

Крянев А.В.
Лукин Г.В.

Математические
методы обработки
неопределенных
данных



**А. В. Крянев
Г. В. Лукин
Д. К. Удумян**

МЕТРИЧЕСКИЙ АНАЛИЗ И ОБРАБОТКА ДАННЫХ

*Рекомендовано УМО
в качестве учебного пособия
по направлению «Ядерные физика и технологии»*



МОСКВА
ФИЗМАТЛИТ®
2012

УДК 519.2+6
ББК 22.17, 22.19
К 85

Крянев А.В., Лукин Г.В., Удумян Д.К. **Метрический анализ и обработка данных.** — М.: ФИЗМАТЛИТ, 2012. — 308 с. — ISBN 978-5-9221-1068-6.

Основная цель книги — ознакомить читателя с наиболее эффективными и апробированными классическими и новыми стохастическими и детерминированными методами оценки и прогнозирования, научить использовать эти методы при решении конкретных задач обработки данных.

Изложены основные понятия параметрической и непараметрической статистики, а также некоторые новые методы робастного оценивания, учета априорной информации и прогнозирования, включая алгоритмы их численной реализации. Представлены основы нового направления обработки данных — метрического анализа, позволяющего решать задачи интерполяции, восстановления и прогнозирования функций одной и многих переменных на основе эффективного использования информации стохастического и детерминированного характеров об исследуемой функциональной зависимости.

Предполагается, что читатель предварительно освоил курс теории вероятностей и математической статистики на базе, например, книги В.С. Пугачева «Теория вероятностей и математическая статистика».

Предназначена преподавателям, научным работникам, аспирантам, студентам старших курсов различных специальностей, использующих математические методы обработки данных.

Учебное издание

КРЯНЕВ Александр Витальевич

ЛУКИН Глеб Владимирович

УДУМЯН Давид Каджикович

МЕТРИЧЕСКИЙ АНАЛИЗ И ОБРАБОТКА ДАННЫХ

Редактор *Н.Б. Бартошевич-Жагель*

Оригинал-макет: *Ю.В. Владимирова*

Оформление переплета: *Н.В. Гришина*

Подписано в печать 12.04.10. Формат 60×90/16. Бумага офсетная. Печать офсетная.

Усл. печ. л. 19,25. Уч.-изд. л. 19,25. Тираж 500 экз. Заказ №

Издательская фирма «Физико-математическая литература»

МАИК «Наука/Интерпериодика»

117997, Москва, ул. Профсоюзная, 90

E-mail: fizmat@maik.ru, fmlsale@maik.ru;

http://www.fml.ru

Отпечатано с электронных носителей издательства

в ГУП «ИПК Чувашия», 428019

г. Чебоксары, пр-т И.Яковлева, 13

ISBN 978-5-9221-1068-6



9 785922 110686

© ФИЗМАТЛИТ, 2012

© А. В. Крянев, Г. В. Лукин, Д. К. Удумян,
2012

ISBN 978-5-9221-1068-6

ОГЛАВЛЕНИЕ

Предисловие	7
Введение	13
Основные обозначения	16
Глава 1. Элементы математической статистики	19
§ 1. Параметрическая статистика. Свойства оценок	20
§ 2. Метод моментов	23
§ 3. Неравенства Фишера–Крамера–Рао	29
§ 4. Метод максимального правдоподобия (ММП)	35
Глава 2. Методы учета априорной информации в рамках параметрической статистики	42
§ 1. Метод Байеса	42
§ 2. Минимаксный метод учета априорной информации	48
§ 3. Обобщенный метод максимального правдоподобия учета априорной детерминированной информации	49
§ 4. Обобщенный метод максимального правдоподобия учета априорной стохастической информации	52
Глава 3. Устойчивые методы оценивания параметра положения	56
§ 1. Минимаксный метод Хьюбера	57
§ 2. Робастные М-оценки	65
Глава 4. Методы непараметрической статистики	70
§ 1. Восстановление функции распределения	70
§ 2. Восстановление плотности распределения методом гистограмм	72
§ 3. Восстановление плотности распределения методом Розенблатта–Парзена	74

§ 4. Восстановление плотности распределения проекционными методами.	77
§ 5. Восстановление плотности распределения регуляризованным методом гистограмм.	80
§ 6. Метод корневой оценки плотности распределения	83
Глава 5. Проверка гипотез о законе распределения.	86
§ 1. Проверка гипотезы о законе распределения в рамках непараметрической статистики.	86
§ 2. Проверка гипотезы о законе распределения в рамках параметрической статистики.	88
Глава 6. Численные методы статистического моделирования.	92
§ 1. Способы моделирования случайных величин	92
§ 2. Применение метода статистического моделирования для решения некоторых прикладных задач.	98
Глава 7. Метод наименьших квадратов для линейных моделей с неопределенными данными.	105
§ 1. Примеры линейных моделей с неопределенными данными	106
§ 2. Классическая схема МНК.	108
§ 3. Обобщения классической схемы МНК.	121
§ 4. Прогнозирование с помощью линейных регрессионных моделей.	127
Глава 8. Робастные методы для линейных моделей с неопределенными данными.	130
§ 1. Робастные М-оценки параметров линейных моделей	130
§ 2. Численные методы нахождения робастных оценок параметров линейных моделей	133
Глава 9. Учет априорной информации в линейных моделях с неопределенными данными.	137
§ 1. Метод Байеса в рамках линейных моделей	139
§ 2. Минимаксный метод учета детерминированной априорной информации	141
§ 3. Обобщенный метод максимального правдоподобия учета априорной стохастической информации в рамках линейных моделей	145

