n)$ сходится при безграничном росте объемов выборок к стандартному нормальному распределению с математическим ожиданием 0 и дисперсией 1.

4. Метод линеаризации

При разработке методов прикладной статистики и, в частности, нечисловой статистики часто возникает следующая задача [3, с. 338]. Имеется последовательность k-мерных случайных векторов $X_n = (X_{1n}, X_{2n}, ..., X_{kn})$, $n = 1, 2, ...$, такая, что $X_n \to a = (a_1, a_2, ..., a_k)$ при $n \to \infty$, и последовательность функций f_n: $R^k \to R^1$. Требуется найти распределение случайной величины $f_n(X_n)$.

Основная идея — рассмотреть главный линейный член функции f_n в окрестности точки a. Из математического анализа известно, что

$$f_n(X_n)-f_n(a)=\sum_{j=1}^{k}\frac{\partial f_n(a)}{\partial x_j}(X_{jn}-a_j)+O_n(\|X_n-a\|^2),$$

где остаточный член является бесконечно малой величиной более высокого порядка малости, чем линейный член. Таким образом, произвольная функция может быть заменена на линейную функцию от координат случайного векто-

ра. Эта замена проводится с точностью до бесконечно малых более высокого порядка. Конечно, должны быть выполнены некоторые математические условия регулярности. Например, функции f_n должны быть дважды непрерывно дифференцируемы в окрестности точки a.

Если вектор X_n является асимптотически нормальным с математическим ожиданием a и ковариационной матрицей $\sum/n$, где $\sum = ||\sigma_{ij}||$, причем $\sigma_{ij} = nM(X_i - a_i)(X_j - a_j)$, то линейная функция от его координат также асимптотически нормальна. Следовательно, при очевидных условиях регулярности $f_n(X_n)$ — асимптотически нормальная случайная величина с математическим ожиданием $f_n(a)$ и дисперсией:

$$\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{k}\frac{\partial f_n(a)}{\partial x_i}\frac{\partial f_n(a)}{\partial x_j}\sigma_{ij}.$$

Для практического использования асимптотической нормальности $f_n(X_n)$ остается заменить неизвестные моменты a и $\sum$ на их оценки. Например, если X_n — это среднее арифметическое независимых одинаково распределенных случайных векторов, то a можно заменить на X_n, а $\sum$ — на выборочную ковариационную матрицу.

Пример. Пусть $Y_1, Y_2, ..., Y_n$ — независимые одинаково распределенные случайные величины с математическим ожиданием a и дисперсией σ^2. В качестве X_n ($k = 1$) рассмотрим выборочное среднее арифметическое:

$$\overline{Y} = \frac{Y_1 + Y_2 + ... + Y_n}{n}.$$

Как известно, в силу закона больших чисел $\overline{Y} \to a = M(Y)$. Следовательно, для получения распределений функций от выборочного среднего арифметического можно использовать метод линеаризации. В качестве примера рассмотрим $f_n(y) = f(y) = y^2$. Тогда

$$\left(\overline{Y}\right)^2 - a^2 = \frac{df(a)}{dy}\left(\overline{Y} - a\right) + O\left(\left(\overline{Y} - a\right)^2\right) = 2a\left(\overline{Y} - a\right) + O\left(\left(\overline{Y} - a\right)^2\right).$$

Из этого соотношения следует, что с точностью до бесконечно малых более высокого порядка:

$$\left(\overline{Y}\right)^2 = a^2 + 2a\left(\overline{Y} - a\right).$$

Поскольку в соответствии с Центральной Предельной Теоремой выборочное среднее арифметическое является асимптотически нормальной случайной величиной с математическим ожиданием a и дисперсией σ^2/n, то квадрат этой статистики является асимптотически нормальной случайной величиной с математическим ожиданием a^2 и дисперсией $4a^2\sigma^2/n$. Для практического использования может оказаться полезной замена параметров (асимптотического нормального распределения) на их оценки, а именно, математического ожидания — на $\left(\overline{Y}\right)^2$, а дисперсии — на $4\left(\overline{Y}\right)^2 s^2/n$, где s^2 — выборочная дисперсия.

Большое внимание (целая глава!) уделено методу линеаризации в классическом учебнике Е. С. Вентцель [7].

