Дж. С. Дэвис

СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ В ГЕОЛОГИИ

STATISTICS AND DATA ANALYSIS IN GEOLOGY

Second edition

John C. Davis

Kansas Geological Survey

John Wiley and Sons

New York • Chichester • Brisbane

Toronto • Singapore

Дж. С. Дэвис

СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ В ГЕОЛОГИИ

Перевод с английского доктора
физико-математических наук
В.А. Голубевой

Под редакцией доктора
геолого-минералогических наук
Д.А. Родионова

В ДВУХ КНИГАХ

КНИГА 1

МОСКВА "НЕДРА" 1990

ББК 26.3
Д 94
УДК 550.8.053:519

Рекомендовано к изданию кандидатом геолого-минералогических наук *Р. И. Коганом.*

Дэвис Дж. С.

Д 94 Статистический анализ данных в геологии: Пер. с англ. В 2 кн./Пер. В. А. Голубевой; Под ред. Д. А. Родионова. Кн. 1. — М.: Недра, 1990. — 319 с.: ил.

ISBN 5-247-02122-3

Изложены методы математической статистики и матричной алгебры, применяемые в современных геологических исследованиях. Рассмотрены известные критерии проверки статистических гипотез: о нормальном распределении, критерии Стьюдента, Фишера, Манна-Уитни и др. Большое внимание уделено непараметрическим методам статистического анализа данных. Описаны процедуры анализа последовательностей данных: интерполяция, ортогональная полиномиальная регрессия, зонирование, классификация, спектральный анализ, вариограммы, фильтрация и тренд.

Для геологов всех специальностей, занимающихся обработкой количественных и качественных результатов наблюдений.

ББК 26.3

Д $\frac{1804010000-288}{043(01)-90}$ 35—90

ISBN 0-471-08079-9
ISBN 5-247-02121-5
ISBN 5-247-02122-3

ПРЕДИСЛОВИЕ

В 1973 г., когда вышло первое издание этой книги, использование геологами вычислительных средств находилось на качественно ином уровне, чем сейчас. Это было время массивных ЭВМ, сосредоточенных в вычислительных центрах, доступ к которым осуществлялся через окошко в закрытых дверях. При этом результат исследователь получал в лучшем случае через несколько дней.

Теперь большинство геологов имеют непосредственный доступ к ЭВМ через терминал или к мини-компьютеру, или имеют даже персональный компьютер. Компьютер стал обыденной вещью в жизни геолога. К нему обращаются как новички, так и профессионалы в надежде повысить эффективность своей работы.

К сожалению, легкий доступ к компьютерам не обеспечивает легкого получения знаний о том, что с ним делать. Для многих геологов анализ поверхностей тренда так же мало понятен, как и 10 лет назад. То же можно сказать и о факторном анализе. Более того, появились еще более экзотические и трудно доступные методы. Необходимость в обучении геологов количественному анализу была очевидна уже в 1973 г., то же верно и сейчас. Вот почему написана эта книга.

В ответ на многие замечания, которые я получил после издания книги в 1973 г., а также, учитывая собственный опыт преподавания, я внимательно пересмотрел книгу для нового издания. Расположение материала сохранено, оно начинается с основных понятий и заканчивается анализом последовательностей, карт, многомерных наблюдений. Так как большинство студентов слушает один или более курсов по ФОРТРАНу, то глава о ФОРТРАНе в этом издании отсутствует.

Изложение начинается с основ теории вероятностей, очень важных в анализе данных. Добавлен новый раздел о непараметрических методах, которые представляются более пригодными для геологических данных. Тема «Собственные значения и собственные векторы» остается трудной для геологов, и поэтому она затрагивается дважды: в разделе «Матричная алгебра» и в разделах, посвященных факторному анализу. Рассмотрена также связь процедур нахождения собственных значений и собственных векторов с методом главных компонент, *R*- и *Q*-методами факторного анализа, анализа соответствия.

Некоторые темы анализа данных в последние несколько лет приобретали все большее изначение в науках о Земле. Теория регионализованных переменных привлекается сейчас для объяснения пространственных свойств геологических переменных многими исследователями. Центральную роль в этой тео-

рии играют полувариограммы и крайгинг. Эти методы представлены в настоящем издании. Геофизики поняли важную роль спектрального анализа; очевидна полезность этих методов при решении многих других задач, начиная от предсказания землетрясений и кончая описанием формы ископаемых остатков. Раздел о рядах Фурье излагается с учетом этих изменений.

Ряд таблиц и рисунков в этой книге воспроизведены с разрешения авторов (владельцев авторских прав). Источник для каждой таблицы и рисунка указан в квадратных скобках, а полная ссылка приводится в списках литературы к каждой главе. Таблицы 2.10, 2.22, 2.25 и 2.26 являются собственностью издательства Джон Уайли и Сыновья Inc., таблицы 2.11, 2.14 и 2.18 — собственностью Пингвин-Бук Ltd, а таблицы 4.30 и 4.31 — Американского химического общества. Все они воспроизведены с соответствующих разрешений. Часть таблиц 5.6 — собственность Американской статистической ассоциации и другая часть — Американского института биологических наук; комбинирование таблиц сделано с их разрешения. Таблицы 5.7 и 5.9 представляют собственность Академик Пресс Inc (Лондон) Ltd и воспроизведены с их разрешения. Рис. 5.24 — собственность Американской статистической ассоциации, а рис. 5.25 — собственность Харкурт Брейс Йованович, Inc, оба рисунка воспроизведены с соответствующих разрешений.

В тексте, в ответ на многочисленные пожелания читателей, сделано много изменений, исправлений, добавлений. Перечислять тех, кто написал мне, нет смысла. Я их благодарю. В дополнение к тем, кто был назван в первом издании, приношу мою благодарность доктору Паулю Брокинстону, доктору Джиму Кемпбеллу и доктору Кейту Тернеру за их помощь. Мои рецензенты, доктор Дейв Бест, профессор Франк Этридж и профессор Джи-эн Фэнг, сделали много ценных исправлений в окончательном тексте.

Многочисленные добавления были сделаны коллегами из Канзасской геологической службы, включая доктора Дэйвида Коллиниза и доктора Калина Фергюсона и моего помощника по первому изданию мистера Роберта Сэмпсона. Трое из моих коллег приняли активное участие в написании книги: доктор Рикардо Олеа — раздел по регионализованным переменным, доктор Зоу Ди — раздел о собственных значениях и доктор Джон Доветон, который любезно предложил многие из упражнений и примеров во всей книге и который помогал мне на всех стадиях работы. Наконец, я особенно признателен моему ассистенту, исследователю и компаньону миссис Джо Энн Де-Греффенрайд, без поддержки которой выход этой книги оказался бы невозможным.

Джон С. Дэвис

Глава 1
ВВЕДЕНИЕ

«...если Вы можете измерить то, о чем говорите, и результат выразить числом, это означает, что Вы кое-что знаете о предмете разговора; но если Вы не можете охарактеризовать этот предмет числом, то из этого следует, что Ваши знания скудны и неудовлетворительны, и они могут быть только отправной точкой процесса познания».

Лорд Кельвин

Еще на заре становления геологии некоторые геологи пользовались математическими методами. Например, горные инженеры и геологи сотни лет назад подсчитывали запасы по результатам опробования и оценивали содержание в руде полезных компонентов. Фишер [2] отмечал, что расчленение третичного периода Лайелем на основе относительной распространенности современных организмов было статистической процедурой. Литологи еще в начале этого столетия изучали размер зерен и их очертания, которые представляли собой важные источники геологической информации. Такие науки о Земле, как геохимия, геофизика и гидрология, требуют прочного математического фундамента, хотя используемые ими приемы первоначально были разработаны не на основе геологии. Точно так же минералоги и кристаллографы используют математический аппарат физической и аналитической химии.

Хотя эти разделы имеют важное значение в специализированных областях, они не являются предметом рассмотрения этой книги. Начиная с конца 50-х годов нашего века вычислительные машины стали широко применяться в университетах и корпорациях, в результате чего геологи значительно чаще стали прибегать к математическим методам анализа данных, которые они заимствовали из различных, особенно технических, наук и применяли к любому разделу наук о Земле, это более общие методики, чем используемые обычно. Геология сама по себе привела к некоторым успехам в вычислительных науках, особенно в области построения графиков, включая карты и оконтуривание. Однако наша наука выгадала больше, чем пожертвовала в обмен на количественные методы.

Нефтяные компании США, ведущие большие геологоразведочные работы (не считая правительственных учреждений), широко используют вычислительные машины. Поэтому огромный интерес, который проявляют эти организации к геоматематическим методам, закономерен. Он выражается также в увеличении роли языков программирования и математической подготовки при обучении геологов. К сожалению, не существует широко распространенных традиций использования математического анализа в геологии; более того, формирование соответствующих программ обучения происходило только в некоторых институтах благодаря усилиям отдельных ученых. Всего лишь несколько школ преуспело в этом направлении настолько, чтобы иметь право считаться пионерами в количественной геологии. Обучение у них основано на прочном фундаменте геологии, математики и статистики.

Многих геологов вычислительная революция застала врасплох: воспитанные в традициях, которые требуют получения качества за счет количества, они оказались плохо подготовленными математически и незнакомыми со статистикой. Но даже они быстро оценили потенциальные возможности аналитических методов, которые вычислительная техника сделала легко доступными. Многие организации, как коммерческие, так и государственные, создали обширные библиотеки программ для ЭВМ, предназначенных для реализации геоматематических процедур. Искушение использовать эти программы слишком сильно, даже несмотря на то что их основа не может быть ясно осознана.

Широкое внедрение персональных компьютеров привело к усилению этих тенденций. В настоящее время небольшие компании, группы консультантов и даже отдельные исследователи-геологи получили доступ к вычислительной технике, применение которой всего лишь несколько лет назад было привилегией больших корпораций и университетов. Многих геологов сейчас можно увидеть за собственным компьютером, даже тех, о которых нельзя было подумать, что они будут нуждаться в нем или просто иметь повод его использовать. Многим из этих геологов, если они умеют применять их в своей профессиональной работе, кажется, что работа с компьютерами обещает им больше, чем оперирование словами и финансовые подсчеты.

Эта книга частично предназначена для того, чтобы помочь геологам, которые сознают, что математические методы могли бы быть им полезны в исследованиях, но недостаточно подготовлены к этому. Конечно, они могли бы прослушать формальный курс теории вероятностей, статистики, численного анализа и программирования с последующей работой под руководством опытного геоматематика. Однако на практике бывает иначе, и большинство исследователей выбирают свой путь наилучшим

возможным для них способом: читая, спрашивая, учась на своих ошибках. Путь, которым эти люди следуют, не является методически прогрессивным из-за того, что при этом в стороне остаются многие важные вопросы. Обычно они возвращаются назад, обращая внимание на те методы, которые, по их мнению, наиболее применимы в их исследованиях, разработках, оперативной работе. Затем они чувствуют пробелы в подготовке и пытаются приспособить для этих же целей технические приемы обработки данных. Это неудовлетворительный и даже опасный метод обучения, возможно сравнимый с обучением врача в процессе работы. Однако он является одним из путей, по которому приходится идти многим геологам. Эта книга может помочь организовать процесс самообучения, а именно дает возможность сделать первые шаги к познанию описанных в ней алгоритмов. Читателю придется освоить внешне менее эффективные темы, составляющие фундамент, на котором построены, например, основы теории поверхностей тренда и факторного анализа.

Эта книга предназначена также для студентов, изучающих статистику и программирование. Такие курсы все чаще становятся обязательными в американских и европейских университетах. К несчастью, они обычно читаются лицами, мало знакомыми с геологией и проблемами наук о Земле. Связь этих предметов с основной тематикой обучения студентов остается неясной. Это чувство осложняется отсутствием математических приложений во многих геологических курсах. В то время как студенты нуждаются в специалисте, их учителя зачастую являются людьми, получившими образование до бурного развития количественных методов и, следовательно, не подготовленными в этом направлении. В настоящей книге читатель найдет не только общий курс вычислительных методов, но также многочисленные примеры их применения в геологии. Конечно, мы надеемся, что и студенты, и преподаватели найдут в этой книге что-либо интересное, и она будет способствовать распространению тех основ знаний, которые мы называем геоматематикой.

ОБ ЭТОЙ КНИГЕ

Читатель вправе знать с самого начала, куда и по какому пути ведет его автор, а также что от него требуется, так как автор делает определенные предположения о подготовке, интересах и возможностях своей аудитории. Эта книга посвящена количественным методам анализа геологических данных, а именно разделу наук о Земле, который в настоящее время называется геоматематикой. Ориентация книги — методологическая, т. е. «как надо делать». Теории уделяется мало внима-

ния по нескольким причинам. Дело в том что многие геологи стремятся быть прагматиками и поэтому интересуются результатами больше, чем теорией, а большинство полезных процедур все еще не имеет соответствующего теоретического обоснования. Теоретически достаточно разработанные методы часто основаны на сильных статистических ограничениях, которые обычно не выполнимы при исследовании геологических данных. Хотя в книге и обсуждаются элементарные аспекты теории вероятностей и описывается большинство статистических критериев, все же подробное изложение геостатистической теории предоставляется другим авторам.

В связи с тем что самые сложные аналитические процедуры можно представить как последовательность простых математических действий, уделим особое внимание операциям. Эти операции зачастую выражаются в терминах матричной алгебры, что в свою очередь приведет к рассмотрению этого предмета.

Первая категория охватывает все классы задач, для которых данные собираются непрерывно по времени или по линии. К ней относятся задачи анализа временных рядов, стратиграфических разрезов и интерпретации графиков. Вторая категория объединяет задачи, учитывающие географические координаты наблюдений: картирование, анализ поверхностей тренда, крайгинг и др. Наконец, третья категория имеет дело с анализом групп (кластер-анализом), классификацией и исследованием внутренних связей внутри наборов данных, в которых положение пробы на карте или профиле не рассматривается. Задачи изучения палеонтологических, геохимических данных часто относятся к этой категории.

Материал в книге изложен по принципу от простого к сложному, причем каждая последующая тема строится на основе предыдущих. Так, вопросы множественной регрессии, излагаемые в гл. 6, основаны на результатах, полученных в гл. 5 (см. кн. 2) применительно к тренд-анализу, которым в свою очередь предшествует описание нелинейной регрессии (см. гл. 4). Основная используемая при этом математическая процедура описана в гл. 3 при изложении методов решения систем уравнений, а статистические основы регрессионного анализа впервые рассматривались в гл. 2. Другие методы изложены по аналогичной схеме.

Первая тема, рассмотренная в этой книге, — элементарное введение в статистику, а последняя — факторный анализ. Между этими темами пропасть, преодоление которой требует нескольких лет изучения соответствующих курсов. Ясно, что в одной книге мы не можем осуществить переход от первой темы к последней, не опуская при этом значительного материала. В связи с этим мы пожертвовали основами статистической теории, деталями математических операций, сохраняя только совершенно необходимые, и всеми усовершенствованиями и уточнениями, которыми обычно сопровождаются основные статистические процедуры. Сохранены фундаментальные алгоритмы, входящие в каждый вид анализа, рассмотрение соотношений между различными количественными методами и простые примеры их применения в решении задач.

Тексты программ не приводятся в этом издании*, так как ими снабжены многие библиотеки программ, предназначенных для любых ЭВМ, начиная с суперкомпьютеров и кончая настольными микрокалькуляторами. Эти библиотеки содержат программы, значительно более совершенные и более гибкие, чем любая из программ, которую мы могли бы привести в этой книге. Однако, чтобы помочь читателю в овладении персональным компьютером, к английскому изданию книги приложена дискета, на которой записаны программы элементарной статистики и матричной алгебры. Дискета предназначена для персонального компьютера фирмы IBM-PC и совместима с машинами, использующими популярную операционную систему MS — DOS. Полная библиотека программ большинства процедур, обсужденных в этой книге, также доступна для большинства персональных компьютеров; информация по этому поводу приведена в Приложении.

Мы считаем, что методы количественного анализа в геологии могут быть весьма полезными в исследовательской работе: они дают не столько доказательства или подтверждения геологических гипотез, выработанных интуицией, сколько критическое исследование явления и проникновение в его сущность. Сбор данных соответствующего качества и в достаточном количестве для целей численного анализа приводит к более полноценному изучению объекта, чем другие способы исследования. Несомненно, что палентолог, тщательно измеряющий сотни образцов некоторого организма, может лучше оценить границы естественного изменения измеряемых характеристик, чем человек, который просто исследует их. Точность и объективность, требуемая количественной методологией, может отчасти компенсировать интуицию и опыт, которые вырабатываются годами работы. В то же время дисциплина, необходимая для выполнения количественных исследований, ускоряет творческий рост и наступление зрелости ученого.

Измерения и анализ данных могут привести к выводам, не вполне понятным или очевидным при использовании других

* В русском переводе этой книги мы сочли целесообразным привести полные тексты программ на языке ФОРТРАН, любезно предоставленные нам автором (Примеч. пер.).

методов исследования. Многомерные методы, например, позволяют объединять объекты в группы, которые находятся в согласии с принятыми классификациями, однако они могут указать на неожиданные соотношения между переменными. Иногда требуемое объяснение этого факта не может быть найдено, а в других случаях, наоборот, могут возникать новые теории, которыми иначе пренебрегли бы.

Возможно, что наибольший эффект от количественных методов заключается не в их способности показать, что верно, а скорее в том, чтобы показать, что неверно. Эти методы могут указать на недостаточность данных, обилие допущений, малое количество информации, на которой базируется большинство геологических исследований. При внимательном и беспристрастном анализе многие геологические выводы превращаются в набор догадок и предположений, основанных на очень незначительном количестве данных, большая часть которых имеет противоречивую и незаконченную форму. Если бы геология была экспериментальной наукой, подобно химии и физике, где наблюдения можно проверить на опыте, то упомянутые противоречия можно было бы устранить. Однако мы имеем дело с описательной наукой, и применение точных количественных методов часто напоминает нам о несовершенстве наблюдателей, каковыми мы являемся. В самом деле, склонность к научному скептицизму — одна из опасностей, которая часто подстерегает геоматематиков. Им часто свойственна подозрительная и противоборствующая позиция по отношению к установившимся в геологии традициям. Однако надо признать, что такой цинизм зачастую имеет оправдание. Геологи обучаются наблюдению образов и структур в природе. Геоматематические методы обеспечивают объективность, необходимую, чтобы избежать существования тех образов, создание которых оправдано только соображениями общего порядка.

ГЕОСТАТИСТИКА

Все методы количественного геологического анализа, рассматриваемые в этой книге, можно отнести к разряду статистических, иногда «квазистатистических» или «протостатистических» процедур. Большая часть их недостаточно разработана и не может быть использована при строгой проверке статистических гипотез. Ни один из этих методов нельзя считать адекватно отвечающим общей теории геологических совокупностей. Однако, подобно статистическим критериям, методы математической геологии основаны на предпосылке, что информацию о явлении можно получить в результате исследования малой выборки, отобранной из значительно большего множества потенциально возможных наблюдений изучаемого явления.

Рассмотрим задачу картирования глубинных структур при поисках месторождений нефти. Изучаемые при этом данные отбираются из скважин, разбросанных на некоторой площади и пронизывающих последовательность стратиграфических горизонтов. Единичное наблюдение представляет собой абсолютную отметку кровли горизонта, замеренную в одной из скважин. Если бы мы могли пробурить неограниченное число скважин, то это позволило бы получить бесконечное множество замеров абсолютной отметки кровли данного горизонта. Однако в действительности мы ограничены уже пробуренными скважинами и, возможно, если это будет оправданно, пробурим небольшое число дополнительных. По этим данным мы должны наилучшим образом описать конфигурацию кровли горизонта между скважинами. Эта задача аналогична статистическому анализу, но, в отличие от статистики, мы не можем составить выборочный план или контролировать способ, которым имеющиеся данные были получены. Однако мы можем использовать методы количественного картирования, которые тесно примыкают к статистическим процедурам, даже в случае если выполнены не все формальные статистические требования.

В противоположность этому можно рассмотреть также горно-проходческие работы и процесс эксплуатации месторождения. В течение многих лет горные инженеры и геологи разрабатывали детальные схемы опробования и бурения и проводили статистический анализ наблюдений. В последнее время число публикаций по теории опробования катастрофически растет. Их авторы, создавая теоретическую базу для применения формальных статистических критериев, предлагают для описания изменчивости содержаний руды ряд сложных статистических распределений. Там, где геологи контролируют отбор проб, они могут быстро выбирать удобную систему отработки. Их успех в разработках месторождений свидетельствует о силе этих методов.

К сожалению, большинство геологов вынуждены брать свои пробы только там, где это возможно. Данные по нефтяным скважинам слишком дороги для того, чтобы отбросить их только потому, что они не укладываются в схему опробования. Палеонтологи вынуждены довольствоваться ископаемыми остатками организмов, взятыми из обнажения, которые, будучи погребенными, никогда бы не были доступны исследованию. Пробы могут быть также отобраны из апикальных частей интрузивов, вскрытых в естественных обнажениях. Пробы из корневых частей тех же тел безнадежно глубоко скрыты в земной коре. Редки случаи, когда в одном месте собрано очень много данных. Чаще их бывает недостаточно. Наши наблюдения, связанные с исследованием Земли, слишком дороги, чтобы ими можно было пренебречь. Мы должны выяснить, какие сведе-

ния мы можем из них извлечь, изучить недостатки этих сведений и выявить имеющиеся тенденции.

Многие опубликованные работы посвящены вопросам планирования статистического эксперимента. Среди них наиболее интересна геологическая часть книги Гриффитса, в которой рассматривается вопрос о влиянии выборки на результаты использования статистических критериев. Хотя примеры Гриффитса взяты из осадочной петрологии, те же методы применимы в равной мере и к другим проблемам в науках о Земле. Книга дает строгую формальную интерпретацию геологических явлений, основанную на использовании статистических методов. Ее можно рекомендовать тем, кто при проведении геологического эксперимента может осуществить строгий контроль над процессом взятия проб. Так как эти вопросы вместе с вычислительной программой подробно освещены в книге Гриффитса [4], а также в руководстве Гриффитса и Ондрика [5], мы не будем касаться планирования экспериментов в этой книге. Вместо этого мы остановимся на более сложных ситуациях, когда схема взятия проб (либо случайно, либо по неведению) находится вне нашего контроля. Однако замечено, что неконтролируемый эксперимент (т. е. такой эксперимент, при котором исследователь не может влиять на места взятия проб), обычно выводит нас за рамки классической статистики. Это область «квазистатистики», или «протостатистики», где допущения формальной статистики не могут быть использованы безоговорочно. В этой области не существует вполне разработанных критериев проверки гипотез, и лучшее, на что мы можем надеяться, — это использование известных процедур во вспомогательных целях, причем в конечном счете выбор решения предоставляется исследователю.

СИСТЕМЫ ИЗМЕРЕНИЙ

Количественные методы в геологии требуют более глубоких знаний, чем поверхностные сведения, используемые при работе с компьютерами. Так как выводы, полученные с помощью количественных методов, основаны хотя бы частично на величинах, полученных в результате измерений, геолог должен иметь представление о природе систем чисел, в которых проводятся измерения. Ученый, исследующий Землю, должен не только понимать геологический смысл записываемых переменных, но также чувствовать математический смысл используемых шкал измерений. Эта тема более сложная, чем может показаться на первый взгляд. Ее подробное изложение и библиография приведены в книге Черчмена и Рэтуша [1]; геологическая сторона вопроса отражена в статье Гриффитса [3].

Измерение — это приписываемое наблюдению число, отражающее величину или значение некоторой характеристики.

Способ, которым наблюдению приписываются численные значения, определяет шкалу измерений. Последняя в свою очередь определяет тип анализа, который может быть осуществлен с этими данными. Существует четыре шкалы измерений, причем каждая последующая более точна, чем предыдущая, и более информативна. Первые две — это номинальная и порядковая шкалы, в которых измерения попросту классифицируются как две взаимно исключающие друг друга категории. Две последние шкалы — интервальная и шкала отношений — являются как раз тем, что мы обычно считаем «измерениями», так как они заключают в себе измерения величин признака.

Номинальная шкала измерений основана на классификации наблюдений во взаимно исключающие друг друга категории одинакового типа. Эти категории могут быть обозначены цветами, например, красный, зеленый, голубой или символами «А», «В», «С», или числами. Однако числа могут использоваться просто как идентификаторы, т. е. может не существовать соотношения «2 вдвое больше 1» или «5 больше 4». Классификация ископаемых остатков по типам — пример номинальных измерений. Отнесение одних ископаемых остатков к брахиоподам, а других — к криноидеям ничего не говорит об относительном значении или величине тех и других.

Можно сосчитать число наблюдений каждого типа в номинальной системе и затем использовать для их обработки какие-либо непараметрические критерии. Классический пример данных этого типа, который мы будем часто рассматривать ниже, — появление герба или решки при бросании монеты. Примером геологического эквивалента этих данных может служить появление зерен полевого шпата или кварца вдоль пересечения шлифа. Кварц и полевой шпат образуют две взаимно исключающие одна другую категории, которые нельзя никаким образом осмысленно ранжировать.

Иногда ранжировку наблюдений можно провести иерархическим способом. Шкала твердости Мооса — четкий пример порядковой шкалы. Хотя твердость минералов в шкале, имеющей десять делений (от 0 до 10), увеличивается с повышением ранга, разности между соседними уровнями различны. Различие между абсолютной твердостью алмаза (ранг 9) больше, чем различие между всеми остальными рангами (от 0 до 9). Аналогично метаморфические породы можно ранжировать по степени метаморфизма, которая отражает интенсивность метаморфического изменения. Однако переходы между разными уровнями не отражают единой закономерности изменения температуры и давления.

Как и для данных номинальной шкалы, количественный анализ порядковых измерений ограничен главным образом подсчетом числа наблюдений в различных состояниях. Однако

мы можем также рассмотреть способ, которым упорядочены различные порядковые классы, следующие один за другим. Это делается, например, при определении того, имеют ли состояния, появляющиеся необычное число раз, тенденцию следовать за наиболее или наименее частыми состояниями порядковой шкалы.

Интервальная шкала — шкала, где длина последовательных интервалов постоянна. Наиболее распространенный пример — температурная шкала. Увеличение температуры в интервале от 10 до 20 °С точно такое же, как увеличение между 110 и 120 °С. Однако интервальная шкала не имеет естественного нуля, или точки, где величина является несуществующей. Таким образом, отрицательные температуры — это просто температуры ниже условного нуля. Начальная точка отсчета для стоградусной шкалы была выбрана произвольно, как точка замерзания воды. В абсолютной шкале Кельвина точка 0 К обозначает температуру, при которой останавливается молекулярное движение. Никакая температура не может быть ниже этой. Таким образом, шкала Кельвина является не интервальной шкалой, а шкалой отношений.

Шкала отношений имеет не только одинаковые приращения между отдельными градациями, но и истинную нулевую точку. Измерения длины относятся к этому типу. Длина в две единицы вдвое превосходит длину в одну единицу. Не существует объектов нулевой длины. Общепринято, что отрицательных длин не может быть.

Шкала отношений — наивысшая форма измерений. С ее помощью можно осуществить все типы математических и статистических операций. Хотя интервальная шкала теоретически менее информативна, чем шкала отношений, для многих целей обе могут быть использованы с одинаковым успехом. Большинство геологических измерений осуществляется на шкале отношений, потому что они состоят из измерений длины, объема, массы и т. д. В следующих главах мы прежде всего будем касаться анализа интервальных и относительных данных. Между ними не будет делаться никакого различия; более того, их можно совместно использовать при решении одной и той же задачи. Такого рода пример встречается в анализе поверхностей тренда, где функция может быть измерена по шкале отношений, тогда как географические координаты, являющиеся аргументами, — по интервальной шкале, так как начало координатной сетки можно выбрать произвольно.

ЛОЖНАЯ УВЕРЕННОСТЬ

Эту главу можно было бы закончить следующим предупреждением. Применяя математические методы, можно запутаться в тех из них, которые имеют некую претензию на точность, в некотором приближении выражают существующие соотношения и основаны, как это принято считать, на непогрешимых процедурах. Заметим, что вычислительные машины можно использовать как очень эффективное средство запугивания. Так, представление численных массивов с точностью до восьми десятичных знаков обычно подавляет умы многих людей и утверждает их природный скептицизм. Геологический доклад, использующий математические термины и переполненный численными данными, обычно отпугивает всех, кроме немногих критиков, и даже те, кто понимает его и может дать объяснения, делают это на непрофессиональном уровне. Итак, и доклад, и критика проходят сквозь умы предполагаемой аудитории. Однако наибольшая опасность для исследователя находится в нем самом. Если ему приходится иметь дело с собственной вычислительной машиной, он может перестать критически анализировать свои данные и методы их интерпретации. Загипнотизированный числами, не видя ничего за пределами вычислительной лаборатории, он может прийти к самым нелепым заключениям. Необходимо всегда иметь в виду изречение, обычно имеющееся на стенах каждого вычислительного центра: «что посеешь, то и пожнешь».

Глава начиналась с одной цитаты, закончим ее другой. Следующие слова были оставлены на моем столе неизвестным критиком.

«Что может быть ограниченнее, чем ввод неверных данных в ЭВМ и наивная надежда получить уточненное наполеоновское решение?».

Майор Александр П. де Северски

СПИСОК ЛИТЕРАТУРЫ

1. *Churchman C. W.* and *P. Ratoosh,* eds., Measurement: definitions and theories: John Wiley and Sons, Inc., New York, 1959, 274 p.
2. *Fisher R. A.* The expansion of statistics: Jour. Royal Statistical Soc., 1953, series A, *116,* p. 1—6.
3. *Griffiths J. C.* Some aspects of measurement in the geosciences: Pennsylvania State Univ. Mineral Industries, 1960, *29,* no. 4, p. *1,* 4, 5, 8.
4. *Griffiths J. C.* Scientific method in analysis of sediments McGraw-Hill, Inc., New York, 1967, 508 p. Имеется русский перевод: *Гриффитс Дж.* Научные методы исследования осадочных пород. М., Мир, 1971.

Глава 2
ЭЛЕМЕНТАРНАЯ СТАТИСТИКА

Несмотря на то, что непосредственные наблюдения геологов ограничены поверхностью земной коры, с их помощью пытаются выяснить природу ядра и мантии, а также глубинных слоев земной коры. Кроме того, изучаются процессы эволюции Земли, такие, как горообразование и развитие континентов, недосягаемые для прямого исследования. Если не считать астрономов, то не существует другой группы ученых, более удаленных от предмета исследования и имеющих меньшие возможности экспериментировать над ним, чем геологи. Геология в значительной степени остается наукой, основанной на наблюдениях, в частности на таких, которые содержат большую долю неопределенности. В связи с этим статистике предназначено играть огромную роль в подобных исследованиях. Несмотря на то, что первоначально термин «статистика» имел отношение к подсчету числовых характеристик, как, например, при игре в бейсбол, сейчас она применяется и для анализа данных, особенно в условиях неопределенности. Статистические задачи скрыто или явно встречаются там, где содержится элемент случайности. Геологи должны иметь представление об этих задачах и о тех статистических приемах, которые позволяют их решать.

ВЕРОЯТНОСТЬ

Несмотря на то, что существует много руководств по теории вероятностей, представляется полезным остановиться на том, как определяется вероятное событие через возможное. В любом явлении существует некоторое множество (иногда бесконечное) возможных исходов, которые находятся в вероятностной зависимости, описываемой частотой встречаемости. Анализируя вероятности этих исходов, можно оценить поведение изучаемого объекта или события в прошлом или будущем.

Каждый из нас имеет интуитивное представление о вероятности. Например, если вас спросят, будет ли завтра дождь, вы ответите с некоторой степенью уверенности, что дождь либо возможен, либо невозможен, или в редком случае, что дождь будет обязательно, или, наоборот, что дождя наверняка не будет. Одним из способов выражения нашей уверенности является числовая шкала, например, процентная. Если мы полагаем, что вероятность того, что пойдет дождь 30%, то вероятность того, что дождя не будет, равна 70%.

Ученые обычно выражают вероятность числами от 0 до 1 или эквивалентным числом процентов от 0 до 100%. Если мы говорим, что вероятность дождя завтра 0, мы абсолютно уверены в том, что дождя завтра не будет. Однако если мы говорим, что вероятность дождя 1, мы абсолютно уверены в том, что дождь будет. Вероятность, выраженная таким способом, подобна утверждению о правдоподобии события. Абсолютной уверенности соответствуют крайние значения шкалы, а различной степени уверенности — промежуточные. Например, если мы говорим, что вероятность дождя завтра равна 1/2 (а значит, и того, что дождя не будет, тоже 1/2), мы выражаем свою точку зрения с наименьшей степенью уверенности. Если мы говорим, что вероятность дождя равна 3/4 (а того, что дождя не будет, 1/4), мы выражаем меньшую степень неуверенности, так как вероятность дождя в 3 раза больше вероятности того, что дождя не будет.

Наши оценки вероятности появления дождя могут быть основаны на многих факторах, а также на субъективных ощущениях. Мы, однако, будем использовать другой подход, позволяющий на основании предшествующего поведения явления, такого, например, как погода, предвидеть его поведение в будущем. Такой подход к определению вероятности через «относительные частоты» интуитивно приемлем геологами, так как эта концепция тесно связана с понятием однородности. В некоторых случаях бывают полезны другие подходы к определению вероятности, но мы не будем на них останавливаться, так как они подробно изложены в книгах Фишера [10] и фон Мизеса [31].

Пример с дождем является дискретным, так как здесь дождь может быть, а может и не быть. Классический дискретный пример, приводимый во всех руководствах по теории вероятностей, связан с бросанием правильной монеты. Одно бросание имеет два исхода: герб и решка, которые равновероятны. Поэтому вероятность выпадения герба равна 1/2. Это, конечно, означает не то, что в каждом втором бросании выпадет герб, а скорее то, что при достаточно большом числе бросаний приблизительно половина исходов — гербы. Бросание монеты — очень хороший пример для иллюстрации определения вероятности. Событие имеет два исхода, и один из них обязательно будет иметь место; если не учитывать очень малую вероятность того, что монета может встать ребром, то всегда выпадет либо герб, либо решка.

На основе схемы бросания монеты можно построить интересную последовательность вероятностей. Если вероятность выпадения герба равна 1/2, то вероятность выпадения двух гербов при двух бросаниях равна $1/2 \cdot 1/2 = 1/4$. Далее вероятность выпадения трех гербов при трех бросаниях равна

$1/2 \cdot 1/2 \cdot 1/2 = 1/8$. Обоснование этой прогрессии простое. В первом бросании вероятность выпадения герба 1/2. Так как исход второго бросания не зависит от первого, наши возможности получить герб во втором бросании снова 1/2. Подобно этому, третье бросание не зависит от предыдущих, и вероятность выпадения герба в третьем бросании снова равна 1/2. Итак, вероятность выпадения трех гербов составляет «половину половины половины».

Предположим, что теперь нам нужно определить вероятность выпадения только одного герба в трех бросаниях. Имеются следующие возможные исходы (Г — герб; Р — решка):

ГГГ	ГРГ	РРР
ГГР	РГГ	[РГР]
[ГРР]	[РРГ]	

В скобках взяты те комбинации, которые удовлетворяют нашим требованиям, т. е. они содержат только один герб. Так как имеется 8 различных комбинаций, вероятность получения только одного герба при трех бросаниях равна 3/8.

Полученное нами число представляет собой число возможных комбинаций из трех по одному. Обобщая этот пример, можно определить число возможных сочетаний из n элементов по r. Это число символически изображается так $\binom{n}{r}$.

Можно доказать, что число сочетаний из n элементов по r равно

$$\frac{n!}{r!(n-r)!}. \qquad (2.1)$$

Восклицательный знак обозначает факториал, и $n!$ есть произведение n последовательных целых чисел:

$$n! = 1 \cdot 2 \cdot 3 \cdots n. \qquad (2.2)$$

Значение 3! равно $3 \cdot 2 \cdot 1 = 6$. В нашей задаче

$$\binom{3}{1} = \frac{3!}{1!(3-1)!} = \frac{3 \cdot 2 \cdot 1}{1 \cdot (2 \cdot 1)} = \frac{6}{2} = 3,$$

т. е. имеется три возможные комбинации, содержащие один герб. Так же вычислим число возможных комбинаций, содержащих два герба:

$$\binom{3}{2} = \frac{3!}{2!(3-2)!} = \frac{3 \cdot 2 \cdot 1}{2 \cdot 1(!)} = \frac{6}{2} = 3$$

ГГГ	[ГРГ]	РРР
[ГГР]	[РГГ]	РГР
ГРР	РРГ	

Требуемые комбинаци взяты в скобки. Наконец, сколько возможных комбинаций при трех бросаниях содержат три герба?

$$\binom{3}{3} = \frac{3!}{3!(3-3)!} = \frac{3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1(1)} = 1.$$

Заметим, что по определению число 0! равно не нулю, а 1. Наконец, остается возможность, когда при трех бросаниях не выпадает ни одного герба, и число таких комбинаций будет равно

$$\binom{3}{0} = \frac{3!}{0!(3-0)!} = \frac{3 \cdot 2 \cdot 1}{1(3 \cdot 2 \cdot 1)} = 1.$$

Таким образом, при трех бросаниях монеты имеется одна возможность не получить ни одного герба, три возможности получить один герб, три возможности получить два герба и одна возможность получить три герба. Этот результат можно изобразить графически, как это сделано на рис. 2.1.

Мы можем подсчитать общее число возможных исходов, которое равно восьми, а затем преобразовать частоты появления отдельных событий в вероятности. Иначе говоря, вероятность отсутствия гербов при трех бросаниях, т. е. вероятность выпадения одной комбинации среди восьми возможных, равна 1/8. Преобразуем теперь гистограмму рис. 2.1, откладывая по вертикальной оси вместо числа комбинаций вероятности соответствующих событий. Мы получим дискретное распределение вероятностей, график которого представлен на рис. 2.2. Общая сумма всех вероятностей равна 8/8 или 1. Таким образом, событие, заключающееся в том, что какая-либо комбинация при трех бросаниях осуществится, является достоверным. Гра-

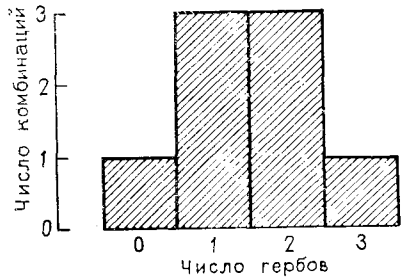


Рис. 2.1. Заштрихованная область на графике показывает число различных способов получения заданного числа гербов при трех бросаниях монеты

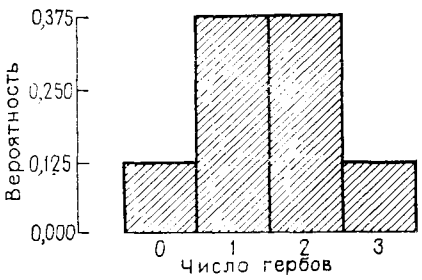


Рис. 2.2. Дискретное распределение, показывающее вероятность получения заданного числа гербов при трех бросаниях монеты

фик функции распределения характеризует вероятность выпадения какой-либо заданной комбинации.

Эксперимент бросания монеты обладает четырьмя особенностями:

1) в каждом испытании (или бросании) имеется только два возможных исхода (назовем их «успех» или «неудача»);

2) исход каждого испытания не зависит от предыдущих исходов;

3) вероятность успеха не меняется от испытания к испытанию;

4) испытания повторяются заданное число раз.

Распределение вероятностей, соответствующее указанному нами типу эксперимента, называется биномиальным распределением.

В качестве примера геологических приложений можно упомянуть предсказание вероятности успеха при бурении нефтяных и газовых скважин. Четыре перечисленных выше характеристики должны предполагаться справедливыми: допущения подобного рода кажутся наиболее приемлемыми при использовании метода «дикой кошки» при разведке в сравнительно мало изученном бассейне. Следовательно, биномиальное распределение часто используется для предсказания результатов бурения на границе области или за пределами концессии.

В предположении о справедливости биномиального распределения каждый результат применения метода «дикой кошки» может быть классифицирован либо как «открытие» («успех») либо как пустая скважина («неудача»). Последовательные испытания предполагаются независимыми, т. е. успех или неудача в одной скважине не влияет на результат опробования в другой скважине. (Это допущение трудно обосновать по многим обстоятельствам, так как открытие обычно влияет на результаты исследования следующих скважин. Появление последовательности пустых скважин также приводит к изменению программы разведки.). Наконец, биномиальное распределение оказывается подходящим, если фиксированное число скважин будет пробурено в процессе разведки или в течение единичного периода времени (например, бюджетного цикла), для которого проведено прогнозирование.

За вероятность p открытия нефтяного или газового месторождения методом «дикой кошки» принимаются широко используемые в промышленности отношения, характеризующие успех бурения, которые были наблюдены в процессы бурения в аналогичных регионах, или аналогичные отношения, полученные компаниями, делающими оценки, или же просто субъективные аналогичные характеристики. Для определения вероятности p в биномиальной модели в разведочном бурении предусмотрены следующие этапы:

1) вероятность «успеха» бурения скважины обозначим через p,

2) вероятность «неудачи» бурения обозначим через $1-p$,

3) вероятность того, что в результате последовательного бурения n скважин методом «дикой кошки» все окажутся пустыми, есть

$$P = (1-p)^n;$$

4) вероятность того, что n-я пробуренная скважина окажется продуктивной, причем $n-1$ предыдущих скважин окажутся пустыми, равна

$$P = (1-p)^{n-1}p;$$

5) вероятность появления одной продуктивной скважины в последовательности n скважин, пробуренных методом «дикой кошки», есть

$$P = n(1-p)^{n-1}p;$$

так как продуктивной может оказаться любая из n скважин;

6) вероятность того, что будет пробурено $n-r$ пустых скважин, за которыми последует r продуктивных, есть

$$P = (1-p)^{n-r}p^r;$$

7) однако $n-r$ пустых скважин и r продуктивных скважин могут быть выбраны $\binom{n}{r}$ способами, или, что эквивалентно, $\frac{n!}{(n-r)!\,r!}$ различными способами. Так, вероятность того, что в программе бурения методом «дикой кошки» среди n скважин получим r продуктивных, равна

$$P = \frac{n!}{(n-r)!\,r!}(1-p)^{n-r}p^r. \qquad (2.3)$$

Это — формула биномиального распределения, задающая вероятность получения r успехов в последовательности n независимых испытаний, вероятность успеха в каждом из которых равна p.

Равенство (2.3) можно понимать как уравнение, которое может быть решено, т. е. может быть определена вероятность появления любой заданной комбинации успехов и неудач при любом заданном числе испытаний и заданной вероятности. Эти вероятности уже вычислены и затабулированы для многих комбинаций n, r и p. Используя это уравнение или опубликованные таблицы, такие, как, например, таблицы в книге Хальда [15], можно дать ответ на многие вопросы. Например, предположим, что мы хотим определить вероятности, соответствующие программе разведки в малоизученном бассейне, содержа-

щем пять скважин, причем отношение, характеризующее успех в этом регионе, равно примерно 10%. Какова вероятность того, что вся программа будет полностью безуспешной и мы не получим ни одной продуктивной скважины? Такой результат называется полным крахом по очевидным причинам, и биномиальное выражение содержит следующие члены

$$n = 5; \quad r = 0; \quad p = 0{,}10;$$

$$P_5 = \binom{5}{0} \cdot 0{,}10^0 \cdot (1 - 0{,}10)^5 = \frac{5!}{5!0!} \cdot 1 \cdot 0{,}90^5 = 1 \cdot 1 \cdot 0{,}59 = 0{,}59.$$

Вероятность того, что в результате разведки не будет получено никаких продуктивных скважин, равна почти 60%.

Если только одна скважина окажется продуктивной, то ею могут окупаться все затраты в процессе разведки. Какова вероятность того, что одна скважина окажется продуктивной в процессе разведки с пятью скважинами?

$$P = \binom{5}{1} \cdot 0{,}10^1 \cdot (1 - 0{,}10)^4 = \frac{5!}{4!1!} \cdot 0{,}10 \cdot 0{,}90^4 = 5 \cdot 0{,}10 \cdot 0{,}656 = 0{,}328.$$

Используя либо биномиальное уравнение либо таблицу биномиального распределения, можно найти вероятности всех возможных исходов в программе бурения с пятью скважинами. Они представлены на рис. 2.3.

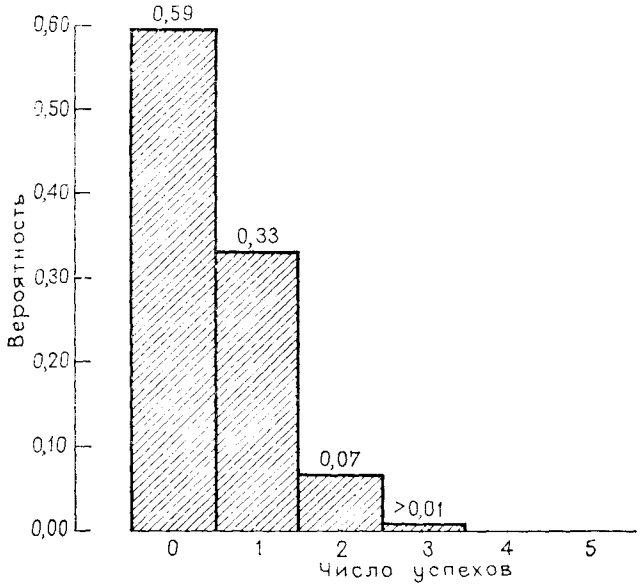


Рис. 2.3. Дискретное распределение, задающее вероятность успеха в программе бурения пяти скважин при условии, что вероятность успеха равна 10%

В других практических задачах, в которых основные допущения отличны от указанных выше, могут быть использованы другие дискретные вероятностные распределения. Предположим, например, что геологоразведочная компания намеревается открыть два новых поля в мало изученном бассейне, предполагаемом перспективным, и намеревается пробурить столько скважин, сколько потребуется для того, чтобы достичь этой цели. Мы можем исследовать вероятность того, что для этого потребуется 2,3,4..., до n разведочных скважин до того как будет обнаружено две продуктивных скважины. Можно допустить, что требуется наложить те же условия, которые были перечислены для биномиального распределения, исключая относящееся к числу испытаний условие, которое теперь не является фиксированным.

Вероятностное распределение, соответствующее такому эксперименту, называется отрицательным биномиальным, и его определение очень напоминает определение биномиального распределения. Как и в предыдущем примере, p есть вероятность появления продуктивной скважины и r есть число «успехов» или продуктивных скважин. Однако n, число испытаний, не задано. Вместо этого можно найти вероятность того, что будет пробурено x пустых скважин раньше, чем будет сделано r открытий. Отрицательное биномиальное распределение имеет вид

$$P = \binom{r + x - 1}{x} (1 - p)^x p^r. \tag{2.4}$$

Заметим сходство между этим уравнением и уравнением (2.3); член $r+x-1$ появляется потому, что последняя из последовательно пробуриваемых скважин, должна быть r-м успехом. Формулу (2.4) можно представить в следующем виде:

$$P = \frac{(r + x - 1)!}{(r - 1)!\, x!} (1 - p)^x p^r. \tag{2.5}$$

Предполагая, что вероятность успеха, соответствующая данному региону, равна 10%, мы можем вычислить вероятность того, что геологоразведочная программа с двумя скважинами будет завершена открытием двух продуктивных скважин. Эта вероятность равна

$$P = \frac{(2 + 0 - 1)!}{(2 - 1)!\, 0!} \cdot (1 - 0{,}10)^0 \cdot 0{,}10^2 =$$

$$= \frac{1!}{1!0!} \cdot 0{,}90^0 \cdot 0{,}10^2 = 1 \cdot 1 \cdot 0{,}01 = 0{,}01.$$

Вероятности, связанные с другими программами бурения, имеют различные числа скважин, и их можно найти аналогичным

образом. Вероятность того, что для достижения успеха в двух испытаниях требуется пять скважин, равна

$$P = \frac{(2+3-1)}{(2-1)!\,3!}\cdot(1-0{,}10)^3\cdot 0{,}10^2 = \frac{24}{1\cdot 6}\cdot 0{,}729\cdot 0{,}01 = 0{,}029.$$

Вычисленные вероятности низки потому, что они связаны с возможностью появления двух продуктивных и в точности x пустых скважин. Полезнее рассмотреть распределение вероятности того, что более x скважин должно быть пробурено до того как цель, состоящая в появлении r продуктивных скважин, будет достигнута. Эту вероятность можно получить, если сначала построить отрицательное биноминальное распределение в кумулятивной форме, дающее вероятность того, что будет достигнуто два успеха при бурении ($x+r$ или менее скважин), как это изображено на рис. 2.4. Вычитая далее каждую из этих вероятностей из 1,0, получим желаемое распределение вероятностей (рис. 2.5). Отрицательное биномиальное распределение будет рассмотрено снова в гл. 5, где с его помощью будет описана важная модель распределения точек в пространстве.

Имеются и другие дискретные распределения, которые применяются в практических задачах, аналогичных тем, для описания которых используется биномиальное распределение. Таково, например, распределение Пуассона, которое следует использовать вместо биномиального, если p, вероятность успеха, очень мала. Пуассоновское распределение рассматривается в гл. 4, где оно применено к анализу редких случайных событий во времени (землетрясений или вулканических изменений), и в гл. 5, оно использовано для построения моделей объектов, случайно размещенных в пространстве. Геометрическое распределение, являющееся частным случаем отрицательного биномиального, используется в тех случаях, когда интерес направлен на число испытаний, предшествующих первому успеху. Полиномиальное распределение — это обобщение биномиального распределения на случай, когда имеется более двух взаимно исключающих исходов эксперимента. Эти вопросы широко представлены в большинстве книг по теории вероятностей, таких, как книги Парзена или Аша [3].

Важным свойством всех упомянутых выше дискретных распределений является то, что вероятность успеха остается постоянной в процессе испытаний. В статистике рассматриваются простые эксперименты, называемые выборкой с возвращением, в которых это условие строго выполнено. Наиболее типичный эксперимент такого рода можно провести с урной, заполненной красными и белыми шарами; если наугад выбирается один шар, то вероятность того, что он будет красным, равна отношению числа таких шаров в урне к общему числу шаров. Если затем этот шар возвращается в урну, то процентное отношение двух цветов в урне остается неизменным, и вероятность появления красного шара при повторном выборе остается неизменной. Эта вероятность также остается приблизительно постоянной, если имеется очень большое число шаров в урне, даже в том случае, если делается выбор без возвращения, так как выбор шара приводит к бесконечно малому изменению пропорций среди остающихся шаров. Это последнее условие обычно предполагается выполненным при геологических поисках, когда естественно использовать дискретные распределения. В нашем примере «урна» — это геологический бассейн, в котором проводятся работы, а красным и белым шарам соответствуют неоткрытые залежи и пустые площади. До тех

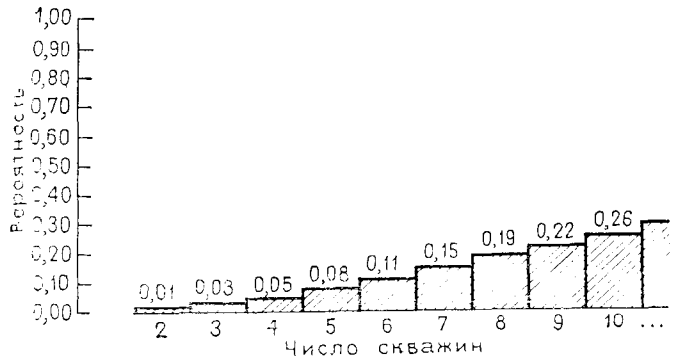


Рис. 2.4. Дискретное распределение, задающее кумулятивную вероятность того, что будут наблюдаться два успеха во время бурения (или до него) данной скважины, если вероятность успеха равна 10%

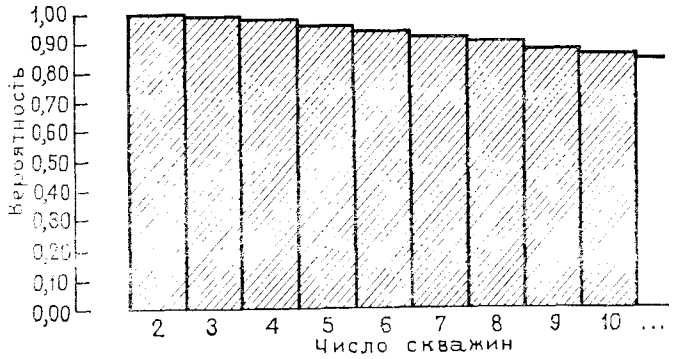


Рис. 2.5. Дискретное распределение, дающее вероятность того, что будет пробурено больше заданного числа скважин, для того, чтобы получить два успеха, при условии, что вероятность успеха равна 10%

пор, пока число не подвергнутых бурению участков велико, а число оправдавших себя в результате бурения прогнозов (шаров, извлеченных из урны) мало, предположение о постоянстве вероятности открытия кажется оправданным. Однако если выборочный эксперимент осуществляется с малым исходным числом раскрашенных шаров в урне, и они выбираются из урны без возвращения, то очевидно, что вероятность изменяется после каждого выбора. Такой эксперимент, называемый выбором без возвращения, описывается дискретным гипергеометрическим распределением. Геологические задачи, в которых целесообразно его использование, весьма специфичны, и Мак Грей [25] приводит пример, связанный с геофизическими методами разведки нефтяных месторождений.