§ 4. Обобщенный метод максимального правдоподобия учета априорной детерминированной информации в рамках линейных моделей	160
§ 5. Регуляризованный метод наименьших квадратов	163
 Глава 10. Метод наименьших квадратов для нелинейных моделей с неопределенными данными	170
§ 1. МНК-оценки параметров нелинейных моделей и их свойства	170
§ 2. Численные методы нахождения МНК-оценок параметров нелинейных моделей	173
 Глава 11. Методы выделения детерминированных и хаотических компонент временных рядов	178
§ 1. Метод сглаживающих ортогональных полиномов	178
§ 2. Метод сглаживающих линейных сплайнов	184
§ 3. Робастные линейные сглаживающие сплайны	189
§ 4. Метод сглаживающих кубических сплайнов	193
§ 5. Метод вейвлетов	205
 Глава 12. Методы прогнозирования хаотических временных рядов	211
§ 1. Методы прогнозирования с использованием априорной информации и ортогональных полиномов	211
§ 2. Методы прогнозирования с использованием априорной информации и линейных сплайнов	214
§ 3. Методы прогнозирования с использованием сингулярно-спектрального анализа	216
 Глава 13. Прогнозирование с помощью метрического анализа	231
§ 1. Матрицы метрической неопределенности и их свойства	233
§ 2. Прогнозирование значения функции многих переменных в условиях наличия стохастических погрешностей	236
 Глава 14. Динамический детерминированный хаос	245
§ 1. Примеры процессов детерминированного хаоса	245
§ 2. Свойства детерминированных хаотических процессов	248

Глава 15. Планирование оптимальных измерений при восстановлении функциональных зависимостей	260
§ 1. Постановка задач планирования оптимальных измерений	261
§ 2. Методы решения задач планирования оптимальных измерений	268
 Список литературы	 275

Предисловие

Представляемая нами книга является расширенным вариантом ранее изданной книги «Математические методы обработки неопределенных данных».

При подготовке настоящей книги были внесены существенные дополнения, включающие некоторые последние достижения, в том числе авторов книги, и прежде всего по новому направлению — метрическому анализу, что и определило название настоящей книги.

Книга в первую очередь предназначена лицам, использующим математические методы обработки данных при решении прикладных задач различного содержания. Кроме того, большинство разделов книги используются студентами НИЯУ МИФИ, выполняющими учебно-исследовательские работы, студентами-дипломниками и аспирантами, тематика исследований которых требует применения различных математических методов обработки данных.

В гл. 1 представлены базовые элементы математической статистики. Анализируются основные свойства оценок и соотношения между ними, в частности эффективность и робастность оценок. Изложен традиционный материал, в том числе интервальное оценивание математического ожидания и дисперсии нормально распределенной случайной величины. Рассматриваются два универсальных метода в рамках параметрической статистики: метод моментов и метод максимального правдоподобия (ММП). Метод моментов изложен в обобщенном виде, позволяющем существенно расширить рамки его применения. С помощью неравенства Фишера–Крамера–Рао анализируется эффективность оценок. Представлены основные свойства ММП-оценок. Изложение теоретических результатов сопровождается примерами, при рассмотрении которых конструируются функции правдоподобия и находятся информация Фишера и информационная матрица Фишера, ММП-оценки и их характеристики. Особое внимание уделяется многомерному нормальному распределению.

Во 2-й главе рассматриваются методы учета дополнительной априорной информации в рамках параметрической статистики. Представлены четыре метода учета априорной информации в зависимости от ее вида. Если априорная информация об искомым

параметрах имеет стохастический характер, то предлагается использовать два метода: метод Байеса и обобщенный метод максимального правдоподобия (ОММП) с заданием априорной выборки. Если же априорная информация об искомым параметрах u задается в виде принадлежности u априорному множеству R_a , то предлагается использовать два метода — минимаксный и ОММП с учетом соотношения $u \in R_a$, оба из которых используют понятия метрического характера, связанные прежде всего с использованием расстояния в пространстве искомого вектора u . Анализируются схемы применения методов учета априорной информации и алгоритмы их численной реализации. Рассматриваются примеры применения методов, в частности случай, когда члены исходной выборки подчиняются нормальному закону.

В 3-й главе представлены робастные методы оценивания параметра положения в условиях наличия больших выбросов. Подчеркивается, что схема робастного оценивания параметра положения может быть без существенных качественных изменений обобщена на многомерный вариант оценивания параметров регрессионной модели. За основу робастного оценивания взят минимаксный подход Хьюбера. Предлагается итерационный метод численного нахождения робастной оценки параметра положения, основанный на геометрическом положении робастной оценки искомого параметра. Рассматривается общая схема получения робастных M -оценок, основанная на использовании функции влияния.

В гл. 4 представлены некоторые часто используемые для решения прикладных задач методы непараметрической статистики восстановления функции и плотности распределения. Подчеркивается, что задача восстановления плотности распределения по выборке, в отличие от аналогичной задачи для функции распределения, принадлежит к классу некорректно поставленных задач и поэтому может быть эффективно решена только при использовании дополнительной априорной информации об искомой плотности. Форма и объем априорной информации могут быть различными, и в зависимости от этого используются различные методы восстановления плотности распределения.

В главе описаны 5 методов восстановления плотности распределения, использующие различные виды и уровни априорной информации. В методах гистограмм, Розенблатта–Парзена и корневой оценки плотности априорная информация используется для определения подходящих значений коэффициентов, аналогичных параметру регуляризации. В проекционных методах априорная информация может быть использована в виде задания априор-

ной реперной плотности, в регуляризованном методе гистограмм априорная информация об искомой плотности задается в виде априорного класса плотностей.

В гл. 5 дается краткое изложение схем проверки гипотез о восстанавливаемом законе распределения в рамках параметрической и непараметрической статистик. Представлены критерии согласия Колмогорова, ω_2 , χ_2 .

Глава 6 посвящена численным методам статистического моделирования. В этой главе представлены способы моделирования случайных величин, в частности датчики равномерно распределенной нормированной случайной величины γ , основанные на методах Лемера и Неймана. Дано применение метода статистического моделирования для вычисления определенных интегралов и решения интегральных уравнений Фредгольма второго рода.

В гл. 7 изложен метод наименьших квадратов (МНК) для линейных моделей с неопределенными данными. Представлена классическая схема МНК и ее обобщения. Приводятся наиболее значимые свойства МНК-оценок и их обобщений. В конце главы дается линейная прогнозная модель, использующая МНК на этапе ее обучения.