5. Принцип инвариантности

Пусть $Y_1, Y_2, \ldots, Y_n$ — независимые одинаково распределенные случайные величины с непрерывной функцией распределения $F(x)$. Многие используемые в прикладной статистике функции от результатов наблюдений выражаются через эмпирическую функцию распределения $F_n(x)$, при каждом x равную доле наблюдений, не превосходящих x. К ним относятся статистики Колмогорова, Смирнова, омега-квадрат, обсуждаемые в главе 2. Отметим, что и другие статистики выражаются через эмпирическую функцию распределения, например:

$$\overline{Y} = \int_{-\infty}^{+\infty} x dF_n(x).$$

Полезным является преобразование Н. В. Смирнова $t = F(x)$. Тогда независимые случайные величины $Z_j = F(Y_j)$, $j = 1, 2, .., n$, имеют равномерное распределение на отрезке [0; 1]. Рассмотрим построенную по ним эмпирическую функцию распределения $F_n(t)$, $0 \le t \le 1$. *Эмпирическим процессом* называется случайный процесс:

$$\xi_n(t) = \sqrt{n}(F_n(t) - t).$$

Рассмотрим критерии проверки согласия функции распределения выборки с фиксированной функцией распределения $F(x)$. Статистика критерия Колмогорова записывается в виде:

$$K_n = \sup_{0 \le t \le 1} |\xi_n(t)|,$$

статистика критерия Смирнова — это

$$S_n = \sup_{0 \le t \le 1} \xi_n(t),$$

а статистика критерия омега-квадрат (известного также как критерий Крамера — Мизеса — Смирнова) имеет вид:

$$\omega_n^2 = \int_0^1 \xi_n^2(t)dt\,.$$

Случайный процесс $\xi_n(t)$ имеет нулевое математическое ожидание и ковариационную функцию $M\xi_n(s)\xi_n(t) = \min(s, t) - st$. Рассмотрим гауссовский случайный процесс $\xi(t)$ с такими же математическим ожиданием и ковариационной функцией. Он называется броуновским мостом. (Напомним, что гауссовским процесс именуется потому, что вектор $(\xi(t_1), \xi(t_2), \ldots, \xi(t_k))$ имеет многомерное нормальное распределение при любых наборах моментов времени $t_1, t_2, \ldots, t_k$.)

Пусть f — функционал, определенный на множестве возможных траекторий случайных процессов. *Принцип инвариантности* [1] состоит в том, что последовательность распределений случайных величин $f(\xi_n)$ сходится при $n \to \infty$ к распределению случайной величины $f(\xi)$. Сходимость по распределению обозначим символом =>. Тогда принцип инвариантности кратко записывается так: $f(\xi_n) => f(\xi)$. В частности, согласно принципу инвариантности статистика Колмогорова и статистика омега квадрат сходятся по распределению к распределениям соответствующих функционалов от случайного процесса ξ:

$$K_n = \sup_{0 \le t \le 1} |\xi_n(t)| => \sup_{0 \le t \le 1} |\xi(t)|,\ \omega_n^2 = \int_0^1 \xi_n^2(t)dt => \int_0^1 \xi^2(t)dt\,.$$

Таким образом, от проблем прикладной статистики сделан переход к теории случайных процессов. Методами этой теории найдены распределения случайных величин:

$$\sup_{0 \le t \le 1} |\xi(t)|,\ \int_0^1 \xi^2(t)dt\,,$$

т.е. предельные распределения статистик Колмогорова и омега-квадрат, а также и многих иных. Следовательно, принцип инвариантности — инструмент получения предельных распределений функций от результатов наблюдений, используемых в прикладной статистике.

Обоснование принципу инвариантности может быть дано на основе теории сходимости вероятностных мер в функциональных пространствах [8]. Более простой подход, позволяющий к тому же получать необходимые и достаточные условия в предельной теории статистик интегрального типа (принцип инвариантности к ним нельзя применить), рассмотрен в главе 2.

Почему «принцип инвариантности» так назван? Обратим внимание, что предельные распределения рассматриваемых статистик не зависят от их функции распределения $F(x)$. Другими словами, предельное распределение инвариантно относительно выбора $F(x)$.

В более широком смысле термин «принцип инвариантности» применяют тогда, когда предельное распределение не зависит от тех или иных характеристик исходных распределений [1]. В этом смысле наиболее известный «принцип инвариантности» — это Центральная Предельная Теорема, поскольку предельное стандартное нормальное распределение — одно и то же для всех возможных распределений независимых одинаково распределенных слагаемых (лишь бы слагаемые имели конечные математическое ожидание и дисперсию).

Литература

1. Вероятность и математическая статистика : энциклопедия / главный редактор Ю. В. Прохоров. — Москва : Большая Российская энциклопедия, 1999. — 910 с.