В некоторых случаях можно определить размер совокупности, в пределах которой ведется поиск. Предположим, что концессия в открытом море обнаружила десять явных аномалий, причиной которых являются, возможно, перемещения соли на глубине. Из опыта предыдущих исследований следует, что около 40% таких сейсмических аномалий указывают на наличие продуктивных структур. При реализации программы разведочных работ в силу ограниченности средств нет возможности осуществить бурение в местах проявления аномалий. Для оценки вероятности того, что будет обнаружено заранее заданное число залежей, если только некоторые из предсказанных перспективных площадей будут разбурены, может быть использовано гипергеометрическое распределение.

Биномиальное распределение для решения этой задачи непригодно, так как вероятность обнаружения изменяется при каждом случайном испытании. Если имеется четыре залежи, каждая из которых входит в число десяти площадей, предсказанных сейсмическими методами, то открытие одной из них уменьшает шансы открытия другой, так как остается меньше объектов, которые должны быть обнаружены. Наоборот, обнаружение пустой скважины на основе сейсмических данных увеличивает вероятность того, что оставшиеся непроверенными структуры будут продуктивными, потому что одна непродуктивная структура уже исключена из совокупности.

Вычисление гипергеометрической вероятности состоит просто в нахождении всех возможных комбинаций продуктивных и пустых структур внутри совокупности и затем в перечислении тех комбинаций, которые дают требуемое число обнаружений. Вероятность осуществления x обнаружений при бурении n скважин, если делается выборка из совокупности N элементов, S из которых, возможно, содержат залежь, равна

$$P = \binom{S}{x}\binom{N-S}{n-x} \Big/ \binom{N}{n}. \qquad (2.6)$$

Это есть произведение числа комбинаций из числа залежей по числу обнаружений на число комбинаций из пустых аномалий по числу пустых скважин, деленное на число комбинаций из всех предсказанных объектов по общему числу скважин в программе бурения.

Гипергеометрическое вероятностное распределение может быть применено к нашей концессии в открытом море, которая содержит десять сейсмических аномалий, четыре из которых, возможно, содержат залежи. К сожалению, мы заранее не знаем, бурение каких из четырех аномалий даст положительный результат. Если в текущем сезоне бюджетные ограничения позволяют осуществить бурение только трех скважин, мы можем определить вероятности, связанные с каждым из возможных исходов.

Какова вероятность того, что программа бурения потерпит полный провал, т. е. не будет получено ни одной продуктивной скважины среди трех пробуренных?

$$P = \frac{\binom{4}{0}\binom{6}{3}}{\binom{10}{3}} = \frac{1 \cdot 20}{120} = 0{,}167.$$

Вероятность полного краха равна примерно 17%.

Какова вероятность того, что будет сделано одно обнаружение?

$$P = \frac{\binom{4}{1}\binom{6}{2}}{\binom{10}{3}} = \frac{4 \cdot 15}{120} = 0{,}50.$$

Вероятность того, что будет найдена одна продуктивная скважина, равна 50%.

Можно построить гистограмму, на которой будут представлены все вероятности, связанные со всеми возможными исходами в этой разведочной задаче (рис. 2.6). Заметим, что вероятность успеха равна (1,00—0,17), или 83%.

Предыдущий пример относился к случаю, когда имеется только два возможных исхода: пустая скважина или обнаружение нефти. Если нефть найдена, скважина не может быть пустой, и обратно. События, в которых частота появления одного исхода позволяет предсказать частоту появления другого исхода, называются взаимно исключающими друг друга. Вероятность того, что произойдет одно или другое событие, есть сумма вероятностей появления каждого из них; т. е. p (обнаружения или пустой скважины) $= p$ (обнаружения) $+ p$ (пустой скважины). Это — правило сложения вероятностей.

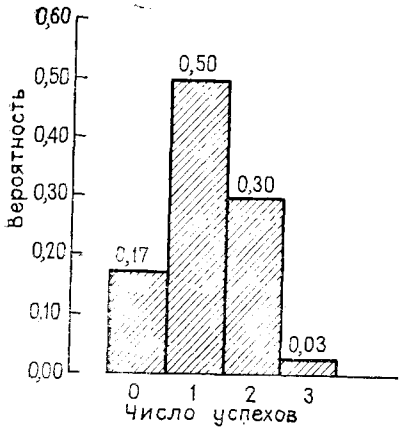


Рис. 2.6. Дискретное распределение, задающее вероятность n успехов в трех скважинах из десяти, если четыре из десяти пересекли нефтяную залежь

События не обязательно являются взаимно исключающими. Например, при разведочном бурении на нефть и газ появление залежи с пористым песчаником можно интерпретировать как антиклинальную структуру, исходя из сейсмических данных. Эти два исхода, пористый песчаник и бурение по антиклинали, не являются взаимно исключающими друг друга, так как мы встречаем их одновременно. Поскольку наличие песчаника управляется факторами, которые действовали во время образования месторождения, и поскольку появление антиклинальной складки предполагается связанным с тектоническими условиями в более поздний период, эти два исхода являются несвязанными или независимыми. Если два события не являются исключающими, но являются независимыми, то совместная вероятность того, что они появятся одновременно, есть произведение вероятностей их появления. То есть p (песчаника или антиклинали) $=p$ (песчаника) $\times p$ (антиклинали). Это — правило умножения вероятностей.

Два события могут быть связаны некоторым образом, так что исход одного отчасти зависит от исхода другого. Совместная вероятность таких событий называется условной. Такие события очень важны в геологии, так как одно из событий может быть прямо доступно наблюдению, а другое при этом может быть скрыто. Если они обусловлены друг другом, то частота наблюдаемого события говорит нам кое-что о вероятном состоянии скрытого объекта. Например, извержение магмы из камеры вулкана, аналогичного вулкану Святой Елены в Вашингтоне, по общему мнению, приводит к гармоническим колебаниям, напоминающим землетрясение. Мы не можем прямо наблюдать активность магматической камеры, однако мы можем наблюдать записи сейсмической активности, связанной с вулканом. Если между этими явлениями существует взаимная

обусловленность, то появление гармонических колебаний может помочь предсказать извержение. Если p (колебаний) — вероятность того, что появятся гармонические колебания, и p (извержения) — вероятность последующего вулканического извержения, то p (колебаний или извержения) $\neq p$ (колебаний) $\times p$ (извержения), если два события имеют взаимную обусловленность.

Условная вероятность того, что произойдет извержение, заданная тем, что гармонические колебания были записаны, обозначается p (извержения/колебания). В этом примере условная вероятность извержения больше, чем безусловная вероятность или p (извержения), которая есть просто вероятность того, что произойдет извержение без какой-либо дополнительной информации о других событиях. Другие условные вероятности могут быть меньше, чем соответствующие безусловные вероятности (вероятность нахождения некоторого ископаемого на площади вулканического происхождения значительно ниже, чем безусловная вероятность нахождения ископаемого). Очевидно, геологи используют условные вероятности на всех стадиях проведения своих работ, делая это сознательно или нет.

Связь между условными и безусловными вероятностями можно выразить теоремой, названной в честь английского священника, жившего в XVIII в. и открывшего ее, теоремой Байеса. Он исследовал изменение вероятности по мере увеличения количества информации. Основное уравнение Байеса имеет вид

$$P(A, B) = P(B|A)P(A). \tag{2.7}$$

Оно устанавливает, что совместная вероятность $P(A, B)$ того, что оба события A и B произойдут, равна вероятности $P(B/A)$ появления события B при условии, что событие A уже произошло, умноженной на вероятность $P(A)$ появления события. $P(B/A)$ называется условной вероятностью, так как выражает вероятность того, что произошло событие B, при условии, что событие A уже произошло. Если события A и B связаны (или зависимы), то тот факт, что событие A уже произошло, говорит нам нечто о правдоподобности появления события B.

Наоборот, также справедливо равенство

$$P(A, B) = P(A|B)P(B).$$

Приравнивая правые части, получаем

$$P(B|A)P(A) = P(A|B)P(B),$$

что можно переписать в виде

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}. \tag{2.8}$$

Это очень полезное соотношение, так как иногда мы знаем условную вероятность в одной форме, но интересуемся и другой. Например, известно, что рудные поля часто характеризуются наличием геомагнитных полей, отличных от нормальных. Однако мы больше интересуемся обратным, а именно, какова вероятность того, что некоторое поле окажется минерализованным, причем последнее обусловлено наличием магнитной аномалии. Нам проще оценить условную вероятность p (аномалия/минерализация) и безусловную вероятность p (минерализации), исходя из изучения известных рудных районов, однако труднее прямо оценить p (минерализации/аномалии), так как это может потребовать исследования геомагнитных аномалий которые, возможно, еще не были предсказаны.

Предположим, что мы имеем n несовместимых событий B_1, B_2, ..., B_n, которые обусловливают событие A, тогда вероятность осуществления события A есть попросту сумма условных вероятностей $P(A/B_j)$, умноженных на вероятности событий B_j, то есть

$$P(A) = \sum_{j=1}^{n} P(A \mid B_j) P(B_j). \qquad (2.9)$$

Если (2.9) подставить вместо $P(A)$ в уравнение Байеса, в форме (2.8), то мы получим более общее уравнение

$$P(B_j \mid A) = \frac{P(A \mid B_j)P(B_j)}{\sum_{j=1}^{n} P(A \mid B_j)P(B_j)}. \qquad (2.10)$$

Простой пример с двумя возможными независимыми исходами иллюстрирует теорему Байеса.

В русле потока в Западном Канзасе был найден фрагмент до сих пор неизвестного вида мезозавров, и специалист по палеонтологии позвоночных счел целесообразным послать студенческую полевую партию для поиска более полного набора ископаемых остатков. К сожалению, место обнаружения уже найденного фрагмента нельзя было идентифицировать с полной уверенностью, так как ископаемое было найдено ниже слияния двух сухих русел. Площадь дренажного бассейна более крупного потока около 18 условных единиц, в то время как площадь бассейна, дренируемого меньшим руслом, около 10 условных единиц. На основе только такой информации, мы можем постулировать, что вероятность того, что фрагмент принесен из какого-либо одного дренажного бассейна, пропорциональна площади этого бассейна или

$$P(B_1) = \frac{18}{28} = 0{,}64, \quad P(B_2) = \frac{10}{28} = 0{,}36.$$

Однако исследование геологического доклада и карты региона дает дополнительную информацию о том, что около 35% обнажений меловых пород в большем бассейне являются морскими, в то время как в меньшем бассейне на их долю приходится почти 80% обнажений меловых пород. Мы можем поэтому постулировать условную вероятность того, что если ископаемый фрагмент принесен из бассейна B, он является морским ископаемым, в соответствии с процентным соотношением меловых морских обнажений в бассейне, т. е. имеет морское происхождение, или для бассейна B_1 $P(A|B_1) = 0{,}35$ и для бассейна B_2 $P(A|B_2) = 0{,}80$.

Используя эти вероятности и теорему Байеса, мы можем подсчитать условную вероятность того, что ископаемый фрагмент принесен из бассейна B_1, т. е. того, что его происхождение морское, по формуле

$$P(B_1 \mid A) = \frac{P(A \mid B_1)P(B_1)}{P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2)} = \frac{(0{,}35)(0{,}64)}{(0{,}35)(0{,}64) + (0{,}80)(0{,}36)} = 0{,}44.$$

Аналогично, вероятность того, что ископаемый фрагмент принесен из меньшего бассейна, можно подсчитать по формуле

$$P(B_2 \mid A) = \frac{(0{,}80)(0{,}36)}{(0{,}35)(0{,}64) + (0{,}80)(0{,}36)} = 0{,}56.$$

К счастью для студентов, которые должны исследовать площадь, представляется несколько более правдоподобным, что фрагмент морского ископаемого мезозавра принесен из меньшего бассейна, а не из большего. Однако различия в вероятностях очень малы, и, конечно, зависят от обоснованности допущений, используемых для оценки вероятностей.

Для того чтобы перейти к следующей теме, мы должны вернуться к биномиальному распределению. На рис. 2.2 представлено вероятностное распределение для всех возможных чисел выпадения гербов и решек при трех бросаниях монеты. Аналогичный эксперимент можно осуществить при большем числе испытаний. На рис. 2.7, например, представлены вероятности получения заданного числа «успехов» (гербов) в десяти бросаниях монеты, а на рис. 2.8 — вероятностное распределение, которое описывает результаты эксперимента, состоящего из 50 бросаний монеты. Все эти вероятности были получены из таблиц биномиального распределения или могут быть вычислены из биномиального уравнения.

В каждом из этих экспериментов вычислялись все возможные числа гербов, которые можно получить, начиная с 0 до 3, 10, 50. Никакие другие комбинации гербов и решек не могут

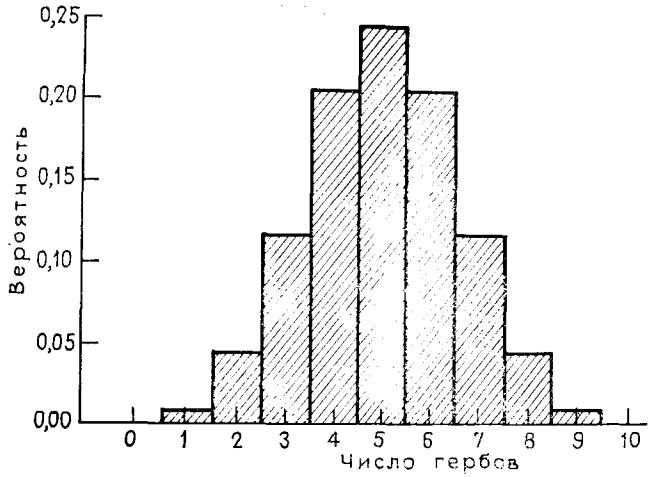


Рис. 2.7. Дискретное распределение, показывающее вероятность получения заданного числа гербов при десяти бросаниях монеты

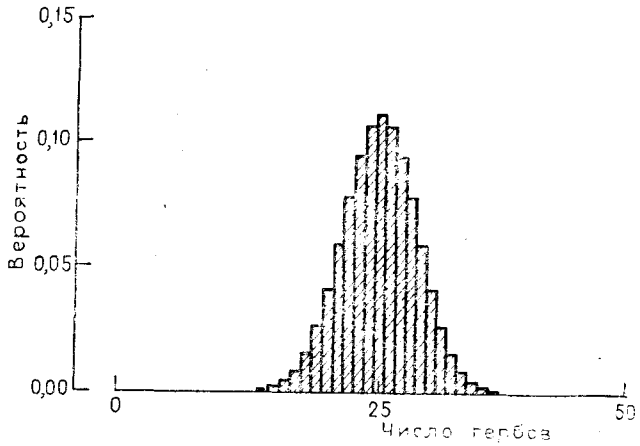


Рис. 2.8. Дискретное распределение, показывающее вероятность получения заданного числа гербов при 50 бросаниях монеты

встретиться. Так как мы обязательно получаем какой-либо результат из перечисленных выше, то сумма всех вероятностей в каждом эксперименте должна равняться 1,00. Это удобно представить, полагая площади под гистограммами на рис. 2.7 и 2.8 равными 1,00, как это сделано на гистограмме, изображенной на рис. 2.2. При таком условии увеличивающееся число бросаний монеты будет сопровождаться только суженнем ширины

полос. Гистограмма становится все более напоминающей гладкую и непрерывную кривую. Можно представить себе эксперимент, состоящий в бесконечном числе бросаний монеты, в результате которого будет получено бесконечное число полос бесконечно малой ширины. Тогда гистограмма превратится в непрерывную кривую, и горизонтальная ось будет представлять скорее непрерывную, чем дискретную переменную.

В эксперименте бросания монеты мы имеем дело с дискретными исходами, т. е. со специфической комбинацией гербов и решек. Однако в большинстве экспериментов, встречающихся на практике, возможные исходы не являются дискретными. Обычно имеется бесконечный континуум возможных исходов, которые могут быть получены. Множество возможных исходов может быть конечным и на самом деле ограниченным, но в пределах этого множества результат, который может быть получен, нельзя предсказать точно. В данном случае мы имеем дело с непрерывными случайными переменными. Предположим, например, что измеряется длина замочного края раковины брахиоподы и установлено, что она равна 6 мм. Однако если провести измерение, используя бинокулярный микроскоп, то можно получить длину 6,2 мм, а при использовании компаратора — 6,23 мм, и, наконец, используя сканирующий электронный микроскоп, получим 6,231 мм. Непрерывная переменная теоретически может бесконечно подразделяться. Это является следствием того факта, что всегда можно найти разность между двумя измерениями, если проводить измерения в достаточно мелкой шкале. Следствие этого утверждения таково, что каждый исход в непрерывной шкале измерений уникален и что вероятность получения конкретного точного результата равна нулю.

Если это так, то кажется невозможным определить вероятность на основе относительных частот встречаемости. Однако хотя невозможно наблюдать исходы, которые в точности соответствуют 6,000...000 мм, вполне доступно получить ряд измерений, попадающих внутрь некоторого интервала, включающего это значение. При том, что индивидуальные изменения в точности не идентичны, они достаточно близки и можно считать их принадлежащими одному классу. В итоге разобьем непрерывную шкалу на дискретные сегменты, и тогда можно подсчитать число событий, попавших внутрь каждого интервала. Сужая границы класса, мы уменьшим число событий в нем, и снизим оценки вероятностей появления события.

Если мы имеем дело с дискретными событиями, то определяем значения с абсолютной точностью. Непрерывные переменные, однако, измеряются с помощью некоторых физических процедур, которые ограничивают точность их измерения. В повторных измерениях, сделанных на одном и том же объ-

екте, возникают малые отклонения, величина которых отражает как естественные изменения объекта, так и изменения в условиях проведения измерений и, кроме того, изменения, обусловленные деятельностью исследователя, производящего измерения. Единственное, точное, «истинное» значение не может быть определено; иными словами, мы наблюдаем непрерывное распределение возможных значений. Такие свойства присущи непрерывной случайной переменной.

Для иллюстрации непрерывных случайных величин рассмотрим задачу определения показателя проницаемости образцов из керна скважины. Проницаемость определяется временем, требуемым для проникновения заданного количества флюида при стандартных условиях через образец породы. Допустим, что в результате одного определения получена проницаемость, равная 0,108 мкм^2. Является ли это число «истинной» проницаемостью пробы? Другие определения на этом же образце могут дать проницаемость, равную 0,093 и 0,112 мкм^2. На проницаемость, записываемую приборами в ходе любого эксперимента, влияет ряд условий, которые внутри прибора неизбежно изменяются от одного определения к другому в результате капризов потока и его турбулентности и не зависят от действий оператора. Ни одно из полученных значений нельзя взять в качестве абсолютной меры истинной проницаемости. В итоге различные источники флуктуации порождают непрерывную случайную величину, которую мы подвергаем опробованию, делая повторные измерения.

Изменчивость, обусловленная неточностью инструментов, более очевидна, когда делаются повторные измерения на единичном объекте, т. е. испытания повторяются без изменений. Такую изменчивость называют ошибками эксперимента. Кроме этого, изменчивость может проявляться в последовательности измерений или результатов экспериментов, проводимых на ряде изучаемых объектов. Обычно именно эта изменчивость и представляет научный интерес. Довольно часто оба эти типа изменчивости так перепутаны или совмещены, что экспериментатор не может определить, какая часть изменчивости возникает в силу различий между условиями испытаний, а какая является следствием ошибок измерения.

Предположим, что у нас не образец породы, а значительной длины керн, взятый из скважины, проходящей через слой песчаника. Мы хотим определить проницаемость песчаника, но не можем ввести керн длиной в 20 усл. ед. в аппарат для измерения проницаемости. Вместо этого мы вырежем из керна несколько малых частей (интервалов) и определим проницаемость каждого из них. Наблюдаемая изменчивость явится следствием различий как между испытуемыми частями керна, так и между условиями эксперимента. Разработка методов

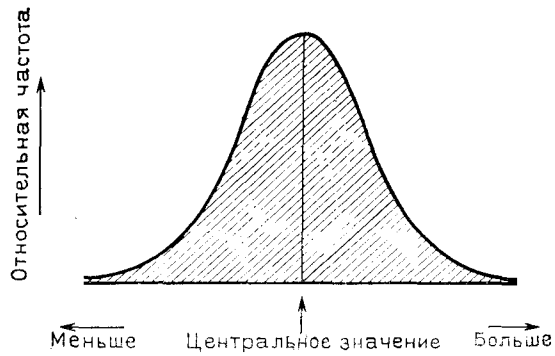


Рис. 2.9. График плотности нормального распределения

оценки величины отдельных источников изменчивости — одна из важнейших задач статистики.

Повторные измерения, проводимые на больших выборках, взятых из естественных совокупностей, дают возможность охарактеризовать распределение частот. Обычно большая часть значений группируется около некоторого центрального значения, при удалении от которого частоты убывают. График, представленный на рис. 2.9, имеет один максимум и называется нормальным распределением. В приложениях часто делается допущение, что случайные величины распределены нормально, и многие статистические критерии основаны на этом допущении.

Общую площадь, заключенную между графиком нормального распределения и горизонтальной осью, можно считать равной 1,00 (или 100%). Поэтому, используя график, можно вычислить вероятность соответствующего события. Читатель уже заметил, наверное, сходство одновершинной непрерывной кривой, изображенной на рис. 2.9, с гистограммой, представленной на рис. 2.8. Однако поскольку в случае непрерывного распределения число подразделений по горизонтальной оси можно считать бесконечным, вероятность получения какого-либо конкретного значения равна нулю. Вместо этого мы рассмотрим вероятность появления значений в пределах некоторого заданного интервала. Эта вероятность равна площади под кривой частот, заключенной между заданными пределами. Если указанный промежуток велик, то осуществление события в этом промежутке представляется более правдоподобным. Если интервал очень мал, то появление события маловероятно.

Выше были введены без определения два важных статистических понятия — «совокупность» и «выборка». Совокупность состоит из вполне определенного множества (либо конечного,

либо бесконечного) элементов. Вообще эти элементы можно рассматривать как измерения, выполненные на объектах заданного типа. Выборка — это подмножество элементов, выбранных из некоторой совокупности.

Примером конечной совокупности могут служить все нефтяные скважины, пробуренные в Канзасе в 1963 г., а набор всевозможных шлифов песчаника Тэнслип — примером бесконечной геологической совокупности. Заметим, что в последнем примере совокупность включает в себя не только ограниченное число испытаний, которые были сделаны, но также и все возможные результаты испытаний. Испытания, которые были действительно осуществлены, можно рассматривать как выборку из совокупности всех потенциально возможных испытаний.

Если наблюдения с заданными свойствами систематически исключаются из выборки, то такую выборку называют смещенной. Предположим, например, что нас интересует пористость данного слоя песчаника. Если из выборки исключить все рыхлые и раздробленные породы, так как их пористость трудно измерить, то результат изменится. Вероятно, полученный интервал значений пористости будет усечен справа, что даст смещение выборки в сторону более низких значений, и потому мы получим ошибочно заниженную оценку изменчивости пористости в слое.

Обычно выборки извлекаются из совокупности наудачу. Это значит, что все элементы совокупности имеют равные возможности быть включенными в выборку. Случайная выборка будет несмещенной, и по мере возрастания ее объема она будет точнее описывать рассматриваемую совокупность. К сожалению, получение истинно случайной выборки практически невыполнимо, так как при опробовании геологических объектов не все их части доступны. Пробы из глубинных объектов не имеют такой же возможности попасть в выборку, как пробы из поверхностных обнажений. Задача опробования в подобных условиях весьма сложна. В конце этой главы рассматриваются эффекты, возникающие из различия выборочных схем, и проведено сравнение последних. Однако при решении многих геологических задач анализируются данные, собранные без предварительного выборочного плана. Ярким тому примером является интерпретация погребенных структур по данным скважин.

СТАТИСТИКИ

Распределения имеют ряд характеристик, например, такие, как средняя точка, меры разброса и меры симметрии. Эти характеристики называются параметрами, если они описывают совокупности, и статистиками, если они относятся к выборкам. Статистики можно использовать для оценки параметров ис-

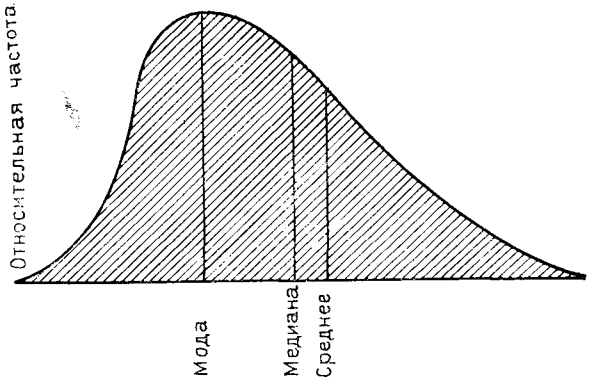


Рис. 2.10. Соотношение между мерами центральной тенденции в асимметричном частотном распределении

ходных совокупностей и для проверки гипотез, сформулированных относительно этих совокупностей.

Наиболее очевидная характеристика совокупности или выборки — ее среднее значение. Существуют различные виды среднего значения, но только некоторые из них используются на практике. Мода — значение, которое соответствует наибольшей частоте. Например, в распределении, приведенном на рис. 2.10, мода соответствует наивысшей точке кривой частот, а медиана — средняя точка распределения. На рис. 2.10 показано, что половина площади под кривой распределения находится справа от медианы, а другая половина — слева. Среднее значение — это, иными словами, среднее арифметическое, которое определяется как сумма всех результатов наблюдений, деленная на их число. В условиях асимметричных кривых распределения медиана расположена между средним значением и модой, а в случае симметричных кривых, подобных нормальной, все три меры совпадают.

Некоторые символы традиционно используются в качестве характеристик кривых распределения. Обычно для обозначения характеристик теоретических распределений используются греческие буквы, а для выборочных — латинские. Так, например, выборочное среднее обозначается $\bar{X}$, а теоретическое среднее значение всей совокупности μ. Основная задача обычно заключается в том, чтобы оценить некоторые параметры изучаемого распределения. Статистика, которую мы вычисляем по выборке из взятой совокупности, используется как оценка требуемого параметра. Применение греческих и латинских букв подчеркивает разницу между параметрами и соответствующими им статистиками.

Среднее арифметическое, вычисленное по данным выборки, имеет два в высшей степени желательных свойства, которые делают его более полезным для оценки среднего или центрального значения распределения, чем любая из двух других выборочных характеристик: медиана или мода. Во-первых, среднее арифметическое является несмещенной оценкой истинного среднего значения совокупности. Необходимо отметить, что статистика — это несмещенная оценка соответствующего параметра, если ее среднее значение, взятое по большому набору выборок, равно этому параметру. Во-вторых, можно показать, что для симметричных распределений, подобных нормальному, среднее арифметическое характеризуется тенденцией лучшего приближения к среднему значению совокупности, чем любая другая несмещенная оценка (такая, как медиана), построенная по той же выборке. Это равносильно тому, что выборочные средние имеют меньшую дисперсию, чем выборочные медианы, и, следовательно, являются более эффективными.

В практике геохимического анализа принято проводить серию определений на одном образце. В табл. 2.1 указано пять значений содержания хрома, полученных в результате спектрографического анализа образца глинистого сланца пенсильванского возраста из юго-восточного Канзаса. Найдите среднее арифметическое по этим данным.

Другая характеристика распределения — мера разброса отдельных значений относительно среднего, или дисперсия. Известны различные меры этого свойства, но только две из них широко используются. Одна из них — уже упомянутая дисперсия, а другая — квадратный корень из дисперсии, называемый стандартным отклонением. Дисперсию можно рассматривать как среднее значение квадратов отклонений всех возможных значений случайной величины от истинного среднего совокупности, которая определяется по формуле

Таблица 2.1

Содержание хрома в сланцах Канзаса

Номер	Содержание Cr, г/т
1	205
2	255
3	195
4	220
5	235
Сумма	1110
Среднее значение	1110/5 = 222

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2. \qquad (2.11)$$

Этим равенством определяется истинная дисперсия совокупности σ^2. Выборочная дисперсия определяется символом s^2. Если наблюдения $X_1, \ldots, X_n$ — случайная выборка из совокупности с нормальным распределением, то s^2 является эффективной оценкой для σ^2.

Причина использования среднего значения квадратов отклонений может оказаться не совсем очевидной. Может показаться, что целесообразнее охарактеризовать изменчивость просто как среднее значение отклонения от среднего, но простая проверка показывает, что такая величина всегда равна нулю, т. е.

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}) = 0. \qquad (2.12)$$

Конечно, можно оценить абсолютное отклонение от среднего, или так называемое среднее отклонение (MD):

$$MD = \frac{1}{n}\sum_{i=1}^{n}|X_i - \bar{X}|. \qquad (2.13)$$

Вертикальные черточки обозначают абсолютное значение (т. е. значение, взятое без знака) заключенной в них величины. Однако можно доказать, что эта статистика менее эффективна, чем выборочная дисперсия. Хотя это интуитивно и непонятно, необходимо подчеркнуть, что дисперсия имеет свойства, которые делают ее намного более полезной, чем другие меры изменчивости.

Так как дисперсия является средним значением квадратов отклонений от среднего, то ее размерность характеризуется квадратами единиц, которыми измерялись исходные наблюдения. Порода, например, может содержать кристаллы полевого шпата, большие оси которых имеют среднюю длину 13,2 мм и дисперсию 2,0 мм^2. Многие не считают площадь мерой дисперсии длин. В большинстве примеров, где мы имеем дело с дисперсией, она используется в стандартизированном безразмерном виде, т. е. в виде, не зависящем от выбранных единиц измерения. Этот вопрос будет еще подробно рассмотрен в данной главе.

Для того чтобы получить статистику, которая характеризует дисперсию, или разброс данных относительно среднего зна-

чения и обладает той же размерностью, что и исходные данные, можно воспользоваться стандартным отклонением. Оно определяется как квадратный корень из дисперсии и обозначается символом σ, являющимся параметром совокупности; соответствующая выборочная статистика обозначается через s.

Малое значение стандартного отклонения указывает, что наблюдения хорошо группируются около центрального значения. Наоборот, большое стандартное отклонение показывает, что наблюдения широко рассеяны относительно среднего значения и имеют слабую тенденцию к централизации. Это проиллюстрировано на рис. 2.11, где изображены две симметричные кривые распределения, имеющие различные стандартные отклонения. Кривая *A* характеризует насыщение нефтью (в процентах) образцов керна из продуктивной зоны северо-восточной Оклахомы. Кривая *B* представляет те же величины для нефтеносной области западного Техаса. Среднее насыщение нефтью в этих двух регионах различно, но наибольшее различие между кривыми заключается в том, что для Техаса характерна значительно более высокая изменчивость насыщения.

Весьма полезное свойство нормального распределения состоит в том, что площадь под кривой в пределах некоторого заданного интервала может быть точно вычислена. Например, более 2/3 наблюдений (68,27%) попадают в интервал с центром в среднем значении и длиной, равной двум стандартным отклонениям. Примерно 95% всех наблюдений заключается в интервале от —2 до +2 стандартных отклонений и более 99% содержится в интервале от —3 до +3 стандартных отклонений. Это показано на рис. 2.12.

Распределение, указывающее степень насыщения нефтью пород северо-восточной Оклахомы (см. рис. 2.11, кривая *A*), имеет среднее значение 20,1% и стандартное отклонение 4,3%. Если предположить, что распределение нормально, то следует ожидать, что около 2/3 исследуемых образцов будет иметь насыщение нефтью от 16 до 24%. Изучение исходных данных показало, что 1145 проб характеризуются насыщением, которое находится в указанных пределах, что составляет около 68% всех данных. Только 101 образец, т. е. около 6%, имеет насыщение вне интервала 2σ (12—29%).

Те, кто не имел дела со статистическим анализом, обычно с трудом развивают интуитивное восприятие численного значения дисперсии или стандартного отклонения. Является ли дисперсия, равная 10, большой или малой? Что значит стандартное отклонение 23? Оказывается, для интерпретации как дисперсии, так и стандартного отклонения не требуется приписывать каждому из них численного значения, а требуется сравнивать одну дисперсию с другой. Выборка, имеющая наибольшую дисперсию или стандартное отклонение, характеризуется

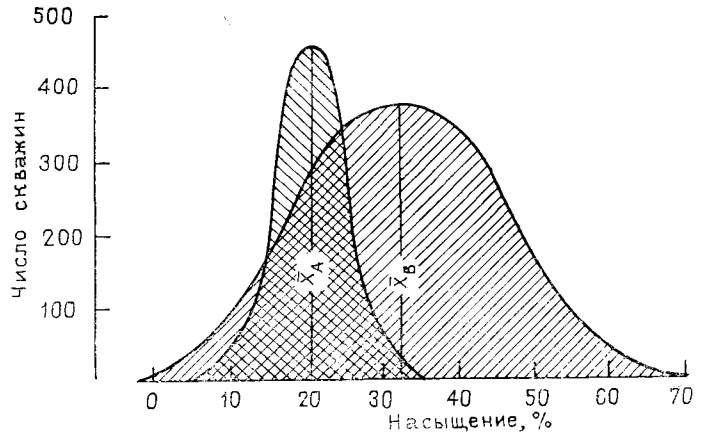


Рис. 2.11. Частотное распределение процентного содержания насыщения нефтью в нефтяном поле Оклахомы (А) и Техаса (В)

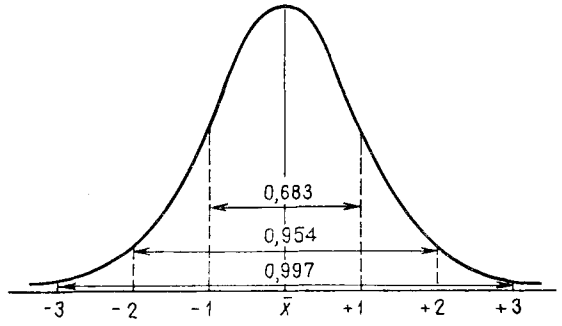


Рис. 2.12. Площади стандартного нормального распределения, заключенные в пределах интервалов, кратных стандартному отклонению

большим разбросом наблюдаемых значений при условии, что все измерения сделаны в одних и тех же единицах.

Равенство (2.11), хотя и определяет дисперсию, обычно не используется для вычислений, так как содержит n операций вычитания, n — умножения и n — сложения. Вместо этой формулы для вычисления оценки дисперсии используется другая формула, которая имеет следующий вид:

$$s^2 = \frac{\sum_{i=1}^{n} X_i^2 - \bar{n}X^2}{n-1} \qquad (2.14)$$

или

$$s^2 = \frac{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}{n(n-1)}. \qquad (2.15)$$

С помощью настольной вычислительной машины ΣX_i и ΣX_i^2 можно подсчитать одновременно, что позволяет уменьшить число требуемых операций на число n. На вычислительной машине формула (2.15) может быть использована для одновременного нахождения среднего и дисперсии, что позволяет избежать необходимости дважды использовать одни и те же данные.

Для вычисления оценок дисперсии и стандартных отклонений введем некоторые промежуточные величины, которые часто будут использоваться во многих процедурах, излагаемых в последующих главах. Нецентрированная сумма квадратов — это просто ΣX_i^2; центрированная сумма квадратов (SS) определяется по формуле

$$SS = \sum_{i=1}^{n} (X_i - \bar{X})^2 \qquad (2.16)$$

или, что алгебраически эквивалентно,

$$SS = \sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2. \qquad (2.17)$$

Оценку дисперсии вычисляют путем деления этой величины на $n-1$, т. е.

$$s^2 = \frac{1}{n-1} SS = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2\right] =$$

$$= \frac{1}{n(n-1)}\left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right]. \qquad (2.18)$$

Величина $n-1$, которая содержится как в формуле (2.14), так и в формуле (2.15), требует некоторого пояснения. Дисперсия определяется как среднее значение квадратов отклонений от среднего. Однако, имея дело лишь с выборкой, мы не знаем истинного среднего значения совокупности μ, но можем оценить его с помощью выборочного среднего $\bar{X}$, которое вычисляется так, чтобы минимизировать квадраты отклонений от него. Иначе говоря, операция $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, дает значение $\bar{X}$, для которого $\sum_{i=1}^{n}(X_i - \bar{X})^2$ имеет минимальное значение среди всех возможных. В силу этого свойства выборочного среднего оценка дисперсии будет занижена, если использовать формулу (2.11). Иными словами, $s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ является смещенной оценкой для $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$. Для того чтобы устранить смещение, мы используем в качестве знаменателя в формуле для выборочной дисперсии $n-1$, увеличивая таким образом оценку дисперсии.

Вычисление этих величин можно показать на примере геохимических данных по содержанию хрома в глинистых сланцах, приведенных в табл. 2.1. Переписав эту таблицу так, чтобы она содержала столбец квадратов, получим табл. 2.2.

Допуская, что содержания хрома распределены приблизительно по нормальному закону, можно ожидать, что около двух третей значений расположено в пределах 198—246 г/т. Анализ таблицы показывает, что три значения из пяти, т. е. 60%, действительно попадают в этот интервал.

Заметим, что при вычислении сумм квадратов геохимических данных появляются числа, содержащие семь знаков. Эта

Таблица 2.2

Вычисление сумм квадратов и дисперсий по данным табл. 2.1

X	X^2
205	42 025
255	65 025
195	38 025
220	48 400
235	55 225
$\Sigma X_i = 1{,}110$	$\Sigma X_i^2 = 248\,700$

$$(\Sigma X)^2 = 1\,232\,100;$$

$$SS = 248\,700 - \frac{1232100}{5} = 2280;$$

$$s^2 = \frac{2280}{4} = 570;$$

$$s = \sqrt{570} = 23{,}88.$$

Таблица 2.3

Содержание хрома, никеля и ванадия в сланцах Канзаса, г/т

	Cr	Ni	V
	205	130	180
	255	165	215
	195	100	135
	220	135	200
	235	145	205
Суммы	1110	675	935
Средние значения	222	135	187

тенденция к возникновению в процессе вычисления очень больших чисел приводит иногда к возникновению затруднений в ЭВМ, приспособленных для работы с числами, содержащими мало значащих цифр. Это также приводит к возникновению трудностей при выводе данных, если поля формата недостаточно широки для того, чтобы вмещать числа, которые должны быть напечатаны.

Для большинства геологических исследований характерно, что на каждом изучаемом объекте измеряется более одной переменной. В качестве примеров можно привести результаты измерений коллекции кораллов, последовательности проб из ряда скважин или же определения параметров пород в коллекции образцов песчаника. Такие данные обычно записываются в виде таблицы порядка $n \times m$, где n — число наблюдений, а m — число изучаемых переменных. Так, например, полные анализы, из которых извлечены данные табл. 2.1, содержат 17 переменных. В табл. 2.3 представлены только три из них, а именно содержания никеля, ванадия и хрома. Для каждого столбца можно подсчитать соответствующие суммы и оценить среднее значение и стандартное отклонение.

Однако различные переменные могут не быть независимыми, между ними может существовать некоторая форма условной связи. Важно, что мы в состоянии оценить природу и силу этих условных связей, так же как было важно определить условные вероятности появления дискретных событий.

Согласованное изменение двух переменных

Вычислительные процедуры, используемые для получения оценки дисперсии одной переменной, можно расширить для вычисления меры взаимной изменчивости пары переменных.

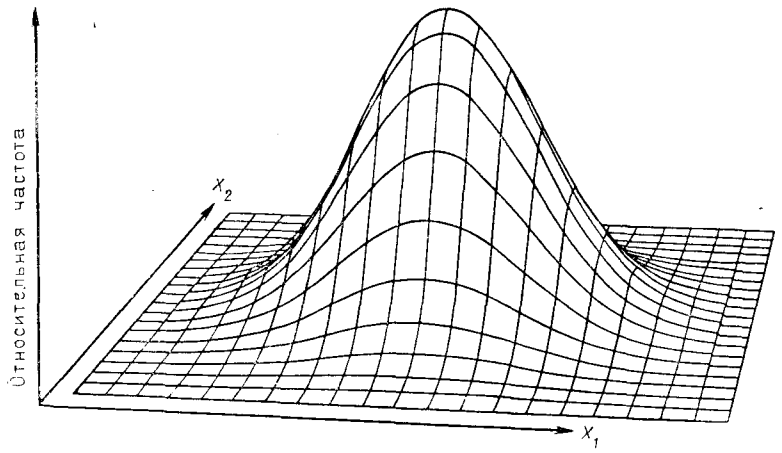


Рис. 2.13. Совместное вероятностное распределение двух независимых нормальных случайных величин. X_1 и X_2 нормально распределены

Эта мера, называемая ковариацией, является характеристикой совместного изменения двух переменных по отношению к их общему среднему значению. Это соотношение показано на рис. 2.13, где изображены формы поверхностей распределения вероятностей, порожденных двумя кривыми нормального распределения.

Пусть X_1 и X_2 имеют кривые распределения вероятностей, аналогичные изображенным на рис. 2.12. Точно так же, как дисперсия характеризует разброс значений относительно центральной точки, как это показано на рис. 2.12, ковариация является мерой разброса значений распределения относительно общего среднего.

Для вычисления оценки ковариации мы снова введем величину, аналогичную сумме квадратов. Эта величина называется центрированной суммой смешанных произведений (SP) и определяется по формуле

$$SP_{jk} = \sum_{i=1}^{n} (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k), \qquad (2.19)$$

где X_{ij} — i-е значение j-й переменной, а X_{ik} — i-е значение k-й переменной. Символ SP_{jk} — сумма произведений центрированных j-й и k-й переменных. Запишем это выражение в форме, удобной для вычисления:

$$SP_{jk} = \sum_{i=1}^{n} X_{ij} X_{ik} - \frac{1}{n} \sum_{i=1}^{n} X_{ij} \sum_{i=1}^{n} X_{ik}. \qquad (2.20)$$

Величина $\Sigma(X_{ij}X_{ik})$ называется нецентрированной суммой смешанных произведений. Связь величины SP_{jk} с суммой квадратов можно легко установить, если выбрать $j=k$.
Тогда получаем

$$SP_{jj}=\sum_{i=1}^{n}X_{ij}X_{ij}-\frac{1}{n}\sum_{i=1}^{n}X_{ij}\sum_{i=1}^{n}X_{ij}=$$

$$=\frac{1}{n}\left[\sum_{i=1}^{n}X_{ij}^2-\left(\sum_{i=1}^{n}X_{ij}\right)^2\right]=SS_j. \qquad (2.21)$$

Если мы вычислим суммы смешанных произведений и суммы квадратов для всех возможных комбинаций наших трех переменных из табл. 2.3, то получим следующую таблицу порядка 3×3:

	Cr	Ni	V
Cr	SS_{Cr}	SP_{Cr-Ni}	SP_{Cr-V}
Ni	SP_{Ni-Cr}	SS_{Ni}	SP_{Ni-V}
V	SP_{V-Cr}	SP_{V-Ni}	SS_V

Легко заметить, что некоторые из величин встречаются в этой таблице дважды: например, сумма произведений для ванадия и никеля такая же, как и сумма произведений для никеля и ванадия. Обобщая этот факт, можно написать $SP_{jk}=SP_{kj}$. Это равенство будет нами использовано в следующих главах.

Подобно тому как мы при вычислении дисперсии делили величину SS на $n-1$, вычислим оценку ковариации, также разделив величину SP на $n-1$:

$$\mathrm{cov}_{jk}=\frac{SP_{jk}}{n-1}=\frac{1}{n-1}\left[\sum_{i=1}^{n}X_{ij}X_{ik}-\frac{1}{n}\sum_{i=1}^{n}X_{ij}\sum_{i=1}^{n}X_{ik}\right]=$$

$$=\frac{1}{n(n-1)}\left[n\sum_{i=1}^{n}X_{ij}X_{ik}-\sum_{i=1}^{n}X_{ij}\sum_{i=1}^{n}X_{ik}\right]. \qquad (2.22)$$

Теперь, возвращаясь к геохимическим данным, приведенным в табл. 2.3, можно вычислить оценки ковариаций для всех трех элементов. Обозначая содержания хрома и никеля соответственно через X_1 и X_2, можно вычислить величины, приведенные в табл. 2.4. Мы знаем теперь дисперсию X_1 (хрома) и оценку ковариации между X_1 и X_2 (хромом и никелем). У нас

Таблица 2.4

Вычисление оценки ковариаций между хромом (X_1) и никелем (X_2)

X_1^2	X_1	X_1X_2	X_2	X_2^2
42 025	205	26 650	130	16 900
65 025	255	42 075	165	27 225
38 025	195	19 500	100	10 000
48 400	220	29 700	135	18 225
55 225	235	34 075	145	21 025
$\Sigma X_1^2=248\,700$	$\Sigma X_1=1110$	$\Sigma X_1X_2=152\,000$	$\Sigma X_2=675$	$\Sigma X_2^2=93\,375$

$$SP_{1,2}=152\,000-\frac{(1110)\ (675)}{5}=2150.$$

$$\mathrm{cov}_{12}=\frac{2150}{4}=537{,}5.$$

есть также все необходимые данные для вычисления дисперсии X_2 (никеля) по формуле 2.12. Читатель может попытаться вычислить это значение, заполнив таблицу порядка 2×2, приведенную ниже.

	X_1	X_2
Хром (X_1)	570	537,5
Никель (X_2)	537,5	s_2^2

Чтобы закончить анализ геохимических данных, приведенных в табл. 2.3, остается вычислить дополнительно три величины. Это оценки ковариаций для хрома и ванадия (cov_{13}), никеля и ванадия (cov_{23}) и дисперсию ванадия (s_3^2). Следуя процедурам, использованным при построении табл. 2.4, вычислите величину (cov_{13}).

На рис. 2.14 приведена диаграмма совместного распределения двух переменных, которые тесно связаны и имеют довольно высокое значение ковариации. Распределения двух переменных, изображенные на рис. 2.15, имеют те же дисперсии, что и приведенные на рис. 2.14, но не зависят одно от другого, о чем свидетельствует относительно низкое значение ковариации. Интерпретация значений оценок ковариаций должна проводиться таким же образом, как и дисперсий, но при этом следует помнить, что рассматриваемые значения не слишком содержательны, так как они зависят от единиц измерения.

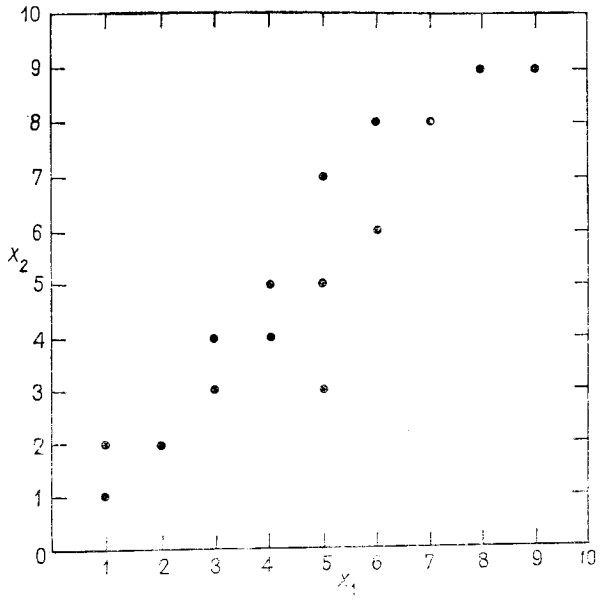


Рис. 2.14. Диаграмма рассеяния двух переменных с высоким коэффициентом ковариации

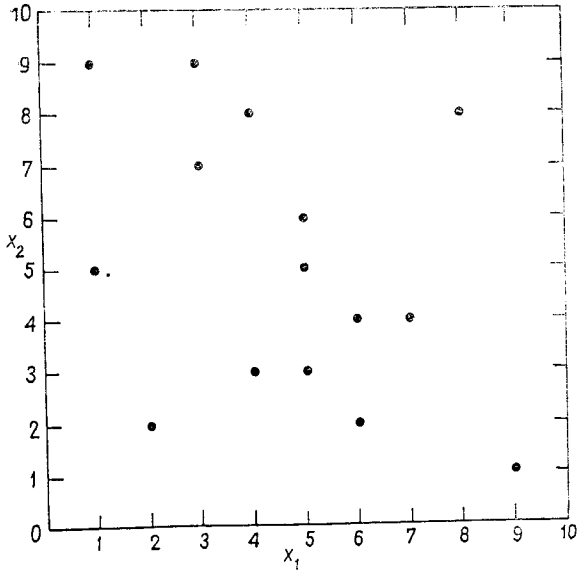


Рис. 2.15. Диаграмма рассеяния двух переменных с низким коэффициентом ковариации

Для оценки степени взаимной связи между переменными, не зависящей от единиц измерения, используется коэффициент корреляции r, который представляет собой отношение ковариации двух переменных к произведению их стандартных отклонений:

$$r_{jk} = \mathrm{cov}_{jk}/(s_j s_k). \qquad (2.23)$$

Так как коэффициент корреляции является отношением, то эта величина безразмерная. При этом ковариация может равняться величине произведения стандартных отклонений рассматриваемых переменных, но не может превышать ее. Поэтому коэффициент корреляции принимает значения в интервале от -1 до $+1$. Если коэффициент корреляции равен $+1$, это указывает на прямую линейную связь между двумя переменными. Если же коэффициент корреляции равен -1, это указывает на то, что одна переменная изменяется в противоположном направлении по отношению к другой. Между двумя упомянутыми крайними случаями находится спектр менее сильных связей, включающий случай равенства коэффициента корреляции нулю, что указывает на полное отсутствие любого типа линейных зависимостей.

На рис. 2.16, *а* изображена ситуация, когда сильная корреляция между переменными очевидна и коэффициент корреля-

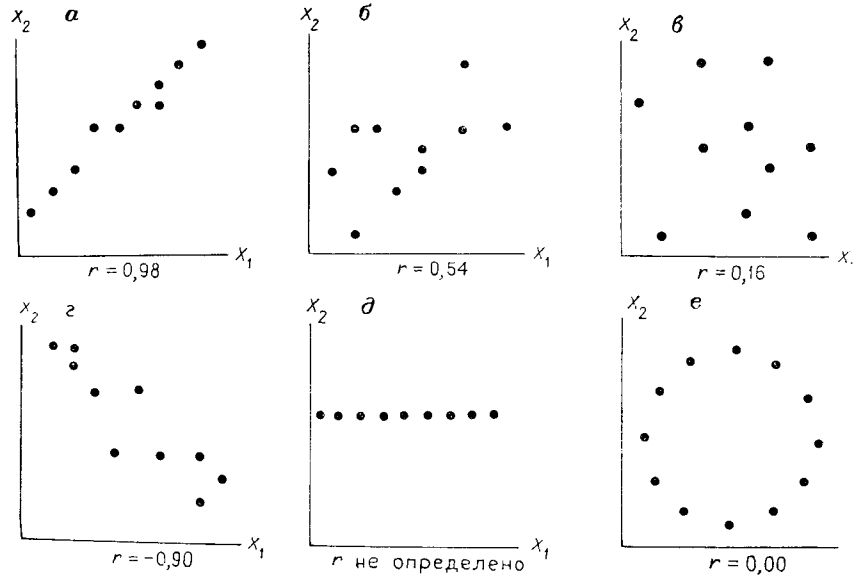


Рис. 2.16. Точечные диаграммы, иллюстрирующие различные коэффициенты корреляции между двумя переменными

ции почти равен +1,00. Менее явная корреляция изображена на рис. 2.16, *б*. В этом случае коэффициент корреляции равен только +0,54. Положение точек на рис. 2.16, *в* определено по таблице случайных чисел, и поэтому значения двух переменных совсем не имеют связи друг с другом, о чем свидетельствует коэффициент корреляции, близкий к нулю. Отрицательная корреляционная зависимость с коэффициентом корреляции, равным —0,90, изображена на рис. 2.16, *г*, который иллюстрирует тот случай, когда одна переменная уменьшается, в то время как другая увеличивается. Интересный предельный случай представлен на рис. 2.16, *д*: одна переменная инвариантна, т. е. ее значения не изменяются. Попытка вычислить коэффициент корреляции приводит к необходимости деления на нуль; в этом случае коэффициент корреляции не определен. В примере, изображенном на рис. 2.16, *е*, очевидна взаимная зависимость между двумя переменными. Наблюдения X_1 и X_2 расположены на окружности, поэтому соотношение между двумя переменными можно представить в виде

$$X_2 = \sqrt{a^2 - X_1^2}$$

в предположении, что центром окружности является начало координат. Радиус окружности равен a. Однако если вычислить корреляцию между X_1 и X_2, она окажется равной нулю. Это происходит потому, что коэффициент корреляции есть мера линейной зависимости между двумя переменными, а указанное соотношение нелинейно. Существует много возможных нелинейных соотношений, которые могут возникнуть между двумя переменными. В подобной ситуации коэффициент корреляции нельзя считать удовлетворительной мерой степени таких зависимостей.