В гл. 8 представлены робастные схемы для линейных моделей с неопределенными данными. Все робастные схемы оценивания для линейных моделей строятся на основе функций влияния и М-оценивания. Особое внимание уделяется М-оценке Хьюбера. Предлагаются итерационные численные схемы нахождения нелинейных робастных оценок параметров линейных моделей, в частности итеративный МНК и метод вариационно-взвешенных квадратических приближений. Для нахождения робастной оценки Хьюбера предлагается эффективная итерационная процедура, сходящаяся к робастной оценке за конечное число итераций.

В гл. 9 изложены схемы учета дополнительной априорной информации в линейных моделях с неопределенными данными. Отмечается, что в условиях, когда информационная матрица Фишера исходной линейной модели вырождена или близка к вырожденной, задача оценивания параметров линейной модели принадлежит к классу некорректно поставленных задач и без учета дополнительной априорной информации невозможно получить приемлемые по точности оценки искомых параметров. В главе представлены различные схемы учета априорной информации, основанные на методах Байеса, минимаксном и ОММП. Для ОММП даны две схемы оценивания в зависимости от вида априорной информации. При применении первой схемы ОММП

можно учитывать априорную информацию стохастического характера, задаваемую в виде априорной выборки. В этом случае предполагается, что совместная плотность вероятностей членов выборки зависит от искомого вектора u линейной модели. При применении второй схемы ОММП учитывается априорная детерминированная информация, задаваемая в виде априорного детерминированного множества, которому заведомо принадлежит искомым вектор u . В п. 9.5 рассматривается также регуляризованный метод наименьших квадратов, позволяющий учитывать погрешность в элементах матрицы A исходной линейной модели $Au = f - \varepsilon$.

В гл. 10 в сжатой форме изложен метод наименьших квадратов для нелинейных моделей. Приведен итерационный метод Ньютона–Гаусса численного нахождения МНК-оценок решений нелинейных моделей. Даны регуляризованные модификации Левенберга–Марквардта итерационных процессов Ньютона–Гаусса.

Главы 11, 12 посвящены анализу и прогнозированию временных рядов. В гл. 11 представлены методы выделения детерминированных компонент временных рядов. Все представленные в главе методы основаны на представлении детерминированной компоненты в виде разложения по базисной системе функций и оценке коэффициентов разложения с помощью МНК или робастной схемы. В качестве базисной системы функций берутся: 1) система полиномов, ортогональных на множестве фиксированных значений аргумента; 2) линейные сплайны; 3) кубические сплайны; 4) вейвлеты. Предлагаются эффективные численные схемы расчета искоемых коэффициентов детерминированных компонент. В гл. 12 представлено несколько сравнительно новых методов прогнозирования временных процессов. Первые два из них основаны на учете априорных экспертных оценок прогнозируемых величин и применении схем выделения детерминированных компонент, изложенных в гл. 11. Третий из представленных в гл. 12 методов прогнозирования базируется на сингулярно-спектральном анализе и выделении главных компонент исследуемого временного ряда. Указывается на соответствие между прогнозированием с помощью метода главных компонент и прогнозированием на основе линейной регрессионной модели. Значительная часть гл. 12 посвящена нестационарному сингулярно-спектральному анализу.

В гл. 13 рассматриваются основы нового направления обработки данных — метрического анализа, позволяющего решать задачи интерполяции, экстраполяции и прогнозирования функ-

ций одной и многих переменных. Вводится понятие матрицы метрической неопределенности, с помощью которой определяются интерполяционные и прогнозируемые значения исследуемой функциональной зависимости. В рамках вычислительных схем метрического анализа можно оптимальным образом учитывать неопределенности как метрического, так и стохастического характера, получая эффективные прогнозные значения. Содержание гл. 13 соответствует содержанию работы авторов [51].

Глава 14 посвящена интенсивно развиваемому в последние годы детерминированному хаосу. В главе приведены широко цитируемые в научной литературе, посвященной детерминированному хаосу, примеры процессов с детерминированным хаосом. Приведены наиболее значимые известные к настоящему времени свойства детерминированных хаотических процессов, использующие понятия и инструменты метрического характера. Читателям, желающим более обстоятельно ознакомиться с содержанием теории детерминированного хаоса, можно рекомендовать книгу С.П. Кузнецова [53] и книги [59, 60].

В гл. 15 дано краткое введение в планирование оптимальных измерений при восстановлении функциональных зависимостей. Рассматриваемая задача планирования особенно важна при проведении физических экспериментов, когда экспериментатор имеет возможность выбора значений аргумента, при которых производятся измерения восстанавливаемой функции. Вводятся различные типы оптимальности планов экспериментов и изложены методы решения задач оптимизации планов измерений для рассматриваемых типов оптимальности, включая алгоритмы их численного решения. Ссылки на используемую литературу даны в квадратных скобках, а сам список литературы представлен в алфавитном порядке и включает в себя лишь цитируемые источники. Представленный список не претендует на полноту, а отражает лишь некоторые стороны изложенного в книге материала. Для читателя, желающего углубить свои знания по теории вероятностей и традиционным главам математической статистики, рекомендуем монографии В.С. Пугачева [73], Ж. Закса [38], М. Кендалла и А. Стюарта [43], С.Р. Рао [76], энциклопедию «Вероятность и математическая статистика» [22]. Теория матриц достаточно полно изложена в монографии Ф.Р. Гантмахера [26]. С численными методами, используемыми при обработке неопределенных данных, можно ознакомиться в книгах [23, 41, 61, 66, 81, 88].

Авторы благодарят заведующего кафедрой «Прикладная математика» МИФИ, д.ф.-м.н., профессора Н.А. Кудряшова,

д.ф.-м.н., профессора этой же кафедры Т.И. Савёлову и рецензентов за просмотр рукописи и высказанные замечания, учтенные в окончательном варианте книги.

Авторы будут благодарны за все замечания и предложения, которые можно направлять по адресу: avkryanev@mephi.ru. Материалы, использованные в данной монографии, а также информацию о направлениях деятельности авторов можно найти на сайте www.kryanev.ru.