2. *Гнеденко, Б. В.* Курс теории вероятностей : учебник / Б. В. Гнеденко. — 7-е изд., испр. — Москва : Эдиториал УРСС, 2001. — 320 с.

3. *Рао, С. Р.* Линейные статистические методы и их применения / С. Р. Рао. — Москва : Наука, 1968. 548 с.

4. *Келли, Дж.* Общая топология / Дж. Келли. — Москва : Наука, 1968. — 384 с.

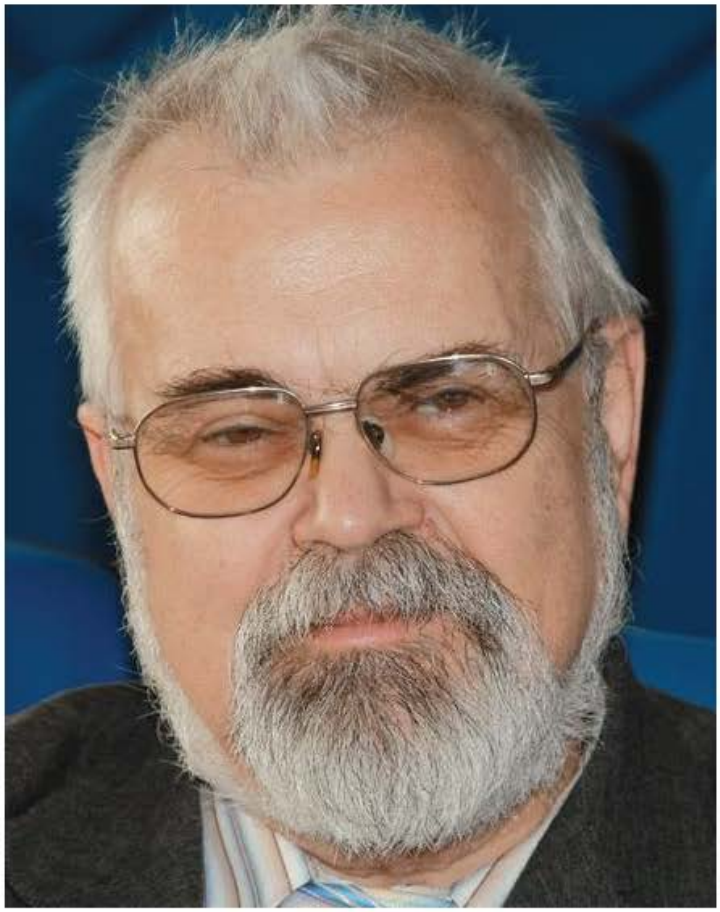
5. *Орлов, А. И.* Устойчивость в социально-экономических моделях / А. И. Орлов. — Москва : Наука, 1979. — 296 с.

6. *Орлов, А. И.* Асимптотическое поведение статистик интегрального типа / А. И. Орлов // Вероятностные процессы и их приложения : межвузовский сборник. — Москва : МИЭМ, 1989. — С. 118–123.

7. *Вентцель, Е. С.* Теория вероятностей / Е. С. Вентцель. — Москва : Наука, 1964. — 576 с.

8. *Биллингсли, П.* Сходимость вероятностных мер / П. Биллингсли. — Москва : Наука, 1977. — 352 с.

ОБ АВТОРЕ



Орлов Александр Иванович (1949 г.р.) — профессор (1995 г. — по кафедре математической экономики), доктор экономических наук (2009 г. — по математическим и инструментальным методам экономики), доктор технических наук (1992 г. — по применению математических методов), кандидат физико-математических наук (1976 г. — по теории вероятностей и математической статистике).

Профессор кафедр «Экономика и организация производства» (факультет «Инженерный бизнес и менеджмент») и «Вычислительная математика и математическая физика» (факультет «Фундаментальные науки») Московского государственного технического университета им. Н. Э. Баумана, руководитель секции «Организационно-экономическое моделирование, эконометрика и статистика», директор Института высоких статистических технологий и эконометрики, заведующий Лабораторией экономико-математических методов в контроллинге.

Член редколлегий журналов «Заводская лаборатория. Диагностика материалов», «Контроллинг», «Инновации в менеджменте», «Социология: методология, методы, математическое моделирование», «Управление большими системами: сборник трудов». Главный редактор электронного еженедельника «Эконометрика».

Академик Международной академии исследований будущего, Российской Академии статистических методов. Вице-президент Всесоюзной Статистической Ассоциации, президент Российской ассоциации статистических методов.