На практике выборочный коэффициент корреляции r вычисляется по формуле

$$r_{jk} = \frac{SP_{jk}}{\sqrt{SS_j \cdot SS_k}} = \frac{\sum_{i=1}^{n} X_{ij}X_{ik} - \frac{1}{n}\left(\sum_{i=1}^{n} X_{ij} \sum_{i=1}^{n} X_{ik}\right)}{\sqrt{\left\{\sum_{i=1}^{n} X_{ij}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_{ij}\right)^2\right\}\left\{\sum_{i=1}^{n} X_{ik}^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_{ik}\right)^2\right\}}}. \quad (2.24)$$

Так как r измеряет линейное соотношение между двумя переменными, можно определить прямую линию, характеризующую зависимость между ними. Это важный вопрос статистического корреляционного и регрессионного анализа, являющегося основой большинства методов аппроксимации поверхностей.

Таблица 2.5

Размеры раковин брахиопод рода *Composita*, мм

Длина	Ширина
18,4	15,4
16,9	15,1
13,6	10,9
11,4	9,7
7,8	7,4
6,3	5,3

Детально этот вопрос будет рассмотрен в следующих главах, а здесь мы остановимся на процедуре вычисления величины r.

При биологических исследованиях обычно наблюдается сильная корреляция свойств в пределах одной биологической группы организмов, так как результаты измерения отдельных характеристик в значительной степени зависят от общих размеров особи. Так, например, в табл. 2.5 приведены результаты измерения длины и ширины раковин брахиопод рода *Composita*. Как легко установить, имеется сильная связь между этими двумя характеристиками, о чем свидетельствует вычисленное значение выборочного коэффициента корреляции.

Для вычисления оценки коэффициента корреляции между двумя столбцами измерений подсчитаем соответствующие квадраты и смешанные произведения. Это сделано в табл. 2.6, где X_1 — длина, X_2 — ширина. Коэффициент корреляции, равный 0,99, оказывается очень высоким, что подтверждает подозрение в том, что имеется прямая связь между длиной и шириной раковины. Столь сильные зависимости встречаются не всегда; в действительности весьма обычны задачи, в которых требуется определить, существует ли хоть какая-нибудь корреляция. К этому вопросу мы еще вернемся.

Наведенная корреляция

Некоторые корреляции между переменными не отражают соотношений между ними, но они индуцированы операциями или преобразованиями, которым были подвергнуты переменные. Две независимые случайные величины обычно имеют нулевую корреляцию. Однако некоторые операции над переменными могут привести к корреляции, отличной от нуля, хотя между ними не существует никакого линейного соотношения. Существующие корреляции могут быть изменены или даже обращены такими операциями.

Таблица 2.6

Вычисление сумм квадратов, смешанных произведений и коэффициента корреляции по данным табл. 2.5

X_1^2	X_1	X_1X_2	X_2	X_2^2
338,56	18,4	283,36	15,4	237,16
285,61	16,9	255,19	15,1	228,01
184,96	13,6	148,24	10,9	118,81
129,96	11,4	110,58	9,7	94,09
60,84	7,8	57,72	7,4	54,76
39,69	6,3	33,39	5,3	28,09
$\Sigma X_1^2=1039.62$	$\Sigma X_1=74,4$	$\Sigma X_1X_2=888,48$	$\Sigma X_2=63,8$	$\Sigma X_2^2=760,92$

$$SP_{12}=(888,48)-\frac{(74,4)(63,8)}{6}=97,37.\quad \mathrm{cov}_{12}=\frac{97,37}{5}=19,47.$$

$$SS_1=(1039,62)-\frac{(74,4)^2}{6}=117,06.\quad SS_2=(760,92)-\frac{(63,8)^2}{6}=82,51.$$

$$s_1^2=\frac{117,06}{5}=23,41.\quad s_1=\sqrt{23,41}=4,84.\quad s_2^2=\frac{82,51}{5}=16,50.$$

$$s_2=\sqrt{16,50}=4,06.$$

$$r_{1,2}=\frac{19,47}{(4,84)(4,06)}=0,991.$$

Предположим, что образцы гальки случайно выбираются с галечного пляжа и измеряются три ортогональные оси на каждом из них. Никаких попыток измерить самую длинную или самую короткую ось гальки не предпринимается ни для одного из образцов. Можно было бы предположить, что эти измерения будут коррелированы, так как наиболее вероятно, что большая галька будет иметь большие размеры по всем трем осям, а малая галька наоборот будет иметь малые размеры по всем трем осям.

В табл. 2.7 приведены замеры, сделанные на коллекции гальки, и корреляции между переменными. Данные также представлены в виде скалярных диаграмм на рис. 2.17. Однако если теперь оси выбрать в соответствии с соглашением — по определению наибольшая ось гальки — a, наименьшая — c; средняя — b, то такое упорядочение приведет к измерению корреляций (табл. 2.8). Это особенно хорошо видно на скалярной диаграмме (рис. 2.17), так как такое определение приводит к смещению всех точек в пределах сектора диаграммы с углом 45°. В силу этого всегда должна существовать положительная корреляция между любой парой осей, или между отношениями

Таблица 2.7

Длины (в см) осей образцов гальки, собранной на пляже, Оси перечислены в порядке измерений

Образец	Ось 1	Ось 2	Ось 3
a	3	7	8
b	16	5	8
c	10	12	9
d	13	5	12
e	14	16	5
f	9	8	14
g	16	13	13
h	6	3	11
i	9	15	9
j	13	10	9
Суммы	109	94	98
Средние	10,9	9,4	9,8

Корреляции $r_{1,1}=0,279$ $r_{1,3}=-0,021$ $r_{2,3}=-0,349$

двух осей и третьей осью (например, между b/a в отношении к c).

Наведенные корреляции, причиняющие наибольшее беспокойство, — это ложные отрицательные корреляции, которые появляются в замкнутых множествах данных. Замкнутое множество данных — это такое множество, в котором сумма всех переменных, измеренных на индивидуальных представителях множества, равна 1,00 или 100%, что означает, что эти переменные представляют собой определенные пропорции от целого. Так как сумма переменных есть фиксированное число, то

Таблица 2.8

Длины (в см) осей образцов гальки, собранной на пляже (a — наибольшая ось, b — промежуточная ось, c — наименьшая ось)

Образец	*a*	*b*	*c*
a	8	7	3
b	16	8	5
c	12	10	9
d	13	12	5
e	16	14	5
f	14	9	8
g	16	13	13
h	11	6	3
i	15	9	9
j	13	10	9
Суммы	134	98	69
Средние	13,4	9,8	6,9

Корреляции $r_{ab}=0,597$ $r_{ac}=0,499$ $r_{bc}=0,467$

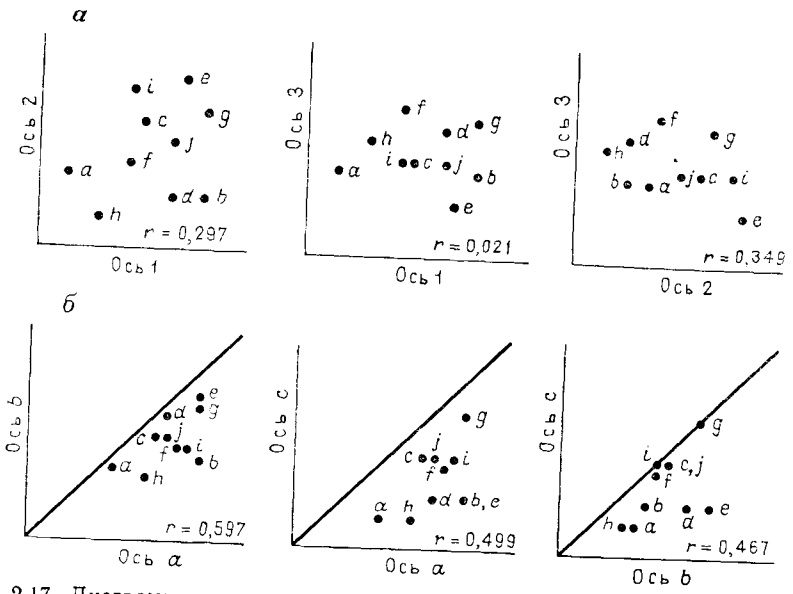


Рис. 2.17. Диаграмма рассеяния длин осей (в см) образцов гальки, отобранных на галечном пляже:

а — исходные данные собраны наудачу; *б* — измерения рассортированы по осям *a*, *b*, *c*, что привело к смещению всех точек ниже диагонали диаграммы

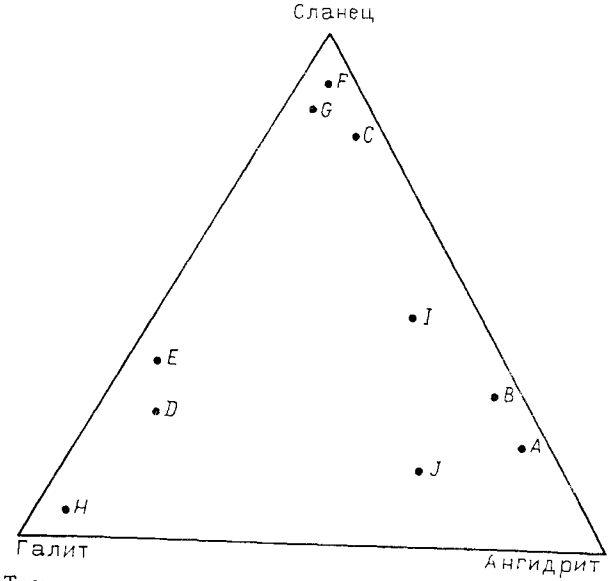


Рис. 2.18. Треугольная диаграмма галит—ангидрит—сланец как композиционная система. Точки характеризуют средние содержания пород Веллингтоновской формации (пермь), извлеченных из скважин в Центральном Канзасе с интервалом 1,5 м

увеличение доли одной переменной может лишь привести к сокращению доли других переменных.

В открытом множестве, в котором измерения не представляются в виде пропорций двух линейно независимых переменных, будет существовать корреляция, которая незначимо отличается от нуля. Если открытое множество данных замкнуть преобразованием измерений в пропорции, то появятся давно значимые отрицательные корреляции, хотя исходные данные представляли собой совершенно независимые переменные. В специальном случае замкнутой таблицы данных для трех переменных корреляции между замкнутыми переменными определяются только через дисперсии в соответствии со следующим соотношением

$$r_{1,2} = \frac{s_3^2 - (s_1^2 + s_2^2)}{2s_1 s_2}. \qquad (2.25)$$

Такие взаимные корреляции присущи любым геологическим данным, которые нанесены на треугольные диаграммы, например, диаграммы песчаник — глина — известняк или трехфазные диаграммы. Эти обратные соотношения проистекают из того факта, что по мере увеличения пропорций одной составляющей пропорции двух других составляющих должны уменьшаться.

Рис. 2.18 — это треугольная диаграмма хлорид натрия — ангидрит — глина — составляющих компонент осадочных горных пород. Нанесенные на рисунке точки представляют осадочные литологические пропорции в 5-футовом интервале солевого члена Хатчисона пермской формации Веллингтона в скважине, пробуренной на территории Центрального Канзаса. Композиции были вычислены по результатам γ-, нейтронного, плотностного и акустического каротажа, используемого для измерения петрофизических свойств интервала. Была пробурена контрольная скважина для определения потенциальных возможностей размещения радиоактивных отходов.

В табл. 2.9 приведены композиции 10 интервалов, нанесенных на рис. 2.18. В ней также представлены дисперсии трех минералогических компонент и корреляции, вычисленные на основе этих дисперсий. Заметим, что ковариации не обязательны для вычисления корреляций, как это было предопределено дисперсиями и эффектом замкнутости.

Так как данные с постоянными суммами наиболее распространены в геологии, было предпринято много попыток придать смысл статистическим связям между ними. Кох и Линк [18, т. II гл. 11] приводят некоторое число специальных статистических критериев, пригодных для таких данных, Чейес [5] написал книгу, посвященную проблеме замкнутости. К сожалению, предлагаемые статистические процедуры не универсальны

Таблица 2.9

Литологический состав (с точностью до 5%) 1,5-метровых интервалов в пермской формации Веллингтона в Центральном Канзасе; оценки основаны на петрографических измерениях по результатам каротажа скважин

Интервал	Ангидрит	Сланец	Хлорит натрия
a	75	20	5
b	65	30	5
c	15	80	5
d	10	25	65
e	5	35	60
f	5	90	5
g	5	85	10
h	5	5	90
i	45	45	10
j	60	15	25
Суммы	290	430	280
Средние	29	43	28
Дисперсии	832,22	962,22	1001,11
Стандартные отклонения	28,25	31,02	31,64

$$r_{1,2}=\frac{1001,11-(832,22+962,22)}{2\cdot 28,85\cdot 31,02}=\frac{-793,47}{1789,85}=-0,44;$$

$$r_{1,3}=\frac{962,22-(832,22+1001,11)}{2\cdot 28,25\cdot 31,64}=\frac{-871,11}{1787,66}=-0,48;$$

$$r_{2,3}=\frac{832,22-(962,22+1001,11)}{2\cdot 31,02\cdot 31,64}=\frac{-1131,11}{1962,95}=-0,58.$$

[19, 1] и в настоящее время нет вполне удовлетворительного метода вычисления силы связей между переменными в замкнутых множествах данных.

ПРОВЕРКА ГИПОТЕЗЫ О НОРМАЛЬНОМ РАСПРЕДЕЛЕНИИ

Прежде чем продолжить изложение, возвратимся немного назад к распределениям частот и, в частности, к нормальному распределению. Если вместо того чтобы рассматривать выборку только из шести значений, представленных в табл. 2.5, измерить длины раковин очень большой коллекции *Composita*, то мы увидим, что частотная диаграмма будет выглядеть аналогично графику, изображенному на рис. 2.19. Среднему значению длины, в данном случае равному 14,2 мм, будет соответствовать наибольшая частота, а постепенно уменьшающимся и увеличивающимся значениям будут отвечать уменьшающиеся частоты. Приблизительно две трети раковин попадают в пределы интервала $(\mu-s,\ \mu+s)$ с центром в точке $\mu=14,2$, причем оценка стандартного отклонения приблизительно равна 4,7 мм.

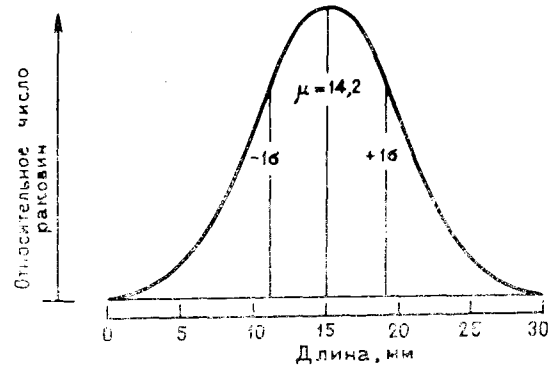


Рис. 2.19. Гипотетическое распределение значений длины особей рода *Composita*

Теперь рассмотрим измерения ширины, которые были сделаны при исследовании этой очень большой коллекции *Composita*. Распределение этого показателя по форме напоминает распределение длины, но его среднее значение и стандартное отклонение в этом случае иные. Оно может выглядеть, например, подобно графику, изображенному на рис. 2.20, со средним значением 10,3 мм и стандартным отклонением 3,6 мм.

Можем ли мы сравнивать два распределения друг с другом? Измерения проведены в одних и тех же единицах, что облегчает проблему сравнения распределений длины и ширины. Оба эти распределения можно изобразить в одном и том же масштабе, в результате чего получим рис. 2.21.

Конечно, сравнение было бы проще, если бы оба распределения имели один и тот же центр, т. е. равные средние значения. Мы можем центрировать их по отношению к общему среднему значению, вычитая подходящее число из всех значений совокупности (или прибавляя некоторое число к значениям другой совокупности) таким образом, чтобы средние обеих совокупностей совпали. Вместо этого вычтем соответствующее среднее значение из каждого наблюдения в каждой из двух совокупностей. Получим новые значения. Это преобразование сдвигает каждое из распределений вдоль горизонтальной оси до тех пор, пока их центры не совпадут со значением 0, являющимся средним для обоих преобразованных распределений, изображенных на рис. 2.22.

В рассмотренном примере мы связаны размерностью результатов измерений, выраженной в миллиметрах. При этом никаких проблем не возникает, если мы будем сравнивать распределения длины и ширины, но если мы захотим сравнивать эти распределения с распределениями, характеризующими мас-

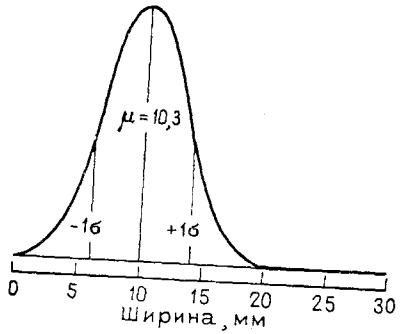


Рис. 2.20. Гипотетическое распределение значений ширины особей рода *Compositа*

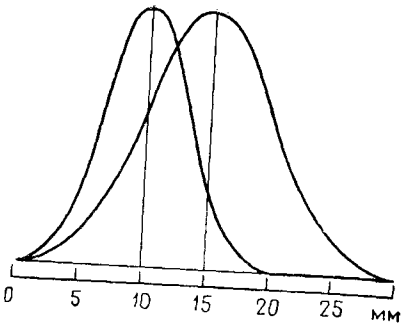


Рис. 2.21. Диаграмма распределения значений длины и ширины особей рода *Compositа*

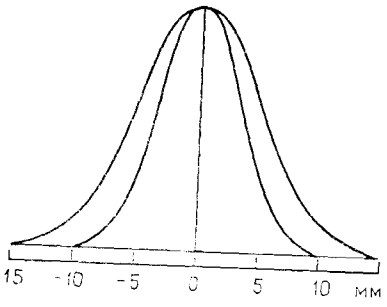


Рис. 2.22. Распределения значений длины и ширины особей рода *Compositа*

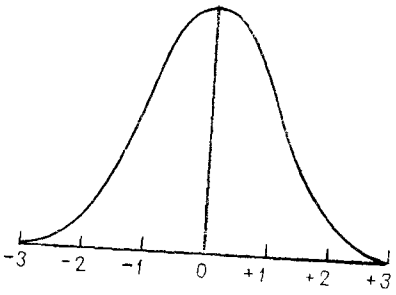


Рис. 2.23. Распределения значений длины и ширины особей рода *Compositа* после стандартизации имеют нулевое среднее значение и стандартное отклонение, равное 1,0

су раковин, то нам это сделать не удастся. Существует ли какое-либо дополнительное преобразование, которое позволяет сделать наши распределения не зависящими от единиц измерения? Одно из таких чрезвычайно полезных преобразований называется стандартизацией: в результате его применения новые значения переменных имеют не только нулевое среднее значение, но также измеряются в единицах стандартных отклонений. Это делается просто с помощью вычитания среднего значения распределения из каждого наблюдения и деления каждой полученной разности на стандартное отклонение распределения. Эта новая переменная имеет стандартную нормальную форму

$$Z_i = (X_i - X)/s. \qquad (2.26)$$

Таблица 2.10

Значения кумулятивной функции распределения стандартного нормального распределения [15]

Стандартные отклонения от среднего значения	Кумулятивная вероятность	Стандартные отклонения от среднего значения	Кумулятивная вероятность	Стандартные отклонения от среднего значения	Кумулятивная вероятность
—3,0	0,0014	—0,9	0,1841	+1,1	0,8643
—2,9	0,0019	—0,8	0,2119	+1,2	0,8849
—2,8	0,0026	—0,7	0,2420	+1,3	0,9032
—2,7	0,0035	—0,6	0,2743	+1,4	0,9192
—2,6	0,0047	—0,5	0,3085	+1,5	0,9332
—2,5	0,0062	—0,4	0,3446	+1,6	0,9452
—2,4	0,0082	—0,3	0,3821	+1,7	0,9554
—2,3	0,0107	—0,2	0,4207	+1,8	0,9641
—2,2	0,0139	—0,1	0,4602	+1,9	0,9713
—2,1	0,0179	—0,0	0,5000	+2,0	0,9773
—2,0	0,0228	+0,0	0,5000	+2,1	0,9821
—1,9	0,0287	+0,1	0,5398	+2,2	0,9861
—1,8	0,0359	+0,2	0,5793	+2,3	0,9893
—1,7	0,0446	+0,3	0,6179	+2,4	0,9918
—1,6	0,0548	+0,4	0,6554	+2,5	0,9938
—1,5	0,0668	+0,5	0,6915	+2,6	0,9953
—1,4	0,0808	+0,6	0,7257	+2,7	0,9965
—1,3	0,0968	+0,7	0,7580	+2,8	0,9974
—1,2	0,1151	+0,8	0,7881	+2,9	0,9987
—1,1	0,1357	+0,9	0,8159	+3,0	0,9981
—1,0	0,1587	+1,0	0,8413		

Теперь, как это показано на рис. 2.23, наши кривые частот различных совокупностей рода *Compositа* идентичны. Характеристики стандартного нормального распределения очень хорошо известны, а таблицы площадей, ограниченных указанными сегментами кривой, можно найти почти во всех учебниках по статистике. Напомним, что площади выражаются прямо через вероятности. Используя сокращенную таблицу (например, табл. 2.10), можно найти любую вероятность, связанную со случайной выборкой из нормальной совокупности, значения которой расположены в некотором заданном интервале. Однако для этого нужно знать дисперсию совокупности.

Давайте сделаем нереальное предположение, что мы исследовали всю совокупность рода *Compositа*. Это значит, что мы знаем среднее значение длин ее элементов, равное 14,2 мм, и их стандартное отклонение, равное 4,7 мм. Какова вероятность появления при случайном выборе образца, меньшего 3 мм? Для получения ответа на этот вопрос приведем 3 мм к единицам стандартного отклонения и затем обратимся к табл. 2.10:

$$Z = (3{,}0 - 14{,}2)/4{,}7 = -2{,}4.$$

Вероятность получения представителя совокупности рода *Composita*, длина которого меньше — 2,4 стандартных отклонений, есть кумулятивная вероятность в этой точке: по нашей табл. 2.10 найдем значение 0,0082, которое в действительности очень мало. Теперь вычислим вероятность появления представителя, длина которого превышает 20 мм.

Снова требуемую величину преобразуем в стандартную нормальную форму:

$$Z = (20{,}0 - 14{,}2)/4{,}7 = 1{,}2.$$

Так как суммарная площадь под кривой нормального распределения равна 1,00, то вероятность получения величины x, равной или большей 1,2 стандартных отклонений, т. е. большей, чем среднее, равна разности 1,00 и кумулятивной вероятности получения значений, не превосходящих 1,2. Иначе говоря,

$$P(x \geqslant 1{,}2) = 1{,}0 - P(x < 1{,}2).$$

Табл. 2.10 дает нам кумулятивные вероятности вплоть до 1,2, и вычитаемая вероятность равна 0,8849. Поэтому вероятность появления особей *Composita* длиннее 20 мм равна 1,0000 — 0,8849 = —0,1151, или немногим больше одной десятой. Теперь вычислим вероятность случайного выбора *Composita*, длина которой попадает в интервал от 15 до 20 мм:
для 15 мм

$$Z = (15{,}0 - 14{,}2)/4{,}7 \approx 0{,}2,$$

для 20 мм

$$Z = (20{,}0 - 14{,}2)/4{,}7 \approx 1{,}2,$$

$$P(x \leqslant 1{,}2) = 0{,}8849,$$

$$P(x \leqslant 0{,}2) = 0{,}5793,$$

$$P(0{,}2 \leqslant x \leqslant 1{,}2) = 0{,}3056,$$

т. е. примерно одна треть образцов попадает в заданный интервал.

Центральная предельная теорема

В этом примере предполагалось, что выборка была сделана из нормально распределенной совокупности. К сожалению, мы обычно не знаем, какой вид имеет распределение, и иногда подозреваем, что оно значительно отличается от нормального. Из этого не следует, что нормальное распределение бесполезно, так как имеет место замечательная центральная предельная теорема. Она утверждает, что если выборки извлечены случайно из любой совокупности, то средние, вычисленные для этих данных, а именно выборочные средние, являются случай-

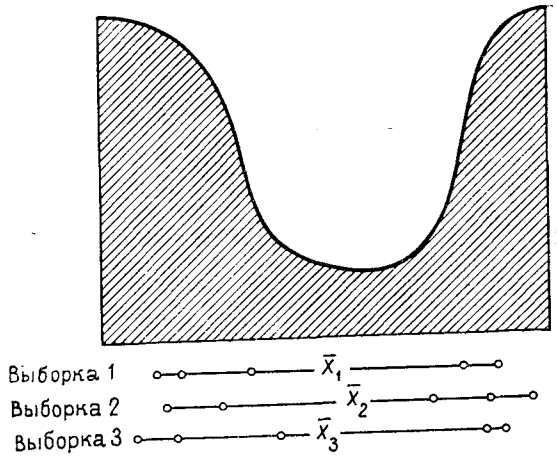


Рис. 2.24. Три выборки из пяти наблюдений, взятые наудачу из совокупности с U-образным распределением. Средние значения выборок обозначены через $\bar{X}$

ными величинами, распределение которых стремится к нормальному при увеличении объема выборки.

Центральная предельная теорема кажется на первый взгляд не вполне понятной; трудно понять, почему средние выборок должны подчиняться нормальному распределению, если образцы были выбраны из совокупности совершенно другого типа. Однако моделирование позволяет убедиться в том, что эта теорема на самом деле верна. Предположим, что мы производим выборку из совокупности, имеющей совершенно отличное от нормального распределения U-образного вида, как это показано на рис. 2.24. Большая часть индивидуальных наблюдений в выборке будет получена из двух краев распределения, которые содержат пик совокупности. Когда эти значения усредняются с целью нахождения среднего арифметического, большие значения погашаются низкими значениями, в результате получается среднее, близкое к центру распределения. Только в очень редких обстоятельствах, когда все случайно выбранные наблюдения окажутся близкими либо к высоким значениям, либо к самым низким, при вычислении среднего мы получим значение, которое сильно отличается от центрального.

Заметим, что выборочные средние значения кластеризуются (собираются в пучки) вблизи центрального значения гипотетического распределения на рис. 2.24. Если этот эксперимент повторить тысячу раз и больше, то окажется, что выборочные средние будут располагаться наподобие хорошо известной колоколообразной нормальной кривой. По существу, те же самые

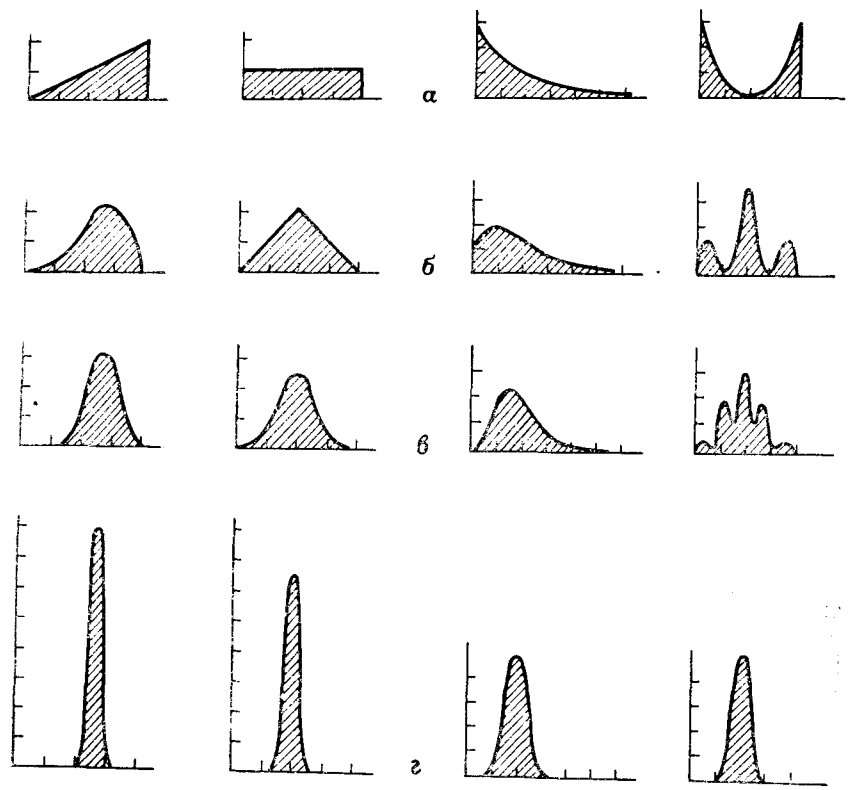


Рис. 2.25. Распределение среднего $\overline{X}$ для большого числа выборок объема n, взятых наудачу из совокупностей, с распределением, отличающимся от нормального.

Центральная предельная теорема позволяет утверждать, что распределение $\overline{X}$ по мере увеличения n стремится к нормальному. *а* — распределения исходных совокупностей, из которых взяты выборки; *б—г* — распределение $\overline{X}$ для выборов объема $n=2$ (*б*), $n=4$ (*в*) и $n=25$ (*г*) [22]

результаты будут получены, если начать почти с любого другого исходного распределения, как, например, изображено на рис. 2.25, взятом из книги Л. Л. Лапина [22].

Так как распределение выборочных средних стремится к нормальному, то его можно описать только двумя статистиками — средним и дисперсией. Как теоретическое, так и эмпирическое исследования показывают, что среднее значение выборочных средних равно среднему совокупности, т. е. $\overline{X}_{\overline{X}} = \mu$. Дисперсия выборочной средней равна дисперсии совокупности, деленной на объем выборки, или $s^2_{\overline{X}} = \sigma^2/n$. Стандартное отклонение выборочных средних есть квадратный корень из этого числа и называется стандартной ошибкой оценки среднего, или

просто стандартной ошибкой. Оно описывает изменчивость, которую можно ожидать от средних выборок при повторном случайном выборе из той же совокупности. Стандартная ошибка равна

$$s_e = \sqrt{\sigma^2/n},$$

или эквивалентно

$$s_e = \sigma\sqrt{1/n}. \qquad (2.27)$$

Центральная предельная теорема позволяет сформулировать статистические критерии, основанные на характеристиках нормальной кривой, и применять их даже в тех случаях, когда совокупность, из которой взята выборка, не распределена нормально. Предположим, что палеонтолог, который занимался исследованием коллекции *Compositа*, нашел очень большую плиту, покрытую брахиоподами. Ископаемые выглядят аналогично *Composita*, но по размерам очень велики, средняя длина десяти образцов составляет примерно 30,0 мм. Напомним, что мы «знаем», что среднее и стандартное отклонения совокупности *Composita* соответственно равны примерно 14,2 и 4,7 мм. Можно ли считать, что новая выборка брахиопод была извлечена из этой совокупности?

Мы можем определить разность между средним значением нашей новой выборки и средним значением совокупности. Эту разность затем можно сравнить с изменчивостью, которую мы бы хотели иметь для средних значений выборок, случайно извлеченных с заданной совокупности. Эта изменчивость задается стандартной ошибкой и является функцией как дисперсии совокупности, так и объема выборки.

Сравнение между разностью средних и стандартной ошибкой можно осуществить по следующей формуле

$$Z = \frac{\overline{X} - \mu}{s_e} = \frac{\overline{X} - \mu}{\sigma\sqrt{\frac{1}{n}}}. \qquad (2.28)$$

Заметим, что проверяемая статистика вычисляется таким образом, что она в точности эквивалентна критерию, используемому для стандартизации переменной (см. уравнение 2.26). Проверяемая статистика Z нормально распределена со средним значением, равным нулю, и стандартным отклонением, равным единице, если выборочное среднее действительно было получено для гипотетической совокупности. Если Z крайне велико, то мы вправе заключить, что наша выборка не была взята из этой совокупности. Формальное решение, однако, требует, чтобы мы установили соответствующую процедуру для вычисления проверяемой статистики.

Первый шаг в статистической проверке гипотез — формулировка подходящей гипотезы об исследуемой переменной. Обычно такая гипотеза называется нулевой, обозначается H_0 и, в сущности, является гипотезой об отсутствии различия. Например, можно предположить, что данная выборка взята из совокупности, имеющей заданное среднее значение. Нулевая гипотеза выражается в форме

$$H_0 : \mu_1 = \mu_0, \tag{2.29}$$

которая означает, что среднее значение μ_1 изучаемой совокупности, из которой была взята выборка, равно заданному среднему значению.

В нашем примере мы должны будем предположить, что среднее значение совокупности, из которой были взяты брахиоподы, находящиеся на плите, совпадает со средним значением совокупности рода *Composita*.

Сформулировав нулевую гипотезу, мы должны указать и альтернативу к ней. Подходящая альтернатива в этой ситуации может быть следующей:

$$H_1 : \mu_1 \neq \mu_0, \tag{2.30}$$

т. е. что среднее значение совокупности, из которой была взята выборка, не равно заданному значению μ_0. Теперь рассмотрим процедуры проверки гипотез при заданном уровне значимости. Если две изучаемые совокупности окажутся различными, следует сделать вывод, что ископаемые остатки были взяты не из совокупности рода *Composita*, а из совокупности некоторого другого рода.

Как только гипотеза сформулирована, можно на основании нашего статистического критерия принять ее или отвергнуть. Гипотеза также может быть истинной или ложной. Это приводит к тому, что возникает четыре комбинации возможных исходов, две из которых приводят к правильному выводу, а две — к неправильному. Это можно проиллюстрировать следующим образом:

	Гипотеза верна	Гипотеза неверна
Гипотеза принимается	Правильное решение	Ошибка второго рода
Гипотеза отвергается	Ошибка первого рода	Правильное решение

Только принятие правильной или отклонение неправильной гипотезы можно считать верным решением. Если нулевая гипотеза отвергается, а на самом деле она верна, то возникает ошибка, называемая ошибкой первого рода. Наоборот, если ошибочная гипотеза принимается, то совершается ошибка второго рода. Возвращаясь к нашему примеру, проиллюстрируем введенные понятия:

Гипотеза	В действительности	
	Особи с плиты принадлежат совокупности	Особи с плиты не принадлежат совокупности
μ плиты $=\mu_0$	Правильное решение	Ошибка второго рода
μ плиты $\neq\mu_0$	Ошибка первого рода	Правильное решение

Здесь «μ плиты» относится, конечно, к среднему значению совокупности, к которой принадлежат особи, собранные с плиты.

В распространенных статистических процедурах вероятность появления ошибки первого рода обозначается через α и называется уровнем значимости; эту вероятность можно задать до применения критерия. Для того чтобы минимизировать вероятность появления ошибки второго рода, запишем нулевую гипотезу при условии, что она будет отклонена. Если гипотеза отклоняется, то вероятность появления ошибки второго рода равна нулю, тогда как вероятность появления ошибки первого рода известна, так как она задается заранее. Если, однако, критерий не приводит к отклонению нулевой гипотезы (т. е. нулевая гипотеза принимается), то появляется некоторая вероятность сделать ошибку второго рода. Эта вероятность β, вообще говоря, неизвестна. Таким образом, если гипотеза о равенстве средних отвергается, мы делаем вывод о том, что две изучаемые совокупности имеют различные средние значения и вероятность того, что принято ошибочное решение, равна α. С другой стороны, если H_0 не отвергается, утверждение о том, что средние двух совокупностей совпадают, может оказаться ложным с неизвестной вероятностью β.

Все статистические критерии основаны на предположении, что нулевая гипотеза и альтернатива к ней взаимно исключают друг друга и вместе образуют полное множество событий. Так как нулевая гипотеза записывается в явном виде, то альтернатива должна быть довольно общей. Если H_0 отвергается, то мы считаем, что заданное соотношение, описываемое нулевой гипотезой, не выполняется. Более того, истинное соотношение в этом случае содержится в обширном множестве альтернатив, заключенных в общей альтернативе. Мы не можем определить, какое из соотношений истинно; мы можем только установить, какое из соотношений не выполняется. Иногда в математической статистике применение статистических критериев позволя-

ет говорить об «опровержении нулевой гипотезы» против альтернативы о неуспехе опровержения. Неуспех опровержения, которому соответствует неизвестная вероятность принятия ошибочного решения, не служит эквивалентом принятия гипотезы. Статистические критерии в некотором смысле не могут сказать нам, что именно имеет место, а только могут сказать, чего нет.

Возвращаясь к нулевой гипотезе и альтернативе, определенной формулами (2.29) и (2.30), предположим, что мы сочли уровень значимости (т. е. вероятность ошибки первого рода) $\alpha = 0{,}05$ подходящим для наших целей. Иными словами, мы допускаем возможность приблизительно 5 раз на 100 испытаний ошибочно отвергнуть проверяемую гипотезу в случае, когда она верна.

Предположим, что дисперсия совокупности, по отношению к которой ведется проверка, нам известна. Палеонтолог определил, что дисперсия значений длины для совокупности особей рода *Compositа* равна 22,1 (напомним, что стандартное отклонение было 4,7). Теперь можно формально записать статистический критерий следующим образом:

1) пусть проверяемая гипотеза и альтернатива имеют вид

$$H_0 : \mu_1 = \mu_0,$$
$$H_1 : \mu_1 \neq \mu_0;$$

2) принимаем уровень значимости: $\alpha = 0{,}05$;
3) вычисляем статистический критерий:

$$Z = \frac{\overline{X} - \mu_0}{\sigma\sqrt{1/n}}. \qquad (2.31)$$

Если выборка взята наудачу из нормальной совокупности с известной дисперсией, то статистический критерий Z будет распределен нормально со средним значением, равным нулю, и дисперсией, равной единице. Мы приняли соглашение о том, что приблизительно один раз на 20 испытаний допускается ошибочное отклонение гипотезы о равенстве средних, в то время как она верна. Иными словами, мы принимаем 5%-ный уровень риска или вероятность ошибки первого рода равную 0,05. Определим для стандартизованного нормального распределения область, заключающую 5% площади под кривой нормального распределения. Эта область называется критической. Если вычисленное значение статистического критерия попадает в эту область, мы вынуждены отклонить нулевую гипотезу.

Так как альтернатива — просто одно из неравенств, то гипотеза будет отклонена, если значение критерия слишком велико или слишком мало. Это значит, что существуют три возможные ситуации: $\mu_1 = \mu_0$; $\mu_1 > \mu_0$ или $\mu_1 < \mu_0$. В данном случае нас не

интересует различие между двумя последними неравенствами. Критическая область охватывает крайние значения оси абсцисс, причем каждая подобласть занимает 2,5% площади, ограниченной кривой нормального распределения.

Сказанное можно резюмировать следующим образом: мы знаем характеристики нормальной кривой, которые получены из теоретических соображений, и поэтому их эмпирическое использование вполне оправданно. Если дисперсия нормально распределенной совокупности известна, то мы знаем также процентное содержание особей, размеры которых заключены в различных пределах (например, две трети особей приходится на интервал с центром в среднем значении, имеющий длину, равную двум стандартным отклонениям). Если особи извлечены из этой совокупности случайным образом, вероятность получения выборки в заданном интервале кривой распределения равна площади, заключенной под соответствующей частью этой кривой. Если выборка взята из области, соответствующей очень малой вероятности, то это значит, что наша выборка не является выборкой из совокупности, указанной проверяемой гипотезой, которую мы отвергаем. Однако имеется некоторая вполне определенная вероятность извлечь выборку из критической области совокупности, равная площади этой критической области.

Возвращаясь к примеру рода *Compositа*, напишем:

1) $H_0 : \mu$ плиты $= 14{,}2$ мм;
   $H_1 : \mu$ плиты $\neq 14{,}2$ мм;
2) $\alpha = 0{,}05$;
3) $Z = \dfrac{30{,}0 - 14{,}2}{4{,}7\sqrt{1/6}} \approx 8{,}2.$

Мы уже знаем, что гипотеза о равенстве средних отвергается, если выборочное среднее либо слишком велико, либо слишком мало. Это приводит к двустороннему критерию, представленному на рис. 2.26. Критическая область, которая по соглашению должна содержать 5% площади нормального распреде-

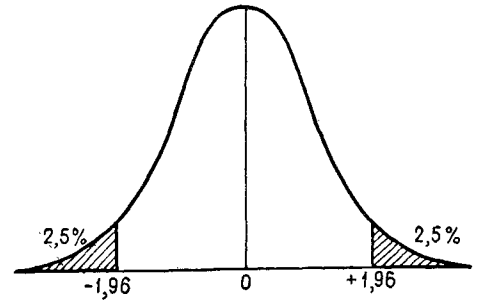


Рис. 2.26. Кривая нормального распределения с двумя заштрихованными критическими областями, охватывающими 5% площади под кривой

ления, распадается на две части, причем каждая из них содержит 2,5% общей площади. Если вычисленное значение Z попадает в левую половину, то мы делаем вывод, что выборка извлечена из совокупности, имеющей меньшее среднее значение, чем данная совокупность. Наоборот, если оно попадает в правую половину, то среднее выборочной совокупности больше, чем среднее заданной совокупности. Из табл. 2.10 мы находим, что приблизительно 2,5% площади под кривой находится слева от значения Z, равного $-1{,}9$, и 97,5% (100% — 2,5% = 97,5%) — справа от значения $+1{,}9$. Вычисленное значение критерия 8,2 превышает 1,9, из чего мы делаем вывод, что средние значения двух совокупностей не равны между собой, и коллекция ископаемых остатков на плите должна принадлежать к роду, отличному от рода *Composita*.

Необходимо отметить те допущения, которые делаются при использовании указанного критерия. Критерий Z основан на предположениях:

1) выборка брахиопод извлечена случайным образом;

2) совокупность длин остатков *Composita* распределена нормально;

3) дисперсия длин остатков *Composita* известна и равна 22,1 мм.

Если в частном примере какое-либо из указанных предположений является необоснованным, результаты, полученные с применением Z-критерия, могут показаться сомнительными. Тогда следует обратиться к другой процедуре принятия решений, основанной на предположениях, более отвечающих случаю.

Значимость

Прежде чем продолжать перечень статистических критериев, полезно сделать несколько комментариев относительно выбора уровня значимости. Во многих статистических руководствах, в частности тех, которые касаются вопросов сельского хозяйства или промышленного контроля качества, в примерах и упражнениях обычно используются уровни значимости один к двадцати ($\alpha=0{,}05$) или один к тысяче ($\alpha=0{,}001$). Казалось, что подобная практика могла бы помочь обосновать целесообразность такого выбора, однако это не так. Определение уровня значимости находится целиком в компетенции исследователя, он должен решить, какой риск при отклонении истинной гипотезы является допустимым.

В геологии мы часто имеем дело с обстоятельствами большой неопределенности, и кажется мало реальным, что мы можем позволить себе сделать ошибку только в одном случае из тысячи или даже в одном случае из двадцати. Если выбрать очень стеснительные уровни значимости, мы увидим, что нуле-

Рис. 2.27. Распределение статистического критерия с критической областью, определяющей отклонение гипотезы о том, что разведка безрезультатна:

а — критическая область для уровня значимости $\alpha=0{,}05$ *б* — критическая область для $\alpha=0{,}40$. *1* и *1'* — соответственно редкая и частая сети бурения; *2* и *2'* — соответственно редкие и частые ошибки первого рода; *3* и *3'* — соответственно большая и малая вероятности пропуска залежи

вую гипотезу мы никогда не отвергнем, и будем нуждаться во все большем и большем объеме данных, которые не имеем возможности получить. Выбирая более скромные уровни значимости, можно быстрее прийти к заключению, хотя вероятность получить ошибочные выводы может оказаться очень высокой в сравнении со стандартами, принятыми в других областях.

На рис. 2.27 проиллюстрирован эффект от принятия различных уровней значимости для некоторого гипотетического статистического критерия в прогнозе нефтеносности. Представим себе, что компания нашла некоторые количественные переменные, позволяющие определить приоритеты при бурении, цель которого — убедиться в правильности прогноза продуктивности скважин. Компания применяет статистический критерий к этим переменным с целью решить, продолжать бурение скважины или лучше оставить ее. Нулевая гипотеза состоит в том, что образцы взяты из совокупности бесперспективных объектов; альтернатива состоит в том, что они взяты из совокупности перспективных продуктивных объектов.

Если согласиться принять уровень значимости, например $\alpha=0{,}05$, и нанести его на рисунок так, как это сделано на рис. 2.27, *а*, то очень немногие прогнозы окажутся отличающимися от неперспективной нулевой совокупности. Если же окажется, что они отличны от нее, то это почти наверняка даст открытие при бурении. Компания получит очень высокое отношение для числа успехов, но при этом пропустит много объектов, которые могли бы оказаться продуктивными. В итоге компания будет редко бурить, редко ошибаться и оставит много резервуаров неоткрытыми.

Теперь представим на рис. 2.27, *в* такой уровень значимости, как $\alpha=0{,}40$. Тогда многие прогнозные участки придется бурить,

однако вероятность отрицательного исхода будет значительно выше. Полагаясь на это правило принятия решения, компания будет часто бурить, часто ошибаться, но и намного меньше нефти останется неоткрытой.

В нефтяной промышленности рассматриваются последствия получения отрицательного результата при бурении, если нефть существует (ошибка второго рода), значительно чаще, чем последствия получения положительного результата при бурении пустых скважин (ошибка первого рода). Причина этого состоит в том, что финансовый успех одного большого открытия часто может покрыть стоимость десятков или даже сотен пустых скважин. Оценка вероятности успеха в нефтяной промышленности США при бурении методом «дикой кошки» равна примерно 10%. Если бы эти скважины были пробурены на основе применения статистических критериев, то эта оценка соответствовала бы уровню значимости примерно $\alpha = 0{,}90$.

Это, вероятно, крайний случай, но он показывает, что уровень значимости следует выбирать в соответствии с конкретными обстоятельствами, при которых используется критерий. Значения уровня α основываются на оценке последствий, которые возникнут, если сделать ошибку первого рода. Эти последствия могут быть осязаемыми и привести к потере денег, времени и даже жизней, или они могут быть неосязаемыми и приводить к ущербу профессиональной репутации или личной гордости. Для того чтобы сохранить интеллектуальную честность, исследователь должен принимать решения на границе области риска и соответствующим образом выбирать уровень значимости. Выбор уровня значимости после проведения проверки критерия, когда результаты уже известны, — это бесстыдное искажение фактов. Полученные таким образом значения уровней могут отражать лишь желание исследователя принять или отклонить гипотезу, а не дать беспристрастную оценку имеющегося риска.

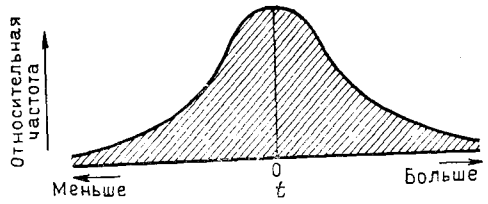
t-КРИТЕРИЙ

Для того чтобы применить описанный выше критерий, нужно выполнить ряд условий, которые редко осуществимы на практике. Мы обычно не знаем истинных значений параметров изучаемого распределения, так как не можем изучить всей совокупности рода *Compositа*, и ясно, что это нельзя сделать.

Так как μ и σ неизвестны, то лучшее, что можно сделать, — это оценить их по выборке. Однако такие оценки допускают некоторую степень неопределенности, поэтому решения, принимаемые на их основе, нельзя считать точными.

Неопределенность, возникающую как следствие применения оценок, построенных по выборке, можно учесть, если использовать распределение с более широкой областью значений, чем у

Рис. 2.28. t-распределение Стьюдента



нормального распределения. Одно из распределений такого типа называется t-распределением Стьюдента. Оно похоже на нормальное, но зависит от объема взятой выборки. Типичная кривая этого распределения изображена на рис. 2.28. Форма кривой меняется в зависимости от числа наблюдений. Когда число наблюдений в выборке бесконечно, то t-распределение совпадает с нормальным.

Степени свободы

Для того чтобы подсчитать значения статистического критерия, нужно по выборочным данным оценить параметры изучаемой совокупности. Интуитивно кажется невозможным решить сразу две задачи: оценить параметры совокупности и применить критерий, используя одну и ту же выборку без какой-либо компенсации, связанной с двукратным обращением к имеющемуся набору наблюдений. В связи с этим вводится величина, называемая числом степеней свободы, которую можно определить как разность между числом наблюдений в выборке и числом параметров, которые требуется оценить по выборочным данным. Иными словами, число степеней свободы — превышение числа наблюдений над числом оцениваемых параметров распределения. Числа степеней свободы обозначаются греческой буквой ν, это всегда целые положительные числа.

В качестве примера рассмотрим рис. 2.29, на котором представлено вычисление среднего и дисперсии выборки. Среднее

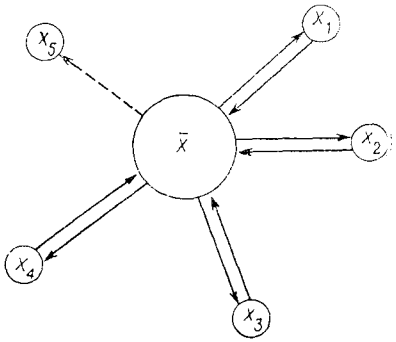


Рис. 2.29. Представление в виде диаграммы метода вычисления среднего и дисперсии по пяти наблюдениям.

Среднее $\bar{X}$ вычисляется по всем наблюдениям; дисперсия — по разности между наблюдениями и средним. Когда четыре разности найдены, пятая разность известна

оценивается по пяти независимым наблюдениям и поэтому имеется пять степеней свободы. Дисперсия оценивается по пяти квадратам разностей $(\overline{X}-X_i)^2$. Однако заметим, что если мы определили четыре из этих разностей, то автоматически можно вычислить пятую, так как

$$\overline{X} - X_5 = 5\overline{X} - (X_1 + X_2 + X_3 + X_4).$$

Поэтому имеется только четыре независимых источника информации, по которым вычисляется дисперсия.

К сожалению, понятие степеней свободы редко объясняется в начальных курсах статистики, скорее оно представляется как очевидное произвольное число, например $n-1$. Отличное общее изложение этого вопроса как в физическом, так и в статистическом контексте содержится в книге Уолкера [32]. Мы вкратце будем дальше рассматривать причины различия чисел степеней свободы, ассоциированных с различными статистическими критериями по мере их появления в тексте.

Таблицы t-распределения (и других выборочных распределений) используются точно таким же образом, как и таблицы кумулятивного стандартного нормального распределения; отличие состоит лишь в том, что для нахождения требуемой вероятности в таблице t-распределения надо знать два числа: α — заданный уровень значимости (вероятность ошибки первого рода) и число степеней свободы ν. Табл. 2.11 является сокращенным вариантом таблицы значений t-распределения; более подробные таблицы можно найти во многих руководствах по математической статистике.

Так называемые t-критерии, которые основаны на распределении Стьюдента, полезны для проверки гипотезы о том, что данная выборка взята из совокупности с заданными характеристиками или же для проверки гипотезы об однородности двух выборок. Проблемы такого типа рассматриваются во вводных курсах в математическую статистику и являются основными в экспериментальных науках и в области контроля качества продукции.

Пусть, например, нужно проверить гипотезу, заключающуюся в том, что ряд образцов песчаника Тенслип, результаты анализов которых приведены в табл. 2.12, взят из совокупности, имеющей среднюю пористость более 18%. Допустив, что образцы были взяты наудачу из нормальной совокупности, вычислим t-критерий:

$$t = \frac{\overline{X} - \mu_0}{s_e} = \frac{\overline{X} - \mu_0}{s\sqrt{1/n}}, \qquad (2.32)$$

где $\overline{X}$ — среднее арифметическое, вычисленное по данным выборки; μ_0 — гипотетическое среднее, равное 18%; n — число

Таблица 2.11

Критические значения t-критерия при ν степенях свободы и заданном уровне значимости [17]

ν	Уровень значимости, α, %					
	10	5	2,5	1	0,5	0,1
1	3,078	6,314	12,706	31,821	63,657	318,310
2	1,886	2,920	4,303	6,965	9,925	22,327
3	1,638	2,353	3,182	4,541	5,841	10,215
4	1,533	2,132	2,776	3,747	4,604	7,173
5	1,476	2,015	2,571	3,365	4,032	5,893
6	1,440	1,943	2,447	3,143	3,707	5,208
7	1,415	1,895	2,365	2,998	3,499	4,785
8	1,397	1,860	2,306	2,896	3,355	4,501
9	1,383	1,833	2,262	2,821	3,250	4,297
10	1,372	1,812	2,228	2,764	3,169	4,144
11	1,363	1,796	2,201	2,718	3,105	4,025
12	1,356	1,782	2,179	2,681	3,055	3,930
13	1,350	1,771	2,160	2,650	3,012	3,852
14	1,345	1,761	2,145	2,624	2,977	3,787
15	1,341	1,753	2,131	2,602	2,947	3,733
16	1,337	1,746	2,120	2,583	2,921	3,686
17	1,333	1,740	2,110	2,567	2,898	3,646
18	1,330	1,734	2,101	2,552	2,878	3,610
19	1,328	1,729	2,093	2,539	2,861	3,579
20	1,325	1,725	2,086	2,528	2,845	3,552
21	1,323	1,721	2,080	2,518	2,831	3,527
22	1,321	1,717	2,074	2,508	2,819	3,505
23	1,319	1,714	2,069	2,500	2,807	3,485
24	1,318	1,711	2,064	2,492	2,797	3,467
25	1,316	1,708	2,060	2,485	2,787	3,450
26	1,315	1,706	2,056	2,479	2,779	3,435
27	1,314	1,703	2,052	2,473	2,771	3,421
28	1,313	1,701	2,048	2,467	2,763	3,408
29	1,311	1,699	2,045	2,462	2,756	3,396
30	1,310	1,697	2,042	2,457	2,750	3,385
40	1,303	1,684	2,021	2,423	2,704	3,307
60	1,296	1,671	2,000	2,390	2,660	2,232
120	1,289	1,658	1,980	2,358	2,617	3,160
∞	1,282	1,645	1,960	2,326	2,576	3,090

наблюдений, s — оценка стандартного отклонения; s_e — стандартная ошибка определения среднего значения. Заметим, что этот критерий совпадает с критерием (2.31), исключая то, что нужно оценить стандартную ошибку по формуле $s_e = s\sqrt{1/n}$, а не по формуле $\sigma\sqrt{1/n}$, так как истинная дисперсия совокупности неизвестна.