Москва, март 2010 г.

А. В. Крянев

Г. В. Лукин

Введение

Статистические методы все более широко применяют при решении различных прикладных задач. В первую очередь традиционно они используются при обработке данных экспериментов, причем круг задач обработки непрерывно расширяется прежде всего за счет расширения областей, где используются экспериментальные методы, и усложнения самих экспериментов.

В последние годы в связи с запросами практики интенсивно развиваются методы исследования, основанные на измерениях косвенной информации об изучаемых системах и объектах. Часто задачи математической обработки и интерпретации неопределенных данных по косвенным проявлениям изучаемых объектов принадлежат к классу некорректных задач. Разработка статистических методов решения некорректных задач обработки и интерпретации неопределенных данных еще далека от завершения, но и в этой области уже создано много подходов и методов, а на их основе — апробированных эффективных численных алгоритмов решения поставленных задач. Знать эти методы и уметь их использовать — насущная необходимость для любого экспериментатора и специалиста, обрабатывающего неопределенные данные.

Вторая значительная особенность последнего периода развития статистических методов исследования — широкое применение современной вычислительной техники при обработке неопределенных данных. Применение компьютеров при обработке данных дает возможность решать такие задачи, решение которых совсем недавно было невозможно. Например, большинство задач обработки результатов измерений косвенных проявлений изучаемых объектов невозможно решить без применения компьютеров. Кроме того, сокращение времени обработки при применении компьютеров дает возможность создания оперативных систем управления сложными объектами, в которых слежение за поведением объекта осуществляется с помощью экспериментальных методов (например, с помощью систем различных датчиков и приборов измерения). Бурное внедрение компьютеров и вычислительных методов в статистику, и, в частности, в статистическую обработку данных, заставило заново переосмыслить место и значение многих классических методов математической статистики под углом возможности их использования с применением компьютеров и потребовало

разработки самих численных реализаций, новых методов и создания эффективных численных алгоритмов обработки данных. Знать основные из этих методов и алгоритмов, владеть ими не только полезно, но и необходимо обработчику неопределенных данных.

В последние годы в самой математической статистике, в силу ее внутреннего развития и потребности практики, созданы новые подходы и методы, более полно и адекватно учитывающие реальные ситуации, возникающие при получении неопределенных данных. Прежде всего это относится к критическому переосмыслению основных предположений, упрощений и гипотез, которые берутся в качестве основы получения тех или иных статистических подходов и методов. Всякий разумный метод обработки должен обладать свойством устойчивости в том смысле, что любое возможное (физически реализуемое) малое отклонение от базовых предположений приведет лишь к малым изменениям в конечных результатах обработки данных. Такие устойчивые методы обработки в математической статистике получили название *робастных*.

При решении прикладных задач обработку неопределенных данных, как правило, проводят в условиях, когда у исследователя помимо исходных неопределенных данных имеется дополнительная априорная информация об изучаемом объекте. Ясно, что учет этой дополнительной информации позволяет в ряде случаев существенно повысить точность конечных результатов обработки. Более того, для целого класса задач обработки данных без учета дополнительной априорной информации вообще невозможно получить приемлемый конечный результат.

В математической статистике разработаны различные методы учета дополнительной информации об изучаемом объекте и ее «соединения» в процессе обработки с неопределенными данными, ознакомить с такими методами — также одна из целей настоящей книги.

В главах 7–13 книги представлены методы и алгоритмы восстановления регрессионных зависимостей и их применения для прогнозирования.

Задачи восстановления регрессионных зависимостей встречаются во всех областях естествознания при обработке неопределенных данных.

Обширный класс регрессионных зависимостей составляют линейные задачи. К линейным регрессионным задачам относятся, например, многие задачи восстановления функциональной зависимости, задачи исследования изучаемых объектов и си-

стем по измерениям косвенной информации об этих объектах и системах (линейные задачи идентификации). Математические модели линейных регрессионных задач — системы линейных уравнений, правая часть которых имеет случайный характер. В некоторых задачах идентификации не только правая часть исходной системы уравнений, но и оператор левой части имеет неопределенный характер и часто задается со случайными погрешностями. Из-за наличия погрешностей исходная система уравнений, как правило, является противоречивой, т.е. не существует ни одного точного решения исходной системы при заданной правой части с погрешностями. Поэтому попытки решения таких систем «в лоб» с помощью традиционных методов, ориентированных на получение точного решения, часто приводят к неудовлетворительному результату. В то же время, учитывая случайный характер погрешностей, для получения приемлемого приближенного решения этих систем естественно применять статистические методы.

Применение статистических подходов и методов для решения задач восстановления регрессионных зависимостей имеет ряд особенностей и специфических моментов, требующих качественного анализа и разработки эффективных численных алгоритмов получения приближенных решений. Ознакомить читателя с основными, наиболее эффективными и апробированными методами и численными алгоритмами решения задач восстановления регрессионных зависимостей — одна из целей настоящей книги.

В книге рассматриваются также нелинейные задачи идентификации, численное решение которых требует разработки специальных методов и алгоритмов.

Часто исследователь является не только пассивным обработчиком неопределенных данных, но и может повлиять на процесс их получения. В таких условиях естественно добиваться такого набора неопределенных данных (в рамках тех ограничений и возможностей, в которых проводится набор), при котором возможно получение наилучшего конечного результата по идентификации изучаемого объекта или системы. В книге изложены некоторые основные методы и алгоритмы решения задач планирования оптимального набора при восстановлении функциональных зависимостей.