Основные направления научной и педагогической деятельности: теория принятия решений, прикладная статистика и другие статистические методы, эконометрика, экономико-математические методы, экспертные оценки, менеджмент, экономика предприятия, макроэкономика, экология.

Автор более 1 100 научных и методических публикаций в России и за рубежом, в том числе более 50 книг. Один из наиболее цитируемых математиков и экономистов России.

Более подробная информация приведена на сайте «Википедия», в статье «Орлов, Александр Иванович (ученый)».

Основные публикации профессора А. И. Орлова

1. *Орлов, А. И.* Устойчивость в социально-экономических моделях / А. И. Орлов. — Москва : Наука, 1979. — 296 с.

2. *Орлов, А. И.* Задачи оптимизации и нечеткие переменные / А. И. Орлов. — Москва : Знание, 1980. — 64 с.

3. Анализ нечисловой информации (препринт) / А. И. Орлов, Ю. Н. Тюрин, Б. Г. Литвак [и др.]. — Москва : Научный Совет АН СССР по комплексной проблеме «Кибернетика», 1981. — 80 с.

4. *Гусев, В. А.* Внеклассная работа по математике в 6–8 классах / В. А. Гусев, А. И. Орлов, А. Л. Розенталь. — Москва : Просвещение, 1977. — 288 с.

5. *Гусев, В. А.* Внеклассная работа по математике в 6–8 классах / В. А. Гусев, А. И. Орлов, А. Л. Розенталь. — 2-е изд., перераб. — Москва : Просвещение, 1984. — 286 с.

6. *Орлов, А. И.* Пакет программ анализа данных «ППАНД» : учебное пособие / А. И. Орлов, И. Л. Легостаевая, О. М. Черномордик. — Москва : Сотрудничающий центр Всемирной организации здравоохранения по профессиональной гигиене, 1990. — 93 с.

7. *Орлов, А. И.* Математическое моделирование процессов налогообложения (подходы к проблеме) / А. И. Орлов, В. Г. Кольцов, Н. Ю. Иванова. — Москва : Изд-во ЦЭО Министерства общего и профессионального образования РФ, 1997. — 232 с.

8. *Орлов, А. И.* Экология : учебное пособие / А. И. Орлов, С. А. Боголюбов. — Москва : Знание, 1999. — 288 с.

9. *Орлов, А. И.* Менеджмент : учебное пособие / А. И. Орлов, С. А. Боголюбов, Ж. В. Прокофьева. — Москва : Знание, 2000. — 288 с.

10. *Орлов, А. И.* Управление качеством окружающей среды : учебник / А. И. Орлов, С. А. Боголюбов. — Т. 1. — Москва : Изд-во МГИЭМ(ту), 2000. — 283 с.

11. *Орлов, А. И.* Системы экологического управления : учебник / А. И. Орлов, С. А. Боголюбов. — Москва : Европейский центр по качеству, 2002. — 224 с.

12. *Орлов, А. И.* Эконометрика : учебник / А. И. Орлов. — Москва : Экзамен, 2002. — 576 с.

13. *Орлов, А. И.* Эконометрика : учебник / А. И. Орлов. — 2-е изд., перераб. и доп. — Москва : Экзамен, 2003. — 575 с.

14. *Орлов, А. И.* Эконометрика : учебник / А. И. Орлов. — 3-е изд. — Москва : Экзамен, 2004. — 573 с.

15. Управление промышленной и экологической безопасностью : учебное пособие / А. И. Орлов, В. Н. Федосеев, В. Г. Ларионов, А. Ф. Козьяков. — Москва : Изд-во УРАО, 2002. — 220 с.

16. Управление промышленной и экологической безопасностью : учебное пособие / А. И. Орлов, В. Н. Федосеев, В. Г. Ларионов, А. Ф. Козьяков. — 2-е изд. — Москва : Изд-во УРАО, 2003. — 220 с.

17. *Орлов, А. И.* Менеджмент в техносфере : учебное пособие / А. И. Орлов, В. Н. Федосеев. — Москва : Академия, 2003. — 384 с.

18. *Орлов, А. И.* Теория и методы разработки управленческих решений : учебное пособие / А. И. Орлов. — Москва : МарТ ; Ростов-на-Дону : МарТ, 2005. — 496 с.

19. *Орлов, А. И.* Прикладная статистика : учебник / А. И. Орлов. — Москва : Экзамен, 2006. — 672 с.

20. *Орлов, А. И.* Теория принятия решений : учебник / А. И. Орлов. — Москва : Экзамен, 2006. — 576 с.