Формально мы проверяем гипотезу

$$H_0 : \mu_1 \leqslant \mu_0$$

при множестве альтернатив

$$H_1 : \mu_1 > \mu_0.$$

Таблица 2.12

Результаты измерения пористости десяти образцов песчаников Тенслип пенсильванского возраста, владина Бигхорн, Вайоминг

Номер образца	Пористость, %
01	13
02	17
03	15
04	23
05	27
06	29
07	18
08	27
09	20
10	24
Сумма	213
Среднее	21,3

$s^2 = 30{,}46$, $s = 5{,}52$, $s_e = 0{,}57$

Проверяемая гипотеза заключается в том, что среднее значение пористости совокупности, из которой была взята выборка, меньше или равно заданному значению 18%. Множество альтернатив заключается в том, что изучаемая совокупность имеет среднюю пористость, превосходящую 18%.

Для определения критического значения t по табл. 2.11 требуется задать два числа: уровень значимости и число степеней свободы. В данном примере предполагается, что один параметр (μ) известен, а другой требуется оценить (оценкой для σ является величина s, т. е. выборочное стандартное отклонение). Поэтому выборке, содержащей десять измерений пористости, соответствуют девять степеней свободы.

Нулевая гипотеза отвергается только в том случае, когда средняя пористость существенно превышает 18%, и поэтому попадающими в критическую область можно считать только очень большие значения критерия, как это показано на рис. 2.30. Такой критерий называется односторонним, так как его критическая область расположена только с одной стороны области

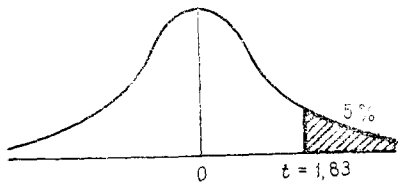


Рис. 2.30. Распределение Стьюдента с девятью степенями свободы

значений распределения. Если же нам нужно проверить эту гипотезу при уровне значимости $\alpha = 0{,}05$, то вычисленное значение статистики t для одностороннего критерия должно превышать значение 1,83. Статистический критерий имеет тот же вид, что и в предыдущем случае:

1) $H_0 : \mu_1 \leqslant 18\%$,
   $H_1 : \mu_1 > 18\%$;
2) $\alpha = 0{,}05$;
3) $t = \dfrac{21{,}3 - 18{,}0}{5{,}52\sqrt{1/10}} = 1{,}89$.

Вычисленное значение 1,89 превышает табличное, соответствующее девяти степеням свободы и 5%-ному уровню значимости, т. е. попадает в критическую область. Это значит, что мы должны отклонить нулевую гипотезу и принять альтернативу, заключающуюся в том, что пористость совокупности, из которой были извлечены образцы песчаников Тенслип, больше 18%. Если бы вычисленная величина t оказалась меньше чем 1,83, то не было бы никаких оснований предполагать, что выборочное среднее больше 18%. Заметим, что мы при этом не утверждаем, что среднее меньше 18%, а только говорим, что нет оснований считать, что оно больше. Ранее было установлено, что эта неопределенность лежит в основе статистических критериев. Они могут показать с некоторой вероятностью, чего нет, но не позволяют установить, что же имеет место.

С другой площади в Вайоминге были получены десять дополнительных измерений значений пористости в песчаниках Тенслип, которые приведены в табл. 2.13. Можно ли средние двух выборок считать равными? В отличие от предыдущей задачи, где мы сравнивали выборочное среднее с заданным выборочным средним значением совокупности, в данном случае проверяется гипотеза, имеющая следующий вид:

$$H_0 : \mu_1 = \mu_2.$$

Проверяемая гипотеза заключается в том, что среднее значение совокупности, из которой взята первая выборка, равно среднему значению совокупности, из которой взята вторая выборка. Множество альтернатив для гипотезы

$$H_1 : \mu_1 \neq \mu_2$$

утверждает, что средние значения двух совокупностей не равны. Снова мы должны задать уровень значимости, и пусть он будет равен 10% ($\alpha = 0{,}10$). Теперь статистический критерий имеет следующий вид:

$$t = (\bar{X}_1 - \bar{X})/s_e, \qquad (2.33)$$

Таблица 2.13

Результаты измерения пористости десяти образцов песчаников Тенслип пенсильванского возраста, бассейн реки Уинд, Вайоминг

Номер образца	Пористость, %
11	15
12	10
13	15
14	23
15	18
16	26
17	24
18	18
19	19
20	21

Сумма 189

Среднее 18,9

$s^2 = 28{,}21$, $s = 4{,}82$

где s_e — оценка стандартного отклонения разности между $\bar{X}_1$ и $\bar{X}_2$, полученная по двум объединенным выборкам. Эту оценку s_e можно вычислить по формуле

$$s_e = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Здесь s_p — объединенная оценка стандартного отклонения, найденная комбинацией двух выборочных дисперсий:

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}, \qquad (2.34)$$

где индексы соответствуют выборкам из бассейнов Бигхорн и Уинд Ривер. Процесс объединения двух выборок приводит к дополнительным степеням свободы, так как требуется оценить два параметра σ_1^2 и σ_2^2. Число степеней свободы поэтому для t-критерия эквивалентности, заданного формулой (2.32), есть $\nu = n_1 + n_2 - 2$. Является ли различие между двумя средними значимым при десятипроцентном уровне значимости?

$$s_p^2 = \frac{9(30{,}46) + 9(23{,}21)}{10 + 10 - 2} = \frac{483{,}03}{18} = 26{,}84;$$

$$s_p = 5{,}18;$$

$$t = \frac{21{,}3 - 18{,}9}{5{,}18\sqrt{1/10 + 1/10}} = \frac{2{,}4}{2{,}32} = 1{,}03.$$

Так как табличные значения двустороннего критерия с 18 степенями свободы, соответствующие 10%-ному уровню значимости (5% на каждом конце распределения), равны —2,10 и 2,10, то вычисленное значение не попадает в критическую область и нулевую гипотезу нельзя отклонить. (Напомним, что критическая область охватывает 10% площади под кривой t-распределения). Отсюда следует, что нет оснований предполагать, что две изучаемые выборки взяты из совокупностей, имеющих разные средние значения.

Для того чтобы применять этот критерий, необходимо выполнить следующие условия. Во-первых, обе выборки должны быть получены на основании процедуры случайного выбора. Во-вторых, значения случайных величин в совокупностях, из которых были извлечены выборки, должны описываться нормальным распределением. В-третьих, дисперсии этих совокупностей должны быть равны. Выполнение первого условия в большинстве геологических задач проверить трудно. Однако его невыполнение в случае, если выборки имеют сильное и систематическое смещение (как в том случае, когда измерения пористости проводятся только в образцах, взятых из продуктивных зон или нефтяных полей), может явиться серьезным источником ошибок. Конечно, проверку гипотезы о нормальности распределения значений признака изучаемой совокупности можно провести, однако одно только отклонение от нормальности редко приводит к изменению результатов, в особенности если выборочная совокупность достаточно велика. Третье условие — равенство дисперсий двух совокупностей — очень важно, так как почти все статистические критерии основаны на предположении о равенстве дисперсий сравниваемых совокупностей. К счастью, это предположение легко проверяется. Приближенные критерии применимы, если при сравнении двух выборок окажется, что они значимо различаются. Они приводятся в большинстве вводных курсов, включая те, которые перечислены в списке литературы.

Корреляционный критерий

Выше мы ввели коэффициент корреляции как стандартизованную меру линейной связи между двумя переменными, но не рассмотрели вопрос о статистической значимости этого коэффициента. Коэффициент выборочной корреляции r является оценкой параметра ρ, который отражает связь между двумя переменными совокупности. Предполагая, что обе переменные нормально распределены и наблюдения случайно выбраны из некоторой совокупности, мы можем осуществить проверку значимости r.

Проверяемая гипотеза и альтернатива таковы:

$$H_0 : \rho = 0,$$
$$H_1 : \rho \neq 0,$$

т. е. мы можем определить, значимо ли отличается от нуля выборочный коэффициент корреляции. Нулевая гипотеза устанавливает, что две переменные независимы и что любое ненулевое значение r возникло просто из-за случайных флюктуаций при случайном выборе. t-критерий значимости r задается по формуле

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad (2.35)$$

и имеет $(n-2)$ степеней свободы.

В качестве примера можно проверить значимость вычисленных выше на основе данных табл. 2.8. коэффициентов корреляции, которые мы измерили между экземплярами гальки с галечного пляжа. Первый коэффициент корреляции между осями a и b ($r_{ab}=0{,}597$) вычислен по десяти парам измерений. Проверяемая статистика вычисляется по формуле (2.35):

$$t = \frac{0{,}597\sqrt{10-2}}{\sqrt{1-0{,}597^2}} = \frac{1{,}688}{0{,}802} = 2{,}10.$$

Критическое значение t с 8 степенями свободы и 10%-ным уровнем значимости равно 1,860. Напомним, что критерий двусторонний и r может быть значительно больше или меньше нуля, так что наша область отклонения гипотезы распадается на верхнюю и нижнюю части. Так как проверяемая статистика попадает в верхнюю критическую область, мы должны заключить, что на самом деле существует корреляция между длинами наибольшей и средней осей пляжной гальки.

Для двух других корреляций множества данных, приведенных в табл. 2.8, $r_{ac}=0{,}499$ и $r_{bc}=0{,}467$. Соответственно

$$t = \frac{0{,}499\sqrt{10-2}}{\sqrt{1-0{,}499^2}} = \frac{1{,}411}{0{,}866} = 1{,}629,$$

$$t = \frac{0{,}467\sqrt{10-2}}{\sqrt{1-0{,}467^2}} = \frac{1{,}321}{0{,}884} = 1{,}494.$$

Критическое значение остается тем же самым, и мы видим, что ни одна из этих двух корреляций не отличается значимо от нуля. Другими словами, если переменные были полностью независимы друг от друга, то наблюдаемые коэффициенты корреляции возникли случайно при случайном выборе десяти образцов гальки.

F-КРИТЕРИЙ

Критерии для проверки гипотезы о равенстве дисперсий основаны на так называемом F-распределении Фишера. Это теоретическое распределение отношения $F=s_1^2/s_2^2$ двух выборочных дисперсий для выборок, взятых из нормальных совокупностей при условии, что истинные дисперсии равны.

Вполне естественно, что выборочные дисперсии в случае, когда число наблюдений, используемое для их вычисления, мало, изменяются от испытания к испытанию в довольно широком диапазоне. Поэтому вид F-распределения изменяется с изменением объема выборки. Это снова заставляет учитывать степени свободы, но в данном случае F-распределение зависит от двух значений ν, каждое из которых соответствует одной из двух оценок дисперсий F-отношения. Так как F-статистика является отношением двух положительных чисел, то ясно, что случайная величина F не может принимать отрицательных значений. Если выборка велика, то при условии равенства истинных значений дисперсий среднее значение отношения будет близко к 1,0.

Так как F-распределение описывает поведение отношений выборочных дисперсий, полученных по выборкам из одной и той же совокупности, то его можно использовать для проверки гипотезы о равенстве дисперсий.

Можно предположить, что две выборки взяты из совокупностей, характеризующихся равными дисперсиями. После вычисления F-отношения можно определить вероятность получения значения, большего или равного вычисленному для двух случайных выборок из одной нормальной совокупности. Если это значение будет неправдоподобным, то мы вынуждены считать, что выборки извлечены из различных совокупностей, имеющих неравные дисперсии.

Для любой пары оценок дисперсии можно вычислить два отношения s_1/s_2 и s_2/s_1, если принять, что бо́льшая оценка всегда будет располагаться в числителе, это отношение всегда будет больше 1,0 и статистические критерии принимают более простой вид. В этом случае достаточно использовать только односторонние критерии, и альтернативные гипотезы на самом деле являются утверждением о том, что абсолютное различие между двумя выборочными дисперсиями больше, чем можно было бы ожидать в случае, если бы истинные значения дисперсий сравниваемых совокупностей были равны. Типичный график кривой F-распределения с заштрихованной критической областью, или областью отклонения проверяемой гипотезы, изображен на рис. 2.31.

В качестве элементарного примера применения F-распределения рассмотрим две выборки результатов измерений пористо-

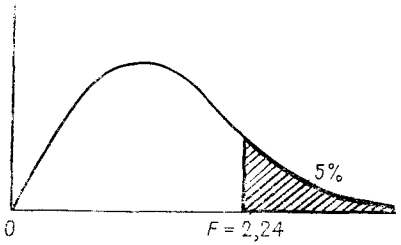


Рис. 2.31. Типичное F-распределение с $v_1=10$ и $v_2=25$ степенями свободы и заштрихованной критической областью, составляющей 5% площади под кривой

сти песчаников Тенслип. Нужно выяснить, одинакова ли дисперсия на двух сравниваемых площадях. С этой целью выберем уровень значимости 5%. Таким образом, принятие неверной гипотезы о том, что пористости различны, в то время как они одинаковы, будет происходить в среднем один раз на двадцать исходов. Оценки дисперсий двух выборок можно вычислить по формуле (2.15). Тогда соответствующее им F-отношение вычисляется по формуле

$$F = s_1^2/s_2^2, \tag{2.36}$$

где s_1^2 — большая выборочная дисперсия, а s_2^2 — меньшая выборочная дисперсия.

По этим данным требуется проверить гипотезу

$$H_0 : \sigma_1^2 = \sigma_2^2$$

против множества альтернатив

$$H_1 : \sigma_1^2 \neq \sigma_2^2.$$

Нулевая гипотеза утверждает, что изучаемые совокупности имеют равные дисперсии; множество альтернатив устанавливает, что это не так. Степени свободы ν_1 и ν_2, отвечающие этому критерию, соответственно равны n_1-1 и n_2-1. Критическое значение F с $\nu_1=9$ и $\nu_2=9$ степенями свободы и уровнем значимости 5% ($\alpha=0{,}05$) можно найти по табл. 2.14. Это значение равно 3,18.

Значение F, вычисленное по формуле (2.36), попадает в одну из двух областей. Если вычисленное значение F превышает 3,18, то нулевая гипотеза отвергается и мы приходим к заключению, что дисперсии пористости можно считать неодинаковыми в двух группах. Если вычисленное значение меньше 3,18, то мы не можем утверждать, что дисперсии различны. В качестве примера вычислим дисперсию пористости двух совокупностей песчаников Тенслип и проверим предположение о равенстве дисперсий при 5%-ном уровне значимости.

В большинстве практических задач мы обычно не знаем истинных значений параметров совокупности и можем лишь по выборке вычислить их оценки. При сравнении двух выборок сначала целесообразно установить, являются ли их дисперсии

статистически эквивалентными. Если они оказываются равными и если выборки были извлечены без смещения из изучаемых совокупностей, то мы можем спокойно перейти к использованию следующих статистических критериев.

В качестве примера рассмотрим такую задачу. Образцы снега и льда, собираемые с участков земли, покрытых многолетними льдами, содержат частички пыли, называемые микрочастицами. Размеры отдельных таких частиц имеют пределы от 0,5 до 3,0 мкм; они попадают в атмосферу разными путями: в результате вулканических извержений, пылевых штормов, падения микрометеоритов. Частицы настолько малы, что могут быть во взвешенном состоянии неограниченное время, но легко поглощаются снегом, так как служат ядрами для кристаллизации льда. Существует гипотеза, что перемешивание атмосферы и пути, которыми микрочастицы попадают в снег из воздуха, приводят к равномерной концентрации микрочастиц в снегу на земной поверхности. Если эта гипотеза соответствует действительности, то выводы из нее имеют значение для предсказания последствий испытания ядерного оружия в атмосфере. Поэтому были собраны две выборки проб снега в снеговых покровах Гренландии и Антарктики. При тщательном контроле снег расплавили и с помощью электрического классификатора определили число содержащихся в нем частиц. Концентрация микрочастиц в талом снегу приведена в табл. 2.15. Можно ли считать обе выборки извлеченными из одной и той же совокупности и приводят ли результаты испытаний к подтверждению или опровержению идеи об атмосферной однородности?

Предполагая, что выборки были взяты без смещения и что распределение микрочастиц подчиняется нормальному закону на всем протяжении снеговых полей, проверим сначала гипотезу о равенстве дисперсий по двум выборкам. Используя формулу (2.36), запишем нулевую гипотезу и альтернативу:

$$H_0 : \sigma_1^2 = \sigma_2^2,$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2.$$

Если дисперсии не сильно отличаются, то следующий шаг процедуры — проверка гипотезы о равенстве средних значений. Используя критерий (2.36), запишем соответствующую нулевую гипотезу и альтернативу:

$$H_0 : \mu_1 = \mu_2,$$
$$H_1 : \mu_1 \neq \mu_2.$$

Такая нулевая гипотеза соответствует предположению: нет никаких оснований считать, что одна область имеет большее среднее, чем другая. Ясно, что уровень значимости, соответствующий этому критерию, не может быть выше, чем уровень зна-

Критические значения F-распределения с ν_1 и ν_2 степенями

ν_2	ν_1						
	1	2	3	4	5	6	7
1	161,45	199,50	215,71	224,58	230,16	233,99	236,77
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01

чимости, использованный при проверке гипотезы о равенстве дисперсий. Если дисперсии и средние окажутся неразличимыми, т. е. нулевая гипотеза не может быть отклонена на основании выборочных данных, то нет и статистических оснований считать, что средние концентрации микрочастиц на двух изучаемых площадях соответствуют различным совокупностям. С другой стороны, если хотя бы один из двух критериев приводит к отклонению нулевой гипотезы, то вопрос об атмосферной однородности можно поставить под сомнение. Более того, если критерий (2.36) приведет к отклонению гипотезы о равенстве дисперсий, то и применение критерия (2.33) в нашей задаче теряет смысл. Приблизительные критерии, аналогичные описан-

Таблица 2.14а

свободы и 5%-ным уровнем значимости ($\alpha=0,05$) [17]

ν_1								
8	9	10	12	15	20	24	∞
238,88	240,54	241,88	243,91	245,95	248,01	249,05	250,10
19,37	19,38	19,40	19,41	19,43	19,45	19,45	19,46
8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62
6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75
4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50
4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81
3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38
3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08
3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86
3,07	3,02	2,98	2,91	2,84	2,77	2,74	2,70
2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57
2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47
2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,48
2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31
2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25
2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19
2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15
2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11
2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07
2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04
2,42	2,37	2,32	2,25	2,18	2,10	2,05	2,01
2,40	2,34	2,30	2,23	2,15	2,07	2,03	1,98
2,37	2,32	2,27	2,20	2,13	2,05	2,01	1,96
2,36	2,30	2,25	2,18	2,11	2,03	1,98	1,04
2,34	2,28	2,24	2,16	2,09	2,01	1,96	1,92
2,32	2,27	2,22	2,15	2,07	1,99	1,95	1,90
2,31	2,25	2,20	2,13	2,06	1,97	1,93	1,88
2,29	2,24	2,19	2,12	2,04	1,96	1,91	1,87
2,28	2,22	2,18	2,10	2,03	1,94	1,90	1,85
2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84
2,18	2,12	2,08	2,00	1,92	1,84	1,79	1,74
2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65
2,02	1,96	1,91	1,83	1,75	1,66	1,61	1,55
1,94	1,88	1,83	1,75	1,67	1,57	1,52	1,46

ному в работе Гюнтера [14], применимы для проверки гипотезы о равенстве средних в условиях неравенства дисперсий, но они мало помогают при решении данной задачи.

ДИСПЕРСИОННЫЙ АНАЛИЗ

До сих пор мы рассматривали только методы сравнения двух выборок, тогда как существует еще ряд задач, касающихся групп наблюдений. Предположим, например, что получено пять образцов песчаника с кальцитовым цементом. Каждый из них обладает своими литологическими особенностями: в одном бросается в глаза его крупнозернистость, другой характеризу-

Критические значения F-распределения с ν_1 и ν_2 степенями

ν_2	ν_1						
	1	2	3	4	5	6	7
1	647,79	799,50	864,16	899,58	921,85	937,11	948,22
2	38,51	39,00	39,17	39,25	39,30	39,33	39,36
3	17,44	16,04	15,44	15,10	14,88	14,73	14,62
4	12,22	10,65	9,98	9,60	9,36	9,20	9,07
5	10,01	8,43	7,76	7,39	7,15	6,98	6,85
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61
13	6,41	4,97	4,35	4,00	3,77	3,60	3,48
14	6,30	4,86	4,24	3,89	3,66	3,50	3,38
15	6,20	4,77	4,15	3,80	3,58	3,41	3,29
16	6,12	4,69	4,08	3,73	3,50	3,34	3,22
17	6,04	4,62	4,01	3,66	3,44	3,28	3,16
18	5,98	4,56	3,95	3,61	3,38	3,22	3,10
19	5,92	4,51	3,90	3,56	3,33	3,17	3,05
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01
21	5,83	4,42	3,82	3,48	3,25	3,09	2,97
22	5,79	4,38	3,78	3,44	3,22	3,05	2,93
23	5,75	4,35	3,75	3,41	3,18	3,02	2,90
24	5,72	4,32	3,72	3,38	3,15	2,99	2,87
25	5,69	4,29	3,69	3,35	3,13	2,97	2,85
26	5,66	4,27	3,67	3,33	3,10	2,94	2,82
27	5,63	4,24	3,65	3,31	3,08	2,92	2,80
28	5,61	4,22	3,63	3,29	3,06	2,90	2,78
29	5,59	4,20	3,61	3,27	3,04	2,88	2,76
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51
120	5,15	3,80	3,23	2,89	2,67	2,52	2,39
∞	5,02	3,69	3,12	2,79	2,57	2,41	2,29

ется наличием глинистых частиц, третий слабо ожелезнен и т. д. Мы хотим определить, одинаковы ли в них содержания карбоната. Для решения этой задачи можно использовать один из статистических методов, называемый дисперсионным анализом.

В общем виде этот метод основан на разделении общей дисперсии изучаемой совокупности на компоненты, соответствующие источникам изменчивости, а применяемые критерии позволят одновременно изучить различия как в средних значениях, так и в дисперсиях.

Экспериментальный подход к этой задаче заключается в дроблении образцов на более мелкие части и определении содержания карбоната в каждой из них путем взвешивания после

Таблица 2.14б

свободы и 2,5%-ным уровнем значимости ($\alpha=0{,}025$)

ν_1							
8	9	10	12	15	20	24	∞
956,66	963,28	968,63	976,71	984,87	933,10	997,24	1001,4
39,37	39,39	39,40	39,41	39,43	39,45	39,46	39,46
14,54	14,47	14,42	14,34	14,25	14,17	14,12	14,08
8,98	8,90	8,84	8,75	8,66	8,56	8,51	8,46
6,76	6,68	6,62	6,52	6,43	6,33	6,28	6,23
5,60	5,52	5,46	5,37	5,27	5,17	5,12	5,07
4,90	4,82	4,76	4,67	4,57	4,47	4,41	4,36
4,43	4,36	4,30	4,20	4,10	4,00	3,95	3,89
4,10	4,03	3,96	3,87	3,77	3,67	3,61	3,56
3,85	3,78	3,72	3,62	3,52	3,42	3,37	3,31
3,66	3,59	3,53	3,43	3,33	3,23	3,17	3,12
3,51	3,44	3,37	3,28	3,18	3,07	3,02	2,96
3,39	3,31	3,25	3,15	3,05	2,95	2,89	2,84
3,29	3,21	3,15	3,05	2,95	2,84	2,79	2,73
3,20	3,12	3,06	2,96	2,86	2.76	2,70	2,64
3,12	3,05	2,99	2,89	2,79	2,68	2,63	2,57
3,06	2,98	2,92	2,82	2,72	2,62	2,56	2,50
3,01	2,93	2,87	2,77	2,67	2,56	2,50	2,44
2,96	2,88	2,82	2,72	2,62	2,51	2,45	2,39
2,91	2,84	2,77	2,68	2,57	2,46	2,41	2,35
2,87	2,80	2,73	2,64	2,53	2,42	2,37	2,31
2,84	2,76	2,70	2,60	2,50	2,39	2,33	2,27
2,81	2,73	2,67	2,57	2,47	2,36	2,30	2,24
2,78	2,70	2,64	2,54	2,44	2,33	2,27	2,21
2,75	2,68	2,61	2,51	2,41	2,30	2,24	2,18
2,73	2,65	2,59	2,49	2,39	2,28	2,22	2,17
2,71	2,63	2,57	2,47	2,36	2,25	2,19	2,13
2,69	2,61	2,55	2,45	2,34	2,23	2,17	2,11
2,67	2,59	2,53	2,43	2,32	2,21	2,15	2,09
2,65	2,57	2,51	2,41	2,31	2,20	2,14	2,07
2,53	2,45	2,39	2,29	2,18	2,07	2,01	1,94
2,41	2,33	2,27	2,17	2,06	1,94	1,88	1,82
2,30	2,22	2,16	2,05	1,94	1,82	1,76	1,69
2,19	2,11	2,05	1,94	1,83	1,71	1,64	1,57

обработки кислотой. Каждая мелкая часть называется повторением. Цель, которую мы преследуем, разбивая первоначальный кусок на части, — определение изменчивости, вызванной погрешностями взвешивания. Очевидно, что если изменчивость между повторными определениями для одного образца велика по сравнению с различиями между образцами, то последние трудно обнаружить.

Предположим, что мы разбили исходный образец на шесть частей и собираемся проанализировать каждую из них. Наблюдаемые изменения возникают по ряду причин: из-за колебаний состава внутри исходного образца, из-за небрежности в получении повторных наблюдений (остатки одного травления могут

Критические значения F-распределения с ν_1 и ν_2 степенями

Число степеней свободы ν_2	ν_1							
	1	2	3	4	5	6	7	8
1	4052,2	4999,5	5403,4	5624,6	5763,6	5859,0	5928,4	5981,1
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99
60	7,08	4,98	4,13	3,85	3,34	3,12	2,95	2,82
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51

быть промыты более тщательно, чем остатки другого), из-за изменения условий взвешивания (повторные образцы могут содержать различные количества влаги либо на результаты взвешивания может повлиять зависимость положения нулевой точки на весах от изменений температуры в течение дня и т. д.) и благодаря влиянию других более тонких факторов. Комбинация всех этих источников изменчивости приводит к возникновению так называемой экспериментальной ошибки или изменчивости, не учитываемой только различиями между образцами.

Для того чтобы избежать возможности появления систематической ошибки в статистическом анализе, повторные наблю-

Таблица 2.14в

свободы и 1%-ным уровнем значимости ($\alpha=0{,}01$)

ν_1							
9	10	12	15	20	24	∞
6022,5	6055,8	6106,3	6157,3	6208,7	6234,6	6260,6
99,39	99,40	99,42	99,43	99,45	99,46	99,47
27,35	27,23	27,05	26,87	26,69	26,60	26,50
14,66	14,55	14,37	14,20	14,02	13,93	13,84
10,16	10,05	9,89	9,72	9,55	9,47	9,38
7,98	7,87	7,72	7,56	7,40	7,31	7,23
6,72	6,62	6,47	6,31	6,16	6,07	5,99
5,91	5,81	5,67	5,52	5,36	5,28	5,20
5,35	5,26	5,11	4,96	4,81	4,73	4,65
4,94	4,85	4,71	4,56	4,41	4,33	4,25
4,63	4,54	4,40	4,25	4,10	4,02	3,94
4,39	4,30	4,16	4,01	3,86	3,78	3,70
4,19	4,10	3,96	3,82	3,66	3,59	3,51
4,03	3,94	3,80	3,66	3,51	3,43	3,35
3,89	3,80	3,67	3,52	3,37	3,29	3,21
3,78	3,69	3,55	3,41	3,26	3,18	3,10
3,68	3,59	3,46	3,31	3,16	3,08	3,00
3,60	3,51	3,37	3,23	3,08	3,00	2,92
3,52	3,43	3,30	3,15	3,00	2,92	2,84
3,46	3,37	3,23	3,09	2,94	2,86	2,78
3,40	3,31	3,17	3,03	2,88	2,80	2,72
3,35	3,26	3,12	2,98	2,83	2,75	2,67
3,30	3,21	3,07	2,93	2,78	2,70	2,62
3,26	3,17	3,03	2,89	2,74	2,66	2,58
3,22	3,13	2,99	2,85	2,70	2,62	2,54
3,18	3,09	2,96	2,81	2,66	2,58	2,50
3,15	3,06	2,93	2,78	2,63	2,55	2,47
3,12	3,03	2,90	2,75	2,60	2,52	2,44
3,09	3,00	2,87	2,73	2,57	2,49	2,41
3,07	2,98	2,84	2,70	2,55	2,47	2,39
2,89	2,80	2,66	2,52	2,37	2,29	2,20
2,72	2,63	2,50	2,35	2,20	2,12	2,03
2,56	2,47	2,34	2,19	2,03	1,95	1,86
2,41	2,32	2,18	2,04	1,88	1,79	1,70

дения должны быть отобраны наудачу. Это так называемая рандомизация наблюдений. Необходимость этой процедуры станет очевидной, если имеется некоторый фактор, который непрерывно изменяется во время эксперимента, например продолжающееся высыхание проб, ожидающих своей очереди взвешивания. Если взвесить все шесть проб, полученных из образца 1, а затем все пробы, полученные из образца 2 и т. д., то при последнем взвешивании могут быть зарегистрированы бо́льшие весовые потери лишь по той причине, что пробы высыхали в течение более продолжительного периода времени. Один из способов решения этой задачи — последовательная нумерация каж-

Таблица 2.15

Концентрация микрочастиц в талой воде, мг/т

Антарктика, $n=16$		Гренландия, $n=18$	
3,7	0,6	3,7	1,6
2,0	1,4	7,8	2,4
1,3	4,4	1,9	1,3
3,9	3,2	2,0	2,6
0,2	1,7	1,1	3,7
1,4	2,1	1,3	2,2
4,2	4,2	1,9	1,8
4,9	3,5	3,7	1,2
		3,4	0,8

дой повторной процедуры и выбор этих номеров в процессе анализа по таблице случайных чисел. Действительно, если процесс протекает поэтапно, то целесообразно приписать наудачу номера каждому образцу на каждом шаге. Тогда различные источники ошибок перемешиваются или совмещаются для всех повторных проб, а не концентрируются в нескольких из них.

Проверку гипотезы эквивалентности пяти образцов можно провести с помощью процедуры, называемой однофакторным дисперсионным анализом, при котором проверяемая гипотеза и альтернатива имеют следующий вид:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5;$$

H_0: по крайней мере одно среднее значение отлично от остальных.

Для проверки этой гипотезы требуется выполнение некоторых условий, а именно: а) каждое множество повторных проб рассматривается как случайная выборка из соответствующей совокупности; б) каждая исходная совокупность характеризуется нормальным распределением; в) все исходные совокупности имеют равные дисперсии.

Данные для рассматриваемой нами задачи приведены в табл. 2.16. В однофакторном дисперсионном анализе общая дисперсия разбивается на две составляющие: дисперсию внутри каждого множества повторных проб (внутривыборочную дисперсию) и дисперсию между сравниваемыми образцами (межвыборочную дисперсию). В математической статистике разработана формализованная процедура дисперсионного анализа, которая приведена в таблице ANOVA (Analysis of Varianse). Последняя содержит перечень источников изменчивости, столбец исправленных сумм квадратов, соответствующих различным источникам, число степеней свободы для каждой из них, столбец средних квадратов, который содержит выборочные оценки дисперсий и значений F-критерия. Соответствующая нашей задаче таблица приведена ниже:

Таблица 2.16

Содержание карбонатного цемента в пяти образцах песчаника, % (числа в скобках обозначают порядковый номер пробы в процессе анализа)

Номер поэтапной пробы	Номер образца				
	1	2	3	4	5
1	19,2(11)	18,7(04)	12,5(28)	20,3(12)	19,3(21)
2	18,7(08)	14,3(19)	14,3(16)	22,5(30)	24,3(06)
3	21,3(09)	20,2(14)	8,7(20)	17,6(24)	17,6(18)
4	16,5(17)	17,6(07)	11,4(29)	18,4(03)	20,2(22)
5	17,3(26)	19,3(05)	9,5(27)	15,9(13)	18,4(12)
6	22,4(15)	16,1(25)	16,5(01)	19,0(02)	19,1(10)

Источник изменчивости	Сумма квадратов	Число степеней свободы	Средний квадрат	F-критерий
Между выборками	SS_A	$m-1$	MS_A	MS_A/MS_W
Внутри выборок	SS_W	$N-m$	MS_W	
Общая изменчивость	SS_T	$N-1$		

Общая изменчивость по всем наблюдениям (по всем повторным пробам и по всем образцам) SS_T характеризуется формулой

$$SS_T = \sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij}^2 - \frac{1}{N}\left(\sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij}\right)^2, \qquad (2.37)$$

где X_{ij} — i-я повторная проба в j-м образце.

В двойной сумме первая указывает, что суммирование проводится по каждому столбцу, содержащему n повторных проб, а затем складываются полученные результаты всех m столбцов. Общее число наблюдений N равно числу повторных проб в выборке, умноженному на число выборок, т. е. $N=nm$. Последний член в правой части выражения (2.37) называется поправочным. Отметим, что такие же члены имеются и в других аналогичных суммах.

Сумму, характеризующуюся межвыборочной изменчивостью, находят по следующей формуле:

$$SS_A = \sum_{j=1}^{m}\frac{1}{n}\left(\sum_{i=1}^{n} X_{ij}\right)^2 - \frac{1}{N}\left(\sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij}\right)^2, \qquad (2.38)$$

где суммирование проводится по всем повторным пробам в каждом образце $\sum_{i=1}^{n} X_{ij}$, а затем каждая из полученных сумм возводится в квадрат и полученный результат делится на число повторений n в каждой выборке; далее полученные результаты суммируются по всем выборкам и, наконец, вычитается поправочный член.

Величина, характеризующая второй источник изменчивости, имеет вид

$$SS_W = \sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij}^2 - \frac{1}{n}\sum_{j=1}^{m}\left(\sum_{i=1}^{n} X_{ij}\right)^2. \qquad (2.39)$$

Заметим, что первый член в правой части здесь такой же, как и первый член формулы (2.37) для SS_T, а последний член совпадает с первым членом формулы (2.38) для SS_A. Поэтому SS_W можно вычислить по формуле

$$SS_W = SS_T - SS_A. \qquad (2.40)$$

Число степеней свободы по всем данным равно $N-1$. Число степеней свободы для величины SS_A равно $m-1$, так как мы оцениваем ее по средним значениям каждого образца. Разность между этими числами степеней свободы равна числу степеней свободы для величины SS_W.

С целью иллюстрации этого метода дисперсионного анализа проведем вычисления по данным табл. 2.16. Используя формулу (2.37) для SS_T, получим $SS_T = 383{,}79$.

Далее мы можем подсчитать величину SS_A по средним значениям для пяти образцов. Используя формулу (2.38), просуммируем сначала все пять столбцов, возведем в квадрат полученные суммы, разделим результаты на 6 и вычтем поправочный член. Это даст нам межвыборочную сумму квадратов $SS_A = 237{,}42$.

Наконец, вычитая SS_A из SS_T, получаем внутривыборочную сумму квадратов $SS_W = 146{,}37$.

Общее число степеней свободы равно $N-1$, или 29. Так как мы оцениваем межвыборочную изменчивость по пяти измерениям (т. е. по средним значениям пяти столбцов), то число степеней свободы для SS_A равно $m-1$, т. е. 4. Разность чисел степеней свободы должна соответствовать остатку сумм квадратов или «мере» ошибки. Эта разность чисел степеней свободы равна $N-m$, или 25. Теперь вычисленные исправленные суммы квадратов SS_T, SS_A и SS_W нужно разделить на соответствующие им числа степеней свободы. В результате мы получаем оценки дисперсий (или средние квадраты, что является просто другим названием тех же величин).

Оценка общей дисперсии

$$\frac{SS_T}{N-1} = \frac{388{,}79}{29} = 13{,}23.$$

Оценка межвыборочной дисперсии

$$MS_A = \frac{SS_A}{m-1} = \frac{237{,}42}{4} = 59{,}35.$$

Оценка внутривыборочной дисперсии

$$MS_W = \frac{SS_W}{N-m} = \frac{146{,}37}{25} = 5{,}85.$$

Сущность проведенного дисперсионного анализа лучше пояснить, рассмотрев тот предельный случай, когда все повторные пробы идентичны. Тогда средние значения столбцов будут такими же, как средние по всем столбцам, и оценка дисперсии, вычисленная по всем наблюдениям, будет совпадать с оценкой, вычисленной только на основании данных одного столбца. Иными словами, мера ошибки обратится в нуль. В этом случае нет дисперсии, возникающей из-за различий между повторными пробами. Такой неправдоподобный результат должен навести на мысль, что первоначальные образцы на самом деле различны и что каждое множество повторных проб было извлечено из различных совокупностей, имеющих нулевые дисперсии.

Рассмотренный пример является менее экстремальным. Вычисляя значение F-критерия, мы получаем следующее критическое значение. Выбрав критическую область, соответствующую заданному уровню значимости и заданному числу степеней свободы, можно теперь принять или отвергнуть проверяемую гипотезу:

$$F = MS_A/(MS_W) = 10{,}14.$$

Однофакторный дисперсионный анализ применяется в тех случаях, когда требуется проверить гипотезу о том, что некоторый набор совокупностей, представленных выборками, состоит из идентичных объектов. Однако для его проведения требуется, чтобы мы могли случайно выбрать повторные пробы внутри выборок и провести их анализ в случайной последовательности. В некоторых ситуациях это может оказаться сильным ограничением и может привести к неполноценному анализу, при котором теряется слишком много информации об изменчивости. Например, предположим, что некоторая измерительная процедура приводит к завышению дисперсии. Используя однофакторную модель, мы не можем оценить величину возникающей по этой причине изменчивости, так как она входит в сумму квадратов вместе с изменчивостью, возникающей от других причин. Однако более кропотливый статистический анализ может дать возможность выделить эту изменчивость и оценить ее.

Двухфакторный дисперсионный анализ

Известен ряд более сложных критериев, подробно описанных в руководствах по дисперсионному анализу и планированию эксперимента. Прекрасные описания некоторых схем, весьма полезных при геологических исследованиях, содержатся в книгах Гриффитса [13] и Крамбейна, и Грейбилла [20]. Здесь мы ограничимся рассмотрением лишь одного дополнительного примера и соответствующей статистической процедуры. Ордовикские песчаники Сент-Питер представлены очень чистыми ортокварцитами, которые распространены в верховьях р. Миссисипи. Так как зерна этих пород хорошо окатаны и отсортированы, то они необыкновенно однородны по своему строению. В связи с этим нефтяные месторождения, приуроченные к песчаникам, при добыче нефти путем откачки ведут себя так, как можно в точности предсказать с помощью теоретических моделей их поведения, хотя последние построены на основе идеализации условий. Отклонения поведения модели от действительности могут указать на ошибочность допущений в структуре модели.

Небольшой нефтяной район в южном Иллинойсе представляется идеально приспособленным для исследования совпадения в поведении модели и реального нефтяного месторождения. Так как этот район арендовался только одной компанией, тщательно хранившей документацию, то данные о добыче нефти из этого месторождения оказались доступными для исследования. Однако, прежде чем выполнить исчерпывающий анализ поведения месторождения, целесообразно проверить на примере вышеупомянутого песчаника предположение об однородности его свойств.

Из множества скважин, пробуренных в процессе разработки, десять были выбраны случайным образом для проведения анализа. В каждой пробе наудачу был высечен 1 куб породы объемом 16 $см^3$ таким образом, чтобы вертикальная ориентация пробы сохранялась. С помощью соответствующего прибора были сделаны два измерения скорости движения флюида сквозь высеченные кубы: в вертикальном направлении по отношению к слоистости и в горизонтальном, параллельно слоистости. Используя эти измерения, вычислили проницаемость образца в квадратных микрометрах.

Двадцать вычисленных значений проницаемости приведены в табл. 2.17. По этим двадцати значениям требуется получить ответ на вопрос: имеются ли значимые различия в проницаемости, зависящие от положения образца в изучаемом районе (т. е. от расположения скважин) или от выбранных направлений измерения?

Эту задачу можно решить с помощью двухфакторного дисперсионного анализа. Рассмотрим два главных источника изменчивости: один, возникающий как следствие различий между пробами, и другой, возникающий из-за различий направлений при измерениях проницаемости. Третий источник изменчивости — остаток, или дисперсия ошибки, соответствующая дисперсии внутри повторных проб в однофакторном анализе. В этом примере проверяются две гипотезы:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_{10},$$

$$H_0 : \mu_{\text{верт}} = \mu_{\text{гориз}}.$$

Соответствующие альтернативные гипотезы заключаются в том, что по меньшей мере одна скважина имеет неравное остальным среднее и что вертикальные и горизонтальные проницаемости неодинаковы. Эта проблема очень напоминает уже рассмотренную задачу изучения содержаний карбонатов с помощью однофакторного дисперсионного анализа, но с одним исключением: вместо того чтобы произвести измерения в образцах на основании случайного выбора, мы провели измерения разного типа. Последние можно назвать эффектами: этот термин означает, что числа, порожденные одним эффектом, могут фундаментально отличаться от чисел, порожденных другим, даже в случае если используются одни и те же пробы. Так как измерения не вполне рандомизированы, а вместо этого разделены в соответствии с воздействием на них различных эффектов, то данные можно проанализировать с целью выявления различий как между эффектами, так и между образцами. Таким образом, изменчивость, возникшая из-за различий в эффектах, не должна совпадать с изменчивостью, возникшей по другим причинам, и может быть отделена на основании статистической процедуры

Для обоснованного применения двухфакторного анализа необходимо выполнение следующих четырех основных положений: а) каждая комбинация эффекта и объекта является случайной выборкой, взятой из различных совокупностей; б) каждая исходная совокупность нормальна; в) все изучаемые сово-

Таблица 2.17

Проницаемость случайно отобранных образцов песчаников Сент-Питер (штат Иллинойс), измеренная в различных направлениях, мкм²

Направления			
Вертикальное	Горизонтальное	Вертикальное	Горизонтальное
1,037	1,124	0,928	0,943
0,963	0,960	1,108	1,165
0,842	0,921	0,821	0,803
1,121	1,202	0,797	0,792
1,043	1,028	0,949	1,004

купности имеют одну и ту же дисперсию; г) нет никакой зависимости между различными эффектами и различными образцами.

Последнее допущение является утверждением того, что частная комбинация эффекта и образца не приводит к большей дисперсии, чем эффекты и образцы в других комбинациях. Если бы мы выполнили однофакторный анализ, используя повторные наблюдения (т. е. выполнили бы более одного измерения проницаемости для каждой комбинации факторов направление — образец), то могли бы обнаружить взаимодействие, однако в этой простой схеме мы предполагаем, что взаимодействие отсутствует. Если все же взаимодействия имеются, то их наличие обесценивает результаты испытания проб. Ошибочное предположение о независимости параметров между собой при наличии такой зависимости довело не одного исследователя до беды. Хорошее введение в теорию эффектов взаимодействия содержится в гл. 6 книги Хикса [16].

Ниже приводится схема двухфакторного дисперсионного анализа без повторений в ячейках. Величина SS_T вычисляется по формуле (2.37); SS_A — по формуле (2.38). Величина SS_B является суммой квадратов по эффектам:

$$SS_B = \sum_{i=1}^{n} \frac{1}{m}\left(\sum_{j=1}^{m} X_{ij}\right)^2 - \frac{1}{N}\left(\sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij}\right)^2, \qquad (2.41)$$

где m — целое положительное число.

Ошибка суммы квадратов SS_e находится по формуле

$$SS_e = SS_T - (SS_A + SS_B). \qquad (2.42)$$

Так как величина SS_B является разностью двух средних значений по образцам в пределах каждого эффекта, то из этих равенств вытекает, что SS_B является мерой изменчивости эффектов. Сумма квадратов SS_e является остатком от общего вклада при вычитании вкладов, зависящих от этих источников изменчивости. Обозначения в этом случае такие же, как и в однофакторном дисперсионном анализе, только теперь n является числом эффектов, а не числом повторных проб.

Источник изменчивости	Сумма квадратов	Число степеней свободы	Средний квадрат	Значения F-критерия
Между выборками	SS_A	$m-1$	MS_A	MS_A/MS_e[а]
Между эффектами	SS_B	$n-1$	MS_B	MS_B/MS_e[б]
Ошибка	SS_e	$(m-1)(n-1)$	MS_e	
Общая изменчивость	SS_T	$N-1$		

[а] Критерий значимости различия между выборками.

[б] Критерий значимости различия между эффектами.

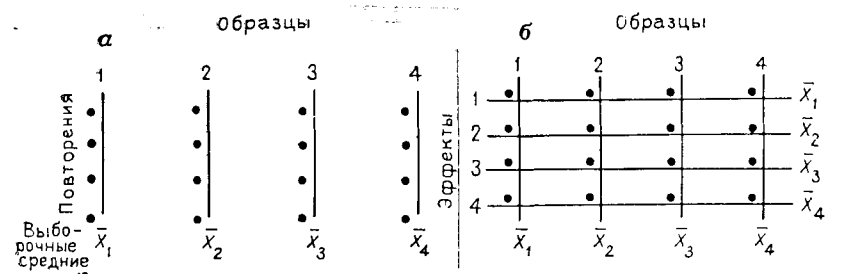


Рис. 2.32. Схемы суммирования в дисперсионном анализе:
а — однофакторный дисперсионный анализ; суммирование для нахождения выборочного среднего производится сверху вниз по столбцам; *б* — двухфакторный дисперсный анализ; суммирование для нахождения выборочных средних производится как по строкам, так и по столбцам

Выбрав некоторый уровень значимости для обеих гипотез, можно использовать эту статистическую процедуру с целью исследования данных табл. 2.17 по проницаемости. Вопросы, на которые дает ответ этот критерий, следующие: а — имеются ли значимые различия в проницаемости в пределах нефтеносного района? б — имеются ли значимые различия в проницаемости по вертикали и горизонтали?

Можно изменить однофакторную процедуру так, чтобы она позволяла находить дополнительные члены MS_B и новый поправочный член MS_e, а затем переходила бы к нахождению двух значений F. Обращение к рис. 2.32 помогает понять, какие изменения должны быть внесены. Учитывая полученные результаты, какие бы вы хотели дать рекомендации относительно целесообразности использования этих методов для проверки моделей залежи?

Дисперсионный анализ — один из наиболее распространенных статистических методов, особенно в таких областях, как контроль качества изделий в промышленности и биологические эксперименты. Следовательно, нужны программы для ЭВМ по дисперсионному анализу практически любой степени сложности.

χ^2-критерий

Рассмотрим еще одно распределение, тесно связанное с нормальным. Если выборка объема n взята из нормальной совокупности, имеющей среднее значение μ и стандартное отклонение σ, то каждое наблюдение в выборке можно преобразовать по формуле (2.16):

$$Z = (X - \mu)/\sigma. \qquad (2.43)$$

Такая величина также распределена нормально с математическим ожиданием 0 и дисперсией, равной 1. Если значения Z

возвести в квадрат и сложить, то сумма будет представлять новую статистику, которую мы обозначим ΣZ^2, т. е.

$$\Sigma Z^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2. \qquad (2.44)$$

Так как эта статистика строится по выборочным данным, то она изменяется от выборки к выборке. Если взять всевозможные выборки объема n из нормальной совокупности и нанести соответствующие им значения ΣZ^2 на график, то они будут подчинены χ^2-распределению. Распределение χ^2 зависит от числа степеней свободы, которое связано с объемом соответствующих выборок. Типичная кривая χ^2-распределения изображена на рис. 2.33, а значения χ^2-распределения для различных чисел степеней свободы и уровней значимости приведены в табл. 2.18. Отметим, что кривая χ^2-распределения, так же как и кривая F-распределения, выходит из начала координат и, проходя через положительные значения, стремится к бесконечности. Подробно χ^2-распределение рассмотрено в книге Ли [23].

Распределение χ^2 имеет большое значение в практике, так как его можно использовать для проверки гипотез, содержащих как номинальные, так и порядковые данные. До сих пор мы рассматривали только методы исследования данных, представляющих собой результаты измерения значений переменных. Теперь же мы обратимся к некоторым методам изучения данных подсчета, например таким, как число морских ежей на единицу площади морского дна, или число кристаллов плагиоклаза, расположенных на пересечении шлифа, или число зерен заданной гранулометрической фракции в раздробленном песчанике.

Общеизвестная задача статистического анализа заключается в сравнении выборочного распределения с некоторым заранее заданным стандартным распределением. Так, статистические критерии можно применить для проверки гипотезы, заключающейся в том, что имеющиеся данные извлечены из совокупности с заранее известным распределением, возможно, нормальным или логнормальным. Для того чтобы убедиться в том, что это предположение не противоречит действительности, надо сравнить выборочное и теоретическое распределения. В большинстве случаев геолог задает соответствующие классы размеров частиц и затем проверяет, согласуется ли распределение размеров частиц в естественных скоплениях гравия или на выходе из камнедробилки с некоторым заданным теоретическим распределением. В обеих рассмотренных задачах требуется установить соответствие между формой двух распределений, одно из которых получено по выборке, а другое либо заранее

Таблица 2.18

Критические значения χ^2-распределения с ν степенями свободы при заданном уровне значимости [17]

ν	Уровень значимости α, %				
	$p = 20$	$p = 10$	$p = 5$	$p = 2{,}5$	$p = 1$
1	1—64	2—71	3—84	5—02	6—63
2	3—22	4—61	5—99	7—38	9—21
3	4—64	6—25	7—81	9—35	11—34
4	5—99	7—78	9—49	11—14	13—28
5	7—29	9—24	11—07	12—83	15—09
6	8—56	10—64	12—59	14—45	16—81
7	9—80	12—02	14—07	16—01	18—48
8	11—03	13—36	15—51	17—53	20—09
9	12—24	14—68	16—92	19—02	21—67
10	13—44	15—99	18—31	20—48	23—21
11	14—63	17—28	19—68	21—92	24—72
12	15—81	18—55	21—03	23—34	26—22
13	16—98	19—81	22—36	24—74	27—69
14	18—15	21—06	23—68	26—12	29—14
15	19—31	22—31	25—00	27—49	30—58
16	20—47	23—54	26—30	28—85	32—00
17	21—61	24—77	27—59	30—19	33—41
18	22—76	25—99	28—87	31—53	34—81
19	23—90	27—20	30—14	32—85	36—19
20	25—04	28—41	31—41	34—17	37—57
21	26—17	29—62	32—67	35—48	38—93
22	27—30	30—81	33—92	36—78	40—29
23	28—43	32—01	35—17	38—08	41—64
24	29—55	33—20	36—42	39—36	42—98
25	30—68	34—38	37—65	40—65	44—31
26	31—79	35—56	38—89	41—92	45—64
27	32—91	36—74	40—11	43—19	46—96
28	34—03	37—92	41—34	44—46	48—28
29	35—14	39—09	42—56	45—72	49—59
30	36—25	40—26	43—77	46—98	50—89
40	47—27	51—81	55—76	59—34	63—69
50	58—16	63—17	67—50	71—42	76—15
60	68—97	74—40	79—08	83—30	88—38
70	79—71	85—53	90—53	95—02	100—43
80	90—41	96—58	101—88	108—63	112—33
90	101—05	107—57	113—15	118—14	124—12
100	111—67	118—50	124—34	129—56	135—81

известно, либо предполагается имеющим определенный тип. Требуется в вероятностных терминах получить ответ на вопрос: можно ли два указанных распределения отнести к одному типу?

Аналогичная задача возникла при опробовании в заливе Уайтуотер (штат Флорида), где было проведено 48 измерений солености поверхностных вод (табл. 2.19). Предварительный визуальный анализ этих данных позволяет предположить, что значения содержаний соли в выборке с некоторой площади распределены нормально. Если эта гипотеза верна, то из нее

Таблица 2.19

Измерения солености вод в заливе Уайтуотер, штат Флорида

Соленость, мкг/т									
46	53	58	60	60	49	59	48	46	78
37	58	46	46	47	48	42	50	63	48
62	49	47	36	40	39	61	43	53	42
59	60	52	34	40	36	67	44	40	
40	56	51	51	35	47	53	49	50	

следует, что происходят свободное перемешивание и обмен между открытыми морскими водами и пресной водой, втекающей в залив. С другой стороны, если бы существовал какой-либо механизм, который стремился бы разделить соленую и пресную воду в заливе, то распределение содержаний соли позволило бы его обнаружить. Это дало бы возможность получить представление о циркуляции воды и предсказать тип распределения донных осадков. С помощью соответствующего критерия согласия можно проверить, насколько хорошо выборочное распределение согласуется с нормальным.