Основные обозначения

$X_1, \dots, X_n$ — элементы выборки

$p(x)$ — плотность вероятностей

$F(x)$ — функция распределения вероятностей

$p_\xi(x)$ — плотность вероятностей случайной величины ξ

$p(x; u)$ — плотность вероятностей, зависящая от параметра u

$p(x | u)$ — условная плотность вероятностей при фиксированном значении u

$\hat{p}(x)$ — оценка плотности вероятностей $p(x)$

$\hat{F}(x)$ — оценка функции распределения $F(x)$

$\hat{u}$ — оценка параметра u

$F_n(x)$ — эмпирическая функция распределения

$M\hat{u}$ — математическое ожидание случайной величины $\hat{u}$

$P_{\text{дов}}$ — доверительная вероятность

α_k — начальный момент k -го порядка

μ_k — центральный момент k -го порядка

$\sigma^2(\xi), D(\xi)$ — дисперсия случайной величины ξ

s^2 — эмпирическая дисперсия

$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_0^t e^{-\frac{s^2}{2}} ds$ — интеграл вероятностей

$\Gamma(t)$ — гамма-функция

χ_n^2 — случайная величина, распределенная по закону хи-квадрат с n степенями свободы

$I_n, I_n(u)$ — информация Фишера или информационная матрица Фишера

$L_n, L_n(u)$ — функция правдоподобия

$l_n, l_n(u)$ — логарифмическая функция правдоподобия

K_ξ — ковариационная матрица случайного вектора ξ

$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ — средняя арифметическая

$N(m, \sigma^2)$ — нормальный закон распределения с математическим ожиданием m и дисперсией σ^2

$\hat{u}_B$ — оценка Байеса параметра u

K_a — ковариационная матрица априорного распределения

u_a — центр априорного распределения параметра u

$p_{\text{апр}}(u)$ — плотность априорного распределения

U_a, R_a — априорное множество

$\hat{u}_{\text{ММП}}$ — оценка метода максимального правдоподобия параметра u (ММП-оценка)

$\hat{u}_{\text{ОММП}}$ — оценка обобщенного метода максимального правдоподобия параметра u (ОММП-оценка)

Δu — смещение оценки u

$\arg \min_u (\max) f(u)$ — значение u , для которого целевая функция $f(u)$ достигает минимума (максимума)

K — параметр Хьюбера

$\rho(s)$ — функция влияния M -робастной оценки параметра положения

I_0, I_+, I_- — множество индексов, соответствующих средней части, правому и левому «хвостам» распределения Хьюбера

$K(u)$ — функция распределения Колмогорова

$K_n(u)$ — функция распределения Масси

$\delta(x)$ — обобщенная функция Дирака

$H_k(x)$ — полином степени k Чебышева–Эрмита

$L_k^\alpha(x)$ — полином степени k Чебышева–Лагерра

$L(c)$ — функция Лагранжа

γ — случайная величина, равномерно распределенная на промежутке $[0, 1]$

ε — доля засорения в модели смеси Хьюбера или вектор погрешностей правой части регрессионной модели

$\hat{u}_{\text{МНК}}$ — оценка метода наименьших квадратов параметра u

$\lambda_{\min}(C)(\lambda_{\max}(C))$ — минимальное (максимальное) собственное значение квадратной матрицы C

$\text{cov}(X, Y)$ — ковариация случайных величин X и Y

$\text{grad}_u f(u)$ — градиент функции $f(u)$

A^T — транспонированная матрица матрицы

C^{-1} — обратная матрица невырожденной квадратной матрицы C

$\hat{u}_{\text{ММ}}$ — оценка минимаксного метода параметра u

(u, v) — скалярное произведение векторов u и v

$\|u\|$ — евклидова норма вектора u

$S_n(x)$ — кубический интерполяционный сплайн

$S_{n\alpha}(x)$ — кубический сглаживающий сплайн

$S_\alpha(t)$ — линейный сглаживающий сплайн

$\psi(t)$ — базисный вейвлет

$\mathcal{E}(N)$ — план эксперимента

$\varepsilon(N)$ — нормированный план эксперимента

$M(\varepsilon)$ — информационная матрица, соответствующая плану эксперимента $\varepsilon(N)$

$\mathcal{D}(\varepsilon) = M^{-1}(\varepsilon)$ — обратная матрица информационной матрицы $M(\varepsilon)$

ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Математическая статистика, и, в частности, математическая обработка экспериментальных данных как ее составная часть, базируются на многих разделах математики и прежде всего на теории вероятностей. Предполагается, что читатель знаком с такими основными понятиями теории вероятностей, как случайные события и величины, вероятность события, операции над случайными событиями, случайными величинами и вероятностями, законы распределения случайных величин, предельные теоремы теории вероятностей. В качестве базовой книги, где изложены теория вероятностей и математическая статистика, рекомендуем использовать монографию [73]. Некоторые дополнительные сведения из теории вероятностей, используемые в математической статистике, будут вводиться в соответствующих местах монографии по мере необходимости.

Одной из основных, базовых задач математической статистики является задача восстановления неизвестного закона распределения случайной величины по конечному числу ее реализаций. Более подробно, рассматривается случайная величина ξ , для которой известна (из эксперимента) совокупность ее случайных реализаций

$$X_1, \dots, X_n. \quad (1.1)$$

Требуется с максимально возможной точностью определить закон распределения ξ , например плотность вероятностей $p_\xi(x)$.

Совокупность случайных реализаций (1.1) называется *выборкой*, число реализаций n — *объемом выборки*. Если случайные реализации, рассматриваемые как случайные величины, независимы, то выборка называется *повторной*, в противном случае — *бесповторной*. Для повторной выборки, и только для нее, совместная плотность вероятностей совокупности случайных величин (1.1) (если она существует) имеет вид $\prod_{i=1}^n p_\xi(x_i)$.

В дальнейшем, если специально не оговорено, предполагается, что (1.1) — повторная выборка. В математической статистике лю-

бую измеримую функцию (скалярную или векторную) от членов выборки (1.1) принято называть *статистикой*.

Иногда исследователя интересует не сам закон распределения случайной величины ξ , а лишь некоторые его характеристики, например *математическое ожидание* и *дисперсия* (соответственно, вектор математических ожиданий и ковариационная матрица, если ξ — векторная случайная величина) или некоторые моменты случайной величины более высокого порядка.

При восстановлении плотности вероятности по выборке различают два подхода — параметрический и непараметрический.