21. Проектирование интегрированных производственно-корпоративных структур: эффективность, организация, управление / А. И. Орлов, С. Н. Аниси-

мов, А. А. Колобов [и др.] ; под редакцией А. А. Колобова, А. И. Орлова. — Москва : Изд-во МГТУ им. Н. Э. Баумана, 2006. — 728 с.

22. *Колобов, А. А.* Менеджмент высоких технологий. Интегрированные производственно-корпоративные структуры: организация, экономика, управление, проектирование, эффективность, устойчивость / А. А. Колобов, И. Н. Омельченко, А. И. Орлов. — Москва : Экзамен, 2008. — 621 с.

23. *Орлов, А. И.* Организационно-экономическое моделирование : учебник : в 3 ч. Ч. 1. Нечисловая статистика / А. И. Орлов. — Москва : Изд-во МГТУ им. Н.Э. Баумана, 2009. — 542 с.

24. *Орлов, А. И.* Эконометрика : учебник для вузов / А. И. Орлов. — 4-е изд., доп. и перераб. — Ростов-на-Дону : Феникс, 2009. — 572 с.

25. *Орлов, А. И.* Менеджмент: организационно-экономическое моделирование : учебное пособие для вузов / А. И. Орлов. — Ростов-на-Дону : Феникс, 2009. — 475 с.

26. *Орлов, А. И.* Вероятность и прикладная статистика: основные факты : справочник / А. И. Орлов. — Москва : КноРус, 2010. — 192 с.

27. *Орлов, А. И.* Организационно-экономическое моделирование: теория принятия решений : учебник / А. И. Орлов. — Москва : КноРус, 2011. — 568 с.

28. *Орлов, А. И.* Организационно-экономическое моделирование : учебник : в 3 ч. Ч. 2. Экспертные оценки / А. И. Орлов. — Москва : Изд-во МГТУ им. Н. Э. Баумана, 2011. — 486 с.

29. *Орлов, А. И.* Устойчивые экономико-математические методы и модели. Разработка и развитие устойчивых экономико-математических методов и моделей для модернизации управления предприятиями / А. И. Орлов. — Саарбрюккен : Lambert Academic Publishing, 2011. — 436 с.

30. *Орлов, А. И.* Организационно-экономическое моделирование : учебник : в 3 ч. Ч. 3. Статистические методы анализа данных / А. И. Орлов. — Москва : Изд-во МГТУ им. Н.Э. Баумана, 2012. — 624 с.

31. *Орлов, А. И.* Проблемы управления экологической безопасностью. Итоги двадцати лет научных исследований и преподавания / А. И. Орлов. — Саарбрюккен : Palmarium Academic Publishing, 2012. — 344 с.

32. *Орлов, А. И.* Системная нечеткая интервальная математика : монография / А. И. Орлов, Е. В. Луценко. — Краснодар : Изд-во КубГАУ, 2014. — 600 с.

33. *Орлов, А. И.* Перспективные математические и инструментальные методы контроллинга : монография / А. И. Орлов, Е. В. Луценко, В. И. Лойко ;

под научной редакцией профессора С. Г. Фалько. — Краснодар : Изд-во КубГАУ, 2015. — 600 с.

34. *Орлов, А. И.* Организационно-экономическое, математическое и программное обеспечение контроллинга, инноваций и менеджмента : монография / А. И. Орлов, Е. В. Луценко, В. И. Лойко ; под общей редакцией С. Г. Фалько. — Краснодар : Изд-во КубГАУ, 2016. — 600 с.

35. *Лойко, В. И.* Современные подходы в наукометрии : монография / В. И. Лойко, Е. В. Луценко, А. И. Орлов ; под научной редакцией профессора С. Г. Фалько. — Краснодар : Изд-во КубГАУ, 2017. — 532 с.

36. *Орлов, А. И.* Методы принятия управленческих решений : учебник / А. И. Орлов. — Москва : КНОРУС, 2018. — 286 с.

37. *Лойко, В. И.* Современная цифровая экономика / В. И. Лойко, Е. В. Луценко, А. И. Орлов. — Краснодар : Изд-во КубГАУ, 2018. — 508 с.

38. *Лойко, В. И.* Высокие статистические технологии и системно-когнитивное моделирование в экологии : монография / В. И. Лойко, Е. В. Луценко, А. И. Орлов. — Краснодар : Изд-во КубГАУ, 2019. — 258 с.

39. *Агаларов, З. С.* Эконометрика : учебник / З. С. Агаларов, А. И. Орлов. — Москва : Дашков и К°, 2021. — 380 с.