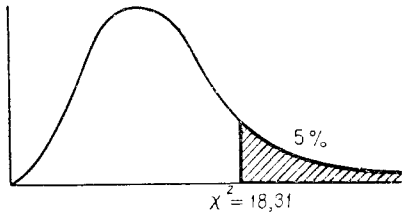
Предположим, что совокупность определений солености, из которой взята наша выборка, характеризуется нормальным распределением с неизвестным средним значением μ и дисперсией σ^2. Альтернативой этой гипотезе, конечно, является предположение, что это распределение не согласуется с нормальным законом. Значение статистического критерия можно вычислить путем подразделения области определения стандартного нормального распределения на некоторое число отрезков. Вероятность того, что одно случайное наблюдение, извлеченное из стандартного нормального распределения, попадает в один из отрезков, равна площади под кривой в пределах отрезка. Используя эти вероятности, можно вычислить ожидаемое число наблюдений в каждом отрезке. Ожидаемые частоты в каждом отрезке затем сравниваются с соответствующими выборочными частотами. Если эти числа значительно отклоняются от ожидаемых, то маловероятно, чтобы выборка была извлечена из нормальной совокупности. Используя χ^2-распределение, можно придать вероятностный смысл словам «значимый» и «маловероятный».

Указанный статистический критерий вычисляется по формуле

$$\chi^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j}, \qquad (2.45)$$

где O_j — число наблюдений в j-м классе, F_j — ожидаемое число

Рис. 2.33. χ^2-распределение при десяти степенях свободы с заштрихованной критической областью, ограничивающей 5% площади под кривой



наблюдений в этом классе. Предполагается, что имеется k различных классов (или интервалов).

В этой задаче значение статистического критерия вычисляется с помощью подразделения области определения наблюдаемых значений на некоторое число отрезков, например четыре, но так, чтобы им соответствовали равные площади под нормальной кривой, которым, следовательно, отвечают равные вероятности попадания в соответствующий интервал. Границами этих интервалов, соответствующих равным вероятностям, будут $-\infty$; $-0{,}67$; $0{,}0$; $+0{,}67$ и ∞. Если наши данные стандартизованы, то можно ожидать, что приблизительно одна четверть значений попадает в каждый из интервалов. Далее подсчитывается число проб, попадающих в каждый из этих интервалов, находится разность между ожидаемыми и реальными числами, а результат возводится в квадрат. Квадрат разности делится на ожидаемое число попаданий в данный интервал. Полученные значения суммируются по всем четырем интервалам. Если сумма превышает критическое значение, то нулевая гипотеза отклоняется и делается вывод, что распределение значений солености не согласуется с нормальным.

На рис. 2.33 представлен типичный пример χ^2-распределения. Следует отметить, что его конкретный вид зависит от числа степеней свободы. Однако число степеней свободы в нашей задаче не зависит от числа наблюдений, как это было в предыдущих критериях. В этом случае «выборки» — это четыре категории, сравниваемые с соответствующими категориями стандартного нормального распределения. В нашем примере число степеней свободы равно числу категорий без трех, или в нашем примере — единице. Мы потеряли две степени свободы потому, что для неизвестных параметров μ и σ^2 использовали их оценки $\bar{X}$ и s^2 соответственно, а еще одну степень свободы за счет того, что сумма частот по интервалам равна единице. Критическое значение χ^2-распределения, соответствующее 10%-ному уровню значимости ($\alpha = 0{,}10$) и одной степени свободы, равно 2,71 (см. табл. 2.18).

Как и в случае F-распределения, значения χ^2-распределенных случайных величин не имеют центра в нуле и всюду положительны. Так как отклонения ожидаемых частот от наблюдае-

Таблица 2.20

Стандартизованные значения солености в заливе Уайтуотер

Номер образца	Выборочные значения	Стандартизованные значения	Номер образца	Выборочные значения	Стандартизованные значения
1	46,00	—0,38	25	35,00	—1,57
2	37,00	—1,35	26	49,00	—0,06
3	62,00	1,34	27	48,00	—0,17
4	59,00	1,02	28	39,00	—1,14
5	40,00	—1,03	29	36,00	—1,46
6	53,00	0,37	30	47,00	—0,27
7	58,00	0,91	31	59,00	1,02
8	49,00	—0,06	32	42,00	—0,81
9	60,00	1,13	33	61,00	1,24
10	56,00	0,70	34	67,00	1,88
11	58,00	0,91	35	53,00	0,37
12	46,00	—0,38	36	48,00	—0,17
13	47,00	—0,27	37	50,00	0,05
14	52,00	0,27	38	43,00	—0,71
15	51,00	0,16	39	44,00	—0,60
16	60,00	1,13	40	49,00	—0,06
17	46,00	—0,38	41	46,00	—0,38
18	36,00	—1,46	42	63,00	1,45
19	34,00	—1,68	43	53,00	0,37
20	51,00	0,16	44	40,00	—1,03
21	60,00	1,13	45	50,00	0,05
22	47,00	—0,27	46	78,00	3,07
23	40,00	—1,03	47	48,00	—0,17
24	40,00	—1,03	48	42,00	—0,81

мых в каждой категории возводится в квадрат, то значения статистического критерия не могут быть отрицательными. Следовательно, χ^2-критерий всегда является односторонним, и область отклонения гипотезы расположена справа.

В нашем примере область наблюдаемых значений должна быть разбита на четыре части с равными вероятностями. Если значения солености распределены нормально, то приблизительно 12 нормализованных значений должно попасть в каждую из четырех категорий. По выборке вычисляем действительное число наблюдений (частот попадания), содержащихся в каждой из этих групп. Так как групп всего четыре, то ожидаемые значения числа наблюдений равны 12. Первый шаг — стандартизация данных по формуле (2.26), повторяемой здесь:

$$Z_i = (X_i - \overline{X})/s. \qquad (2.46)$$

Выборка данных, полученных при опробовании в заливе Уайтуотер, имеет оценку среднего $\overline{X} = 49{,}54$ и оценку стандартного отклонения $s = 9{,}27$. Поэтому нормализация наблюдений осуществляется по формуле

$$Z_i = (X_i - 49{,}54)/9{,}27.$$

Таблица 2.21

Стандартизованные значения солености, сгруппированные для проверки гипотезы о нормальном распределении таким образом, что каждой группе соответствует вероятность 0,25

Категория от $-\infty$ до $-0{,}67$		Категория от $-0{,}67$ до $0{,}0$	
—1,35	—1,14	—0,38	—0,17
—1,03	—1,46	—0,06	—0,27
—1,46	—0,81	—0,38	—0,17
—1,68	—0,71	—0,27	—0,60
—1,03	—1,03	—0,38	—0,06
—1,03	—0,81	—0,27	—0,38
—1,47		—0,06	—0,17
Общее число наблюдений 13		Общее число наблюдений 14	
Категория от 0,0 до $+0{,}67$		Категория от $+0{,}67$ до ∞	
0,37	0,37	1,34	1,13
0,27	0,05	1,02	1,02
0,16	0,37	0,91	1,24
0,16	0,05	1,13	1,88
		0,70	1,45
		0,91	3,07
		1,13	
Общее число наблюдений 8		Общее число наблюдений 13	

Стандартизованные значения приведены в табл. 2.20. В табл. 2.21 приведены результаты разбиения всей выборки на четыре категории. Если выборку можно считать извлеченной из нормальной совокупности, то следует ожидать приблизительно 12 наблюдений на категорию. Вычисляя значения критерия χ^2, получим следующие промежуточные результаты:

$$\chi^2 = \frac{(13-12)^2}{12} + \frac{(14-12)^2}{12} + \frac{(8-12)^2}{12} + \frac{(13-12)^2}{12} = \frac{22}{12} = 1{,}83.$$

Вычисленное значение χ^2 меньше критического 2,71 10%-ного уровня значимости и одной степени свободы. Поэтому нет оснований считать, что распределение значений солености в поверхностных водах существенно отклоняется от нормального закона.

Конечно, статистика χ^2 позволяет проверить гипотезу не только о нормальном распределении. Мы можем применить этот критерий для проверки гипотезы о любом другом законе рас-

пределения, например, таком, как логнормальный, экспоненциальный и т. д. При этом процедура проверки не изменяется, хотя число степеней свободы в каждом случае зависит от числа оцениваемых параметров. Кохран [7] подробно рассматривает эти вопросы.

ЛОГНОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ И ДРУГИЕ ПРЕОБРАЗОВАНИЯ

Многие геологические переменные очевидно не подчиняются нормальному распределению. Если, например, нанести на график объемы нефтяных полей (рис. 2.34), то полученное распределение будет в высшей степени асимметричным. Большинство полей малы по размеру, но имеется убывающая последовательность более крупных полей, немного редких гигантов, которые значительно превосходят другие по объему. Геохимические переменные, подвергнутые опробованию в процессе геохимического исследования, например, такие, как концентрация селена в растительном материале или концентрация иода, обнаруженная в пробах грунтовых вод, также подчиняются асимметричным распределениям. В распределениях размеров зерен в осадках ярко выражена асимметрия, и целая система классификации основывается на этом факте [21]. На рис. 2.35 представлена гистограмма, показывающая концентрацию меди в осадочном русле на Юконе. На ней представлено асимметричное распределение, типичное для многих других геологических переменных.

Если наблюдения, представленные на рис. 2.34 и 2.35, преобразовать в логарифмическую форму (т. е. вместо переменных X используются переменные $Y_i = \log X_i$, то мы убедимся, что их распределения станут приблизительно нормальными (рис. 2.36 и 2.37). Такие переменные называются логнормальными. Так как они часто встречаются в геологии, то логнормальное распределение в высшей степени важно. Однако если ограничиться рассмотрением преобразованных переменных Y_i, а не самих X_i, то свойства логнормального распределения можно просто охарактеризовать ссылкой на нормальное распределение.

Среднее и дисперсия логарифмически преобразованной переменной Y_i находятся обычным способом:

$$\overline{Y} = \Sigma Y_i/n \quad \text{и}$$

$$s_Y^2 = \frac{\Sigma (Y_i - \overline{Y})^2}{n-1}. \qquad (2.47)$$

Однако в терминах исходной непреобразованной переменной X_i

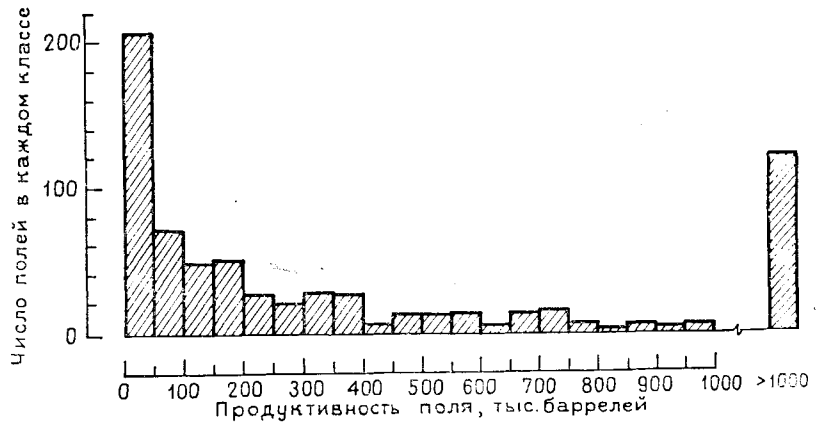


Рис. 2.34. Гистограмма распределения продуктивности нефтеносных полей, открытых в Денверском бассейне в 1969 г.

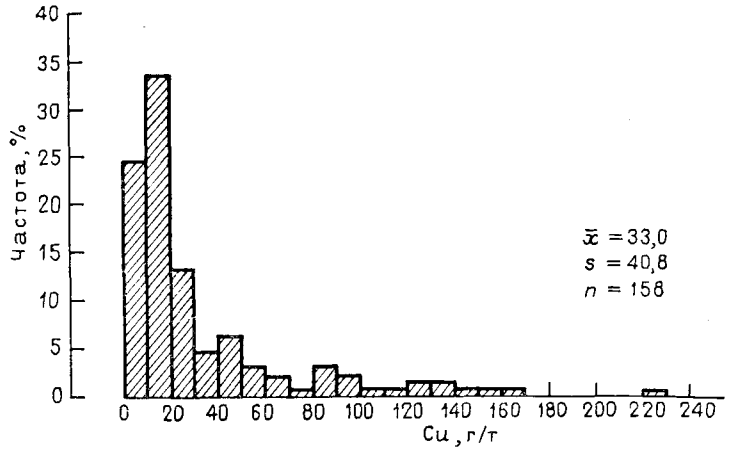


Рис. 2.35. Гистограмма содержания меди в осадочных породах на площади Нансен в Юконе [29]

среднее $\overline{Y}$ соответствует корню n-й степени из произведений X_i

$$\overline{Y} = GM = \sqrt[n]{\Pi X_i}, \qquad (2.48)$$

который называется геометрическим средним, GM. Символ Π аналогичен Σ, только он означает, что элементы указанного вслед за ним ряда перемножаются, а не складываются. Π также имеет пределы, аналогичные тем, которые используются при знаке суммирования Σ. Иногда эти пределы, если они очевид-

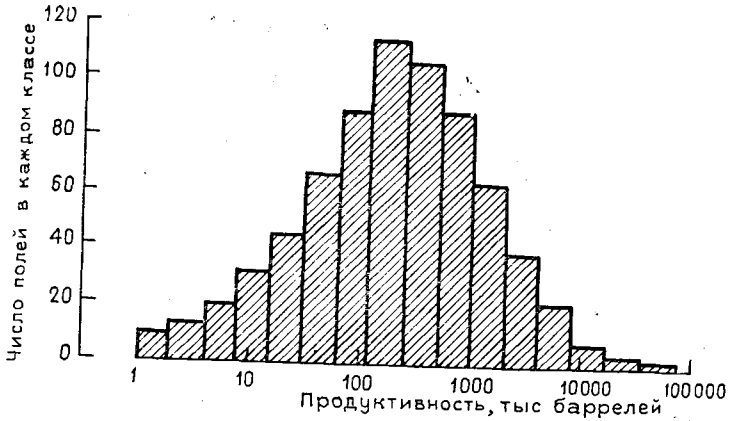


Рис. 2.36. Гистограмма распределения продуктивности нефтеносных полей, открытых в Денверском бассейне в 1969 г.
Масштаб логарифмический. Взято из Харбуха, Давтока, Девиса

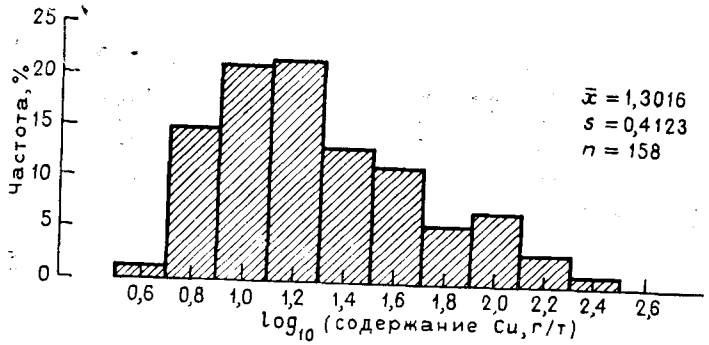


Рис. 2.37. Гистограмма содержаний меди в осадочных породах на площади Нанси в Юконе.
Представлена в логарифмическом масштабе [29]

ны, опускаются. Так, $\prod_{i=1}^{3} X_i$, где $X_1=2$, $X_2=3$, $X_3=4$, равно $\Pi X_i=2\times3\times4=24$.

Дисперсия логарифмически преобразованной переменной называется геометрической дисперсией и эквивалентна

$$s_Y^2 = s_g^2 = \sqrt[n-1]{\Pi 2\left(\frac{X_i}{GM}\right)}. \qquad (2.49)$$

На практике, конечно, проще преобразовать наши наблюдения, взяв их логарифмы, и затем вычислить их среднее и дисперсию. Если требуется найти геометрическое среднее и дисперсию, то надо взять антилогарифмы от $\bar{Y}$ и s^2_Y. До тех пор пока

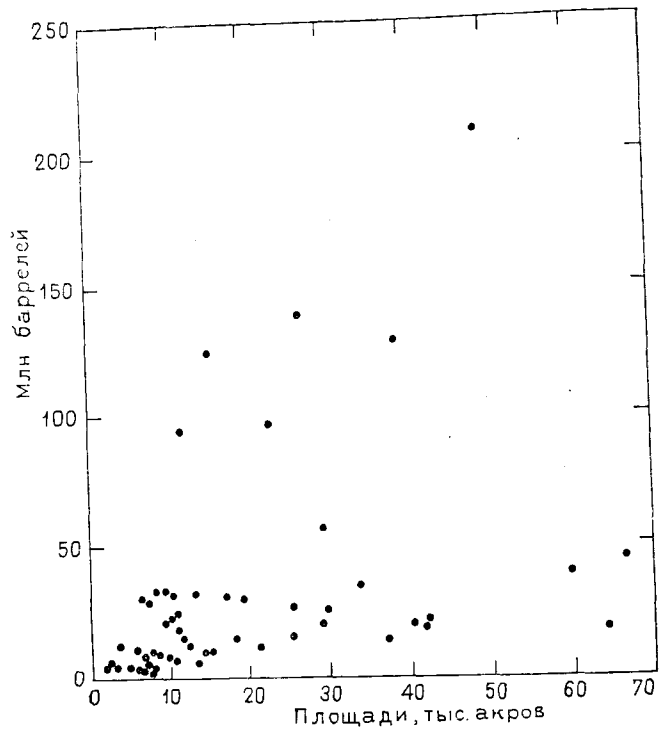


Рис. 2.38. Объемы нефтяных залежей, связанных с соляными куполами в Луизиане на внешнем континентальном шельфе (Мексиканский залив). По сейсмическим данным

мы работаем с данными в преобразованном виде, все статистические процедуры, применяемые к обыкновенным переменным, пригодны и для логарифмически преобразованных переменных. Добавим еще, что логарифмическое преобразование переменных полезно, если требуется стабилизировать дисперсию и тем самым привести переменную с асимметричным распределением к более симметрическому виду. На рис. 2.38 представлены объемы залежей нефти, ассоциированных с соляным куполом на площади внешнего континентального шельфа Луизианы. Эти объемы залежей нанесены на оси ординат графика; на оси абсцисс отложены площади структур, в которых расположены залежи. В общем случае имеется положительная связь между двумя переменными; т. е. более крупные структуры обычно содержат более крупные залежи. Однако размер залежи также увеличивается по мере увеличения размеров структур, или, другими словами, наблюдаемая дисперсия пропорциональна среднему.

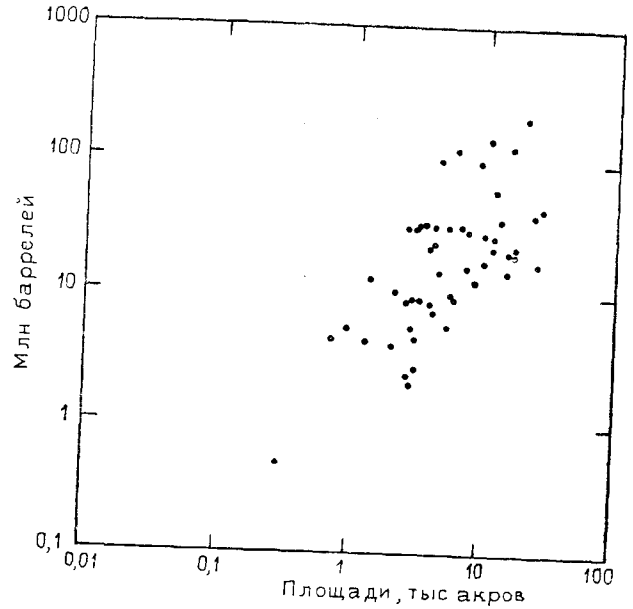


Рис. 2.39. Данные рис. 2.38, представленные в логарифмическом масштабе

Логарифмическое преобразование помогает скорректировать это условие, как это легко увидеть на рис. 2.39. Здесь как объем залежи, так и размер структуры преобразуется взятием их логарифмов, в результате получается график двойного логарифма. Дисперсия логарифмов объемов залежей остается почти постоянной для всех значений логарифмов размеров структур.

Характеристики логнормального распределения рассмотрены в монографии Ачисона и Брауна [2] и в геологическом контексте — Кохом и Линком [18]. Нормальное распределение обычно возникает, когда производятся повторные замеры некоторой фиксированной величины μ. Каждое индивидуальное измерение претерпевает флуктуацию в силу многих случайных воздействий, которые складываются с измерением, действуя иногда в одном, а иногда в противоположных направлениях. Обычно эти случайные воздействия взаимно уничтожаются, и окончательное измерение близко к истинному значению. Однако в редких случаях большинство случайных отклонений может иметь один и тот же знак, и тогда возникают экстремальные значения. Это явление отражено в колоколообразном виде нормального распределения.

Логнормальное распределение может возникнуть при тех же обстоятельствах, если случайные воздействия не аддитивны,

а мультипликативны. Большинство случайных возмущений, перемноженных вместе, дают промежуточное значение произведения, близкое к геометрическому среднему. В редких случаях при случайном выборе все возмущения могут оказаться очень малыми, и их произведение будет близко нулю. Так же редко все возмущения могут оказаться большими, и их произведение будет экстремально большим значением. Результатом многих случайных реализаций будет распределение, которое начинается в нуле и возрастает до своего максимума, и затем спускается вниз, достигая экстремально больших значений.

Биологи часто ссылаются на «закон пропорционального эффекта», состоящий в том, что изменение переменной в течение процесса есть случайная величина, пропорциональная исходному значению этой переменной. Например, вероятность изменения размера колоний микробов во временной промежуток пропорциональна размеру колоний в предшествующий отрезок времени. Большие колонии стремятся расшириться (или уменьшиться) в большей степени, чем малые колонии. Возможно, нефтяные месторождения формировались таким же образом, так что в течение миграции углеводородов большие скопления стремились увеличиться с пропорционально большей скоростью, чем это происходило с малыми скоплениями. Такие процессы подчиняются логнормальному распределению.

Геологи, возможно, менее знакомы с «теорией дробления», которая предсказала заранее и объяснила логнормальное распределение для размеров частиц, которые наблюдаются в естественных осадках и в измельченном материале, производимом мельницами и дробилками. Предположим, что взят набор частиц одинакового размера и затем каждая из них разделена случайным образом. В общем случае в результате один из обломков каждой исходной частицы будет больше, другой — меньше. Если затем каждый из этих обломков снова раздробить случайным образом, то из малых кусков получатся еще меньшие, в то время как каждый большой обломок даст снова больший и меньший куски. Если этот процесс повторять снова и снова, то в результате получим очень большое число очень малых частиц и немного «избранных» зерен, размеры которых близки к исходным размерам частиц. Другими словами логнормальное распределение часто наблюдается при изучении осадков.

Другие преобразования

Для получения приблизительно нормального распределения можно над переменной X произвести также некоторые другие преобразования, которые преобразуют дисперсию к более при-

емлемому виду или дадут некоторые другие статистические полезные результаты. Хотя ничего не было сказано об исходной шкале измерения, но мы должны чувствовать себя свободными в случае, если окажется полезным изменить ее. Однако следует постоянно помнить, что наш статистический анализ имеет целью проверку различных статистических гипотез, которые могут иметь место для характеристик преобразованных переменных и не обязательно справедливых для исходных переменных. Конечно, следует позаботиться о том, чтобы используемые преобразования не были настолько экзотическими, чтобы за ними терялась природа исходных переменных, свойства которых исследуются.

Если наши данные представляют собой, например, число продуктивных скважин в регионе или число зерен циркона в шлифе, эти числа могут подчиняться распределению Пуассона. Вместо того чтобы считать эти данные дискретными, удобнее привести их к приблизительно нормальному виду, извлекая из них квадратный корень, т. е. каждое значение X_i заменяется на $Y_i=\sqrt{X_i}$. Это преобразование сделает дисперсии более однородными и приведет к сокращению длинного хвоста пуассоновского распределения. Если наблюдаемые значения X_i меньше, чем примерно 10, то удобнее использовать преобразование $Y_i=\sqrt{X_i+1/2}$, особенно в тех случаях, когда некоторые наблюдения равны нулю.

Робинсон [28] рекомендует использовать степенное преобразование, например $Y_i=X_i^2$, $Y_i=X_i^3$ и так далее, при выявлении петрофизических свойств, по данным каротажа в скважинах. Возведение в степень приводит к большему увеличению больших значений, чем малых. Если возведение в степень применять после изменения масштаба так, что множество значений $(Y_{max}-Y_{min})$ будет таким же, как и множество значений исходных переменных $(X_{max}-X_{min})$, то эффект будет заключаться в том, что на каротажных диаграммах будут подчеркиваться области высоких значений и подавляться участки, где значения низки. Степенное преобразование имеет такое же влияние на распределение данных и может быть использовано для исправления отрицательной асимметрии. Однако оно может также привести к увеличению дисперсии и сделать ее неоднородной.

Отрицательно асимметричное распределение иногда может быть приближенно преобразовано в нормальное, если применить преобразование $Y_i=\arcsin X_i$. При этом исходные переменные должны быть предварительно преобразованы в числа в пределах интервала (0,00—1,00). Другое преобразование арксинуса, описанное в гл. 4, можно использовать для преобразования биномиального распределения в нормальное.

НЕПАРАМЕТРИЧЕСКИЕ МЕТОДЫ

Все предыдущие статистические методы являются параметрическими, т. е. они основаны на характеристиках распределений, параметры которых известны. Все используемые критерии (t, F и χ^2) строятся для выборок из нормальных совокупностей. Для обоснования возможности использования этих критериев в тех случаях, когда исследуемая совокупность не является нормальной, при условии, что объем выборки велик и совокупность не очень сильно отличается от нормальной, следует обратиться к центральной предельной теореме. Иногда, однако, исследуемая совокупность может сильно отличаться от нормальной, или же объем выборки нельзя увеличить. В таких случаях следует обратиться к категории критериев, называемых непараметрическими статистическими критериями. Их можно применять для обработки информации более низких шкал, таких, как номинальные и порядковые данные, в отличие от метрических данных, используемых в параметрической статистике. Не требуется никаких допущений о виде исходного распределения, отсюда и название — непараметрические критерии. Вообще, в тех случаях, когда выборочная совокупность имеет характеристики, необходимые в параметрическом анализе, непараметрические критерии оказываются менее мощными, чем эквивалентные параметрические. Однако если выборочная совокупность не имеет специфических характеристик, непараметрические методы оказываются более мощными.

Непараметрические критерии в геологии широко не использовались и обычно не приводятся в элементарных учебниках статистики. Однако есть много прекрасных книг, в которых описаны непараметрические эквиваленты параметрических процедур, уже рассмотренных нами. Среди них можно назвать книги Зигеля [30], Бредли [4] и Коновера [8] и др.

Критерий Манна — Уитни

Критерий Манна — Уитни можно использовать как непараметрический эквивалент t-критерия для проверки гипотезы о равенстве средних двух выборок. Предположим, что мы имеем две выборки объема m и n и хотим проверить гипотезу о том, что они являются выборками из одной и той же совокупности. Объединим обе выборки и расположим значения наблюдений в порядке возрастания от меньшего к большему. Каждому наблюдению припишем его ранг, т. е. наименьшему значению припишем ранг 1, следующему по величине — ранг 2 и так далее, до наибольшего наблюдения, которое будет иметь ранг $(m+n)$. Если обе выборки были взяты из одной и той же совокупности наудачу, то можно ожидать, что наблюдения одной

Таблица 2.22

Критические значения для критерия Манна — Уитни [8]; значения приведены для нижнего критического предела; соответствующий предел для верхней критической площади дается значением $T_{1-\alpha}=nm-T_\alpha$ [8]

n	α	$m=2$	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	.01	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	2	2
	.05	0	0	0	1	1	1	2	2	2	2	3	3	4	4	4	4	5	5	5
	.10	0	1	1	2	2	3	3	4	4	5	5	5	6	6	7	7	8	8	8
3	.01	0	0	0	0	0	1	1	2	2	2	3	3	3	4	4	5	5	5	6
	.05	0	1	1	2	3	3	4	5	5	6	6	7	8	8	9	10	10	11	12
	.10	1	2	2	3	4	5	6	6	7	8	9	10	11	11	12	13	13	15	16
4	.01	0	0	0	1	2	2	3	4	4	5	6	6	7	9	8	9	10	10	11
	.05	0	1	2	3	4	5	6	7	8	9	10	11	12	13	15	16	17	18	19
	.10	1	2	4	5	6	7	8	10	11	12	13	14	16	17	18	19	21	22	23
5	.01	0	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
	.05	1	2	3	5	6	7	9	10	12	13	14	16	17	19	20	21	23	24	26
	.10	2	3	5	6	8	9	11	13	14	16	18	19	21	23	24	26	28	29	31
6	.01	0	0	2	3	4	5	7	8	9	10	12	13	14	16	17	19	20	21	23
	.05	1	3	4	6	8	9	11	13	15	17	18	20	22	24	26	27	29	31	33
	.10	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	32	35	37	39
7	.01	0	1	2	4	5	7	8	10	12	13	15	17	18	20	22	24	25	27	29
	.05	1	3	5	7	9	12	14	16	18	20	22	25	27	29	31	34	36	38	40
	.10	2	5	7	9	12	14	17	19	22	24	27	29	32	34	37	39	42	44	47
8	.01	0	1	3	5	7	8	10	12	14	16	18	21	23	25	27	29	31	33	35
	.05	2	4	6	9	11	14	16	19	21	24	27	29	32	34	37	40	42	45	48
	.10	3	6	8	11	14	17	20	23	25	28	31	34	37	40	43	46	49	52	55
9	.01	0	2	4	6	8	10	12	15	17	19	22	24	27	29	32	34	37	39	41
	.05	2	5	7	10	13	16	19	22	25	28	31	34	37	40	43	46	49	52	55
	.10	3	6	10	13	16	19	23	26	29	32	36	39	42	46	49	53	56	59	63
10	.01	0	2	4	7	9	12	14	17	20	23	25	28	31	34	37	39	42	45	48
	.05	2	5	8	12	15	18	21	25	28	32	35	38	42	45	49	52	56	59	63
	.10	4	7	11	14	18	22	25	29	33	37	40	44	48	52	55	59	63	67	71
11	.01	0	2	5	8	10	13	16	19	23	26	29	32	35	38	42	45	48	51	54
	.05	2	6	9	13	17	20	24	28	32	35	39	43	47	51	55	58	62	66	70
	.10	4	8	12	16	20	24	28	32	37	41	45	49	53	58	62	66	70	74	79
12	.01	0	3	6	9	12	15	18	22	25	29	32	36	39	43	47	50	54	57	61
	.05	3	6	10	14	18	22	27	31	35	39	43	48	52	56	61	65	69	73	78
	.10	5	9	13	18	22	27	31	36	40	45	50	54	59	64	68	73	78	82	87
13	.01	1	3	6	10	13	17	21	24	28	32	36	40	44	48	52	56	60	64	68
	.05	3	7	11	16	20	25	29	34	38	43	48	52	57	62	66	71	76	81	85
	.10	5	10	14	19	24	29	34	39	44	49	54	59	64	69	75	80	85	90	95
14	.01	1	3	7	11	14	18	23	27	31	35	39	44	48	52	57	61	66	70	74
	.05	4	8	12	17	22	27	32	37	42	47	52	57	62	67	72	78	83	88	93
	.10	5	11	16	21	26	32	37	42	48	53	59	64	70	75	81	86	92	98	103
15	.01	1	4	8	12	16	20	25	29	34	38	43	48	52	57	62	67	71	76	81
	.05	4	8	13	19	24	29	34	40	45	51	56	62	67	73	78	84	89	95	101
	.10	6	11	17	23	28	34	40	46	52	58	64	69	75	81	87	93	99	105	111
16	.01	1	4	8	13	17	22	27	32	37	42	47	52	57	62	67	72	77	83	88
	.05	4	9	15	20	26	31	37	43	49	55	61	66	72	78	84	90	96	102	108
	.10	6	12	18	24	30	37	43	49	55	62	68	75	81	87	94	100	107	113	120
17	.01	1	5	9	14	19	24	29	34	39	45	50	56	61	67	72	78	83	89	94
	.05	4	10	16	21	27	34	40	46	52	58	65	71	78	84	90	97	103	110	116
	.10	7	13	19	26	32	39	46	53	59	66	73	80	86	93	100	107	114	121	128

Продолжение табл. 2.22

n	α	$m=2$	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
18	.01	1	5	10	15	20	25	31	37	42	48	54	60	66	71	77	83	89	95	101
	.05	5	10	17	23	29	36	42	49	56	62	69	76	83	89	96	103	110	117	124
	.10	7	14	21	28	35	42	49	56	63	70	78	85	92	99	107	114	121	129	136
19	.01	2	5	10	16	21	27	33	39	45	51	57	64	70	76	83	89	95	102	108
	.05	5	11	18	24	31	38	45	52	59	66	73	81	88	95	102	110	117	124	131
	.10	8	15	22	29	37	44	52	59	67	74	82	90	98	105	113	121	129	136	144
20	.01	2	6	11	17	23	29	35	41	48	54	61	68	74	81	88	94	101	108	115
	.05	5	12	19	26	33	40	48	55	63	70	78	85	93	101	108	116	124	131	139
	.10	8	16	23	31	39	47	55	63	71	79	87	95	103	111	120	128	136	144	152

из выборок будут более или менее равномерно рассеяны в последовательности рангов.

Пусть X_i — i-е наблюдение первой выборки, а Y_i — i-е наблюдение второй выборки. Ранг наблюдения X_i будет обозначаться через $R(X_i)$, а ранг Y_i — через $R(Y_i)$. Критерий Манна — Уитни имеет вид

$$T=\sum_{i=1}^{n}R(X_i)-\frac{n(n+1)}{2}. \qquad (2.50)$$

Первый член — просто сумма рангов наблюдений из первой выборки. Критические значения T, взятые из Коновера [8], приведены в табл. 2.22.

В качестве примера использования критерия Манна — Уитни рассмотрим данные табл. 2.23. В районе Галф Коаст (США) нефть и газ добываются из структурных ловушек, связанных с соляными куполами. Прогнозы делаются на основе картирования подстилающих горизонтов, которые регистрируются сейсмическими методами. Потенциальные углеводородные ловушки включают наклонные блоки на флангах соляных куполов и замкнутые антиклинали на их осях. В табл. 2.23 перечислены прогнозные площади, полученные на основе сейсмических карт двух регионов побережья Луизианы. Нужно узнать, есть ли различия в распространении перспективных участков между двумя регионами.

Табл. 2.23 также содержит ранги объединенной выборки наблюдений перспективных площадей в двух регионах. Сумма рангов первой группы — $\sum_{i=1}^{n}R(X_i)=67$. Поэтому статистика Манна — Уитни равна

$$T=67-\frac{8(8+1)}{2}=31.$$

Таблица 2.23

Перспективные площади на нефть и газ солянокупольных структур двух регионов побережья Луизианы (по картам Морской сейсмической службы)

Прогноз	Площадь, усл. ед.	Ранг
Восточный регион		
X_1	802	16
X_2	174	8
X_3	158	6
X_4	140	4
X_5	166	7
X_6	328	13
X_7	239	10
X_8	99	3
Западный регион		
Y_1	312	12
Y_2	55	2
Y_3	220	9
Y_4	276	11
Y_5	154	5
Y_6	37	1
Y_7	478	14
Y_8	666	15

Критические значения T_α, приведенные в табл. 2.22, принимаются за нижние пределы, т. е. они пригодны для проверки нулевой гипотезы: $H_0: E(X) \geqslant E(Y)$ при альтернативе $H_1: E(X) < E(Y)$. (Символ E обозначает «математическое ожидание» и характеризует центр распределения, за который может быть принято как среднее значение, так и медиана.) Соответствующие пределы для верхнего критического значения можно найти по формуле

$$T_{1-\alpha} = nm - T_\alpha. \qquad (2.51)$$

Последние пригодны для проверки нулевой гипотезы $H_0: E(X) \leqslant E(Y)$ при альтернативе $H_1: E(X) > E(Y)$.

В нашем примере, однако, мы хотим проверить двустороннюю альтернативу и отклонить гипотезу о равенстве прогнозных площадей в двух регионах, если они для первого региона либо значительно больше, либо значительно меньше, чем те же величины для второго региона. На формальном языке наши гипотезы и альтернатива имеют вид

$$H_0: E(X) = E(Y), \quad H_1: E(X) \neq E(Y).$$

Значения границы двух критических областей находятся по

табл. 2.22, как и $T_{\alpha/2}$ и $nm - T_{\alpha/2}$. Если задать уровень значимости $\alpha = 10\%$, эти пределы будут 16 и 56, так как n и m оба равны 8. Проверяемое нами значение критерия не попадает ни в одну из критических областей, так что этот критерий не позволяет утверждать, что объемы прогнозов в двух регионах различаются.

Критерий Манна — Уитни появляется в немного различающихся формах в статистической литературе. Среди них можно назвать критерий Вилкоксона, критерий Зигеля — Тьюки, критерий Фестинджера. Вариант Зигеля — Тьюки наиболее интересен, так как он может быть использован для проверки равенства дисперсий в двух выборках и является, таким образом, непараметрическим аналогом простого F-критерия.

Иногда случается, что два или более наблюдений при ранжировании получают одинаковые ранги. Это явление носит название связанных рангов. Тогда связанным наблюдениям приписываются одинаковые ранги, равные среднему значению рангов, которые были бы приписаны, если бы наблюдения не были в точности совпадающими. Например, в табл. 2.23 мы можем положить наблюдение X_2 равным 220, а не 174 усл. ед. Тогда наблюдения X_2 и Y_3 будут связанными. Каждому будет приписан ранг $(8+9)/2 = 8{,}5$. Критерий, использующий процедуры со связанными рангами, имеет тот же вид, что и ранее.

Критерий Краскла — Уэллиса

Непараметрический критерий эквивалентности нескольких выборок был рекомендован Красклом и Уэллисом. Действительно, он является непараметрическим аналогом однофакторного дисперсионного анализа. Процедура его вычисления очень напоминает соответствующую процедуру для критерия Манна — Уитни; наблюдения из выборок комбинируются или объединяются и затем ранжируются от наименьшего к наибольшему. Для каждой выборки находим сумму рангов:

$$R_k = \sum_{i=1}^{n_k} R(X_{ik}), \qquad (2.52)$$

где $R(X_{ik})$ — ранг i-го наблюдения в k-й выборке. Общее число наблюдений есть $N = \Sigma n_k$, где n_k — число наблюдений в k-й выборке.

Нулевая гипотеза, которую требуется проверить, такова: все совокупности, из которых взяты выборки, имеют одинаковые распределения. Альтернатива состоит в том, что по меньшей мере одна из совокупностей имеет центральное значение, отличное от других. Предполагается, что все наблюдения были

ссбраны случайным образом и что выборки независимы друг от друга.

Для суммы рангов можно вычислить статистику Краскла — Уэллиса

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{[R_k - n_k(N+1)/2]^2}{n_k}. \qquad (2.53)$$

Более легкая для вычислений алгебраически эквивалентная форма имеет вид

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_k^2}{n_k} - 3(N+1). \qquad (2.54)$$

Критические значения H табулированы только для трех выборок, каждая из которых состоит более чем из пяти наблюдений (см., например, [30, табл. 0]). К счастью, статистика H распределена приблизительно как χ^2 с $k-1$ степенями свободы, так что этот критерий можно с успехом применять для решения более сложных задач.

Непараметрическая корреляция

В предыдущих параграфах подробно приведено вычисление коэффициента корреляции, а параметрический критерий значимости выборочного коэффициента корреляции описан в параграфе, посвященном t-критерию. Однако во многих случаях, когда обычный коэффициент корреляции (называемый иногда пирсоновским коэффициентом корреляции) непригоден, желательно иметь и другие меры связей между переменными.

Считается, что структурные свойства песчаника отражают условия осадконакопления. Например, высоко энергетическая фациальная обстановка в окрестностях пляжа приводит к выветриванию и разрушению, в результате которых образуются хорошо рассортированные и хорошо окатанные грубозернистые отложения. При низко энергетических условиях отложения будут представлены менее тщательно отсортированным материалом с более мелкими угловатыми зернами. Фолк [11] определяет структурную зрелость как степень отсортированности и окатанности песчаных зерен. Предполагается, что эти два свойства тесно связаны; песчаник, показывающий аномалии этой связи (например, хорошо отсортированный песчаник с угловатыми зернами), называют структурно обращенным.

Казалось бы, определить понятие структурной зрелости, собрав образцы песчаника и изучив их структурные свойства, а также вычислив затем меры связи между этими свойствами, — простое дело. К сожалению, округлость и степень отсортированности измеряются обычно не в интервальной шкале или шкале отношений, а в порядковой. Степень отсортированности обычно выражается терминами — плохая, умеренная, хорошая, а степень окатанности классифицируется терминами — угловатая, полуугловатая, полуокатанная, окатанная. Если использовать эти термины, то обычный коэффициент корреляции для характеристики силы связи между степенью окатанности и степенью отсортированности непригоден.

В таких случаях используется ранговый коэффициент корреляции Спирмена, который, как это видно из самого названия, выражает степень сходства между двумя наборами рангов. Пусть мы имеем два набора порядковых измерений на некотором множестве объектов, которые обозначим через X_i (для первого набора) и через Y_i (для второго). Затем ранжируем каждый набор измерений, обозначив их ранги соответственно через $R(X_i)$ и $R(Y_i)$. Коэффициент корреляции Спирмена измеряет сходство между двумя наборами рангов:

$$r' = 1 - \frac{6\sum_{i=1}^{n} [R(X_i) - R(Y_i)]^2}{n(n^2-1)}. \qquad (2.55)$$

В скобках под знаком суммы содержатся разности рангов объектов двух наборов.

В табл. 2.24 приведены структурные характеристики двенадцати песчаников, являющихся нефтяными коллекторами. Каждый из них сравнивался с другим для того, чтобы ранжировать их от наиболее плохо до наиболее хорошо отсортированного и от наиболее угловатого до наиболее окатанного. Результаты ранжирования приведены в таблице. Как и в критерии Манна — Уитни, наблюдениям со связанными рангами приписываются усредненные значения тех рангов, которые были бы приписаны этим наблюдениям в случаях, если бы они не были связанными.

Коэффициент корреляции Спирмена между двумя переменными равен

$$r' = 1 - \frac{6(162)}{12(12^2-1)} = 1 - \frac{972}{1716} = 0{,}43.$$

Область изменения коэффициента ранговой корреляции r' такая же, как и у обычного коэффициента корреляции — от $+1{,}0$ (полное соответствие между рангами) до $-1{,}0$ (полная обратная связь между рангами). Равенство коэффициента ранговой корреляции нулю означает, что два множества рангов независимы. Равенство коэффициента ранговой корреляции, например, 0,43, означает, что имеется слабая положительная связь

Таблица 2.24

Окатанность и сортированность зерен песчаника в нефтяной залежи

Формация	Возраст	Сортированность	Ранг	Окатанность	Ранг	$[R(X_i)-R(Y_i)]$	$[R(X_i)-R(Y_i)]^2$
Lakota	Меловой	Б	4	ПК	11	—7	49
Brea	Ранний карбон	Х	10	ПУ	9	1	1
Boise	Плиоцен	Б	2	У	1	1	1
Big Clifty	Ранний карбон	У	8	ПУ	4	4	16
Clear Creek	Средний и поздний карбон	У	6	ПУ	6	0	0
Bromide	Ордовик	Х	9	ПК	12	—3	9
Noxie	Средний и поздний карбон	Б	3	ПУ	8	—5	25
Green River	Эоцен	У	7	ПУ	3	4	16
Reagan	Кембрий	Х	11	ПУ	7	4	16
Peru	Девон	Х	12	ПК	10	2	4
Bartlesville	Средний и поздний карбон	У	5	У	2	3	9
Mt. Simon	Кембрий	Б	1	ПУ	5	24	16
							Сумма 162

Категории сортированности: Б — бедный; У — умеренный; Х — хорошо рассортированный.
Категории окатанности: У — угловатый; ПУ — полууглова тый; ПК — полуокатанный.

Таблица 2.25

Критические значения рангового коэффициента корреляции Спирмена для проверки гипотезы $\rho=0$

n	α=.10	.05	.025	.01	.005	.001
4	.8000	.8000	.9000	.9000		
5	.7000	.8000	.8286	.8857	.9429	
6	.6000	.7714	.8571	.8929	.9643	
7	.5357	.6786	.7450	.8095	.8571	.9286
8	.5000	.6190	.7143	.7667	.8167	.9000
9	.4667	.5833	.6833	.7333	.7818	.8667
10	.4424	.5515	.6364	.7000	.7455	.8364
11	.4182	.5273	.6091	.6713	.7273	.8182
12	.3986	.4965	.5804	.6429	.6978	.7912
13	.3791	.4780	.5549	.6220	.6747	.7670
14	.3626	.4593	.5341	.6000	.6536	.7464
15	.3500	.4429	.5179	.5824	.6324	.7265
16	.3382	.4265	.5000	.5637	.6152	.7083
17	.3260	.4118	.4853	.5480	.5975	.6904
18	.3148	.3994	.4716	.5333	.5825	.6737
19	.3070	.3895	.4579	.5203	.5684	.6586
20	.2977	.3789	.4451	.5078	.5545	.6455
21	.2909	.3688	.4351	.4963	.5426	.6318
22	.2829	.3597	.4241	.4852	.5306	.6186
23	.2767	.3518	.4150	.4748	.5200	.6070
24	.2704	.3435	.4061	.4654	.5100	.5962
25	.2646	.3362	.3977	.4564	.5002	.5856
26	.2588	.3299	.3894	.4481	.4915	.5757
27	.2540	.3236	.3822	.4401	.4828	.5660
28	.2490	.3175	.3749	.4320	.4744	.5567
29	.2443	.3113	.3685	.4251	.4665	.5479
30	.2400	.3059	.3620			

между степенью окатанности зерен и степенью их отсортированности.

Привычный t-критерий значимости коэффициента корреляции не может быть применен к r', так как t-критерий применяется в случаях, когда выборка сделана из совокупности, имеющей двумерное нормальное распределение. К счастью, существуют таблицы критических значений, позволяющих прямо осуществить проверку рангового коэффициента корреляции Спирмена; частично эти таблицы воспроизведены в табл. 2.25. Как и в случае t-критерия, нулевая гипотеза состоит в том, что две переменные независимы или что $\rho'=0$. Наиболее общая альтернатива — $\rho'\neq 0$, так что критерий двусторонний, и как очень большие, так и очень малые (отрицательные) значения приводят к отклонению проверяемой гипотезы. Предположим, что принят уровень значимости $\alpha=0{,}05$, достаточный, по нашему мнению, для проверки значимости корреляции между окатанностью и отсортированностью. Тогда верхнее критическое зна-

чение будет соответствовать $1-\alpha/2=0{,}975$ и будет равно 0,5804 для $n=12$. Нижнее критическое значение, соответствующее $\alpha/2=0{,}025$, равно —0,5804. Вычисленный коэффициент корреляции $r'=0{,}43$ не выходит за эти границы, и гипотезу о том, что степень округлости и степень сортировки не зависят друг от друга, отклонить нельзя. Если все же между этими свойствами имеется связь, то выборка, состоящая из 12 наблюдений, не является достаточно представительной для того, чтобы установить это при 5%-ном уровне значимости.

Критерии Колмогорова — Смирнова

Одна очень полезная группа непараметрических критериев включает критерий Колмогорова — Смирнова. Наряду с другими приложениями их можно использовать для проверки соответствия выборочного и гипотетического распределений, и таким образом, они могут служить альтернативой описанного выше χ^2-критерия. Хотя критерии соответствия типа χ^2 также являются непараметрическими в том смысле, что их можно применять к наблюдениям с любым типом распределения, критерии Колмогорова—Смирнова являются наилучшими в некоторых обстоятельствах. Самое значительное их преимущество состоит в том, что для их использования не обязательно группировать наблюдения в произвольные категории; по этой причине они являются более чувствительными при проверке отклонений в хвостах распределений с низкими частотами, чем критерий χ^2.

На рис. 2.40 проиллюстрировано применение критерия Колмогорова—Смирнова. Из некоторой неизвестной совокупности взята выборка и нужно проверить соответствие выборочного распределения гипотетической модели. Как выборочная, так и гипотетическая модели нанесены на один и тот же график в кумулятивной форме, площади под графиками приведены к 1,0. Находим наибольшую разность между значениями распределения; это и есть статистика Колмогорова—Смирнова (К—С). В табл. 2.26 приведены критические значения для статистики К—С; они могут быть использованы для проверки как односторонней, так и двусторонней гипотез. Двусторонняя нулевая гипотеза утверждает, что классы распределений, из которых получена выборка, для всех значений x совпадают с классами распределений гипотетической модели. Односторонняя нулевая гипотеза утверждает, что все классы выборочного распределения равны или меньше, чем классы гипотетической модели, если мы используем максимальную положительную разность проверяемой статистики, или что все классы выборочного распределения равны или больше, чем классы гипотетической модели, если мы используем максимальную отрицательную разность

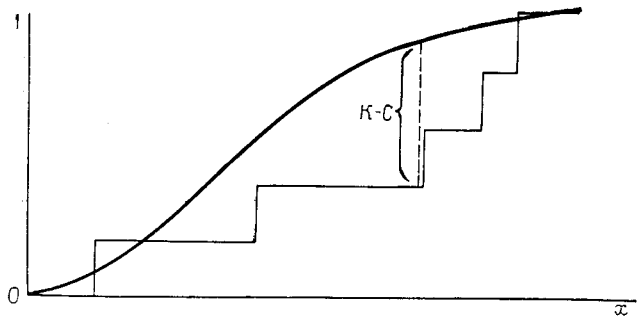


Рис. 2.40. Процедура проверки гипотезы о соответствии выборочного распределения (тонкая линия) гипотетической модели (жирная линия) с помощью критерия Колмогорова—Смирнова.

Оба графика даны в кумулятивной форме, область значения от 0,0 до 1,0. Максимальная разность — это проверяемая статистика К—С

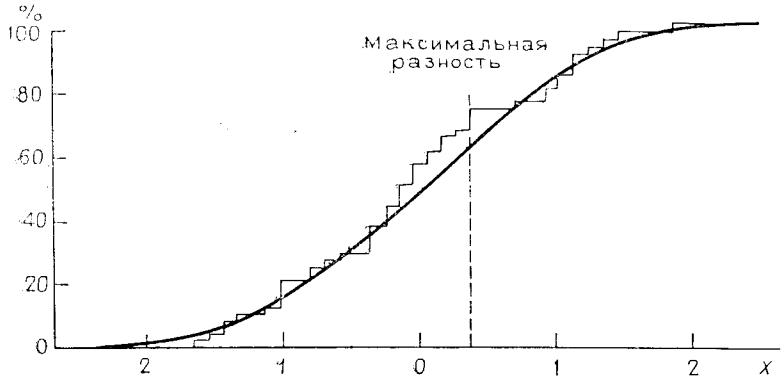


Рис. 2.41. Кумулятивный график распределения значений солености в бассейне Уайтуотер, стандартизованных и приведенных к масштабу от 0,0 до 1,0.

Максимальная разность между кумулятивным выборочным распределением и кумулятивным нормальным распределением — К—С равна 0,06

проверяемой статистики. В большинстве примеров мы будем использовать двустороннюю гипотезу и альтернативу.

Обычно критерий Колмогорова—Смирнова используется в тех случаях, когда гипотетическая модель полностью определена, т. е. параметры распределения известны (или предполагаются известными) из каких-либо дополнительных соображений, отличных от данных, содержащихся в самой выборке. Некоторое усовершенствование критерия, принадлежащее Лиллиефорсу [24], однако, позволяет использовать метод Колмогорова—Смирнова для проверки близости выборочного распределения к нормальному с заранее не определенными средним и диспер-

Таблица 2.26

Критические значения критерия Колмогорова — Смирнова для одностороннего и двустороннего критерия

				Односторонний критерий							
$\alpha=$	.10	.05	.025	.01	.005	$\alpha=$	.10	.05	.025	.01	.00
				Двусторонний критерий							
$\alpha=$	.20	.10	.05	.02	.01	$\alpha=$	.20	.10	.05	.02	.01
$n=1$	.900	.950	.975	.990	.995	$n=21$	.226	.259	.287	.321	.344
2	.684	.776	.842	.900	.929	22	.221	.253	.281	.314	.337
3	.565	.636	.708	.785	.829	23	.216	.247	.275	.307	.330
4	.493	.565	.624	.689	.734	24	.212	.242	.269	.301	.323
5	.447	.509	.563	.627	.669	25	.208	.238	.264	.295	.317
6	.410	.468	.519	.577	.617	26	.204	.233	.259	.290	.311
7	.381	.436	.483	.538	.576	27	.200	.229	.254	.284	.305
8	.358	.410	.454	.507	.542	28	.197	.225	.250	.279	.300
9	.339	.387	.430	.480	.513	29	.193	.221	.246	.275	.295
10	.323	.369	.409	.457	.489	30	.190	.218	.242	.270	.290
11	.308	.352	.391	.437	.468	31	.187	.214	.238	.266	.285
12	.296	.338	.375	.419	.449	32	.184	.211	.234	.262	.281
13	.285	.325	.361	.404	.432	33	.182	.208	.231	.258	.277
14	.275	.314	.349	.390	.418	34	.179	.205	.227	.254	.273
15	.266	.304	.338	.377	.404	35	.177	.202	.224	.251	.269
16	.258	.295	.327	.366	.392	36	.174	.199	.221	.247	.265
17	.250	.286	.328	.355	.381	37	.172	.196	.218	.244	.262
18	.244	.279	.309	.346	.371	38	.170	.194	.215	.241	.258
19	.237	.271	.301	.337	.361	39	.168	.191	.213	.238	.255
20	.232	.265	.294	.329	.352	40	.165	.189	.210	.235	.252
				Аппроксимация для $n>40$			$\frac{1.07}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.52}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

сией. Это усовершенствование позволяет использовать изложенный метод аналогично тому, как мы использовали процедуру χ^2 для проверки гипотез о данных, приведенных в табл. 2.19.