Параметрический подход основан на гипотезе о принадлежности искомой плотности вероятностей $p_\xi(x; u)$ классу плотностей, определяемых совокупностью параметров u (если u фиксирован, то плотность $p_\xi(x; u)$ определена однозначно). Например, часто вводится гипотеза о нормальности распределения скалярной случайной величины ξ , т. е.

$$p_\xi(x; m, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-m)^2/(2\sigma^2)}$$

с неизвестными параметрами m, σ^2 .

В рамках параметрического подхода задача восстановления плотности вероятностей сводится к задаче восстановления значений тех неизвестных параметров u , от которых зависит плотность вероятностей. При решении прикладных задач методами математической статистики исследователя чаще всего интересуют именно значения восстанавливаемых параметров, а не сама плотность вероятностей, поэтому методам параметрической статистики будет уделено наибольшее внимание.

Непараметрические методы восстановления законов распределения наблюдаемых случайных величин не предполагают такой жесткой фиксации класса законов распределения при помощи конечного числа параметров, хотя довольно часто в рамках непараметрического подхода используют метод «погружения» в параметрическое семейство с потенциальной возможностью неограниченного увеличения числа искомых параметров.

§ 1. Параметрическая статистика. Свойства оценок

Интуитивно ясно, что по выборке конечного объема принципиально невозможно точное восстановление закона распределения в рамках как параметрического, так и непараметрического подходов. В результате восстановления мы получаем приближенные значения закона распределения (например, плотности веро-

ятностей) или приближенные значения параметров (если использован параметрический подход). В математической статистике приближенные значения параметров плотности вероятностей или самой плотности вероятностей, полученные на основе выборки (1.1), принято называть *оценками*.

Оценки параметра u и плотности p_{ξ} будем обозначать символами $\hat{u}$ и $\hat{p}_{\xi}$ соответственно.

Любая оценка $\hat{u}$ зависит от членов выборки (1.1) и поэтому является случайной величиной.

Оценка $\hat{u}$ называется *несмещенной*, если $M\hat{u} = u_T$, где u_T — искомое точное значение параметра u . Если $M\hat{u} \neq u_T$, то оценка $\hat{u}$ *смещена*.

Поскольку оценка $\hat{u} = \hat{u}(X_1, \dots, X_n)$ зависит от членов выборки, то она, вообще говоря, будет изменяться при увеличении объема выборки (добавлении новых членов выборки). Интуитивно ясно, что при $n \rightarrow \infty$ информация о законе распределения в выборке неограниченно увеличивается, и поэтому достаточно разумная оценка параметра должна сближаться (в том или ином смысле) с его точным значением.

Оценка $\hat{u} = \hat{u}(X_1, \dots, X_n)$ называется *состоятельной по вероятности* (в среднем квадратичном), если $\hat{u}(X_1, \dots, X_n)$ сходится по вероятности (в среднем квадратичном) к u_T при $n \rightarrow \infty$. Если тип сходимости безразличен или ясен из контекста, то будем говорить о *состоятельности* оценки без указания типа сходимости.

Конечно, состоятельность является хорошим свойством оценки, но все же скорее успокоительного, а не конструктивного характера, поскольку на практике объем выборки всегда конечен. Для практики наиболее важным является вопрос о близости (в том или ином смысле) оценки к искомому значению параметра при фиксированном объеме выборки. Мету близости можно вводить по-разному. Например, если u — скаляр и оценка $\hat{u}$ несмещена, то в качестве меры близости чаще всего выбирают дисперсию оценки (если $\hat{u}$ — несмещенная оценка вектора, то мерой близости может служить ковариационная матрица оценки). Оценки, наиболее близкие к истинным значениям, называют *эффективными*. Одна из основных задач математической статистики — нахождение эффективных оценок.

Как уже указывалось выше, основой параметрического подхода является жесткое постулирование класса распределений. К сожалению, на практике реальные законы распределения случайных величин, образующих выборку (1.1), как правило, от-

личны от постулированного. Хотя такое отличие может быть и малым, оно всегда имеется. Если оценка малочувствительна к малым изменениям (в том или ином смысле) исходного постулированного закона распределения, то такая оценка называется *робастной*.

Любая оценка $\hat{u}$ n -мерного вектора u представима в виде

$$\hat{u} = S(X_1, \dots, X_n), \quad (1.2)$$

где S — оператор из конечномерного пространства элементов $(X_1, \dots, X_n)$ в m -мерное пространство E^m .

Если S — линейный оператор, то оценка (1.2) называется *линейной*, в противном случае $\hat{u}$ — *нелинейная оценка*. В частности, если X_i — векторы размерности k , то линейная оценка представима в виде $\hat{u} = SX + u_0$, где S — матрица размерности $(m \times nk)$, X — совокупность компонент векторов выборки (1.1), u_0 — фиксированный детерминированный вектор.

Класс линейных оценок играет важную роль в математической статистике, поскольку в этом классе часто удается найти «хорошие» оценки, а линейность оценки позволяет в большинстве случаев вычислять ее с меньшими вычислительными затратами по сравнению с нелинейными оценками.

Часто встречаются оценки, которые вместо свойств несмещенности обладают более общим свойством *асимптотической несмещенности*: $M\hat{u}(X_1, \dots, X_n) \rightarrow u_T$ при $n \rightarrow \infty$. В частности, состоятельные в среднем квадратичном оценки асимптотически несмещены.

Выше были рассмотрены некоторые основные свойства оценок. Конкретные оценки могут обладать одними из этих свойств и не обладать другими. Спрашивается, каким из перечисленных свойств оценок следует отдавать предпочтение при выборе подходящей оценки? Универсального ответа на этот вопрос не существует. Конечно, всегда желательно, чтобы оценка была эффективной. Но, к сожалению, часто эффективные оценки неустойчивы и тогда, в условиях опасности нарушения основных предположений, следует отдавать предпочтение робастным оценкам, эффективность которых максимальна. С другой стороны, и эффективные (или почти эффективные), и робастные оценки, как правило, даже для простых параметрических моделей могут быть вычислены только с помощью компьютеров. Линейные оценки наиболее легки для вычислений, но в большинстве случаев не обладают свойством робастности.

До сих пор мы рассматривали точечное оценивание параметров, в результате которого вместо неизвестного точного значения u_T получаем одно его приближенное значение — оценку $\hat{u}$.

В математической статистике существует еще один способ оценивания — *интервальный*. При интервальном оценивании находят две оценки скалярного параметра u u_n, u_b такие, что определена вероятность $P\{u \in (u_n, u_b)\} = p(u)$. Величину $p_{\text{дов}} = \inf_u p(u)$ называют доверительным уровнем, а $1 - p_{\text{дов}}$ уровнем риска. В приложениях часто доверительный уровень и уровень риска выражают в процентах, а границы u_n, u_b подбирают так, чтобы доверительный уровень равнялся одному из стандартных значений: 0,9; 0,95; 0,99; 0,999. Например, если $p_{\text{дов}} = 0,95$, то говорят о 95 % доверительном уровне и 5 % уровне риска.

Интервальные оценки более информативны, чем точечные, но их получение требует больших вычислительных затрат. Наиболее часто интервальные оценки получают, используя подходящую точечную оценку $\hat{u}$, для которой удастся построить плотность вероятностей $p_{\hat{u}}(u)$. Тогда доверительный интервал $[\hat{u} - \delta_1, \hat{u} + \delta_2]$ находят из равенств

$$\int_{\hat{u}-\delta_1}^{\hat{u}} p_{\hat{u}}(u) du = \frac{p_{\text{дов}}}{2}, \quad \int_{\hat{u}}^{\hat{u}+\delta_2} p_{\hat{u}}(u) du = \frac{p_{\text{дов}}}{2}.$$

В частности, если $p_{\hat{u}}(u)$ симметрична относительно $\hat{u}$, доверительный интервал $[\hat{u} - \delta, \hat{u} + \delta]$ симметричен относительно $\hat{u}$, а $\delta > 0$ находят из равенства

$$\int_{\hat{u}-\delta}^{\hat{u}+\delta} p_{\hat{u}}(u) du = p_{\text{дов}}.$$

С методами математической статистики можно ознакомиться в книгах [2–5, 38, 43, 46, 56, 76].

§ 2. Метод моментов

Пусть (1.1) — повторная выборка скалярной случайной величины ξ , у которой существуют моменты достаточно большого порядка.

В качестве оценки $\hat{\alpha}_k$ начального момента k -го порядка $\mathbf{M}\xi^k$ возьмем статистику

$$\hat{\alpha}_k = \frac{1}{n} \sum_{i=1}^n X_i^k. \quad (1.3)$$

В частности, оценка математического ожидания $\hat{m}$ дается равенством

$$\hat{m} = \frac{1}{n} \sum_{i=1}^n X_i. \quad (1.4)$$

В качестве оценки $\hat{\mu}_k$ центрального момента k -го порядка $\mathbf{M}(\xi - \mathbf{M}\xi)^k$ возьмем статистику

$$\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{m})^k; \quad (1.5)$$

в частности, оценка дисперсии $\hat{\sigma}^2$ дается равенством

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{m})^2. \quad (1.6)$$

Оценки (1.3) несмещены, так как $\mathbf{M}\hat{\alpha}_k = \frac{1}{n} \sum_{i=1}^n \mathbf{M}X_i^k = \mathbf{M}\xi^k$, состоятельны в среднем квадратичном в силу закона больших чисел (теорема Чебышева).

Оценки (1.5) также состоятельны в среднем квадратичном, но смещены, и, следовательно, обладают свойством асимптотической несмещенности. Например, для оценки (1.6)

$$\mathbf{M}\hat{\sigma}^2 = \frac{n-1}{n} \sigma^2, \quad \mathbf{D}(\hat{\sigma}^2) = \frac{\mu_4 - \sigma^4}{n} - \frac{2(\mu_4 - 2\sigma^4)}{n^2} + \frac{\mu_4 - 3\sigma^4}{n^3},$$

где μ_4 — центральный момент 4-го порядка случайной величины ξ .

Часто вместо асимптотически несмещенной оценки (1.6) используют несмещенную оценку дисперсии:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \hat{m})^2, \quad (1.7)$$

которую называют *эмпирической дисперсией*.

Дисперсии несмещенных оценок (1.4), (1.7) равны $\mathbf{D}(\hat{m}) = \sigma^2/n$, $\mathbf{D}s^2 = n^2(n-1)^{-2}\mathbf{D}(\hat{\sigma}^2)$. В частности, если ξ распределена по нормальному закону, то $\mu_4 = 3\sigma^4$ и $\mathbf{D}(s^2) = 2\sigma^4/(n-1)$.

На оценках моментов (1.3) случайной величины ξ основан метод моментов оценки параметров u и искомой плотности вероятностей $p_\xi(x; u)$.

Пусть задано m линейно независимых функций $a_1(x), \dots, a_m(x)$ (m — размерность вектора u). Имеем

$$\mathbf{M} a_j(\xi) = \int_{-\infty}^{\infty} p_\xi(x; u) a_j(x) dx = b_j(u),$$

где $b_j(u)$ — известные функции искомых параметров u . В качестве состоятельных оценок $\mathbf{M} a_j(\xi)$ возьмем статистики

$$\hat{a}_j(X) = \frac{1}{n} \sum_{i=1}^n a_j(X_i).$$

Тогда согласно методу моментов в качестве оценок параметров берется решение системы уравнений

$$b_j(\hat{u}) = \hat{a}_j(X), \quad j = 1, \dots, m. \quad (1.8)$$

В частности, если $a_j(X) = X^j$, $j = 1, \dots, m$, то

$$\hat{a}_j(x) = \hat{\alpha}_j = \frac{1}{n} \sum_{i=1}^n X_i^j.$$

Пример 1.1. $p_\xi(x; u) = \begin{cases} ue^{-ux}, & x \geq 0, \\ 0, & x < 0, \end{cases}$ где $u > 0$.