Следуя Лиллиефорсу, сначала приведем данные к стандартной форме, используя преобразование $Z_i=(X_i-\bar{X})/s$.

Среднее и дисперсия находятся обычным образом. Далее стандартное нормальное распределение и значения Z наносятся на график в кумулятивной форме, как это сделано на рис. 2.41 для значений солености в бассейне Уайт-Уотер. Максимальная абсолютная разность между двумя кривыми соответствует пробе с номером 35 и значению $Z=0{,}37$, где соленость равна 53 части/млн. Критические значения критерия Колмогорова—Смирнова приведены в табл. 2.26, хотя таблица доведена лишь до значения $n=40$. Приближенные значения для больших n можно найти по формулам, указанным в таблице. Для $n=48$ и уровня значимости $\alpha=0{,}10$ критическое значение К—С равно 0,17. Вычисленная статистика К—С равна (0,70—0,64) = 0,06. Это значение не попадает в критическую область. Поэтому отклонить нулевую гипотезу о том, что выборка получена из совокупности с нормальным распределением, нельзя.

На этом мы заканчиваем изложение элементарных статистических процедур, но не потому что рассмотрели их во всей полноте, а попросту для того, чтобы перейти к другим вопросам. Рассматриваемый материал является очень кратким изложением полугодового вводного курса в математическую статистику, к которому добавлены избранные главы из более специальных курсов. Автор не рассчитывал на то, что сумеет в столь малом объеме дать строгое обоснование обсуждаемых вопросов. Однако представляется, что изложенного будет достаточно для первого знакомства с основными понятиями теории проверки статистических гипотез.

Чрезвычайно краткое изложение процедур проверки статистических гипотез может служить введением в более сложный круг вопросов, рассматриваемых в последующих главах. Этот материал не исчерпывает содержания математической статистики. Уэллис и Робертс во введении к своей книге [33] указывают: «статистика — живой и увлекательный предмет, но ее изучение очень часто протекает слишком вяло». Изложив эту науку в контексте геологических задач, можно надеяться, что эта книга представит интерес для геологов, имеющих дело со статистической обработкой данных.

СПИСОК ЛИТЕРАТУРЫ

1. *Aichison J.*, A new approach to null correlations of proportions: Jour. of Int'l. Assoc. Mathematical Geology, 1981, *13*, p. 175—189.

2. *Aichison J.*, and *Brown A. C.*, 1969, The lognormal distribution: with special reference to its uses in economics: Cambridge Univ. Press, Cambridge, 1969, p. 176.

3. *Ash R. B.*, Basic probability theory: John Wiley and Sons, Inc., New York, 1970, 337 p.
Содержит полное изложение философских основ теории вероятностей и математической статистики.

4. *Bradley J. V.*, Distribution-free statistical tests: Prentice-Hall, Inc., Englewood Cliffs, N. J., 1968, 338 p.
Эта книга содержит наиболее полный перечень таблиц для непараметрических критериев.

5. *Chayes F.*, Ratio correlation: Univ. Chicago Press, Chicago, 1971, 99 p. Книга содержит подробное изложение результатов статистических исследований отношения и других способов определения замкнутых данных. Приводятся примеры из геохимии. Построенная на основе курса лекций, она легко доступна.

6. *Cheeney R. F.* Statistical methods in geology: George Allen and Union Ltd, London, 1983, 169 p. Имеется русский перевод: *Чини Р. Ф.* Статистические методы в геологии. М., Мир, 1987, 187 с. Четыре главы этой книги посвящены непараметрическим статистическим критериям, которые автор рекомендует, так как они облегчают вычисления.

7. *Cochran W. G.*, The test of goodness-of-fit: Annals of Mathematical Statistics, 1952, *3*, p. 315—345.

8. *Conover W. J.* Practical nonparametric statistics. 2nd ed.: John Wiley and Sons, Inc., New York, 1980, 493 p.
Большинство из упомянутых в этой главе непараметрических критериев рассмотрено в этой книге. В частности в ней содержится очень полное изложение критериев скачков.
9. *Duckworth W. E.* Statistical techniques in technological research. Methuen and Co., London, 1968, 303 p.
Одна из лучших и самых распространенных книг по статистике, предназначена для исследователей, желающих применять статистические процедуры. Большое внимание уделено планированию эксперимента. Несмотря на сложность этих вопросов, они становятся доступными благодаря блестящему изложению.
10. *Fisher R. A.,* Statistical methods and scientific inferences. Oliver and Boyd, Edinburgh, 1973, 175 p.
В книге излагается эволюция статистической мысли. Гл. 2 и 5 посвящены изложению основ теории вероятностей, ее понятий и следствий, вытекающих из них.
11. *Folk R. L.* Stages of textural maturity in sedimentary rocks. Jour. Sedimentary petrology, 1951, *21*, p. 127—130.
12. *Freund J. E., Williams F. J.* Dictionary/outline of basis statistics, McGraw-Hill, Inc., New York, 1966, 195 p.
Необходимое пособие для всех, кто применяет статистику. Вполне доступный и хорошо составленный словарь большинства геологических терминов, сводка общеизвестных (и некоторых неизвестны) формул и статистических критериев и таблиц.
13. *Griffiths J. C.* Scientific method in the analysis of sediments: McGraw-Hill, Inc., New York, 1967, 508 p. Имеется русский перевод: *Гриффитс Дж.* Научные методы исследования осадочных пород, М., Мир, 1971, 422 с.
Главы 13—22 являются введением в прикладную статистику. Особый интерес в этой книге представляет изложение вопросов, связанных с корректными схемами статистических экспериментов.
14. *Guenther W. C.* Concepts of statistical inference. 2nd ed., McGraw-Hill, Inc., New York, 1973, 512 p.
Сжатое введение в статистику для тех, кто хочет получить максимум информации в малом объеме. В книге затронут широкий круг вопросов, обычно не содержащихся в книгах такого объема.
15. *Hald A.,* Statistical tables and formulas, John Wiley and Sons, Inc., New York, 1952, 97 p.
16. *Hicks C. R.* Fundamental concepts in the design of experiments, 2nd ed. Holt, Rinehart and Winston, New York, 1973, 293 p.
Руководство по планированию эксперимента и дисперсионному анализу. В гл. 6 приводятся графическая интерпретация взаимодействий для двухфакторных планов и примеры.
17. *Kellaway F. W.,* ed., Penguin-Honeywell book of tables, Penguin Books Ltd., Harmondsworth, England, 1968, 75 p.
Математические таблицы, приведенные в этой книге, вычислены на ЭВМ фирмой Electronic Data Processing Division of Honeywell Control Ltd. Результаты были записаны на магнитной ленте, а затем напечатаны. Поэтому о них можно сказать, что их не касалась рука человека и что они не содержат ошибок, обычно вкрадывающихся в такие материалы. Табл. 2.11, 2.14 и 2.18 2-й гл. книги Дж. Дэвиса взяты из книги [17] и напечатаны с разрешения автора.
18. *Koch G. S.,* Jr. and Link R. F. Statistical analysis of geological data. Dover, Inc., New York, 1981, 850 p. Это однотомное издание представляет собой перепечатку оригинального двухтомника. Содержит подробное изложение статистического анализа геологических данных, в частности данных опробования. Включает подробное изложение выборочных схем, методов дисперсионного анализа и множественной регрессии. Подробно рассмотрены некоторые специальные вопросы и примеры. Особый интерес представляет гл. 11, в которой речь идет об интерпретации данных с постоянной суммой.
19. *Kork J. O.* Examination of the Chayes-Kruskal procedure for testing correlations between proportions. Jour. of Int'l. Assoc. for Mathematical Geology, 1977, *9*, p. 543—562.
20. *Krumbein W. C.* and *Graybill F. A.* An introduction to statistical models in geology: McGraw-Hill, Inc., New York, 1965, 475 p.
Имеется русский перевод: *Крамбейн У., Грейбилл Ф.* Статистические модели в геологии. М., Мир, 1969, 408 с.
Эта классическая книга будет полезна любому серьезному исследователю в области геостатистики. Особое внимание следует обратить на гл. 6—10.
21. *Krumbein W. C.* and *Pettijohn F. J.* Manual of sedimentary petrography. Appleton-Centry-Crofts, Inc., New York, 1938, 549 p.
22. *Lapin L. L.* Statistics for modern business decisions, 3rd ed. Harcourt, Brace Jovanovich, Inc., New York, 1982, 887 p.
В гл. 7 представлены эксперименты по моделированию доказательств центральной предельной теоремы.
23. *Li J. C. R.* Statistical inference, v. 1. Edwards Bros., Inc., Ann Arbor, Mich, 1964, 658 p.
Первый том представляет собой энциклопедию по элементарной статистике. Изложение в высшей степени наглядное. Особое внимание следует обратить на гл. 7, в которой рассматривается χ^2-распределение.
24. *Liliefors H. W.* On the Kolmogorov—Smirnov test for normality with mean and variance unknown, Jour. of the American Statistical Association, 1967, *62*, p. 399—402.
25. *McGray A. W.* Petroleum evaluations and economic decisions. Prentice-Hall, Inc., Englewood Cliffs, N. J., 1975, 448 p.
В гл. 3 представлено подробное изложение вероятностных распределений, встречающихся при разведке нефти.
26. *Parzen E.,* Modern probability theory and its application. John Wiley and Sons, Inc., New York, 1960, 464 p.
27. *Reyment R. A.* Introduction to quantitative paleontology Elsevier Publ. Co., Amsterdam, 1971, 226 p.
Настоятельно рекомендуемое руководство по приложению элементарной статистики к задачам экологии и палеоэкологии. В гл. 5 содержатся таблицы продолжительности жизни, т. е. темы, не затрагиваемые в этой книге. Описаны также многие непараметрические методы.
28. *Robinson J. E.,* Computer applications in petroleum geology. Hutchinson Ross Publ. Co., New York, 1982, 164 p.
Небольшая книга, в которой основное внимание обращено на построение карт.
29. *Saager R.* and *Sinclair A. J.,* Factor analysis of stream sediment geochemical data from the Mount Nansen area, Yukon Territory, Canada, Mineralogica Deposita, 1974, *9*, p. 243—252.
30. *Siegel S.,* Nonparametric statistics for the behavioral sciences. McGraw-Hill. Inc., New York, 1956, 312 p.
Одна из первых наиболее читаемых книг по непараметрическим методам.
31. *von Mises R.,* Probability, statistics and truth. Dover, New York, 1981, 224 p.
Этот старый перевод (1939 г.) книги немецкого классика является хорошим введением в философские вопросы статистики. Первые три главы посвящены главным образом определению вероятности.
32. *Walker H. M.* Degrees of freedom. Jour. of Educational Psyhology, 1940, *31*, p. 253—269.
Отличное изложение понятия степеней свободы.
33. *Wallis W. A.* and *Roberts H. V.* Statistics, a new approach. The Free press of Glencoe, New York, 1956, 646 p.
Занимательное введение в статистику с множеством примеров, историй, анекдотов о происхождении статистических процедур и их приложениях. Для тех, кто предпочитает увлекательное повествование математике.

Глава 3
МАТРИЧНАЯ АЛГЕБРА

Эта глава посвящена матричной алгебре. Большинство методов, которые будут рассмотрены в следующих главах, основаны на операциях с матрицами, часто выполняемых вычислительными машинами. Особое внимание будет уделено математическим операциям, которые лежат в основе анализа поверхностей тренда, метода главных компонент, дискриминантного анализа и др. Вычислительные процедуры, связанные с этими методами, почти невыполнимы без вычислительных машин в связи со сложностью и многократностью производимых вычислений. С помощью матричной алгебры их можно представить в доступном и сжатом виде. Таким образом, если читатель владеет основами матричной алгебры, он сможет понять основные процедуры, которые будут рассмотрены ниже.

К сожалению, большинству геологов при обучении не преподают курс матричной алгебры, несмотря на то что этот нетрудный предмет, по-видимому, более полезен, чем ряд других математических курсов. Курс матричной алгебры в колледжах обычно насыщен многочисленными теоремами и их доказательствами. Такое изложение непригодно для этой короткой главы, и мы ограничимся только теоремами, которые потребуются в последующем изложении. При этом предпочтение будет отдано не доказательствам, а примерам.

МАТРИЦА

Матрица — это набор чисел, расположенный в виде прямоугольной таблицы, точно такой же, как таблица данных. В матричной алгебре таблица рассматривается не как набор отдельных элементов, а как единое целое, что приводит к значительным упрощениям ряда процедур и зависимостей. Во второй главе отмечалось, что отдельный элемент матрицы обычно обозначается буквой с приписанными внизу индексами. Элементами матрицы могут быть дисперсии и ковариации, результаты наблюдений, коэффициенты системы уравнений и просто любые числа.

Так, например, в гл. 2 требовалось вычислить оценки дисперсий и ковариаций для содержаний микроэлементов, приведенных в табл. 2.3. Полученные результаты можно представить

в виде следующей матрицы:

$$\begin{bmatrix} s^2_{Cr} & cov_{Cr-Ni} & cov_{Cr-V} \\ cov_{Ni-Cr} & s^2_{Ni} & cov_{Ni-V} \\ cov_{V-Cr} & cov_{V-Ni} & s^2_{V} \end{bmatrix} = \begin{bmatrix} 570 & 537{,}5 & 663{,}75 \\ 537{,}5 & 562{,}5 & 718{,}75 \\ 663{,}75 & 718{,}5 & 1007{,}5 \end{bmatrix}$$

Набор чисел (допустим, значений переменной x), представленный в виде матрицы, обычно обозначается $[X]$, $\mathbf{X}$, (X) или $\|X\|$.

В этой книге будут применяться квадратные скобки и отдельные элементы матрицы будут обозначаться заглавными буквами с индексами, например x_{ij}. Вообще в литературе встречаются и такие условные обозначения, как $[x_{ij}]$, $[x]_{ij}$, или эквивалентные греческие буквы, как χ_{ij}. Символ x_{ij} означает элемент i-й строки и j-го столбца. Например, если $[X]$ — матрица порядка 3×3:

$$\begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix},$$

то $x_{33}=9$, $x_{13}=7$, $x_{21}=2$ и т. д. Если число столбцов равно числу строк, то матрица называется квадратной, а ее элементы, индексы которых равны (т. е. $i=j$), называются диагональными элементами. В вышеприведенной матрице характеристик распределения микроэлементов диагональными являются дисперсии содержаний. Все остальные элементы — ковариации. В той же матрице диагональные элементы равны соответственно 1, 5 и 9. Чаще всего встречаются квадратные матрицы, но неквадратные тоже нередки, несмотря на то что операции с ними подчинены строгим ограничениям. Особенно часто используются две формы неквадратных матриц: вектор-строка, т. е. матрица порядка $1\times m$, и вектор-столбец, т. е. матрица порядка $m\times1$.

Очень часто приходится иметь дело с двумя типами квадратных матриц. Это симметричная матрица, в которой $x_{ij}=x_{ji}$, как, например, матрица вида

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{bmatrix}$$

Другим примером матрицы, симметричной относительно главной диагонали, может служить матрица дисперсий и ковариаций, приведенная выше.

Единичная матрица — это симметричная матрица, в которой все диагональные элементы равны 1, а все остальные равны 0:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

ЭЛЕМЕНТАРНЫЕ ОПЕРАЦИИ С МАТРИЦАМИ

Сложение и вычитание матриц подчиняется законам обычной алгебры, но при одном важном условии: складываемые или вычитаемые матрицы должны иметь равное число строк и равное число столбцов.

При выполнении операции сложения $[C]=[A]+[B]$ к каждому элементу матрицы $[A]$ прибавляется соответствующий элемент матрицы $[B]$. Если матрицы разных порядков, то операция сложения невыполнима. Вычитание $[C]=[A]-[B]$ выполняется таким же способом: каждый элемент матрицы $[B]$ вычитается из соответствующего элемента матрицы $[A]$.

В качестве иллюстрации в табл. 3.1 приведена характеристика добычи бентонитовой глины в трех горнодобывающих районах штата Вайоминг.

Добывается три основных сорта глины: один применяется для приготовления бурового раствора, другой — в качестве литейной глины и последний используется как лекарственное и косметическое средство, для керамических целей и др. Все эти данные можно записать в виде матрицы $[A]$ порядка 3×3:

$$[A]=\begin{bmatrix} 105 & 63 & 5 \\ 218 & 80 & 2 \\ 220 & 76 & 1 \end{bmatrix}$$

Продукцию, добытую в следующем году, также можно охарактеризовать матрицей $[B]$:

$$[B]=\begin{bmatrix} 84 & 102 & 4 \\ 240 & 121 & 1 \\ 302 & 28 & 0 \end{bmatrix}$$

Общая продукция за два года в трех районах будет представлять сумму $[C]$ двух матриц $[A]$ и $[B]$:

$$\underset{[A]}{\begin{bmatrix} 105 & 63 & 5 \\ 218 & 80 & 2 \\ 220 & 76 & 1 \end{bmatrix}} + \underset{[B]}{\begin{bmatrix} 84 & 102 & 4 \\ 240 & 121 & 1 \\ 302 & 28 & 0 \end{bmatrix}} = \underset{[C]}{\begin{bmatrix} 189 & 165 & 9 \\ 458 & 201 & 3 \\ 522 & 104 & 1 \end{bmatrix}}$$

Таблица 3.1

Добыча бентонитовых глин в штате Вайоминг, 1964 г. (в $1\cdot10^5$ т)

Регион	Глина для бурового раствора	Литейная глина	Прочие глины
Восточный	105	63	5
Районы, пограничные со штатом	218	80	2
Монтана	220	76	1
Центральный			

Изменение добычи можно представить как разность матрицы $[A]$ и матрицы $[B]$, т. е.

$$\underset{[B]}{\begin{bmatrix} 84 & 102 & 4 \\ 240 & 121 & 1 \\ 302 & 28 & 0 \end{bmatrix}} - \underset{[A]}{\begin{bmatrix} 105 & 63 & 5 \\ 218 & 80 & 2 \\ 220 & 76 & 1 \end{bmatrix}} = \underset{[C]}{\begin{bmatrix} -21 & 39 & -1 \\ 22 & 41 & -1 \\ 82 & -48 & -1 \end{bmatrix}}$$

Отметим, что положительные элементы матрицы $[B]-[A]$ соответствуют росту добычи глин.

Как и в обычной алгебре, сложение матриц порядка $n\times m$ коммутативно, т. е. $[A]+[B]=[B]+[A]$, и ассоциативно, т. е. $([A]+[B])+[C]=[A]+([B]+[C])$. Порядок матриц при вычитании, конечно, существен.

Умножение матрицы на константу сводится к умножению каждого ее элемента на эту константу. Например,

$$3\cdot\begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} = \begin{bmatrix} 3 & 12 \\ 6 & 15 \\ 9 & 18 \end{bmatrix}$$

Аналогично проводится деление матрицы на постоянное число. В качестве простого примера применения этих операций рассмотрим табл. 3.2, содержащую результаты измерения длины трех осей a, b и c кремниевых галек, взятых из ледниковых отложений. Измерения сделаны в дюймах и их нужно перевес-

Таблица 3.2

Измерения осей галек из ледниковых отложений

Образец	Длина оси, дюймы		
	a	b	c
1	3,4	2,2	1,8
2	4,6	4,3	4,2
3	5,4	4,7	4,7
4	3,9	2,8	2,3
5	5,1	4,9	3,8

ти в миллиметры. Чтобы получить матрицу, содержащую данные в миллиметрах, надо умножить матрицу $[A]$ на постоянную величину 25,4.

$$25{,}4 \cdot [A] = [C]$$

$$25{,}4 \cdot \begin{bmatrix} 3{,}4 & 2{,}2 & 1{,}8 \\ 4{,}6 & 4{,}3 & 4{,}2 \\ 5{,}4 & 4{,}7 & 4{,}7 \\ 3{,}9 & 2{,}8 & 2{,}3 \\ 5{,}1 & 4{,}9 & 3{,}8 \end{bmatrix} = \begin{bmatrix} 86{,}36 & 55{,}88 & 45{,}72 \\ 116{,}84 & 109{,}22 & 106{,}68 \\ 137{,}16 & 119{,}38 & 119{,}38 \\ 99{,}06 & 71{,}12 & 58{,}42 \\ 129{,}54 & 124{,}46 & 96{,}52 \end{bmatrix}$$

УМНОЖЕНИЕ МАТРИЦ

Вспомним задачу с бросанием монеты, рассмотренную в гл. 2, где вычислялась вероятность появления последовательности гербов после ряда бросаний, если вероятность выпадения герба при одном бросании равна $^1/_2$. Вероятность выпадения трех гербов при трех бросаниях равна $^1/_2 \times {}^1/_2 \times {}^1/_2$, или $^1/_2{}^3$. Аналогичную задачу можно сформулировать при изучении литологической характеристики стратиграфического разреза. Предположим, что при изучении разреза в нем выделены три литологические разновидности пород: песок, сланец, известняк. Каждый участок разреза можно отнести к одному из перечисленных типов. В итоге можно построить матрицу частот появления каждого типа пород в разрезе вслед за другими типами:

		В: Песок	Сланец	Известняк
Из	Песок	59	18	2
	Сланец	14	86	41
	Известняк	4	34	51

Эта матрица называется матрицей переходных частот и показывает, например, что сланец следует за песком 18 раз, а известняк за песком — только 2 раза; известняк следует за сланцем 41 раз, повторение известняка встречается 51 раз, а за песком — только 2 раза.

Чтобы получить из этих частот вероятности, нужно разделить каждый элемент матрицы на сумму элементов соответствующей строки. Таким образом получим матрицу переходных вероятностей, приведенную ниже, которая содержит вероятности событий, заключающихся в том, что порода одного типа следует за породой другого типа.

Этот вопрос будет рассмотрен детально в следующей главе при анализе временных рядов. Теперь же нас интересует матрица вероятностей, аналогичных вероятностям в задаче о бросании монеты:

		В: Песок	Глина	Известняк
Из	Песок	0,74	0,23	0,03
	Глина	0,10	0,61	0,29
	Известняк	0,05	0,38	0,57

Так же, как и в задаче о бросании монеты, где была определена вероятность выпадения последовательности гербов с помощью возведения в степень вероятности выпадения герба при одном бросании, можно определить вероятность появления определенных последовательностей пород в заданном интервале. Для этого достаточно возвести в степень матрицу переходных вероятностей и получить после n таких перемножений матрицу вероятностей $[P']$, которая будет равна матрице $[P^n]$. Матрица в степени n — это просто результат умножения матрицы на себя n раз. Однако для выполнения этой операции нужно знать правила перемножения матриц.

Самая простая форма матричного умножения — это умножение двух квадратных матриц $[A]$ и $[B]$ одинакового порядка, произведение которых снова является квадратной матрицей $[C]$. При этом удобно располагать матрицу следующим образом:

$[B]$

$[A]$ [Результат]

Чтобы получить значение элемента c_{ij}, нужно умножить каждый элемент i-й строки матрицы $[A]$ (начиная слева) на соответствующий элемент j-го столбца матрицы $[B]$ (начиная сверху). Для получения искомого элемента c_{ij} все эти произведения суммируются. Ниже на примере двух матриц показана последовательность операций при матричном умножении:

$$\begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \\ 5 & 6 & 7 \end{bmatrix}$$

Сначала умножим a_{11} на b_{11} и в результате получим 1:

$$\begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \\ 5 & 6 & 7 \end{bmatrix}$$

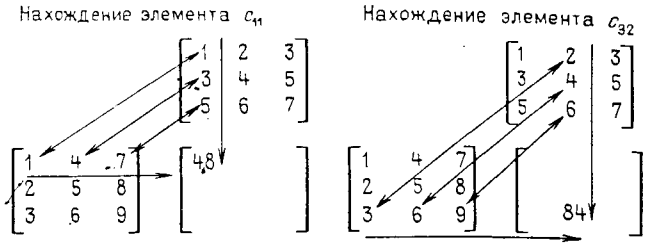
Далее, умножая a_{12} на b_{21}, получаем 12:

$$\begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 5 \\ 5 & 6 & 7 \end{bmatrix}$$

Наконец, умножая a_{13} на b_{31}, получаем 35:

$$\begin{bmatrix}1&4&7\\2&5&8\\3&6&9\end{bmatrix}\cdot\begin{bmatrix}1&2&3\\3&4&5\\5&6&7\end{bmatrix}$$

Значение элемента c_{11} представляет собой сумму трех вычисленных значений $1+12+35=48$. Последовательность этих операций изображена на приведенной ниже диаграмме. Отметим, что каждый элемент c_{ij} произведения матриц получается в результате сложения произведений элементов i-й строки матрицы $[A]$ на элементы j-го столбца матрицы $[B]$.



В итоге результат умножения будет иметь следующий вид:

$$\begin{bmatrix}1&2&3\\3&4&5\\5&6&7\end{bmatrix}$$

$$\begin{bmatrix}1&4&7\\2&5&8\\3&6&9\end{bmatrix}\begin{bmatrix}48&60&72\\57&72&87\\66&84&102\end{bmatrix}$$

Если порядок перемножаемых матриц изменить на обратный $[B]\cdot[A]=[C]$, то получим другой результат:

$$\begin{bmatrix}1&4&7\\2&5&8\\3&6&9\end{bmatrix}$$

$$\begin{bmatrix}1&2&3\\3&4&5\\5&6&7\end{bmatrix}\begin{bmatrix}14&32&50\\26&62&98\\38&92&146\end{bmatrix}$$

Операцию $[A]\cdot[B]=[C]$ называют умножением матрицы $[B]$ на матрицу $[A]$ слева. Аналогично, о матрице $[A]$ говорят, что она умножена на матрицу $[B]$ справа. Это простейший способ определения умножения матриц.

Если перемножаются две квадратные матрицы, то их произведение также будет квадратной матрицей того же порядка. Однако если перемножить матрицы порядков $m\times n$ и $n\times r$, то в результате получим матрицу порядка $m\times r$, т. е. произведение матриц имеет число строк, равное числу строк первой матрицы, и число столбцов, равное числу столбцов второй матрицы. Например, умножая матрицу порядка 5×3 на матрицу порядка 3×2, получим матрицу порядка 5×2:

$$\begin{bmatrix}1&2&4\\2&1&2\\3&1&1\\2&3&1\\1&2&0\end{bmatrix}\cdot\begin{bmatrix}3&4\\2&1\\0&1\end{bmatrix}=\begin{bmatrix}7&10\\8&11\\11&14\\12&12\\7&6\end{bmatrix}$$

Необходимо отметить, что матрицу порядка 3×2 нельзя умножать на матрицу порядка 5×3, так как число столбцов первой матрицы (два) не равно числу строк второй матрицы (пять).

В результате умножения матрицы порядка $m\times n$ на матрицу порядка $n\times m$ получается квадратная матрица, причем порядок перемножаемых матриц определяет порядок результирующей матрицы:

$$\begin{bmatrix}1&2&3\\4&5&6\end{bmatrix}\cdot\begin{bmatrix}9&6\\8&6\\7&5\end{bmatrix}=\begin{bmatrix}46&28\\118&73\end{bmatrix}$$

Однако, поменяв перемножаемые матрицы местами, получим

$$\begin{bmatrix}9&6\\8&5\\7&4\end{bmatrix}\cdot\begin{bmatrix}1&2&3\\4&5&6\end{bmatrix}=\begin{bmatrix}33&48&63\\28&41&54\\23&34&45\end{bmatrix}$$

Общая формула элемента матрицы $[C]$, являющейся произведением двух матриц, имеет следующий вид:

$$c_{ij}=\sum_{k=1}^{n}a_{ik}b_{kj}. \qquad (3.1)$$

Отметим, что при последовательном перемножении более чем двух матриц действует ассоциативный закон, т. е. порядок выполнения несуществен. Так,

$$[A]\cdot[B]\cdot[C]=([A]\cdot[B])\cdot[C]=[A]\cdot([B]\cdot[C]).$$

Матрицу можно возвести в степень, так как возведение в степень состоит попросту в выполнении ряда умножений. Так,

$$[A]^2 = [A]\cdot[A],$$

$$[A]^3 = [A]^2\cdot[A] = [A]\cdot[A]\cdot[A].$$

Матрицы также можно возводить в дробную степень, чаще всего в степень $^1/_2$. Эта операция эквивалентна извлечению квадратного корня из матрицы. $[A]^{1/2}$ есть матрица $[X]$, квадрат которой равен $[A]$:

$$[A]^{1/2} = [X]; \qquad [X]^2 = [A].$$

Нахождение дробной степени матрицы довольно сложно. К счастью, нам ниже понадобятся лишь дробные степени диагональных матриц. Это матрица, все элементы которой, исключая элементы, стоящие на главной диагонали, равны нулю. Их специальные свойства позволяют легко осуществлять возведение в дробную степень. Если возвести диагональную матрицу $[A]$ в полуцелую степень, то в результате получится диагональная матрица с ненулевыми элементами на главной диагонали, равными корням квадратным из соответствующих элементов исходной матрицы $[A]$. Например, если $[A]$ — матрица порядка 3×3, то

$$\begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix}^{\frac{1}{2}} = \begin{bmatrix} \sqrt{a_{11}} & 0 & 0 \\ 0 & \sqrt{a_{22}} & 0 \\ 0 & 0 & \sqrt{a_{33}} \end{bmatrix}$$

Единичная матрица — это диагональная матрица специального вида с очень полезными свойствами. Если некоторая матрица умножается на единичную матрицу, то в результате получается та же исходная матрица:

$$\begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix}\cdot\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix}$$

Таким образом, единичная матрица соответствует единице в обычном умножении. Это свойство особенно важно в операциях, рассматриваемых в следующем разделе.

СВЕРТКА

Свертка — векторная операция, широко используемая в анализе временных рядов, особенно при обработке сейсмических данных. В физическом смысле свертка описывает способ взаимодействия комплексов двух энергетических волн. Рассмотрим

два вектора — $[A] = [a_0\, a_1\, a_2 \ldots]$ и $[B] = [b_0\, b_1\, b_2 \ldots]$. Их свертка обозначается оператором *, $[C]=[A]*[B]$ и приводит к вектору $[C]=[c_0\ c_1\ c_2\ \ldots]$. Компоненты вектора $[C]$ определяются по формуле

$$c_i = \sum_{j=0}^{i} a_j b_{i-j}. \tag{3.2}$$

Например, требуется свернуть два вектора $[A]$ и $[B]$, содержащих только по два члена т. е. $[A]=[a_0\ a_1]$ и $[B]=[b_0\ b_1]$. Первая компонента свертки —

$$c_0 = a_0 b_{0-0} = a_0 b_0,$$

вторая компонента —

$$c_1 = a_0 b_{1-0} + a_1 b_{1-1} = a_0 b_1 + a_1 b_0,$$

третья компонента —

$$c_2 = a_0 b_{2-0} + a_1 b_{2-1} + a_2 b_{2-2} = a_0 b_2 + a_1 b_1 + a_2 b_0.$$

Но ни a_2, ни b_2 не существуют, поэтому третья компонента равна

$$c_2 = a_1 b_1.$$

Таким образом, свертка $[A]$ и $[B]$ в этом случае есть

$$[A]*[B] = [C] = [a_0 b_0 \quad a_0 b_1 + a_1 b_0 \quad a_1 b_1].$$

Заметим, что хотя как $[A]$, так и $[B]$ содержат только по две компоненты, свертка $[C]$ содержит три компоненты. В общем случае, если первый вектор, участвующий в свертке, содержит n элементов, а второй — m элементов то свертка содержит $n+m-1$ элементов.

Робинзон и Трейтель [7] дают интересное графическое объяснение операции свертывания. Используя их метод, свернем два вектора, каждый из которых содержит три элемента. Сначала строим таблицу или матрицу, в которой один вектор определяет строки, а другой вектор — столбцы. Элементы в этой матрице — произведения строк на столбцы

$$\begin{array}{c} \\ a_0 \\ a_1 \\ a_2 \end{array}\begin{array}{c} \begin{array}{ccc} b_0 & b_1 & b_2 \end{array} \\ \begin{bmatrix} a_0 b_0 & a_0 b_1 & a_0 b_2 \\ a_1 b_0 & a_1 b_1 & a_1 b_2 \\ a_2 b_0 & a_2 b_1 & a_2 b_2 \end{bmatrix} \end{array}$$

Полученную матрицу теперь подразделяем на полоски, начиная от нижнего левого угла до верхнего правого, так что каждый элемент полоски есть столбец, если идти справа вниз,

и есть строка, если идти слева вверх, т. е.

$$\begin{bmatrix} a_0b_0 & a_0b_1 & a_0b_2 \\ a_1b_0 & a_1b_1 & a_1b_2 \\ a_2b_0 & a_2b_1 & a_2b_2 \end{bmatrix}$$

Элементы вектора $[C]$ состоят из сумм элементов полоски, начиная с верхнего левого угла. Например,

$$[C] = [a_0b_0 \quad a_1b_0 + a_0b_1 \quad a_2b_0 + a_1b_1 + a_0b_2 \quad a_2b_1 + a_1b_2 \quad a_2b_2].$$

Это в точности тот член, который находится с помощью приведенного выше уравнения. Заметим, что, как и ожидалось, число членов в $[C]$ есть $3+3-1=5$.

Мы снова рассмотрим численный пример, продемонстрировав одновременно, как один вектор может быть свернут с другим иной размерности. Пусть заданы два вектора:

$$[A] = [1 \quad 3 \quad 5 \quad 7 \quad 2],$$
$$[B] = [6 \quad 2 \quad 4].$$

Свертка $[C]$ будет иметь семь членов. Используя графический метод, их можно найти по схеме

$$\begin{array}{c} \\ 1 \\ 3 \\ 5 \\ 7 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 6 & 2 & 4 \end{array} \\ \begin{bmatrix} 6 & 2 & 4 \\ 18 & 6 & 12 \\ 30 & 10 & 20 \\ 42 & 14 & 28 \\ 12 & 4 & 8 \end{bmatrix} \end{array}$$

Таким образом,

$$\begin{aligned} c_0 &= 6, \\ c_1 &= 18 + 2 = 20, \\ c_2 &= 30 + 6 + 4 = 40, \\ c_3 &= 42 + 10 + 12 = 64, \\ c_4 &= 12 + 14 + 20 = 46, \\ c_5 &= 4 + 28 = 32, \\ c_6 &= 8, \end{aligned}$$

или

$$[C] = [6 \quad 20 \quad 40 \quad 64 \quad 46 \quad 32 \quad 8].$$

Свертка не очень широко используется в большинстве статистических операций, исключение составляет лишь анализ временных рядов. Мы еще встретимся с этой процедурой, когда будем рассматривать фильтрацию (см. гл. 4).

ОБРАЩЕНИЕ МАТРИЦ И СИСТЕМЫ УРАВНЕНИЙ

В данном разделе мы определим операцию над матрицами, обратную операции умножения, зная при этом, что прямое деление одной матрицы на другую невыполнимо. Однако, используя правила матричного умножения, можно ввести соответствующее понятие, для чего необходимо решить матричное уравнение $[A]\cdot[X]=[B]$ относительно неизвестной матрицы $[X]$. Необходимо подчеркнуть, что это одна из наиболее важных операций матричной алгебры, которая будет неоднократно использоваться в дальнейшем, в частности для решения систем уравнений в дискриминантном и тренд-анализах.

Приведенное выше уравнение решается с помощью нахождения матрицы $[A]^{-1}$, обратной матрице $[A]$, которая удовлетворяет соотношению $[A]\cdot[A]^{-1}=[1]$. Так как умножение матрицы на единичную матрицу [1] ее не изменяет, мы можем умножить обе части уравнения на $[A]^{-1}$, в результате чего матрица $[A]$ в левой части уравнения исчезнет. В то же время мы находим выражение неизвестной матрицы $[X]$ через матрицы $[B]$ и $[A]^{-1}$. Матрица $[A]$ должна быть квадратной. Исходя из $[A]\cdot[X]=[B]$, умножим обе части на матрицу, обратную к $[A]$, т. е. $[A]^{-1}:[A]^{-1}\cdot[A]\cdot[X]=[A]^{-1}\cdot[B]$.

Так как $[A]^{-1}\cdot[A]=[1]$ и $[1]\cdot[X]=[X]$, получим

$$[X] = [A]^{-1}\cdot[B]. \qquad (3.3)$$

Таким образом, задача, обратная умножению матриц, сводится к нахождению матрицы $[X]$, определенной соотношением (3.3). В некоторых случаях обратную матрицу найти нельзя, так как в процессе обращения иногда приходится выполнять деление на нуль. Необратимая матрица называется вырожденной, но такие матрицы в этой главе мы рассматривать не будем.

Процедуру обращения матриц можно проиллюстрировать, решая систему уравнений в матричной форме. Значения неизвестных в этом случае $x_1=2$ и $x_2=3$. Попытаемся найти их, используя умножение и обращение матриц:

$$4x_1 + 10x_2 = 38,$$
$$10x_1 + 30x_2 = 110.$$

Эту систему уравнений можно записать в матричной форме:

$$[A]\cdot[X] = [B],$$

где $[A]$ — матрица коэффициентов, $[X]$ — вектор-столбец неизвестных, $[B]$ — вектор-столбец правых частей уравнений.

В нашем частном случае

$$\begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 38 \\ 110 \end{bmatrix}$$

Чтобы решить это уравнение, необходимо обратить матрицу $[A]$, тогда произведение матриц $[A]^{-1}$ и $[B]$ даст нам вектор решения $[X]$.

Возможно, читателю не понятно, почему систему уравнений можно представить в указанной матричной форме. Чтобы убедиться в этом, порекомендуем перемножить матрицу $[A]$ и вектор-столбец $[X]$. В результате должен получиться вектор-столбец левых частей уравнений:

$$\begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 4x_1 + 10x_2 \\ 10x_1 + 30x_2 \end{bmatrix}$$

Выполнив это умножение, вы убедитесь, что система записана правильно. Действительно, по правилам умножения матриц получаем

$$\begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 4x_1 + 10x_2 \\ \end{bmatrix}$$

Затем, умножая нижнюю строку, получаем

$$\begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \\ 10x_1 + 30x_2 \end{bmatrix}$$

Теперь решим систему уравнений с помощью обращения матрицы $[A]$. Поместим матрицу $[A]$ слева от единичной матрицы $[I]$ и выполним все необходимые операции одновременно на обеих матрицах. Цель этих операций — свести диагональные элементы матрицы $[A]$ к единицам, а все остальные — к нулям. Это выполняется с помощью деления строк матрицы $[A]$ на константы, а также их сложения или вычитания. Последовательность операций следующая:

1. $\begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ Матрицу $[A]$ помещаем рядом с единичной матрицей $[I]$
2. $\begin{bmatrix} 1 & 2{,}5 \\ 10 & 30 \end{bmatrix} \begin{bmatrix} 0{,}25 & 0 \\ 0 & 1 \end{bmatrix}$ Первую строку делим на 4, получаем при этом 1 на месте a_{11}
3. $\begin{bmatrix} 1 & 2{,}5 \\ 0 & 5 \end{bmatrix} \begin{bmatrix} 0{,}25 & 0 \\ -2{,}5 & 1 \end{bmatrix}$ Вычитаем из второй строки удесятеренную первую, получаем 0 на месте a_{21}
4. $\begin{bmatrix} 1 & 2{,}5 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0{,}25 & 0 \\ -0{,}5 & 0{,}2 \end{bmatrix}$ Делим вторую строку на 5, получаем 1 на месте a_{22}

5. $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1{,}5 & -0{,}5 \\ -0{,}5 & 0{,}2 \end{bmatrix}$ Умножаем вторую строку на 2,5 и вычитаем ее из первой, получаем 0 на месте a_{12}

Мы получили обратную матрицу. Для проверки правильности обращения достаточно умножить исходную матрицу $[A]$ на полученную матрицу $[A]^{-1}$, и в результате должна получиться единичная матрица

$$\begin{bmatrix} 1{,}5 & -0{,}5 \\ -0{,}5 & 0{,}2 \end{bmatrix} \cdot \begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Так как $[A]^{-1} \cdot [A] = [I]$, то выполняются следующие равенства:

$$[A]^{-1} \cdot [A] \cdot [X] = [A]^{-1} \cdot [B],$$
$$[I] \cdot [X] = [A]^{-1} \cdot [B],$$
$$[X] = [A]^{-1} \cdot [B].$$

Умножив матрицу $[A]^{-1}$ на матрицу $[B]$, найдем неизвестную матрицу $[X]$:

$$[A]^{-1} \cdot [B] = [X]$$
$$\begin{bmatrix} 1{,}5 & -0{,}5 \\ -0{,}5 & 0{,}2 \end{bmatrix} \cdot \begin{bmatrix} 38 \\ 110 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

Вектор-столбец X содержит искомые неизвестные $x_1 = 2$, $x_2 = 3$. Вспоминая, что именно эти значения удовлетворяли исходным уравнениям, мы убеждаемся в том, что нашли x_1 и x_2 правильно.

Рассмотрим еще один пример решения системы уравнений с помощью обращения матрицы. Запишем приведенные ниже уравнения в матричной форме и найдем x_1 и x_2, обращая матрицу системы:

$$2x_1 + x_2 = 4,$$
$$3x_1 + 4x_2 = 1.$$

Кратко запишем операции, выполняемые при обращении матрицы:

1. $\begin{bmatrix} 2 & 1 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 1 \end{bmatrix}$
2. $\begin{bmatrix} 2 & 1 \\ 3 & 4 \end{bmatrix}^{-1} = \begin{bmatrix} 4/5 & -1/5 \\ -3/5 & 2/5 \end{bmatrix}$
3. $\begin{bmatrix} 4/5 & -1/5 \\ -3/5 & 2/5 \end{bmatrix} \cdot \begin{bmatrix} 4 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ -2 \end{bmatrix}$

Следовательно, значения неизвестных будут $x_1=3$ и $x_2=-2$.

Заметим, что описанная процедура очень напоминает классический алгебраический метод решения системы двух уравнений. В самом деле, решение системы уравнений — наиболее важное применение процедуры обращения матриц. В данном случае преимущество матричных методов по сравнению с методами обычной алгебры заключается в том, что они более систематизированы и лучше поддаются программированию. Большинство процедур, описанных в следующих главах, приводит к решению систем уравнений. Эти системы обычно записываются в матричном виде и решаются описанным выше методом. Операцию обращения матриц можно применить к квадратным матрицам любого порядка, а не только к матрицам порядка 2×2, рассмотренным в примерах. Убедитесь в этом, обратив приведенную ниже матрицу порядка 3×3:

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 6 & 5 \\ 3 & 5 & 6 \end{bmatrix}$$

Если мы захотим обратить диагональную матрицу, то задача окажется значительно проще. Обратная матрица — это просто диагональная матрица, ненулевые элементы которой — обратные величины соответствующих элементов обращаемой матрицы. Рассмотрим матрицу $[A]$ порядка 3×3:

$$\begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{a_{11}} & 0 & 0 \\ 0 & \frac{1}{a_{22}} & 0 \\ 0 & 0 & \frac{1}{a_{33}} \end{bmatrix}$$

Некоторые комбинации довольно сложных операций становятся очень простыми, если входящие в них матрицы диагональные. Например, рассмотрим умножение $[A]^{-1}\cdot[A]^{1/2}=[A]^{-1/2}$. Если $[A]$ — матрица порядка 3×3, то произведение равно

$$\begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix}^{-\frac{1}{2}} = \begin{bmatrix} \frac{1}{\sqrt{a_{11}}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{a_{22}}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{a_{33}}} \end{bmatrix}$$

В некоторых случаях не обязательно обращать матрицу, если требуется только решить систему уравнений. В рассмотренном примере мы искали значения элементов матрицы $[X]$, удовлетворяющей уравнению

$$\begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 38 \\ 110 \end{bmatrix}$$

С этой целью мы обратили матрицу $[A]$ и нашли матрицу $[X]$, умножив $[A]^{-1}$ на $[B]$. Вместо этого можно было прямо использовать матрицу $[B]$ по мере того, как матрица $[A]$ преобразовывалась в единичную матрицу. С этой целью удобно использовать так называемую расширенную матрицу, состоящую из n строк и $n+1$ столбцов. Вектор-столбец $[B]$ занимает $n+1$-й столбец расширенной матрицы, остальная часть ее — матрица порядка $n\times n$ — обращается. Используя этот способ, решим ту же задачу:

1.	$\left[\begin{array}{cc\|c} 4{,}0 & 10{,}0 & 38{,}0 \\ 10{,}0 & 30{,}0 & 110{,}0 \end{array}\right]$	Из матриц $[A]$ и $[B]$ составляем матрицу порядка $n\times(n+1)$
2.	$\left[\begin{array}{cc\|c} 1{,}0 & 2{,}5 & 9{,}5 \\ 1{,}0 & 3{,}0 & 11{,}0 \end{array}\right]$	Первую строку делим на 4, вторую — на 10
3.	$\left[\begin{array}{cc\|c} 1{,}0 & 2{,}5 & 9{,}5 \\ 0{,}0 & 0{,}5 & 1{,}5 \end{array}\right]$	Первую строку вычитаем из второй
4.	$\left[\begin{array}{cc\|c} 1{,}0 & 0{,}0 & 2{,}0 \\ 0{,}0 & 0{,}5 & 1{,}5 \end{array}\right]$	Умноженную на 5 вторую строку вычитаем из первой
5.	$\left[\begin{array}{cc\|c} 1{,}0 & 0{,}0 & 2{,}0 \\ 0{,}0 & 1{,}0 & 3{,}0 \end{array}\right]$	Вторую строку делим на 0,5

Таким образом, в результате $n+1$-й столбец полученной матрицы содержит решение системы уравнений, а исходная матрица заменена единичной.

Для обращения матриц известен ряд математических процедур, заслуживающих внимания. Десятки методов созданы для решения систем уравнений и существуют сотни программных версий. Некоторые из них предназначены специально для матриц частного вида, таких, как разреженные матрицы (т. е. матрицы, содержащие много нулевых элементов) или матрицы, обладающие некоторыми типами симметрии. Большинство вычислительных центров имеет стандартные программы в своих библиотеках, предназначенные для выполнения матричных операций. Необходимо заранее с ними ознакомиться, прежде чем потребуется выполнять работу, оперируя с большими массивами данных.

ТРАНСПОНИРОВАНИЕ

Операция замены столбцов матрицы строками называется транспонированием.

Например, матрица

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$$

после транспонирования имеет вид

$$\begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

Отметим, что после транспонирования первая строка стала первым столбцом, а вторая строка — вторым столбцом. Если обозначить матрицу через $[A]$, то матрицу, полученную транспонированием матрицы $[A]$, обозначим $[A]^T$. При транспонировании элемент a_{ij} переходит в элемент a_{ji}. Во многих задачах, рассматриваемых в гл. 6, часто будем пользоваться тем, что вектор-строка матрицы $[A]$ при транспонировании превращается в вектор-столбец матрицы $[A]^T$. Вектор-строка и вектор-столбец получаются один из другого транспонированием. Например,

$$[1 \quad 2 \quad 3 \quad 4] \text{ и } \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$$

Позже будем иметь много случаев, когда матрица умножается на матрицу, транспонированную к ней. Если матрица $[X]$ имеет n строк и m столбцов, то транспонированная к ней $[X]^T$ будет иметь m строк и n столбцов. Умножая матрицу $[X]$ слева на транспонированную к ней матрицу, мы получаем квадратную симметрическую матрицу порядка $m \times n$, называемую меньшей матрицей — произведением матрицы $[X]$. Умножая матрицу $[X]$ справа на транспонированную к ней матрицу, получаем квадратную симметрическую матрицу порядка $n \times n$, называемую большой матрицей — произведением матрицы $[X]$.

ОПРЕДЕЛИТЕЛИ

Прежде чем перейти к изложению нашей последней темы, касающейся собственных значений и собственных векторов, остановимся коротко на определителе матрицы. Определитель — это число, приписываемое квадратной матрице в результате

некоторой последовательности операций. Определитель символически обозначается $\det A$, $|A|$ или

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$$

Он вычисляется как сумма $n!$ членов вида

$$(-1)^k a_{1i_1} a_{2i_2} \ldots a_{ni_n}, \tag{3.4}$$

где n — число строк (или столбцов) матрицы; $i_1, i_2, \ldots, i_n$ — некоторая перестановка чисел $1, 2, \ldots, n$; k — число транспозиций пар элементов, необходимых для того, чтобы расположить индексы в порядке $1, 2, \ldots, n$.

Каждый член этой суммы содержит по одному элементу из каждой строки и каждого столбца.

Процесс вычисления определителя начинается с выбора по одному элементу из каждой строки для образования различных комбинаций. Элементы каждого члена суммы выбираются из строк в естественном порядке $1, 2, \ldots, n$, однако каждая комбинация может содержать только по одному элементу из каждого столбца. Например, в матрице порядка 3×3 можно выбрать комбинацию $a_{12}a_{21}a_{33}$. Обратите внимание, что элементы располагаются в порядке возрастания их первых индексов — номеров строк. Каждая комбинация содержит ровно по одному элементу из каждой строки и каждого столбца. Требуется выбрать все возможные комбинации, составленные таким способом. Если матрица имеет порядок $n \times n$, то таких комбинаций будет $n!$

Так как порядок умножения чисел не влияет на результат, т. е. $a_{11}a_{22}a_{33} = a_{22}a_{11}a_{33} = a_{33}a_{22}a_{11}$ и т. д., то можно произвольно изменить порядок элементов в отобранных комбинациях. Попытаемся выбрать элементы так, чтобы их вторые индексы, или номера столбцов, расположились в порядке возрастания. Перестановку можно выполнить, меняя любые два соседних элемента местами. Выполняя операцию, нужно сосчитать, сколько сделано перестановок, необходимых для того, чтобы индексы расположились в нужном порядке. Если для этого потребовалось четное число перестановок (т. е. 0, 2, 4, 6, 8, ...), то произведение этих элементов берется с положительным знаком. Если было использовано нечетное число перестановок (т. е. 1, 3, 5, 7, ...), то произведение берется с отрицательным знаком.

В матрице порядка 2×2

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

можно найти только две комбинации элементов, содержащих

по одному элементу из каждой строки из каждого столбца, — $a_{11}a_{22}$ и $a_{12}a_{21}$. Вторые индексы в члене $a_{11}a_{22}$ расположены в естественном порядке и перестановок не требуют. Число перестановок равно нулю, поэтому знак произведения положительный. Однако элементы в членах a_{12} и a_{21} нужно поменять местами, чтобы их вторые индексы расположились в нужном порядке. Это требует одной перестановки, поэтому результат будет с отрицательным знаком. Итак, определитель матрицы порядка 2×2 равен

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = + a_{11}a_{22} - a_{12}a_{21}.$$

В качестве числового примера рассмотрим матрицу

$$\begin{bmatrix} 2 & 1 \\ 4 & 3 \end{bmatrix}$$

Ее определитель равен

$$\begin{vmatrix} 2 & 1 \\ 4 & 3 \end{vmatrix} = +(2\cdot 3) - (1\cdot 4) = 2.$$

Теперь рассмотрим более сложный пример — определитель матрицы порядка 3×3

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$$

Здесь будет 3!, или $3\cdot 2\cdot 1 = 6$, комбинаций, которые содержат по одному элементу из каждой строки и каждого столбца, индексы которых располагаются в порядке 1, 2, 3. Начиная с верхней, выбираем по одному элементу из каждой строки. Сначала из первой, затем из второй, третьей, ..., n-й, используя при этом не более одного элемента из каждого столбца. Укажем все возможные комбинации, удовлетворяющие этим условиям:

$$a_{11}a_{22}a_{33}, \quad a_{11}a_{23}a_{32},$$
$$a_{12}a_{23}a_{31}, \quad a_{12}a_{21}a_{33},$$
$$a_{13}a_{21}a_{32}, \quad a_{13}a_{22}a_{31}.$$

Для определения знака каждого члена нам нужно выяснить, сколько перестановок необходимо выполнить, чтобы вторые индексы расположились в порядке 1, 2, 3. Для члена $a_{11}a_{22}a_{33}$

перестановок не требуется. Перестановки для других членов приведены ниже:

$a_{11}a_{23}a_{32} = a_{11}a_{32}a_{23}$ $\quad k=1$, знак —
$a_{12}a_{23}a_{31} = a_{12}a_{31}a_{23} = a_{31}a_{12}a_{23}$ $\quad k=2$, знак +
$a_{12}a_{21}a_{33} = a_{21}a_{12}a_{33}$ $\quad k=1$, знак —
$a_{13}a_{21}a_{32} = a_{21}a_{13}a_{32} = a_{21}a_{32}a_{13}$ $\quad k=2$, знак +
$a_{13}a_{22}a_{31} = a_{13}a_{31}a_{22} = a_{31}a_{13}a_{22} = a_{31}a_{22}a_{13}$ $\quad k=3$, знак —

Итак, в определителе три положительных и три отрицательных члена. Складывая члены (с соответствующими знаками), получим

$$+ a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} + a_{12}a_{23}a_{31} -$$
$$- a_{12}a_{21}a_{33} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31}.$$

Теперь рассмотрим матрицу действительных чисел

$$\begin{bmatrix} 4 & 3 & 2 \\ 2 & 4 & 1 \\ 1 & 0 & 3 \end{bmatrix}$$

Шесть членов выглядят так:

$$(4\cdot 4\cdot 3) = 48,$$
$$(4\cdot 1\cdot 0) = 0,$$
$$(3\cdot 1\cdot 1) = 3,$$
$$(3\cdot 2\cdot 3) = 18,$$
$$(2\cdot 2\cdot 0) = 0,$$
$$(2\cdot 4\cdot 1) = 8.$$

Первый, третий и пятый из этих членов требуют четного числа операций, для того чтобы их вторые индексы расположились в порядке возрастания. Для того чтобы индексы остальных элементов расположились в порядке возрастания, требуется нечетное число операций, и поэтому они отрицательны. Складывая, получаем:

$$|A| = 48 - 0 + 3 - 18 + 0 - 8 = 25.$$

Этот метод вычисления определителя описан Петтофреццо [5]. Более общепринятый подход, изложенный в монографии Гири и Уивера [2], основан на так называемом методе алгебраических дополнений, но по существу мало отличается от изложенного.