Возьмем $a_1(x) = x$. Имеем: $\mathbf{M} a_1(x) = \mathbf{M} \xi = 1/u$. Поэтому система (1.8) сводится к одному уравнению $\frac{1}{\hat{u}} = \frac{1}{n} \sum_{i=1}^n X_i$, откуда

$$\hat{u} = n / \sum_{i=1}^n X_i.$$

Замечание 1.1. Если интегралы

$$\int_{-\infty}^{\infty} p_\xi(x; u) a_j(x) dx$$

аналитически не «берутся», то при решении нелинейной системы (1.8) (обычно одним из методов итераций) левые части равенств (1.8) подсчитываем, вычисляя эти интегралы численно.

Если ξ — векторная случайная величина размерности r , то каждый член $X_i = (X_{i1}, \dots, X_{ir})^T$ выборки (1.1) — r -мерный вектор.

В качестве несмещенной и состоятельной оценки $\hat{m}$ вектора математических ожиданий векторной случайной величины ξ можно взять

$$\hat{m} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{n} \left(\sum_{i=1}^n X_{i1}, \dots, \sum_{i=1}^n X_{ir} \right)^T = (\bar{X}_1, \dots, \bar{X}_r)^T.$$

Статистики

$$\hat{h}_{ij} = \frac{1}{n-1} \sum_{k=1}^n (X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j),$$

$$\hat{r}_{ij} = \hat{h}_{ij}(n-1) \left(\sum_{k=1}^n (X_{ki} - \bar{X}_i)^2 \sum_{k=1}^n (X_{kj} - \bar{X}_j)^2 \right)^{-1/2}$$

являются состоятельными оценками ковариации $\text{cov}(\xi_i, \xi_j)$ и коэффициента корреляции $r(\xi_i, \xi_j)$ компонент ξ_i и ξ_j векторной случайной величины ξ . Оценка $\hat{h}_{ij}$ несмещена, оценка $\hat{r}_{ij}$ смещена и потому является асимптотически несмещенной.

Следовательно, состоятельной и несмещенной оценкой ковариационной матрицы K_ξ является симметричная положительно определенная $(r \times r)$ -матрица $\hat{K}_\xi$, (i, j) -й элемент которой равен $\hat{k}_{ij}$. Матрицу $\hat{K}_\xi$ можно представить в виде

$$\hat{K}_\xi = \frac{1}{n-1} X^T X, \quad X = \begin{pmatrix} X_{11} - \bar{X}_1 & X_{12} - \bar{X}_2 & \dots & X_{1r} - \bar{X}_r \\ X_{21} - \bar{X}_1 & X_{22} - \bar{X}_2 & \dots & X_{2r} - \bar{X}_r \\ \dots & \dots & \dots & \dots \\ X_{n1} - \bar{X}_1 & X_{n2} - \bar{X}_2 & \dots & X_{nr} - \bar{X}_r \end{pmatrix},$$

где X — $(n \times r)$ -матрица.

Пусть ξ — скалярная нормально распределенная случайная величина и $p_\xi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-m)^2/(2\sigma^2)}$.

Рассмотрим сначала случай, когда m неизвестен, σ^2 известен. В качестве оценки $\hat{m}$ возьмем статистику (1.4). Случайная величина $\hat{m}$ распределена по нормальному закону $N(m, \sigma/\sqrt{n})$. Поэтому нормированная случайная величина $(\hat{m} - m)/(\sigma/\sqrt{n})$ распределена по нормированному нормальному закону $N(0, 1)$. Имеем

$$\mathbf{P} \left\{ \frac{|\hat{m} - m|}{\sigma/\sqrt{n}} \leq t \right\} = p_{\text{дов}} = 2\Phi(t),$$

где $\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_0^t e^{-s^2/2} ds$ — интеграл вероятностей. В таблице

значений интеграла вероятностей находим $t = t(p_{\text{дов}})$, для которого $2\Phi(t) = p_{\text{дов}}$. Тогда доверительный интервал для параметра m с уровнем доверия $100p_{\text{дов}}\%$ имеет вид

$$\hat{m} - t(p_{\text{дов}}) \frac{\sigma}{\sqrt{n}} \leq m \leq \hat{m} + t(p_{\text{дов}}) \frac{\sigma}{\sqrt{n}}. \quad (1.9)$$

Для 90 %, 95 %, 99 %, 99,9 % уровней доверия коэффициент $t(p_{\text{дов}})$ равен 1,64; 1,96; 2,58; 3,29 соответственно. При увеличении уровня доверия длина доверительного интервала возрастает, например, длина доверительного интервала при уровне риска в 0,1 % почти в два раза больше длины доверительного интервала для уровня риска в 10 %.

Рассмотрим теперь случай, когда m и σ^2 оба неизвестны. Тогда в качестве точечных оценок этих параметров возьмем статистики (1.4) и (1.7). При построении доверительных интервалов для m и σ^2 используется следующее утверждение.

Теорема 1.1. Пусть (1.1) — повторная выборка случайной величины, распределенной по нормальному закону $N(m, \sigma)$. Тогда $\hat{m}$ и s^2 независимы и $s^2(n-1)/\sigma^2$ распределена по закону χ^2_{n-1} .

Доказательство. Совместная плотность вероятностей совокупности независимых центрированных случайных величин $Y_i = X_i - m$, $i = 1, \dots, n$, дается равенством

$$p(y) = p(y_1, \dots, y_n) = \frac{1}{(2\pi)^{n/2} \sigma^n} \exp \left(-\frac{1}{2\sigma^2} \right) \sum_{i=1}^n y_i^2.$$

В n -мерном пространстве $(y_1, \dots, y_n)^T$ совершим поворот вокруг начала координат так, чтобы в новой системе координат $z = (z_1, \dots, z_n)^T$ выполнялось равенство $z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n y_i$. Имеем: $z = Uy$, $Z = UY$, где U — ортогональная матрица поворота. Случайные величины $Z_1, \dots, Z_n$ независимы. Кроме того, $\sum_{i=1}^n Z_i^2 = \sum_{i=1}^n Y_i^2$, поскольку при повороте расстояния не изменяются.