Теперь мы уже умеем вычислять определитель квадратной матрицы, но еще неясно, что это такое. Определители возникают в различных задачах, но чаще всего они неявно участвуют в решении систем уравнений. Читатель мог не заметить их при

рассмотрении этого вопроса, но скрыто определители использовались в процессе обращения матрицы. Решим систему двух уравнений:

$$a_{11}a_1 + a_{12}a_2 = a_1,$$
$$a_{12}a_{11} + a_{22}a_2 = a_2. \tag{3.5а}$$

Представим эту систему в матричной форме:

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}. \tag{3.5б}$$

Мы уже выяснили, что вектор-столбец неизвестных x_1 и x_2 может быть найден с помощью обращения матрицы. Однако мы можем найти неизвестные с помощью обычных алгебраических преобразований. Получим

$$x_1 = \frac{b_1a_{22} - a_{12}b_2}{a_{11}a_{22} - a_{12}a_{21}}, \tag{3.6}$$

$$x_2 = \frac{a_{11}b_2 - b_1a_{21}}{a_{11}a_{22} - a_{12}a_{21}}. \tag{3.7}$$

Нетрудно заметить, что знаменатели у этих выражений одинаковы. Это определитель матрицы $[A]$, т. е.

$$[A] = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{22}a_{21}. \tag{3.8}$$

Далее числители тоже можно представить как определители. Числитель выражения для x_1 можно представить как определитель матрицы

$$| B \cdot A_{i2} | = \begin{vmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{vmatrix} = b_1a_{22} - a_2b_{12}, \tag{3.9}$$

а числитель выражения для x_2 — как определитель матрицы

$$| A_{1i} \cdot B | = \begin{vmatrix} a_{11} & b_1 \\ a_{21} & b_2 \end{vmatrix} = a_{11}b_2 - a_{21}b_1. \tag{3.10}$$

Этот метод можно обобщить на любые системы уравнений. Такой способ решения системы уравнений называется правилом Крамера, которое гласит, что значение любой неизвестной x_i в системе совместных уравнений равно отношению двух определителей. Знаменатель — определитель матрицы коэффициентов (в нашем примере матрицы $[A]$). Числитель — определитель той же матрицы коэффициентов, у которой i-й столбец заменен столбцом правых частей уравнений (вектор-столбец b).

Проверим это правило на приведенном ранее примере:

$$\begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 38 \\ 110 \end{bmatrix}.$$

Знаменатели в обеих дробях равны определителю

$$\begin{vmatrix} 4 & 10 \\ 10 & 30 \end{vmatrix} = (4 \cdot 30) - (10 \cdot 10) = 20.$$

Числитель выражения для x_1 равен определителю

$$\begin{vmatrix} 38 & 10 \\ 110 & 30 \end{vmatrix} = (38 \cdot 30) - (110 \cdot 10) = 40.$$

Числитель выражения для x_2 равен

$$\begin{vmatrix} 4 & 38 \\ 10 & 110 \end{vmatrix} = (4 \cdot 110) - (10 \cdot 38) = 60.$$

Итак, $x_1 = 40/20 = 2$, а $x_2 = 60/20 = 3$. Ранее с помощью обращения матриц мы нашли точно такие значения неизвестных.

СОБСТВЕННЫЕ ЗНАЧЕНИЯ И СОБСТВЕННЫЕ ВЕКТОРЫ

Тема, к которой мы переходим, — собственные значения и собственные векторы — наиболее трудна в матричной алгебре. Трудность состоит не в их вычислении, которое является не более громоздким, чем другие математические процедуры. Скорее всего, трудность заключается в интуитивном понимании этих величин. К несчастью, в большинстве руководств эта тема излагается в строгих математических терминах, не всегда понятных нематематику.

Очень хорошее изложение понятий о собственных векторах и собственных значениях, сопровождающееся геометрической интерпретацией, было подготовлено Гоулдом для студентов географического факультета Пенсильванского университета. Приведенное здесь изложение темы основывается на этой работе, а также на статье Гоулда [3]. Мы будем рассматривать матрицу, составленную из координат точек в некотором пространстве, и интерпретировать свойства собственных значений и ассоциированных с ними собственных функций как геометрические свойства расположения точек. Такой подход требует рассмотрения матриц низких порядков. Однако интуитивные представления, полученные для этого случая, можно экстраполировать на матрицы больших порядков, даже на такие, вычисления для которых практически невыполнимы при ручном счете. Следует заметить, что мы вступаем в такую область, в которой очень часто даже наиболее мощные ЭВМ оказываются не в состоянии

преодолеть вычислительные трудности. Так как мы уже познакомились с определителями, то можем их использовать для введения понятий собственных значений и собственных векторов. Пусть дана система уравнений, записанная в матричной форме:

$$[A]\cdot[X] = \lambda[X]. \tag{3.11}$$

В этом уравнении произведение матрицы коэффициентов (a_{ij}) на вектор-столбец неизвестных $[x]$ равно произведению константы λ на вектор-столбец неизвестных. Задача почти такая же, как и решение системы уравнений

$$[A]\cdot[X] = [B], \tag{3.12}$$

только теперь $[B]=\lambda[X]$.

Нужно найти значения λ, удовлетворяющие этому соотношению. Уравнение (3.11) можно записать в форме

$$([A]-\lambda[I])\cdot[X] = [0], \tag{3.13}$$

где $\lambda[I]$ — единичная матрица $[I]$, умноженная на число λ (порядок матрицы $[I]$ такой же, как и порядок матрицы $[A]$). Матрица $\lambda[I]$ порядка 3×3 имеет вид

$$\lambda[I] = \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix} \tag{3.14}$$

Используя принятую запись, получим следующую систему уравнений:

$$\begin{aligned} (a_{11}-\lambda)x_1 + a_{12}x_2 + a_{13}x_3 &= 0, \\ a_{12}x_1 + (a_{22}-\lambda)x_2 + a_{23}x_3 &= 0, \\ a_{31}x_1 + a_{32}x_2 + (a_{33}-\lambda)x_3 &= 0. \end{aligned} \tag{3.15}$$

Предположим, что у этой системы есть нетривиальные решения, т. е. что существуют $x_i\neq0$. По правилу Крамера для решения системы уравнений неизвестные находятся как отношения двух определителей. В связи с тем что в матрицах, определители которых являются числителями этих отношений, есть нулевой столбец, эти определители равны нулю. Таким образом,

$$[X] = \frac{[0]}{|A|}.$$

Записав это соотношение иначе, получим

$$|A|\cdot[X] = [0]. \tag{3.16}$$

Если $[X]$ — ненулевой вектор, то определитель матрицы $[A]$ равен нулю, т. е.

$$|A-\lambda I| = \begin{vmatrix} a_{11}-\lambda & a_{12} & a_{13} \\ a_{21} & a_{22}-\lambda & a_{23} \\ a_{31} & a_{32} & a_{32}-\lambda \end{vmatrix} = 0. \tag{3.17}$$

Обычно коэффициенты a_{ij} нам известны, и поэтому можно воспользоваться этим соотношением для нахождения значений, удовлетворяющих заданным условиям. Для этого представим определитель как полиномиальное уравнение. Рассмотрим сначала определитель матрицы порядка 2×2:

$$\begin{vmatrix} a_{11}-\lambda & a_{12} \\ a_{21} & a_{22}-\lambda \end{vmatrix} = 0.$$

Представим определитель в виде

$$(a_{11}-\lambda)(a_{22}-\lambda) - a_{21}a_{12} = 0.$$

После перемножения получим

$$(a_{11}-\lambda)(a_{22}-\lambda) = (a_{11}a_{22}) - (a_{11}\lambda) - (a_{22}\lambda) + \lambda^2,$$

а сложив эти два уравнения, имеем

$$(a_{11}a_{22}) - (a_{21}a_{12}) - (a_{11}\lambda) - (a_{22}\lambda) + \lambda^2 = 0.$$

Так как мы знаем значения a_{ij}, то последнее уравнение можно представить в виде

$$\lambda^2 + a_1\lambda + a_2 = 0, \tag{3.18}$$

где a_i — суммы соответствующих значений a_{ij}. Это обычное квадратное уравнение вида $ax^2+bx+c=0$, корни которого вычисляются по формуле

$$x = \frac{-b\pm\sqrt{b^2-4ac}}{2a}. \tag{3.19}$$

Если читатель незнаком с этой формулой, а также с разложением квадратного трехчлена на множители, он может воспользоваться элементарным учебником алгебры.

Вычислим теперь собственные значения матрицы порядка 2×2. Пусть

$$[A] = \begin{bmatrix} 17 & -6 \\ 45 & -16 \end{bmatrix}.$$

Сначала представим матрицу в форме

$$[A]-\lambda[I] = \begin{bmatrix} 17-\lambda & -6 \\ 45 & -16-\lambda \end{bmatrix}.$$

Приравнивая определитель нулю, получаем

$$\begin{vmatrix} 17-\lambda & -6 \\ 45 & -16-\lambda \end{vmatrix} = 0.$$

Перемножение дает: $-272-17\lambda+16\lambda+\lambda^2+270=0$, или $\lambda^2-\lambda-2=0$, или после разложения левой части уравнения на множители $(\lambda-2)(\lambda+1)=0$, что определяет два собственных значения $\lambda_1=+2$, $\lambda_2=-1$.

Этот пример был выбран нами для упрощения вычислений. Теперь же рассмотрим более сложный пример, в котором используется система уравнений, которую мы неоднократно рассматривали выше. Пусть дана матрица порядка 2×2:

$$[A]=\begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix}.$$

Повторяем все проделанные в предыдущем примере действия:

$$\begin{vmatrix} 4-\lambda & 10 \\ 10 & 30-\lambda \end{vmatrix} = 0.$$

Далее

$$\begin{vmatrix} 4-\lambda & 10 \\ 10 & 30-\lambda \end{vmatrix} = (4-\lambda)(30-\lambda)-100=0$$

или

$$\lambda^2-34\lambda+20=0.$$

Используя формулу для определения корней квадратного уравнения, получаем

$$\lambda=\frac{-b\pm\sqrt{b^2-4ac}}{2a}=\frac{-(-34)\pm\sqrt{34^2-4\cdot1\cdot20}}{2\cdot1}=\frac{34\pm\sqrt{1076}}{2},$$

$$\lambda_1=33{,}4, \qquad \lambda_2=0{,}6.$$

Мы можем проверить эти значения, подставив их в определитель. Учитывая ошибки округления, получаем

$$\begin{vmatrix} 4-33{,}4 & 10 \\ 10 & 30-33{,}4 \end{vmatrix} = (-29{,}4)(-3{,}4)-(10)(10)=-0{,}04$$

и

$$\begin{vmatrix} 4-0{,}6 & 10 \\ 10 & 30-0{,}6 \end{vmatrix} = (3{,}4)(29{,}4)-(10)(10)=-0{,}04.$$

Таким образом, собственные значения найдены верно с точностью до двух десятичных знаков.

Прежде чем покончить с вычислением собственных значений матриц порядка 2×2, рассмотрим еще один пример, чтобы показать, какие могут возникнуть дополнительные осложнения. Попытаемся найти собственные значения следующей матрицы:

$$[A]=\begin{bmatrix} 2 & 4 \\ -6 & 3 \end{bmatrix}$$

Приравняем нулю соответствующий определитель

$$\begin{vmatrix} 2-\lambda & 4 \\ -6 & 3-\lambda \end{vmatrix} = 0,$$

вычисляя который, получим

$$\begin{vmatrix} 2-\lambda & 4 \\ -6 & 3-\lambda \end{vmatrix} = (2-\lambda)(3-\lambda)+24=0$$

или

$\lambda^2-5\lambda+30=0$.

Корни этого уравнения вычисляются по формуле

$$\lambda_{1,2}=\frac{5\pm\sqrt{25-120}}{2}.$$

Мы видим, что нахождение корней этого уравнения требует извлечения квадратных корней из отрицательных чисел, тогда

$$\lambda_1=\frac{5+\sqrt{-95}}{2}=2{,}5+4{,}9i,$$

$$\lambda_2=\frac{5-\sqrt{-95}}{2}=2{,}5-4{,}9i,$$

где λ_1, λ_2 — комплексные числа, содержащие как действительную, так и мнимую части, включающие число $i=\sqrt{-1}$. К счастью, симметричные матрицы всегда имеют действительные собственные значения, и в большинстве наших вычислений, связанных с собственными векторами и собственными значениями, используются ковариационные, корреляционные или аналогичные матрицы, которые всегда симметричны.

Теперь рассмотрим процедуру вычисления собственных значений для матрицы порядка 3×3:

$$\begin{bmatrix} 20 & -4 & 8 \\ -40 & 8 & -20 \\ -60 & 12 & -26 \end{bmatrix}$$

Приравниваем нулю соответствующий определитель

$$\begin{vmatrix} 20-\lambda & -4 & 8 \\ -40 & 8-\lambda & -20 \\ -60 & 12 & -26-\lambda \end{vmatrix} = 0,$$

после чего получим $-\lambda^3+2\lambda^2+8\lambda=0$.

Это — кубическое уравнение, имеющее три корня. Разложив его на множители, можем записать $(\lambda-4)(\lambda-0)(\lambda+2)=0$ и получим корни $\lambda_1=4$, $\lambda_2=0$, $\lambda_3=-2$.

Хотя изложенные приемы применимы к матрице любого порядка, нахождение корней полиномов высоких степеней является нелегкой задачей. Обычно собственные значения находятся не с помощью решения полиномиальных уравнений, а с помощью методов матричных преобразований, сущность которых состоит в последовательных приближениях к собственным значениям. Эти методы оказались применимыми только благодаря использованию быстродействующих ЭВМ, позволяющих за весьма короткий промежуток времени найти приближенное решение, а затем уточнить его в течение нескольких минут.

Теперь, когда мы получили представление о процедуре вычисления собственных значений, можно попытаться понять и их сущность. Матрицу можно рассматривать как набор значений координат точек в n-мерном пространстве. Матрица порядка 2×2 соответствует плоской поверхности, такой же, как лист бумаги. Таким образом, матричные операции можно трактовать геометрически. Так, матрица

$$[A]=\begin{bmatrix}4 & 8\\ 8 & 4\end{bmatrix}$$

определяет две точки плоскости с координатами (4,8) и (8,4). Эти точки соответствуют двум векторам на координатной плоскости, выходящим из начала координат, как это показано на рис. 3.1. Отметим, что в качестве координат точек можно использовать и столбцы матрицы, что не внесет существенных изменений в наши рассуждения. Для определенности будем действовать со строками.

Представим себе, что две точки лежат на эллипсе с центром в начале координат. Тогда эллипс будет представлять кривую, проходящую через эти точки. В результате собственные значения матрицы окажутся равными длинам большой и малой полуосей этого эллипса. В нашем примере собственные значения равны $\lambda_1=12$, $\lambda_2=-4$.

Необходимо отметить, что отношения полуосей можно использовать как меру вытянутости эллипса, что графически показано на рис. 3.2. Первое собственное значение характеризует большую полуось, длина которой от центра до самой удаленной от него точки кривой равна 12 единицам. Второе собственное значение характеризует малую полуось, находящуюся во второй четверти координатной плоскости. Ее длина равна 4 единицам и берется со знаком минус. Если бы две данные точки были ближе друг к другу, то отношение полуосей изменилось

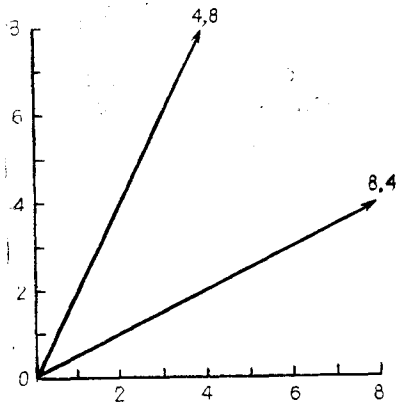


Рис. 3.1. Два вектора, определенные элементами порядка матрицы 2×2

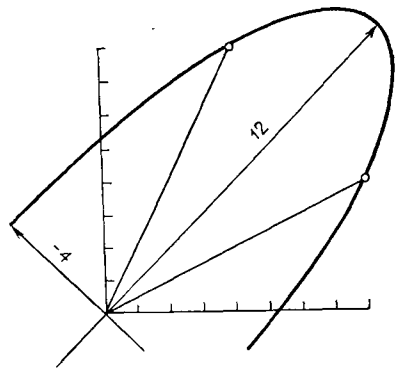


Рис. 3.2. Эллипс, определенный элементами матрицы порядка 2×2. Собственные векторы матрицы соответствуют главным полуосям эллипса

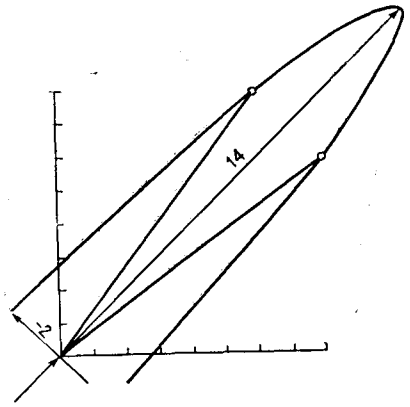


Рис. 3.3. Удлиненный эллипс, соответствующий матрице, элементы которой задают точки с расстояниями между ними, меньшими, чем на рис. 3.2. Собственные векторы являются главными полуосями эллипса

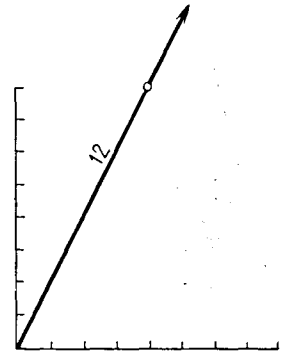


Рис. 3.4. Матрица, содержащая равные строки, приводит к вырожденному эллипсу — прямой

бы. Например, если выбраны точки с координатами, образующими матрицу

$$[A]=\begin{bmatrix}6 & 8\\ 8 & 6\end{bmatrix},$$

то собственными значениями матрицы будут числа $\lambda_1=14$, $\lambda_2=-2$.

Это показано графически на рис. 3.3. Большая полуось данного эллипса намного больше малой. Если две выбранные точки совпадают, т. е. две строки матрицы одинаковы, то второе собственное значение становится равным нулю, и рассматриваемый эллипс вырождается в прямую. Этому случаю отвечает, например, матрица

$$[A] = \begin{bmatrix} 4 & 8 \\ 4 & 8 \end{bmatrix},$$

характеристическое уравнение которой $\lambda^2 - 12\lambda + 0 = 0$ дает собственные значения $\lambda_1 = 12$, $\lambda_2 = 0$.

Такая ситуация изображена графически на рис. 3.4. Две выбранные точки совпадают, через них проходит большая полуось, а перпендикулярная к ней малая ось равна нулю.

Противоположный крайний случай описывается матрицей, соответствующей двум перпендикулярным векторам равной длины. Например, матрица

$$[A] = \begin{bmatrix} -4 & 8 \\ 8 & 4 \end{bmatrix}$$

имеет характеристическое уравнение $\lambda^2 + 0\lambda - 80 = 0$, корни которого $\lambda_1 = +8{,}95$, $\lambda_2 = -8{,}95$.

Этот случай графически представлен на рис. 3.5. Данные векторы являются радиусами окружности, в которую превратился эллипс. Их длина равна собственным значениям.

Заметим одно важное свойство собственных значений, которое читатель может проверить на приведенных ранее примерах. Сумма собственных значений всегда равна сумме диагональных элементов матрицы, которая называется следом матрицы. Это свойство полезно использовать для проверки правильности вы-

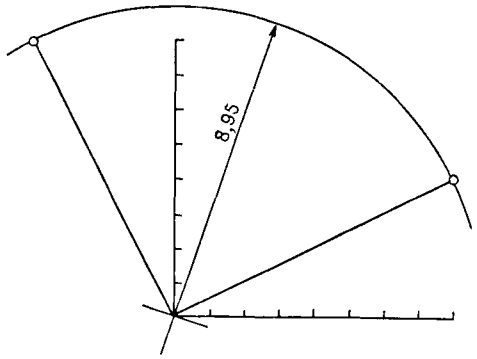


Рис. 3.5. Матрица, элементы которой задают перпендикулярные векторы, определяет эллипс с равными полуосями, т. е. окружность

числения собственных значений на ЭВМ, а также для контроля вычислений при использовании метода главных компонент.

Вспомним, что мы определили собственные значения как величины, входящие в матрицу системы уравнений (3.13). Теперь, когда эти величины найдены, можно вернуться к системам уравнений и вычислить вектор неизвестных x. Для матрицы порядка 2×2 первое собственное значение получим из уравнения

$$\begin{bmatrix} a_{11} - \lambda_1 & a_{12} \\ a_{21} & a_{22} - \lambda_1 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad (3.20)$$

Для второго собственного значения матричное уравнение составляется аналогично. После вычисления диагональных элементов можно найти неизвестный вектор с помощью обращения матрицы. Этот вектор называется собственным вектором (а также характеристическим или главным вектором). Каждому собственному вектору соответствует собственное значение матрицы. Сколько у матрицы собственных значений, или строк (столбцов), столько у этой матрицы и собственных векторов.

Таким образом, чтобы вычислить собственные векторы и собственные значения матрицы порядка $n\times n$, нужно найти n корней ее характеристического уравнения и решить n систем из n совместных уравнений! К счастью, рассматриваемые нами примеры матриц порядка 2×2 не требуют громоздких вычислений.

Начнем с рассмотрения матрицы

$$[A] = \begin{bmatrix} 17 & -6 \\ 45 & -16 \end{bmatrix}$$

Вспомним, что мы уже вычислили собственные значения этой матрицы: $\lambda_1 = +2$, $\lambda_2 = -1$.

Подстановка первого собственного значения в матрицу дает

$$\begin{bmatrix} 17-2 & -6 \\ 45 & -16-2 \end{bmatrix} = \begin{bmatrix} 15 & -6 \\ 45 & -18 \end{bmatrix}$$

Соответствующая система уравнений имеет вид

$$15x_1 - 6x_2 = 0,$$
$$45x_1 - 18x_2 = 0$$

или в матричной форме

$$\begin{bmatrix} 15 & -6 \\ 45 & -18 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Нетрудно видеть, что второе уравнение можно получить из первого умножением на 3, поэтому достаточно решить одно из них, чтобы полученные решения удовлетворяли и другому уравне-

нию. Простой подбор дает следующие решения: $x_1=2$, $x_2=5$. Это координаты первого собственного вектора, соответствующего первому собственному значению. В действительности имеется бесконечное множество решений, так как решением является любой вектор вида

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \beta \begin{bmatrix} 2 \\ 5 \end{bmatrix},$$

где β — любое число. Мы же ограничились только одним значением $\beta=1$. Позже увидим, что интерес представляют только отношения координат вектора, а они при умножении вектора на любое число не изменяются.

Матрица системы уравнений, отвечающая второму собственному вектору, имеет вид

$$\begin{bmatrix} 17-(-1) & -6 \\ 45 & -16-(-1) \end{bmatrix} = \begin{bmatrix} 18 & -6 \\ 45 & -15 \end{bmatrix}$$

Оба уравнения приводятся к одному уравнению: $3x_1-x_2=0$, решением которого является вектор

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \beta \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

Хотя вычисления при увеличении порядка значительно усложняются, рассмотренные методы можно применять к матрицам порядка $n \times n$. Прежде чем перейти к вычислительным аспектам проблемы нахождения собственных значений, рассмотрим некоторые матрицы с целью проведения геометрического анализа этих величин. Возможно, что это позволит понятнее объяснить геометрический смысл собственных значений.

Сначала рассмотрим матрицу

$$[A] = \begin{bmatrix} 4 & 8 \\ 8 & 4 \end{bmatrix}$$

с собственными значениями $\lambda_1=12$, $\lambda_2=-4$.

Подставляя первое из них в исходную матрицу, получим

$$\begin{bmatrix} 4-12 & 8 \\ 8 & 4-12 \end{bmatrix} = \begin{bmatrix} -8 & 8 \\ 8 & -8 \end{bmatrix}$$

Этой матрице соответствует собственный вектор

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Возвращаясь снова к рис. 3.2, убеждаемся, что собственный вектор можно интерпретировать как характеристику угла наклона большой полуоси эллипса. Если рассматривать элементы собственного вектора как координаты точки на плоскости, то первый собственный вектор определяет полуось, являющуюся биссектрисой угла между двумя заданными строками матрицы векторами. Длина полуоси равна первому собственному значению. Подставляя в матрицу второе собственное значение, получим

$$\begin{bmatrix} 4-(-4) & 8 \\ 8 & 4-(-4) \end{bmatrix} = \begin{bmatrix} 8 & 8 \\ 8 & 8 \end{bmatrix}$$

Решением будет второй собственный вектор

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix},$$

который имеет тангенс угла наклона, равный $-1/1=135°$, следовательно, перпендикулярен к главной полуоси эллипса и определяет меньшую полуось (см. рис. 3.2).

Теперь произведем расчеты для второй матрицы, строки которой определяют точки, расположенные ближе друг к другу, чем в предыдущем примере:

$$[A] = \begin{bmatrix} 6 & 8 \\ 8 & 6 \end{bmatrix}$$

Для первого собственного вектора получим

$$\begin{bmatrix} (6-14) & 8 \\ 8 & (6-14) \end{bmatrix} = \begin{bmatrix} -8 & 8 \\ 8 & -8 \end{bmatrix},$$

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Таким образом, угол наклона первого собственного вектора составляет 45°; вектор делит пополам угол, образованный двумя заданными векторами. Длина большой полуоси равна 14, т. е. первому собственному значению. Как и следовало ожидать, этот вектор параллелен вектору, найденному в предыдущем примере, но имеет бóльшую длину. Аналогично можно найти второй собственный вектор, соответствующий второму собственному значению:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

Этот метод применим и к матрицам большего порядка, хотя выкладки становятся более сложными. Для примера рассмот-

рим матрицу

$$[A] = \begin{bmatrix} 5 & 2 & 6 \\ 2 & 4 & 3 \\ 6 & 3 & 2 \end{bmatrix}$$

На рис. 3.6 наглядно изображены три вектора, заданные строками матрицы. Прежде чем начать анализ, остановимся на некоторых известных свойствах собственных значений и собственных векторов. Так как матрица симметрична, все три собственные значения — действительные числа. Первый собственный вектор, соответствующий наибольшему собственному значению, проходит в пространственном углу, в котором лежат три заданных вектора. Сумма собственных значений равна следу матрицы, т. е. 11. Собственные значения и собственные векторы этой матрицы следующие:

	$\lambda_1=11{,}3$	$\lambda_2=-2{,}9$	$\lambda_3=2{,}6$
	Вектор 1	Вектор 2	Вектор 3
λ_1	0,69	—0,56	—0,45
λ_2	0,43	—0,19	0,88
λ_3	0,58	0,81	—0,11

Эти векторы изображены на рис. 3.7. Все указанные выше условия выполняются. Правомерное распространение результа-

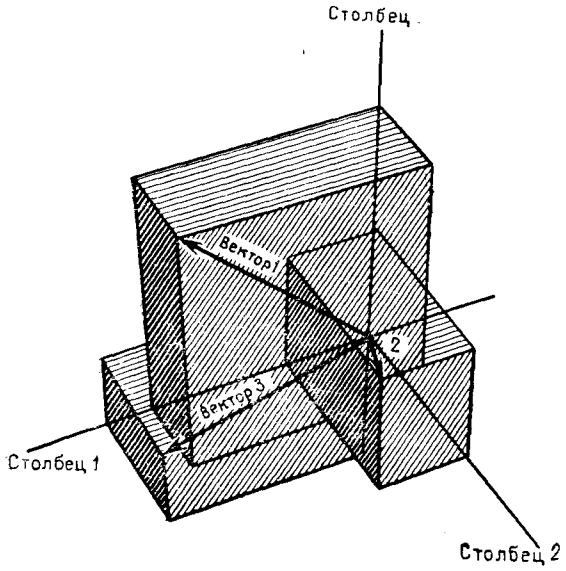


Рис. 3.6. Векторы в трехмерном пространстве, определяемые элементами матрицы порядка 3×3

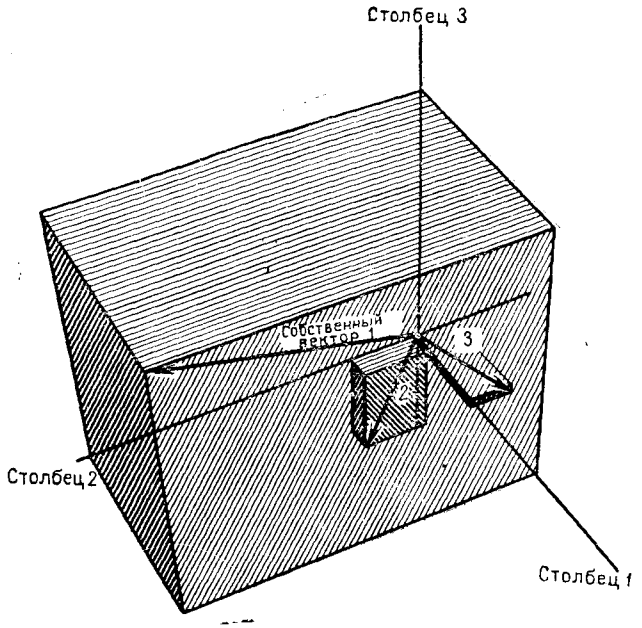


Рис. 3.7. Собственные векторы матрицы порядка 3×3 расположены в том же пространстве, что и векторы рис. 3.6.

Отметим, что первый собственный вектор проходит внутри угла, образованного тремя данными векторами

тов этого анализа на матрицы больших порядков математически очевидно, но наши возможности наглядного представления уже исчерпаны. Однако читатель может себе представить совокупность точек в пространстве более высокой размерности, а также существование в нем требуемого числа взаимно перпендикулярных направлений.

Здесь мы отметим только одно важное свойство собственных векторов матриц общего вида, не являющихся симметричными.

Если вы тщательно изучили эту главу и детально проработали примеры (а также обдумали возможность комплексного применения изложенных методов в более сложных задачах), то уже достаточно подготовлены к тому, чтобы перейти к изучению современных вычислительных методов, применяемых в геологических исследованиях. Мы попытались изложить в упрощенной форме основы матричной алгебры. Как отмечалось в третьей главе, статистика слишком сложная наука, чтобы ее можно было изложить в одной главе или даже в одной книге. Матричная алгебра тоже достаточно сложна, и ее нельзя хорошо изложить на нескольких страницах. Однако нам кажется,

что читатель получил некоторое представление о методах матричной алгебры, что позволит ему без труда усвоить основы вычислительных методов, которые будут изложены далее.

СПИСОК ЛИТЕРАТУРЫ

1. *Davis P. J.*, The mathematics of matrices, 2nd ed. John Wiley and Sons, Inc., New York, 1973, 348 p.
Наиболее распространенный учебник по теории матриц с минимальным количеством терминов и максимумом примеров и приложений.

2. *Gere J. M.* and *Weaver W., Jr.* Matrix algebra for engineers, 2nd ed.: Brooks-Cole Publ. Co., Monterey, Ca., 1982, 175 p.
Эта скромно изданная книга является одним из наилучших руководств по матричной алгебре.

3. *Could P.* On the geographic interpretation of eigenvalies: an initial exploration: Trans. Inst. British Geographers, no. 42, 1967, 53—86 p.
Интуитивный взгляд на теорию собственных значений и векторов с использованием геометрических аналогий. Материал этой отличной книги, предназначенной для студентов, частично использован в гл. 3 настоящего издания.

4. *Maron M. J.*, Numerical analysis—a practical approach: Macmillan Publ. Co., Inc., New York, 1982, 471 p.
Содержит процедуры и алгоритмы матричных операций, в особенности методы обращения матриц, решения систем совместных уравнений и нахождения собственных значений.

5. *Pettofrezzo A. J.* Matrices and transformations: Dover, Inc., New York, 1978, 133 p.
В этом скромном переиздании известного учебника содержится традиционный материал по односеместровому курсу матричной алгебры. Приводятся примеры и проблемы.

6. *Reiner I.* Introduction to matrix theory and linear algebra: Holt, Rinehart, and Winston, Inc., New York, 1971, 154 p.
Вводный курс с подробными объяснениями методов.

7. *Robinson E. A.* and *Treitel S.*, Geophysical signal analysis: Prentice-Hall, Inc., Englewood Cliffs, N. J., 1980, 466 p.
Гл. 3 содержит краткое, но ясное изложение вопросов, связанных с операцией свертки.

Глава 4
АНАЛИЗ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ДАННЫХ

ПОСЛЕДОВАТЕЛЬНОСТИ ГЕОЛОГИЧЕСКИХ ДАННЫХ

В этой главе мы рассмотрим методы исследования данных, которые характеризуются своим положением на прямой. Для таких данных существенную роль играет место, занимаемое некоторым определенным значением в данной последовательности. Множество данных такого типа часто встречается в геологии. Например, они могут представлять собой последовательность значений литологических признаков, геохимических и минералогических характеристик проб, взятых по разрезу или в буровой скважине, значений электрического каротажа нефтяных скважин, зарегистрированных приборами. К этой общей категории можно отнести также данные, изменяющиеся с течением времени, например измерения стока воды в реке, характеристики добычи газа из скважины. Методы изучения последовательностей одномерных данных можно отнести к анализу временных рядов, несмотря на то что последовательности могут характеризовать как временные, так и пространственные зависимости.

Прежде чем переходить к изложению методов исследования таких последовательностей и рассмотрению примеров из геологии, остановимся на различных типах последовательностей данных, с которыми приходится встречаться геологу. Это может быть последовательность абсолютно точных измерений как значений изучаемой переменной, так и значений шкалы, вдоль которой расположены результаты измерений. В качестве примера можно рассмотреть значения каротажной кривой для буровой скважины и изменение продуктивности скважины во времени. В первом примере переменной является признак, измеряемый в омах, а единицами шкалы измерений являются литры. Во втором примере переменная — тоже признак, измеряемый в литрах нефти, а единицами шкалы измерений являются дни, месяцы или годы. При любой форме записи существенны два момента. Во-первых, измеряемая переменная выражается в единицах интервальной шкалы или шкалы отношений; 1000 л нефти в два раза больше, чем 500 л, а сопротивление в 10 Ом в десять раз превышает сопротивление в 1 Ом. Во-вторых, интервалы, вдоль которых располагаются данные, тоже имеют определенную величину. Глубина скважины в 900 м в десять раз превышает глубину в 90 м, а десятилетие между 1940 и 1950 г. имеет ту же продолжительность, что и десятилетие меж-

ду 1950 и 1960 г. На этих замечаниях ввиду их тривиальности не стоило бы останавливаться, но, как мы увидим в дальнейшем, далеко не все геологические последовательности обладают такими свойствами.

В качестве примера рассмотрим последовательность стратиграфических данных, образованную литологическими разновидностями горных пород, слагающих осадочную толщу. Такой последовательностью можно считать серию (снизу вверх) известняк — глинистый сланец — известняк — глинистый сланец — песчаник — уголь — глинистый сланец — известняк. Мы хотим как-то осмысленно описать эту последовательность, однако не можем выбрать для нее шкалу. Очевидно, что это изменение литологических признаков происходит в течение определенного времени, но у нас нет никакого способа выбора соответствующей временной шкалы. Мы могли бы использовать мощность, но она может очень сильно изменяться от места к месту, даже если последовательность пород остается неизменной. Таким образом, использование мощности пород вряд ли поможет нам в наших исследованиях. Тот факт, что известняк в разрезе стоит на третьем месте, а уголь — на шестом, не имеет того значения, которое можно было бы выразить числом (то, что номер 6 вдвое больше номера 3, для нас не имеет смысла). Аналогично литологический состав слоев не может быть выражен на числовой шкале. Можно только закодировать приведенную последовательность, например, таким образом: 1—2—1—2—3—4—2—1, где известняк обозначается цифрой 1, глинистый сланец — 2, песчаник — 3, уголь — 4. Но такая условность совершенно произвольна и не выражает никаких соотношений между разновидностями пород. Очевидно, что эта последовательность ставит перед исследователями большее число различных проблем, чем это было в ранее рассмотренных примерах.

Имеются, однако, другие возможности. Пусть, например, нас интересует некоторая измеряемая характеристика, входящая в последовательность. Предположим, что мы установили значения содержаний бора в каждой литологической разновидности рассматриваемой серии. Мы можем использовать шкалу расстояний между пробами в метрах и считать, что решаем задачу, связанную с изучением глубины или расстояния. Кроме того, мы можем рассматривать зависимость содержания бора от места, занимаемого этим значением в последовательности.

Близко связана с предыдущими задача анализа последовательности, которую можно охарактеризовать присутствием или отсутствием некоторой переменной или переменных в некоторых ее местах. Нас может интересовать, например, повторное появление зависящей от фаций микрофауны в образцах пород, отобранных при бурении скважины. Другой класс задач — это установление последовательности минеральных зерен, наблюдаемых на пересечении шлифа. В этом случае мы можем использовать миллиметровую шкалу, но она не позволяет оценить, что чаще встречается — оливин или плагиоклаз.

Данные, которые можно охарактеризовать непрерывным расположением в пространстве или времени, часто называют рядами, или последовательностями, или цепями. Природа рассматриваемых данных предопределяет те задачи, которые для них можно поставить. Ясно, что мы не можем извлечь информацию о временны́х интервалах из последовательности стратиграфических данных, так как временна́я шкала, соответствующая этой последовательности, неизвестна. В стратиграфических задачах часто используют вместо временно́й шкалы пространственную, но при этом наши заключения не лучше, чем взятое в их основу предположение о том, что время, требуемое для образования осадка определенной мощности, может быть измерено.

В табл. 4.1 приведена классификация различных методов анализа данных, которые будут рассмотрены в этой главе. При этом можно выделить три типа рядов. В первом из них расстояние между наблюдениями изменяется и потому должно быть охарактеризовано в каждой точке. Во втором предполагается,

Таблица 4.1

Классификация рассматриваемых в этой главе методов по типу переменных и их расположению на линии

Тип переменных	Наблюдения, нерегулярно расположенные в пространстве	Наблюдения, равномерно и регулярно расположенные в пространстве	Пространственное размещение не рассматривается
Переменные, измеренные в интервальной шкале или шкале отношений	Интерполяция Полиномиальная регрессия Сплайны	Ортогональная полиномиальная регрессия Скользящие средние Фильтрация и сглаживание Зондирование Автокорреляция и перекрестная корреляция Полувариограммы Спектральный анализ	Автокорреляция и перекрестная корреляция
Переменные, измеренные в номинальной или порядковой шкалах	Ряды событий К—С критерии	Автоассоциация и перекрестная ассоциация Анализ подстановок Цепи Маркова Критерии скачков	Автоассоциация и перекрестная ассоциация Анализ взаимозаменяемости Цепи Маркова Критерии скачков

что точки расположены в пространстве регулярно и равномерно и, кроме единственной постоянной, никакие числовые характеристики пространственного расположения данных не участвуют в анализе. Наконец, в третьем существенна лишь последовательность наблюдений, а их пространственное расположение не имеет значения.

Эти методы можно классифицировать также и по типам необходимых наблюдений. В одних случаях требуется знать интервал между наблюдениями или их отношение: переменная должна быть измерена по некоторой шкале и выражена в вещественных числах. В других используются номинальные или порядковые данные, и наблюдения требуется лишь некоторым образом расклассифицировать. В методах, рассматриваемых в этой главе, классы не ранжированы, т. е. состояние A нельзя считать в некотором смысле более широким или объемлющим, чем состояние B или C. Номинальные данные обычно представляют целыми числами, буквами или символами.

В этой главе мы рассмотрим математические методы, используемые при анализе последовательностей данных. Однако рассматриваемые здесь методы не исчерпывают все существующие возможности. Скорее их можно охарактеризовать как особенно перспективные при количественных исследованиях в науках о Земле. Другие методы могут оказаться более подходящими или более мощными в специфических ситуациях или для некоторых особых последовательностей данных. На наш взгляд, знакомство с описанными ниже методами — хорошее введение в обширную область аналитического исследования геологических данных. Однако многие из этих методов были разработаны специалистами в областях науки, далеких от геологии, и их описание, приспособленное к использованию в инженерном деле, в биологии клетки, к анализу рыночных отношений или речевой терапии, трудно приспособить для решения геологических задач. Некоторые из этих методов оперируют с непараметрическими статистиками, почти не рассматриваемыми во вводных статистических курсах. Так, как большинство геологов не знакомо с основами анализа последовательностей данных, то мы полагаем, что им будет полезно прочесть приведенный здесь обзор разнообразных методов и подходов. Как видно из табл. 4.1, эти методы охватывают последовательности различных типов и предназначены для ответа на ряд вопросов. Ни один из методов не излагается здесь исчерпывающе, однако рассмотренные здесь примеры и приложения могут помочь геологу выбрать наиболее подходящий метод для решения задачи. Список литературы поможет найти детальное изложение специальных вопросов.

Рассмотренные методы дают возможность получить ответы на ряд вопросов, а именно: можно ли считать наблюдения случайными или в них обнаруживается некоторый тренд; если тренд существует, то какова его форма; можно ли обнаружить и измерить циклы и повторения; позволяют ли данные сделать оценки и предсказания; можно ли оценить зависимости между переменными и указать их силу? Хотя вопросы такого рода и не явно ставятся в последующем изложении, читателю рекомендуется продумать сущность каждого метода, а также его возможности при решении задач различного типа. Заметим, что выбранные нами задачи могут помочь в решении многих других.

Геологам приходится иметь дело не только с анализом последовательностей данных, но и сравнивать между собой различные ряды наблюдений. Наглядный пример — стратиграфическая корреляция при изучении разрезов или при электрическом каротаже скважин. Причина, по которой геологи используют корреляцию, — простое желание ускорить получение геологических выводов из закодированной информации, хранящейся в банках данных. Кроме того, геологи сталкиваются с задачами корреляции в тех случаях, когда не могут своими силами решить вопрос об эквивалентности двух рядов наблюдений. Слабое сходство, слишком незначительное, чтобы его можно было обнаружить визуально, может быть выявлено этими методами, даже если это невозможно при использовании других приемов. Количественные методы позволяют геологам рассматривать одновременно несколько переменных, что является мощным средством распознавания изучаемых объектов. Наконец, в силу абсолютной инвариантности операций в вычислительной программе корреляционный анализ бросает вызов геологу. Если корреляционные зависимости, очевидные геологу, не согласуются с результатами, полученными машиной, геолог обязан определить причину этого расхождения. Обычно более тщательное исследование позволяет выявить осложнения и смещения, не замеченные при первоначальном исследовании. Это не означает, что геолог должен изменить свою интерпретацию таким образом, чтобы она согласовывалась с результатами, полученными с помощью вычислительной машины. Совсем наоборот, имеющиеся в настоящее время в нашем распоряжении программы автоматической корреляции довольно грубы и составлены просто в соответствии с ходом мыслей, используемых геологами. Однако по мере продолжения исследований по корреляции можно ожидать, что будут созданы весьма полезные алгоритмы, которые позволят значительно облегчить работу геолога.

Бо́льшую часть методов сравнения двух или более последовательностей можно разбить на две большие категории. В первой из них пары данных могут занимать только одно положение, и наша задача — определить степень сходства между этими двумя последовательностями. В качестве примера можно

рассмотреть сравнение дифрактограммы неизвестного минерала с целью его идентификации с некоторым рядом стандартов. Сравнение со стандартами проводится не только по интенсивности отражения, но и по соответствующему ей углу отражения. Например, никаких выводов нельзя сделать, если сравнивать интенсивность отражения рентгеновских лучей под углом 2θ 20° с интенсивностью отражения под углом 2θ 30°. Даже если величина интенсивности одинакова, сравнение совершенно лишено смысла.

Тот факт, что данные, аналогичные приведенным, записаны в виде последовательности, не имеет значения, так как каждый элемент ряда рассматривается как отдельная и независимая переменная. Интенсивность отражения под углом 2θ 20° характеризуется одной переменной, а интенсивность под углом 2θ в 30° — другой. Методы сравнения таких последовательностей мы рассмотрим подробно в гл. 6 (см. книгу 2), где укажем многомерные критерии сходства и рассмотрим задачи классификации и дискриминантного анализа. В этом классе задач положение наблюдения в последовательности служит для его идентификации с данной переменной и не играет больше никакой роли.

Наоборот, некоторые из описанных в данной главе методов основаны на рассмотрении последовательностей данных как выборок из непрерывного множества наблюдений. Априори нет никаких причин считать, что одна из сравниваемых величин лучше другой. Такие методы, как взаимная корреляция и ассоциация, наиболее близки геологам, но, к сожалению, их применение ограниченно, так как они не допускают изменений масштаба одного разреза при его сравнении с другим. Для многих типов последовательностей, рассматриваемых в этой главе, изменения масштаба не требуется, и поэтому трудностей не возникает. Однако скорости осадконакопления, например, не являются постоянными, поэтому стратиграфические данные трудно изучать с помощью существующих методов корреляционного анализа.

Как мы уже отмечали в гл. 1, вычислительная машина является мощным орудием решения сложных задач. Однако она глупа и может принимать бессмысленные данные и выдавать столь же бессмысленные результаты без колебаний. Множество программ анализа последовательностей результатов наблюдений можно легко получить из различных источников. Если же использовать эти программы как «черный ящик», не понимая производимых в них операций и наложенных ограничений, то легко сбиться с пути. Мы надеемся, что описание методов и примеры этой главы будут достаточными для того, чтобы определить область применимости каждого метода и помочь пониманию программ. Однако в окончательной стадии анализа

исследователь должен полагаться на свою интуицию. Столкнувшись с задачей, в которой данные представлены в виде последовательности, вы можете с целью упорядочения вашего исследования задать себе следующие вопросы: а) на какой вопрос (вопросы) я хочу ответить? б) каков тип моих наблюдений? в) каков тип последовательности, образованной полученными наблюдениями?

Вы можете довольно быстро обнаружить, что ответ на первый вопрос можно получить лишь после четких ответов на второй и третий. Если учесть это до начала исследования, можно избежать части ненужной работы. В противном случае способ сбора данных может предопределить методы, которые нужно использовать для интерпретации, что в свою очередь сильно ограничит область исследования.

ИНТЕРПОЛЯЦИОННЫЕ ПРОЦЕДУРЫ

Многие из описываемых ниже методов пригодны для данных с равномерным расположением в пространстве; наблюдения должны проводиться с сохранением одного и того же интервала на прямой или на пересечении иного типа. Конечно, на практике соблюдение этого условия часто невозможно, например при многих стратиграфических исследованиях или при анализе данных из буровых скважин, а также для выборок, собранных по маршрутам в слабо обнаженном районе. Поэтому необходимо уметь оценивать рассматриваемые переменные в точках с регулярным расположением, используя их значения, полученные для различных интервалов. Эти вопросы излагаются в гл. 5 при рассмотрении построения изолиний. Программы построения изолиний предназначены для регулярной сети контрольных точек на основании наблюдений, нерегулярно расположенных в пространстве. Внешний вид и точность построений карты в большой степени зависят от величины шага применяемой сети и от алгоритма, использованного для получения оценок в ее узлах. Сейчас мы рассмотрим одномерный аналог этой задачи.

В табл. 4.2 приведены содержания магния в пробах из вод реки Стрим Уотер. В связи с трудностью отбора пробы были взяты с различными интервалами. Места отбора проб были тщательно нанесены на аэрофотоснимки, после чего между ними измерялись расстояния.

Из большого числа методов, с помощью которых можно из нерегулярно расположенных данных получить оценки для регулярного расположения, мы подробно разберем только два. Первый — наиболее простой. Это метод простой линейной интерполяции между заданными точками, позволяющий получить оценки в промежуточных точках. Рис. 4.1 иллюстрирует эту

Таблица 4.2

Двадцать измерений содержания магния в водах р. Стрим Уотер (расстояния исчисляются от устья реки до мест взятия проб)

Расстояние, м	Содержание магния, г/т
0,1	6,44
1820	8,61
2542	5,24
2889	5,73
3460	3,81
4586	4,05
6020	2,95
6841	2,57
7232	3,37
10903	3,84
11098	2,86
11922	1,22
12530	1,09
14065	2,36
14937	2,24
16244	2,05
17632	2,23
19002	0,42
20860	0,87
22471	1,26

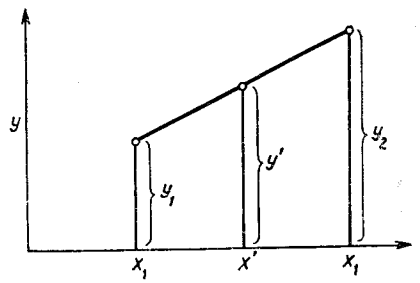
процедуру. Предположим, что Y_1 и Y_2 — наблюдаемые значения в точках X_1 и X_2, и мы хотим оценить значение Y' в точке X'. Считая зависимость в промежуточных точках линейной, значение Y' в точке X' можно вычислить по следующей формуле:

$$Y' = \frac{(Y_2 - Y_1)(X' - X_1)}{X_2 - X_1} + Y_1. \qquad (4.1)$$

Иными словами, мы предполагаем, что разность между значениями в двух соседних точках является линейной функцией разделяющего их расстояния. Значение в точке, лежащей посередине между двумя наблюдениями, находится точно посередине между значениями в двух прилежащих точках. Ближайшей точкой к наблюдению будет точка, значение которой ближе всего к значению наблюдения.

Несмотря на простоту процедуры линейной интерполяции, во многих случаях она обладает некоторыми недостатками. В случае, когда число равномерно расположенных точек приблизительно равно числу точек наблюдений и последние более или менее равномерно распределены, этот метод дает удовлетворительные результаты. Однако если точек наблюдения намного больше точек, получающихся после интерполяции, то бóльшая часть первичных данных не используется, так как для

Рис. 4.1. Линейная интерполяция между точками последовательности данных



определения одной точки в интерполяционной процедуре требуется использовать только два ближайших значения. Если исходные данные подвержены значительному влиянию случайных изменений, то полученные в результате интерполяции новые значения также могут иметь недопустимую случайную изменчивость. Оба эти момента учитываются в методах, предназначенных для использования более двух первоначальных значений и основанных на процедурах усреднения, использующих взвешенные расстояния. Однако методы усреднения, которые будут рассмотрены ниже, при изложении процедур сглаживания данных также имеют недостатки. Например, значение переменной в первоначальной точке при интерполяции не обязательно сохраняется, так как полученное в результате усреднения новое значение может превышать (или быть меньше) значения в близлежащих точках из-за влияния более удаленных точек.

Если первоначальные данные расположены нерегулярно и промежуточные значения должны быть вычислены для каждой пары наблюдений, то применение линейной интерполяции, отражающей равномерность изменения переменной между контрольными точками, вполне оправданно. Однако в любой задаче, в которой приходится использовать интерполяцию наблюдаемых данных, всегда следует иметь в виду, что для получения оценок нельзя использовать любой метод. Точность полученного результата зависит от плотности первоначальных данных, использовавшихся при построении сети, и никакая интерполяция не позволит нам улучшить результаты анализа, если исходных данных недостаточно. Например, мы могли бы оценить содержание магния в реке с интервалом 500 м или даже с интервалом 5 м, однако совершенно ясно, что эти новые значения не дадут нам дополнительной информации о распределении изучаемого элемента в потоке.

Теперь мы рассмотрим метод получения равномерно расположенных в пространстве данных, основанных на использовании всех наблюдений между последовательными точками, в которых были получены оценки. Этот метод называется методом

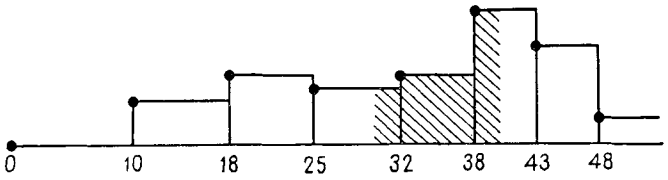


Рис. 4.2. Последовательность данных, рассматриваемая как ступенчатая функция или «кривая прямоугольников»

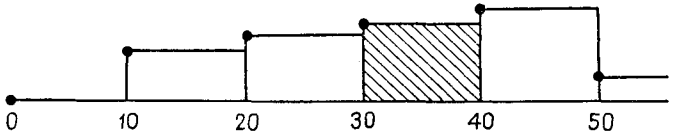


Рис. 4.3. Равномерная пространственная последовательность, полученная из неравномерно расположенных данных путем прямоугольного интегрирования. Площадь заштрихованной области такая же, как и на рис. 4.2

прямоугольников. Исходные данные можно представить в виде ступенчатой функции, имеющей в интервале между последовательными наблюдениями постоянное значение. Пример такой функции представлен на рис. 4.2. Если требуется изобразить распределение с равномерной сетью, надо построить другую ступенчатую функцию, но такую, чтобы площади прямоугольников с равными между собой основаниями были равны соответственно площадям исходных прямоугольников. Такое распределение представлено на рис. 4.3, где получена последовательность равномерно распределенных значений, которые соответствуют данным, представленным на предыдущем рисунке. Обе заштрихованные площади равны между собой. Эта процедура имеет то преимущество, что она учитывает все данные внутри интервала, содержащего точку, для которой требуется получить оценку значения изучаемой величины. Кроме того, поскольку площадь, ограниченная построенным графиком, равна площади, ограниченной первоначальным графиком, то наблюдениям, используемым для получения оценки в точке, должны быть приписаны веса, пропорциональные длине того интервала, который им отвечает.

Несмотря на то что теоретически получение оценок методом прямоугольников очень просто, оно представляет некоторые трудности при программировании. Используя первую полученную оценку, вычисляем произведение расстояния до следующего наблюдения на его величину, что дает нам площадь прямоугольника; далее производим те же вычисления, для всех наблюдений, пока не получим оценку в последней точке. Эта оценка определяется в результате суммирования уже найденных площадей и деления суммы на величину интервала, выбранного для получения оценки. Начальное значение в оцениваемой точке последовательности выбирается таким же, как и в ближайшей предшествующей данной точке.

Различие между двумя программами выступает особенно явно в том случае, когда исходные данные разрознены, и между двумя наблюдениями приходится получать более одной оценки. При линейной интерполяции вычисляются значения, расположенные на прямой линии, соединяющей две ближайшие заданные точки. Наоборот, в методе прямоугольников строятся оценки, равные первому наблюдению.

При изучении зоны метаморфизма вокруг интрузива была опробована скважина, пробуренная перпендикулярно к контакту интрузии с вмещающими породами. Керн раскололи, и все кристаллы граната, расположенные на поверхности, извлекли и проанализировали спектрохимическим методом на содержание железа. Как расстояния между кристаллами, так и содержание железа в них колебались в широких пределах. Данные, полученные в результате этого эксперимента, приведены в табл. 4.3. Желательно было бы получить с их помощью общую картину изменений в составе граната, однако эти данные кажутся довольно незакономерными и не поддаются прямой интерпретации. В качестве первого шага проведения анализа можно построить их оценки для равномерной сети. Желательно в качестве интервала между точками выбрать расстояние 0,5 м. Здесь мы сталкиваемся с положением, отличающимся от рас-

Таблица 4.3

Содержание железа в гранатах из керна скважины, пробуренной алмазной коронкой

Глубина, см	Содержание железа, %	Глубина, см	Содержание железа, %	Глубина, см	Содержание железа, %
0	14,21	121	19,84	419	22,56
3	19,35	130	16,94	425	19,00
10	17,22	163	16,72	429	20,53
14	15,87	168	19,20	443	19,08
23	13,62	205	20,41	447	22,83
30	16,31	239	16,88	465	21,06
36	14,13	251	18,74	474	24,96
48	13,95	283	16,67	493	19,12
59	15,00	297	18,56	502	22,24
66	14,23	322	18,87	522	26,88
68	16,81	335	20,81	550	21,15
81	15,93	351	24,52	558	28,92
94	16,02	370	25,03	571	27,96
96	17,85	408	25,11	586	25,03
102	17,02	416	23,28	596	26,27
115	15,87				

смотренного выше случая изучения содержаний магния в воде; наблюдений значительно больше, чем необходимых оценок, а наша цель — сохранить, насколько это возможно, исходную информацию. В этом случае метод прямоугольников представляется более подходящим для получения оценок, чем метод линейной интерполяции. Однако читателю рекомендуется получить оценки по данным табл. 4.3 и тем, и другим методами и затем сравнить полученные результаты для уяснения различия между этими методами.

В геологии процедуры с равномерным пространственным распределением чаще всего использовались при подготовке стратиграфических данных (исследований разрезов, записей в буровых журналах и др.) для проведения временного тренд-анализа. Этот метод, в сущности, является не чем иным как процедурой сглаживания, и для своего применения требует равномерного расположения данных в пространстве. Такие методы анализа временных рядов, как автокорреляционный и гармонический, также требуют для своего применения равномерного распределения данных. В настоящее время методы анализа временны́х рядов, по-видимому, более мощные, чем другие методы изучения последовательностей наблюдений, и, вероятно, в ближайшее время найдут широкое применение. Однако эти методы требуют больших массивов информации, что ограничивает возможности их применения к стратиграфическим последовательностям и анализу хорошо разведанных рудных месторождений. Некоторые результаты были получены при изучении последовательностей минералов, наблюдаемых в шлифах. Но это будет подробно рассмотрено ниже.

ЦЕПИ МАРКОВА

Многие геологические исследования связаны с изучением последовательностей результатов наблюдений, которые представлены взаимоисключающими друг друга состояниями. В качестве примера можно привести точечный подсчет содержаний минералов по пересечению шлифа, когда наличие того или иного минерала определяет состояние в данной точке. При изучении стратиграфических разрезов или скважины, секущей зоны рудного тела, наблюдаются последовательности литологических типов пород, которым можно дать аналогичную классификацию. Наблюдения по пересечению можно провести с равными интервалами или же расположить их произвольно, как это обычно делается при изучении стратиграфических разрезов. В первом примере мы можем ожидать появления одного и того же состояния, т. е. несколько последовательных наблюдений могут попасть в одну и ту же категорию. Этого, очевидно, не произойдет, если наблюдения делаются только в точках изменения состояния.

Данные, входящие в эти последовательности, в сущности, такие же, как те, которые используются при автокорреляции и взаимной корреляции. Теперь, однако, мы будем изучать не расположение состояний в последовательности, а природу перехода из одного состояния в другое. Мы рассмотрим методы, которые совсем не используют информацию о положении наблюдений внутри последовательности, но зато позволяют получить информацию о тенденции одного состояния следовать за другим.

Данные табл. 4.4 представляют стратиграфический разрез, изображенный на рис. 4.4, в котором осадочные породы определялись в последовательных точках, расположенных на расстоянии 0,3 м. Каждая литологическая разновидность была отнесена к одному из четырех взаимно исключающих друг друга состояний — песчанику, известняку, сланцу и углю, которые условно были обозначены соответственно *A*, *B*, *C*, и *D*. Можно построить матрицу порядка 4×4, в которой указывается, сколько раз данное состояние следует за другим. Матрица такого типа называется матрицей переходных частот. Примеры таких матриц приведены ниже. Исследуемый стратиграфический разрез содержит 63 наблюдения, поэтому имеется $(n-1)=62$ перехода. Матрицу следует читать от строки к столбцу. Это, например, обозначает, что переходу из состояния *A* в состояние *C* соответствует элемент матрицы a_{13}, т. е. если мы хотим знать, сколько раз осуществлялся переход из строки с номером *A* в столбец с номером *C*, мы увидим, что переход из состояния *A* в состояние *C* произошел 5 раз по всей последовательности. Аналогично, имеется пять переходов из состояния *C* в состояние *A* в этой последовательности; это число — элемент матрицы, стоящий на пересечении строки *C* и столбца *A*. Матрица

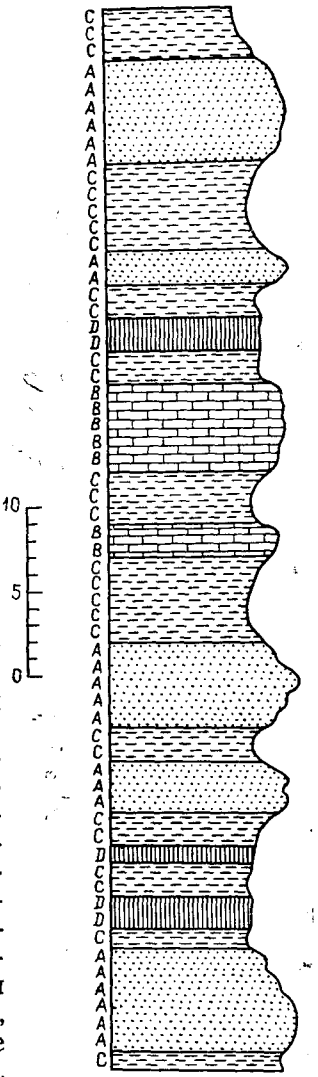


Рис. 4.4. Стратиграфический разрез, в котором литологические состояния взаимно исключают друг друга:
A — песчаник; *B* — известняк; *C* — сланец; *D* — уголь

Таблица 4.4

Стратиграфическая последовательность, представленная на рис. 4.4, закодирована с помощью четырех взаимно исключающих состояний: песчаник (*A*), известняк (*B*), сланец (*C*), уголь (*D*)

(Верх)
C C B C A A
C C B C A A
C C B C A A
A A B C C A
A A B A C A
A C C A D A
A C C A C C
A D C A C (Низ)
A D B A D
C C B C D
C C C C

переходных частот является хорошим способом выражения следования одного состояния за другим:

		В				
		A	*B*	*C*	*D*	Суммы по строкам
Из	*A*	18	0	5	0	23
	B	0	5	2	0	7
	C	5	2	18	2	28
	D	0	0	3	2	5
Суммы по столбцам		23	7	28	5	63 Общая сумма

Заметим, что суммы строк и суммы столбцов одинаковы при условии, что разрез начинается и кончается в одном и том же состоянии: в противном случае две строки и два столбца отличаются на единицу. Заметим также, что в противоположность тем матрицам, которые вычисляли раньше, матрица переходных частот, вообще говоря, если последовательность начинается и кончается в разных состояниях, несимметрична, т. е. $a_{ij} \neq a_{ji}$.

Тенденцию следования одного состояния за другим можно выразить в виде матрицы, превратив частоты в десятичные дроби или проценты. Если каждый элемент i-й строки разделить на сумму элементов i-й строки, то полученная дробь выразит относительную величину, характеризующую, сколько раз за i-м состоянием будет следовать какое-либо другое состояние. В вероятностном смысле эти числа являются оценками условной вероятности $P(j|i)$, т. е. вероятности того, что состояние j появится вслед за данным в настоящий момент состоянием i. (Мы ввели необычное, но эквивалентное обозначение $(P(i \rightarrow j)$, которое можно считать вероятностью того, что за состоянием i будет следовать состояние j.) Это обозначение будет использовано позже.

		В				Общая сумма
		A	*B*	*C*	*D*	
Из	*A*	0,78	0	0,22	0	1,00
	B	0	0,71	0,29	0	1,00
	C	0,18	0,07	0,64	0,11	1,00
	D	0	0	0,60	0,40	1,00

Отсюда, например, видно, что если в некоторой точке мы находимся в состоянии C, то вероятность того, что литологическая разновидность на 0,3 м выше также есть C, равна 0,64. Вероятность того, что эта литологическая разновидность будет состоянием A, равна 18%; она равна 7% для состояния B и 11% — для состояния D. Так как эти четыре состояния взаимно исключают друг друга и исчерпывают все состояния, то литологическая разновидность должна быть одной из четырех. Поэтому сумма, заданная по строкам, равна 100%.

Если мы разделим суммы по строкам переходной матрицы частот на сумму общего числа переходов, то мы получим относительную долю всех четырех литологических разновидностей, представленных в этом разрезе. Этот фактор вероятностей называется маргинальным, или фиксированным вектором вероятностей.

A	0,37
B	0,11
C	0,44
D	0,08

Напомним, что в гл. 2 совместная вероятность осуществления двух событий A и B равнялась

$$P(A, B) = P(B|A)P(A)$$

или

$$P(B|A) = P(A, B)/P(A).$$

Таким образом, вероятность того что состояние B будет следовать за состоянием A, равна вероятности того, что реализуются оба состояния A и B, деленной на вероятность осуществления события A. Если вероятности осуществления событий A и B независимы, то

$$P(A, B) = P(A)P(B) \quad \text{и}$$

$$P(B|A) = [P(A)P(B)]/P(A) = P(B),$$

т. е. вероятность того, что состояние B будет следовать за состоянием A, есть просто вероятность того, что состояние B осуществится в этом разрезе, и эта вероятность задается соответствующим элементом в фиксированном вероятностном векторе.

Если все вероятности реализации состояний в разрезе независимы, то такое же соотношение верно и для всех возможных переходов, например,

$$P(B|A) = P(B|B) = P(B|C) = P(B|D) = P(B).$$

Это позволяет нам предсказать, что матрица вероятностей перехода должна выглядеть так, как если бы частоты литологических состояний в одной точке стратиграфического интервала были бы совершенно независимыми от литологии в точке, расположенной ниже в непосредственной близости. Матрица ожидаемых вероятностей перехода будет состоять из строк, которые идентичны заданному вектору вероятностей. Для нашего стратиграфического примера она имеет вид

		В			
		A	*B*	*C*	*D*
Из	*A*	0,37	0,11	0,44	0,08
	B	0,37	0,11	0,44	0,08
	C	0,37	0,11	0,44	0,08
	D	0,37	0,11	0,44	0,08

Мы можем сравнить эту матрицу ожидаемых вероятностей перехода с наблюдаемой матрицей переходных вероятностей, которую мы действительно наблюдаем при проверке гипотез о том, что все литологические состояния независимы от непосредственно предшествующих состояний. Это делается с помощью критерия χ^2, причем сначала отношения, характеризующие вероятности появления литологических типов, превращаются в частоты, для чего проводится умножение каждой строки на соответствующее общее число событий:

Ожидаемые вероятности перехода				Общее число		Ожидаемые частоты			
0,37	0,11	0,44	0,08	× 23	=	8,5	2,5	10,1	1,8
0,37	0,11	0,44	0,08	× 7	=	2,6	0,8	3,1	0,6
0,37	0,11	0,44	0,08	× 28	=	10,4	3,1	12,3	2,2
0,37	0,11	0,44	0,08	× 5	=	1,9	0,6	2,2	0,4

По виду критерий χ^2 напоминает критерий, описанный в гл. 2. Каждый элемент матрицы частот перехода составляет категорию, в которую входит как наблюденное, так и ожидаемое число переходов. Эти числа сравниваются по формуле

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}, \qquad (4.2)$$

где O — наблюденное число переходов из одного состояния в другое, E — число ожидаемых переходов из одного состояния

в другое, если последовательные состояния независимы. Критерий имеет $(m-1)^2$ степеней свободы, где m — число состояний (одна степень свободы теряется в каждой строке, так как сумма вероятностей в каждой строке равна 1,00). Как и с другими типами критерия χ^2, каждая категория должна иметь ожидаемые частоты по меньшей мере для пяти переходов. В этом примере это не так, но мы можем применить консервативный критерий независимости, вычислив проверяемую статистику для четырех категорий, у которых ожидаемые частоты больше пяти. Остальные категории можно комбинировать до тех пор, пока их ожидаемые частоты превысят пять.

Категории включают переходы $A \to A$, $A \to C$, $C \to A$ и $C \to C$. Комбинируемые категории могут быть образованы из всех элементов строки B, всех элементов строки D и комбинаций переходов $A \to B$, $A \to D$, $C \to B$ и $C \to D$. Получаемая статистика χ^2 имеет вид

$$\chi^2 = \frac{(18-8{,}5)^2}{8{,}5} + \frac{(5-10{,}4)^2}{10{,}4} + \frac{(5-10{,}1)^2}{10{,}1} + \frac{(18-12{,}3)^2}{12{,}3} + \frac{(7-7{,}0)^2}{7{,}0} + \frac{(5-5{,}0)^2}{5{,}0} + \frac{(5-9{,}8)^2}{9{,}8} = 20{,}99.$$

Критическое значение χ^2 для девяти степеней свободы и 5%-ного уровня значимости есть 16,92; проверяемое значение значительно превышает это число, поэтому мы можем заключить, что гипотеза о независимости последовательных состояний неверна, т. е. существует статистически значимая тенденция для некоторых состояний иметь предпочтение в следовании за некоторым другим состоянием.

Последовательность, в которой состояние в одной точке в вероятностном смысле лишь частично зависит от предыдущего состояния, называется марковской цепью. Названная так в честь русского математика А. А. Маркова, она является промежуточным звеном между полностью определенными (детерминированными) и совершенно случайными последовательностями. Наш стратиграфический разрез указывает на цепь Маркова первого порядка, т. е. статистическая зависимость существует между двумя соседними точками. Существуют также свойства марковости высших порядков. Например, последовательность Маркова второго порядка указывает на значимую условную зависимость между точками, предшествующими данной на двух предшествующих шагах.

С помощью матрицы переходных вероятностей мы можем оценить, т. е. предсказать литологию на 0,6 м (на два наблюдения) выше некоторой данной точки. Предположим, что мы начали с известняка (состояние B). Указанные ниже вероят-

ности оценивают литологическое состояние, соответствующее соседней выше расположенной точке.

Состояние	%
A (песчаник)	0
B (известняк)	71
C (глина)	29
D (уголь)	0

Предположим, что следующая точка действительно попадает в глину; мы можем затем определить вероятность появления заданной литологической разновидности в следующей точке:

Состояние	%
A (песчаник)	18
B (известняк)	7
C (глина)	64
D (уголь)	11

Таким образом, вероятность литологической последовательности известняк—глина—известняк оценивается как

$$P(B \rightarrow C) \times P(C \times B) = 29\% \times 7\% \approx 2\%.$$

Однако есть и другой путь достичь в два шага состояния известняка. Возможна также последовательность известняк—известняк—известняк. Вероятность получить эту последовательность такова

$$P(B \rightarrow B) \times P(B \rightarrow B) = 71\% \times 71\% \approx 50\%.$$

Так как другие переходы известняк—песчаник и известняк—уголь имеют нулевые вероятности, то эти две последовательности являются единственно возможными последовательностями, которые приводят от известняка к нему же в два шага. Вероятность того, что порода на два шага выше известняка будет тоже известняком, невзирая на промежуточное литологическое состояние, будет суммой вероятностей, т. е.

$$\begin{aligned} P(B \rightarrow A \rightarrow B) &= 0\% \\ P(B \rightarrow B \rightarrow B) &= 50\% \\ P(B \rightarrow C \rightarrow B) &= 2\% \\ P(B \rightarrow D \rightarrow B) &= 0\% \\ \hline & 52\% \end{aligned}$$

То же рассуждение можно применить для определения вероятности появления любой литологической разновидности спустя два шага, если исходить из любого литологического состояния. Однако нет нужды рассматривать всевозможные последовательности индивидуально, так как используемый при этом процесс умножения и суммирования в точности соответствует матричному умножению. Если матрица вероятностей перехода умножается на себя (т. е. возводится в квадрат), то результат будет матрицей вероятностей переходов второго порядка в последовательности:

$$\begin{bmatrix} 0{,}78 & 0 & 0{,}22 & 0 \\ 0 & 0{,}71 & 0{,}29 & 0 \\ 0{,}18 & 0{,}07 & 0{,}64 & 0{,}11 \\ 0 & 0 & 0{,}60 & 0{,}40 \end{bmatrix}^2 = \begin{bmatrix} 0{,}64 & 0{,}02 & 0{,}31 & 0{,}02 \\ 0{,}05 & 0{,}52 & 0{,}39 & 0{,}03 \\ 0{,}26 & 0{,}09 & 0.54 & 0{,}11 \\ 0{,}11 & 0{,}04 & 0{,}62 & 0{,}23 \end{bmatrix}$$

Заметим, что суммы элементов строк квадрата матрицы также составляют 100%.

Наличие значимой марковости второго порядка можно проверить точно так же, как мы проверяли независимость между последовательными состояниями с помощью критерия χ^2. Если повторить выкладки, проведенные ранее, только используя матрицу переходных вероятностей второго порядка, то мы убедимся в том, что эта последовательность не имеет значимого свойства быть марковской цепью второго порядка.

Мы можем оценить вероятность состояния на любом шаге будущего, просто возводя в степень матрицу вероятностей соответствующее число раз. Если матрица возводится в степень достаточно большое число раз, то она достигает устойчивого состояния, в котором все строки становятся равными фиксированному вероятностному вектору, или, другими словами, получается независимая матрица переходных вероятностей, которая не изменяется при последующих возведениях в степень.

В приведенном выше примере вы можете убедиться в том, что наивысшие переходные вероятности соответствуют переходам из одного какого-либо состояния в себя, например переходам песчаник—песчаник, известняк—известняк и глина—глина. Очевидно, что эти переходные вероятности связаны с мощностью выбираемых стратиграфических единиц и зависят от расстояния между выборочными точками. Например, частоты вдоль главной диагонали матрицы частот переходов будут удваиваться, в то время как внедиагональные частоты остаются неизменными в случае, если наблюдения проводились через каждые 0,15 м. Это в значительной степени усиливает марковское свойство, однако весьма специфическим образом. Выбирая подходящее расстояние между точками опробования, мы сталкиваемся со следующей проблемой: если наблюдения расположены слишком близко, матрица переходов отражает главным образом мощность более массивных стратиграфических единиц; если же расстояния между точками опробования слишком велики, тонкие единицы могут быть полностью пропущены.

Вложенные цепи Маркова

Трудностей выбора подходящего интервала опробования можно избежать, если наблюдения производятся только тогда, когда происходит смена состояний. Стратиграфический разрез, например, может быть представлен как последовательность подстилающих пород, каждая из которых литологически отлична от непосредственно предшествующей. В табл. 4.5 представлены записи последовательности горных пород, пересеченных скважиной, пробуренной в Долине Мидленд в Шотландии. Скважина была пробурена на 480 м в каменноугольных отложениях пенсильванского возраста, состоящих из переслаивающихся толщ глины, ила, песчаника и угольных пластов или корневых зон. Эти осадки интерпретировались как отложенные в дельте плоской равнины, подверженной частым затоплениям, так что можно ожидать, что некоторые литологические разновидности могут находиться в предпочтительных отношениях друг с другом. Данные взяты из большого числа скважин, изучавшихся Даутоном.

Матрица переходных частот для четырех состояний, полученная для разреза Шотландской скважины, приводится ниже. Совершенно очевидна разница между этой и рассмотренной ранее матрицей, состоящая в том, что все ее диагональные члены должны быть нулями, так как состояние не может следовать за самим собой. Матрица переходных вероятностей, получающаяся делением каждого элемента матрицы переходных частот на соответствующую сумму элементов строки, обладает тем же самым свойством. Последовательности, в которых переходы из некоторого состояния в себя запрещены, называются вложенными марковскими цепями. Их анализ представляет специфическую задачу, трудность которой не всегда понятна геологам, изучающим записи стратиграфических данных.

Таблица 4.5

Последовательные литологические состояния, полученные в скважине, пробуренной в угольных пластах Долины Мидленд в Шотландии; взаимно исключающие состояния: *A* — пустой сланец, *B* — сланец с ископаемыми остатками пресноводных раковин; *C* — алевролит, *D* — песчаник и *E* — уголь или коренная зона; читайте вниз по столбцам [15]

(Верх)

B	*D*	*C*	*B*	*E*	*B*	*B*	*A*	*C*	*E*	*A*	*C*	*D*	*A*	*E*	*E*	*A*	*A*	*A*	*D*
E	*C*	*D*	*E*	*A*	*A*	*E*	*C*	*E*	*C*	*B*	*D*	*A*	*C*	*A*	*A*	*C*	*E*	*D*	*C*
A	*A*	*C*	*A*	*C*	*C*	*A*	*D*	*A*	*A*	*E*	*C*	*E*	*A*	*D*	*D*	*E*	*A*	*E*	*D*
E	*E*	*A*	*D*	*D*	*A*	*C*	*C*	*D*	*B*	*A*	*D*	*C*	*B*	*A*	*C*	*D*	*E*	*C*	*C*
A	*D*	*B*	*C*	*C*	*C*	*A*	*D*	*A*	*C*	*B*	*C*	*D*	*C*	*B*	*D*	*A*	*C*	*D*	*A*
D	*C*	*A*	*A*	*D*	*A*	*C*	*A*	*D*	*E*	*A*	*A*	*C*	*D*	*E*	*A*	*B*	*D*	*C*	*D*
A	*A*	*E*	*E*	*C*	*B*	*B*	*B*	*C*	*C*	*B*	*C*	*D*	*A*	*A*	*E*	*D*	*A*	*A*	*A*
C	*D*	*D*	*C*	*D*	*A*	*E*	*E*	*A*	*A*	*E*	*A*	*E*	*E*	*C*	*A*	*B*	*B*	*E*	*B*
D	*C*	*C*	*D*	*C*	*B*	*C*	*A*	*E*	*D*	*A*	*C*	*C*	*A*	*D*	*C*	*A*	*E*	*A*	*A*
C	*A*	*A*	*C*	*A*	*E*	*A*	*C*	*A*	*B*	*B*	*E*	*D*	*C*	*E*	*D*	*E*	*A*	*E*	*B*
D	*E*	*E*	*A*	*B*	*A*	*D*	*D*	*C*	*E*	*A*	*A*	*C*	*D*	*A*	*C*	*A*	*B*	*A*	*E*
C	*C*	*C*	*B*	*E*	*C*	*C*	*C*	*D*	*A*	*B*	*C*	*E*	*C*	*D*	*A*	*C*	*E*	*C*	*A*
A	*D*	*A*	*A*	*A*	*D*	*A*	*A*	*A*	*D*	*E*	*D*	*A*	*E*	*C*	*D*	*A*	*A*	*D*	*D*
B	*C*	*D*	*E*	*B*	*C*	*C*	*B*	*E*	*C*	*C*	*C*	*A*	*A*	*A*	*E*	*E*	*A*	*B*	
E	*B*	*E*	*A*	*A*	*D*	*D*	*A*	*A*	*D*	*A*	*D*	*A*	*C*	*B*	*E*	*C*	*A*	*E*	*A*
A	*E*	*A*	*D*	*B*	*C*	*C*	*B*	*E*	*E*	*C*	*C*	*E*	*B*	*E*	*A*	*D*	*C*	*C*	*E*
D	*A*	*D*	*E*	*A*	*D*	*E*	*E*	*A*	*A*	*D*	*A*	*A*	*E*	*A*	*D*	*C*	*D*	*D*	*A*
C	*D*	*A*	*A*	*B*	*C*	*A*	*A*	*C*	*D*	*A*	*B*	*C*	*A*	*D*	*A*	*D*	*E*	*B*	(Низ)
D	*C*	*C*	*D*	*E*	*A*	*C*	*D*	*D*	*A*	*E*	*E*	*A*	*C*	*C*	*D*	*C*	*A*	*E*	
C	*D*	*A*	*C*	*A*	*C*	*D*	*A*	*C*	*C*	*A*	*A*	*E*	*A*	*D*	*C*	*D*	*D*	*A*	

Из	В *A*	*B*	*C*	*D*	*E*	Суммы по строкам
A	0	11	36	21	52	120
B	28	0	4	4	0	36
C	34	2	0	45	13	94
D	29	1	45	0	3	78
E	28	23	9	8	0	68
Суммы по столбцам	119	37	94	78	68	396 (Общая сумма)

Литологические состояния закодированы буквами: *A* — глины, не содержащие ископаемые организмы и мелкую породу, *B* — глины, содержащие раковины пресноводной фауны, *C* — алеврит, *D* — песчаник и *E* — уголь и коренные породы. Соответствующая матрица переходных вероятностей имеет вид

Из	В *A*	*B*	*C*	*D*	*E*	Суммы по строкам
A	0	0,09	0,30	0,18	0,43	1,00
B	0,78	0	0,11	0,11	0,00	1,00
C	0,36	0,02	0	0,48	0,14	1,00
D	0,37	0,01	0,58	0	0,04	1,00
E	0,41	0,34	0,13	0,12	0	1,00

Маргинальный вероятностный вектор имеет вид

A	0,30
B	0,09
C	0,24
D	0,20
E	0,17

Критерий χ^2, идентичный (4.2), может быть использован для проверки марковского свойства вложенной последователь-

ности. Это делается сравнением наблюдаемой частотной матрицы перехода с ожидаемой матрицей, если последовательные состояния независимы. Однако фиксированный вероятностный вектор не может быть использован для оценки столбцов ожидаемой матрицы переходных вероятностей. Поскольку переходы из некоторого состояния в себя запрещены, мы должны были бы использовать какой-нибудь окольный путь для оценки частот переходов между независимыми состояниями при условии, что такие состояния не могут следовать за самими собой.

Для начала представим себе, что наша последовательность в действительности является некоторой цензурированной выборкой, извлеченной из обычной последовательности, в которой переходы из некоторого состояния в себя могут иметь место. Матрица переходных вероятностей этой последовательности должна иметь вид, аналогичный уже наблюденной, исключая диагональ, которая будет содержать отличные от нуля элементы. Если бы мы вычислили матрицу вероятностей перехода из этой матрицы частот и затем возвели бы ее в достаточно высокую степень, то мы оценили бы матрицу переходных вероятностей последовательности, в которой последовательные состояния независимы. Если затем отбросить диагональные элементы и пересчитать внедиагональные вероятности, то в результате получится ожидаемая матрица переходных вероятностей для вложенной последовательности, у которой состояния независимы.

Как вычислить частоты переходов из каждого состояния в себя, если эта информация недоступна? Мы будем применять для этого метод проб и ошибок, выбирая для них такие значения, которые, будучи вставленными на диагонали матрицы переходных частот, не изменяются при возведении матрицы в степень. Внедиагональные элементы, однако, будут изменяться до тех пор, пока не будет достигнута устойчивая конфигурация, соответствующая модели независимых событий.

На практике совсем нет нужды вычислять внедиагональные элементы. Мы начинаем с того, что приписываем диагональным элементам наблюденной матрицы переходных частот некоторое произвольное большое число, скажем 1000. Просуммировав элементы каждой строки и разделив результаты на общую сумму, используем полученные значения в качестве оценки переходных вероятностей, стоящих на диагонали. Эти вероятности возводятся в квадрат и умножаются на общую сумму, в результате получаются новые оценки диагональных частот. Эти новые оценки вставляются в исходную матрицу переходных частот и затем этот процесс повторяется. Укажем первый цикл этой процедуры.

Шаг 1. Записываем исходную матрицу переходных частот со значением 1000 на каждом месте диагонали:

		В					Суммы по строкам	
		A	*B*	*C*	*D*	*E*		
Из	*A*	1000	11	36	21	52	1120	
	B	28	1000	4	4	0	1036	
	C	34	2	1000	45	13	1094	
	D	29	1	45	1000	3	1078	
	E	28	23	9	8	1000	1068	
							5396	Общая сумма

Шаг 2. Оцениваем соответствующие диагональным элементам переходные вероятности, найденные делением сумм по строкам на общую сумму:

		В					Суммы по строкам
		A	*B*	*C*	*D*	*E*	
Из	*A*	0,208					0,208
	B		0,192				0,192
	C			0,203			0,203
	D				0,200		0,200
	E					0,198	0,198

Шаг 3. Получаем вторую оценку матрицы переходных частот, используя диагональные элементы, полученные умножением диагональных вероятностей на общую сумму 5396. Внедиагональные элементы — исходные наблюденные частоты. Затем находятся новые суммы по строкам и общая сумма.

		В					Суммы по строкам	
		A	*B*	*C*	*D*	*E*		
Из	*A*	233	11	36	21	52	353	
	B	28	199	4	4	0	235	
	C	34	2	222	45	13	316	
	D	29	1	45	215	3	294	
	E	28	23	9	8	212	280	
							1478	Общая сумма

Этот процесс продолжаем до тех пор, пока оценки переходных частот на диагонали не перестанут изменяться. Достижение этой цели требует от 10 до 20 итераций, причем число итераций зависит от того, как близки были выбранные нами наугад значения к стабильным. Окончательный вид матрицы переходных частот с оцененными диагональными частотами приведен ниже.

		В					Суммы по строкам	
		A	*B*	*C*	*D*	*E*		
	A	66	11	36	21	52	186	
	B	28	3	4	4	0	39	
Из	*C*	34	2	29	45	13	123	
	D	29	1	45	17	3	95	
	E	28	23	9	8	13	81	
Сумма по столбцам		185	40	123	95	81	524	Общая сумма

Эту матрицу можно преобразовать в ожидаемую матрицу переходных вероятностей гипотетической марковской последовательности, разделив каждый ее элемент на соответствующую сумму по строке. Однако такая матрица мало интересна, так как она характеризует больше гипотетическую, чем наблюденную вложенную последовательность. Другое дело — маргинальные суммы по строкам. При их вычислении используется маргинальный вероятностный вектор

A	0,355
B	0,074
C	0,235
D	0,181
E	0,155

Мы можем теперь вычислить ожидаемые вероятности и ожидаемые частоты гипотетической последовательности независимых состояний для маргинального вероятностного вектора. Мы проверяем гипотезу о независимости последовательных состояний, замечая, что, например, если состояние A не зависит от состояния B, то $P(A|B)=P(A)P(B)$. Так как $P(A)$ и $P(B)$ заданы соответствующими элементами маргинального вероятностного вектора, то оценка условной вероятности того, что состояние A будет следовать за состоянием B, равна $P(A|B)=$ $=(0{,}355)\,(0{,}074)=0{,}026$. Ожидаемые вероятности для всех переходов приведены ниже:

		В				
		A	B	C	D	E
Из	A	0,125	0,026	0,083	0,064	0,055
	B	0,026	0,006	0,017	0,013	0,012
	C	0,083	0,017	0,055	0,043	0,036
	D	0,064	0,013	0,043	0,033	0,028
	E	0,055	0,012	0,036	0,028	0,024

Ожидаемые частоты находятся умножением этой этой матрицы на общую сумму, равную 524.

		В				
		A	B	C	D	E
Из	A	65,5	13,6	43,5	33,5	28,8
	B	13,6	3,1	8,9	6,8	6,3
	C	43,5	8,9	28,8	22,5	18,9
	D	33,5	6,8	22,5	17,3	14,7
	E	28,8	6,3	18,9	14,7	12,6

Заметим, что эта матрица симметрична и диагональные элементы остаются неизменными (с точностью до ошибок

округления). Внедиагональные элементы есть ожидаемые частоты переходов для вложенной последовательности, если предположить независимость между последовательными состояниями. Если отвлечься от диагональных элементов матрицы, то ее можно сравнить с наблюденной матрицей переходных частот, так как суммы строк и столбцов у них одинаковые (снова с точностью до ошибок округления).

Применяя для сравнения статистику χ^2, получаем значение $\chi^2=172$. Критерий имеет $\nu=(m-1)^2-m$ степеней свободы, где m — число состояний, или в этом примере $\nu=11$. Критическое значение χ^2 с 11 степенями свободы и уровнем значимости $\alpha=$ $=0{,}05$ равно 19,68, т. е. вычисленное значение значительно превышает критическое и поэтому мы можем заключить, что последовательные литологические состояния, зарегистрированные в Шотландии, не являются независимыми, а скорее отражают свойство сильной марковости первого порядка.

Если критерий покажет на наличие частичной зависимости между последовательными состояниями некоторой последовательности, то можно продолжить исследования с целью уточнения этой зависимости. Простые графы наиболее значимых вероятностей напоминают картинки, характеризующие повторения в последовательности; они также могут быть обнаружены с помощью аппарата теории автоассоциаций. Для проверки значимости вероятностей переходов между индивидуальными парами состояний применимы модификации критерия χ^2. Некоторые авторы находят, что собственные значения матрицы переходных вероятностей являются полезными индикаторами цикличности. (Необходимо отметить, однако, что вычисление собственных векторов асимметричной матрицы, какой является матрица переходных вероятностей, совсем не простая задача.) Эти вопросы в нашей книге далее не будут развиваться; интересующийся читатель может обратиться к книге Кемени и Снелла [30] или книге Шварцахера [49], посвященной количественным методам изучения осадконакопления. Критерии типа χ^2 для вложенных последовательностей представлены в книге Гудмена [22]. Соответствующие геологические проблемы рассмотрены в статьях [15] и [16]; можно также рекомендовать работу Гюрка [54].

ПОСЛЕДОВАТЕЛЬНОСТИ СОБЫТИЙ

Мы не рассмотрели еще один интересный тип временны́х рядов, называемых последовательностями событий. Примерами геологических данных такого рода могут служить исторические сведения о землетрясениях в Калифорнии, записи о вулканических извержениях в Средиземном море. Характеристики этих рядов следующие: а) события различаются моментами време-

ни, в которые они произошли; б) события по существу своему мгновенны; в) события настолько редки, что никакие два не происходят в один и тот же временной интервал.

Последовательности событий можно рассматривать как последовательность интервалов между их реализацией. Наши данные могут содержать также продолжительность интервалов между происходящими событиями или состоять из значений суммарных длин временных интервалов, характеризующих события. Данные в одной форме могут быть просто преобразованы в данные другой формы.

Модели последовательностей событий можно использовать для анализа некоторых типов пространственных данных. Например, нас может интересовать частота обнаружения редких материалов, спорадически встречающихся на пересечении шлифа, или же распространенность бентонитовых слоев в вертикальной последовательности разреза осадочных образований. Однако обоснование применимости методов исследования последовательности событий к пространственным данным является очень трудоемким и базируется на предположении о постоянстве скорости образования пространственной последовательности. Это предположение, вероятно, выполняется в первом примере, но во втором требуется установить дополнительное условие, заключающееся в том, что скорость осадконакопления постоянна в пределах данной последовательности.

Исторические записи извержений вулкана Асо в Киушу, Япония, велись с 1229 г. и представлены в табл. 4.6. Асо — это

Таблица 4.6

Даты извержений вулкана Асо в период 1229—1962 гг.

Годы				
1229	1376	1583	1780	1927
1239	1377	1584	1804	1928
1240	1387	1587	1806	1929
1265	1388	1598	1814	1931
1269	1434	1611	1815	1932
1270	1438	1612	1826	1933
1272	1473	1613	1827	1934
1273	1485	1620	1828	1935
1274	1505	1631	1829	1938
1281	1506	1637	1830	1949
1286	1522	1649	1854	1950
1305	1533	1668	1872	1951
1324	1542	1675	1874	1953
1331	1558	1683	1884	1954
1335	1562	1691	1894	1955
1340	1563	1708	1897	1956
1346	1564	1709	1906	1957
1369	1576	1765	1916	1958
1375	1582	1772	1920	1962

сложный стратовулкан, все его извержения относились к взрывному типу, и при этом выбрасывалось огромное количество вулканического пепла. Хотя старые, регулярно проводившиеся записи содержат указания на относительную мощность и продолжительность извержений, для практических целей мы можем считать записи как относящиеся к событиям, произошедшим мгновенно. Анализ истории вулкана может пролить некоторый свет на природу механизма извержений и может привести к построению физической модели вулканов [61]. Конечно, мы можем также надеяться, что такое изучение может привести к появлению технологии предсказания извержений в будущем.

Изучение рядов событий преследует несколько объективных целей. Обычно исследователя интересует средняя частота появления событий, т. е. число событий за некоторый интервал времени. Кроме того, бывает необходимо исследовать ряды событий более детально. Цель такого исследования — выявление какой-либо закономерности, которой могут подчиняться события. Дополнительная информация может быть использована для уточнения частоты появления событий, для определения особенностей выборочной схемы, для обнаружения тренда и для установления других систематических свойств рядов.

Так как ряды событий имеют очень простой вид в том смысле, что они состоят из характеристик типа «да — нет», то для их изучения можно использовать простые и в то же время очень мощные аналитические методы. Кокс и Льюис рассматривают множество графических методов, полезных для исследования рядов событий. Эти методы иллюстрируются в применении к данным извержений вулкана Асо, представленным в табл. 4.6.

На рис. 4.5 изображена кумулятивная кривая общего числа извержений t, произошедших вплоть до момента времени t, она соответствует на графике точке с абсциссой t. Этот рисунок хорошо отражает изменения в средней скорости появления событий. Наклон прямой, соединяющей любые две точки на кумулятивной кривой, равен среднему числу событий за единицу времени, в качестве которой выбран интервал между этими двумя точками.

На рис. 4.6 представлена гистограмма числа извержений, происходящих в последовательные равные интервалы времени. Гистограмма прямо указывает локальные периоды флюктуации относительно средней скорости появления извержений. Из рисунка видно, что гистограмма чувствительна к длине выбранного интервала, поэтому при анализе рядов бывает полезно иметь больше чем одну гистограмму.

Эмпирическая функция деятельности вулкана получается, если представить в процентах зависимость Y (отношения числа временных интервалов длиннее X к общему числу интервалов) от X (длины временного интервала). Полученная функция ха-

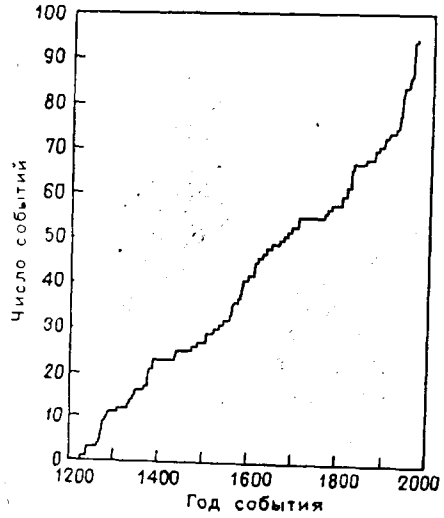


Рис. 4.5. Кумулятивная кривая числа извержений вулкана Асо, отнесенных к году извержения

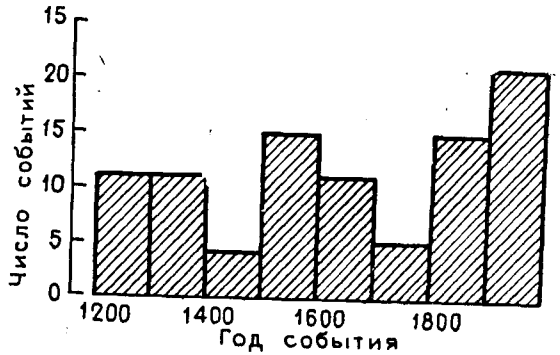


Рис. 4.6. Гистограмма числа извержений вулкана Асо, произошедших в последовательные столетние интервалы

теризует вероятность того, что событие не произошло раньше момента времени X. На рис. 4.7 представлено процентное отношение числа интервалов между извержениями, которые превосходят некоторое заданное число лет. Если события происходят случайно во времени, то функция деятельности будет иметь экспоненциальную форму.

Ту же самую функцию можно изобразить в логарифмическом масштабе, используя log Y как функцию X. Логарифмическая эмпирическая функция деятельности особенно удобна для изучения отклонений от случайной величины, которые представляются на графике как уклонения от прямой линии (рис. 4.8).

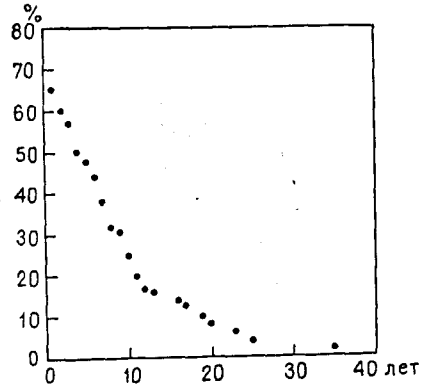


Рис. 4.7. Эмпирическая функция деятельности вулкана Асо.

По вертикальной оси указаны проценты от числа интервалов между извержениями, имеющими продолжительность больше некоторой заданной, по горизонтальной — продолжительность интервала

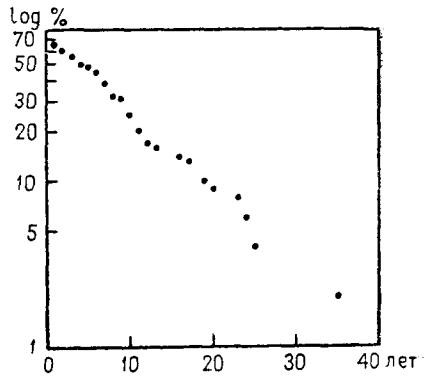


Рис. 4.8. Логарифм эмпирической функции деятельности вулкана Асо.

Вертикальная ось рис. 4.7 здесь представлена в логарифмическом масштабе

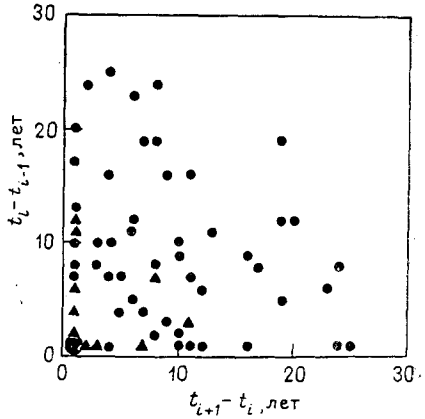


Рис. 4.9. Сериальная корреляция продолжительностей между последовательными изображениями вулкана Асо.

На вертикальной оси представлена продолжительность покоя перед j-м извержением, на горизонтальной оси — после j-го извержения. Треугольниками представлены скопления более чем одной точки в одном и том же месте

Диаграмма рассеяния сериальной корреляции или автокорреляции первого порядка последовательных интервалов между событиями представлена на рис. 4.9. Степень соответствия между длиной некоторого интервала и длиной непосредственно предшествующего интервала представлена на графике точками с координатами $X_i = t_{i+1} - t_i$, $Y_i = t_i - t_{i-1}$, где t_i — время появления события с номером i. На этом графике не удается обнаружить какой-либо закономерности в следовании одного интервала за другим при одинаковой длине. Такая диаграмма рассеяния с

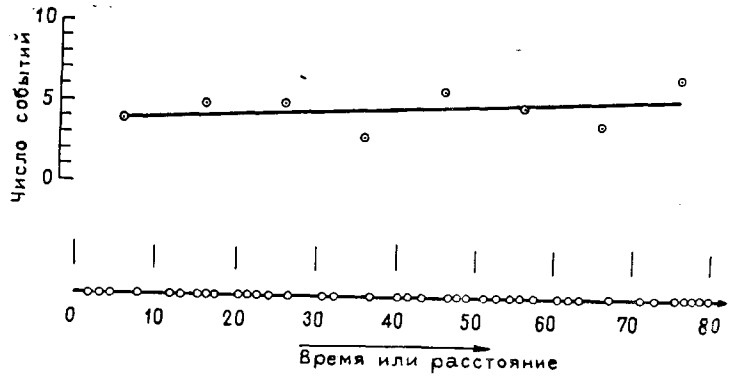


Рис. 4.10. Последовательности событий, которые происходят «мгновенно» в пространственном или временном континууме.

Шкала времени или расстояния разделена на 10 отрезков, каждому из которых поставлено в соответствие число событий. Изображенная вверху линия есть прямая регрессия числа событий на отрезок по отношению к серединам отрезков

большой дисперсией и относительно высокой концентрацией точек вблизи осей типична для рядов случайных событий.

В большинстве исследований последовательностей событий рассчитывают на то, что удастся описать основные черты рассматриваемого явления так, чтобы был вскрыт физический механизм пространственного или временно́го расположения событий. Сначала мы должны рассмотреть возможность проявления тренда в исходных данных, что можно сделать двумя способами.

Последовательность можно разделить на некоторое число участков равной длины, так что каждый из них содержит несколько наблюдений. Число событий в пределах каждого участка ставится в соответствие средней точке рассматриваемого участка. Выбирая в качестве зависимой переменной Y_i значения координат центров участков, а числа событий в пределах сегмента в качестве значений аргумента X_i, можно построить линию регрессии. Ее угловой коэффициент может быть проверен на основе критерия ANOVA, приведенного в табл. 4.12 с целью определения, значимо ли он отличается от нуля. Этот процесс проиллюстрирован на рис. 4.10. К сожалению, этот критерий не очень эффективен, так как при разделении последовательности на отрезки теряется некоторое число степеней свободы.

Имеются критерии, специально предназначенные для обнаружения тренда в скорости осуществления событий, в которых используется метод сравнения средней точки последовательности с ее центроидом. Если последовательность относительно однородна, эти ряды будут очень похожи, но если имеется тренд, то центроид будет смещаться в направлении увеличения

скорости появления событий. Если t_i — время или расстояние от начала ряда до i-го события и N — общее число событий, то мы можем вычислить центроид S по формуле

$$S = \frac{1}{N}\sum_{i=1}^{N} t_i. \qquad (4.3)$$

Эту статистику в свою очередь можно использовать в критерии (4.4):

$$Z = \frac{\sqrt{12N}}{T}\left(S - \frac{1}{2}T\right), \qquad (4.4)$$

где T — общая длина последовательности; Z — стандартизированная нормальная случайная величина. Значимость критерия может быть установлена по таблице нормального распределения, аналогичной табл. 2.10.

Этот критерий очень чувствителен к изменению скорости появления событий. Например, если появление событий можно описать формулой

$$Y_t = e^{\alpha+\beta t}, \qquad (4.5)$$

то нулевую гипотезу можно записать как равенство $\beta = 0$. Если в результате проверки мы устанавливаем, что модель экспоненциальная и что β отлично от нуля, то скорость появления событий V_t изменяется с изменением t. Именно реализацию этой возможности мы и проверяем.

Если в скорости появления событий тренд не обнаружен, то можно сделать вывод, что последовательности событий можно рассматривать как стационарные. Следующее, что следует проверить, — это предположение о независимости последовательных событий. Это можно сделать с помощью вычисления автокорреляционной функции для длин интервалов между событиями. Иными словами, надо рассматривать интервалы между событиями как переменную X_i, принимающую значение в точках с равномерным расположением в пространстве. Если интервалы не являются независимыми, то должна обнаружиться тенденция к тому, чтобы большие значения X (длинные интервалы между событиями) следовали за большими значениями. Аналогично должна быть и тенденция к тому, чтобы малые значения X (короткие интервалы) следовали за другими малыми значениями. Мы можем вычислить автокорреляцию для последовательных значений лага и проверить их значимость. Обычно только первые несколько значений лага представляют интерес. Если изложенные выше методы позволяют установить, что значения автокорреляционной функции несущественно отличаются от нуля, то мы можем заключить, что события происходили независимо во времени или пространстве.

Если мы установили, что последовательности не имеют ни тренда, ни автокорреляции, можно попытаться проверить гипотезу о том, что события подчинены распределению Пуассона.

Напомним, что в гл. 2 распределение Пуассона определялось как дискретное вероятностное распределение, которое можно считать предельным случаем биномиального при условии, что n (число испытаний) становится очень большим, а p (вероятность успеха в одном испытании) становится очень малой. Мы можем представить себе, что наш временный ряд подразделяется на n интервалов равной длины. Если события происходят случайно, то число интервалов, которые содержат в точности нуль, одно, два, ..., x событий, будет подчиняться биномиальному распределению.

Если мы начнем уменьшать длины интервалов, n будет увеличиваться, а вероятности событий будут уменьшаться. Биномиальное распределение в этом случае уже не годится для подсчетов, и удобнее пользоваться распределением Пуассона, так как в нем не требуется точных сведений о величинах n и p. Вместо этого требуется знать произведение $np=\lambda$, которое здесь характеризует скорость появления событий. Пуассоновская вероятностная модель основана на следующих допущениях: а) события происходят независимо; б) вероятность появления события не изменяется с течением времени; в) вероятность появления события в некотором интервале пропорциональна длине этого интервала; г) вероятность того что более чем два события произойдут в одном и том же временном интервале, исчезающе мала.

Уравнение для определения пуассоновского распределения в этом примере имеет вид

$$P(X) = e^{-\lambda}\lambda^X/X!. \tag{4.6}$$

Заметим, что скорость появления событий λ здесь является только параметром распределения. Типичные пуассоновские частотные распределения представлены на рис. 4.11. Распределение Пуассона применяется при решении таких задач, как определение частоты телефонных вызовов на коммутаторе или определение промежутка времени между сбоями в вычислительной системе. Кажется вполне возможным его использование при изучении рядов геологических событий, описанных в начале этого раздела. Если мы смогли установить, что наша последовательность подчиняется пуассоновскому распределению, то мы можем использовать характеристики этого распределения для вероятностного прогнозирования данного ряда.

Критерий Колмогорова—Смирнова обеспечивает нам простой способ проверки соответствия распределения ряда событий пуассоновскому распределению. Сначала ряды переводятся в

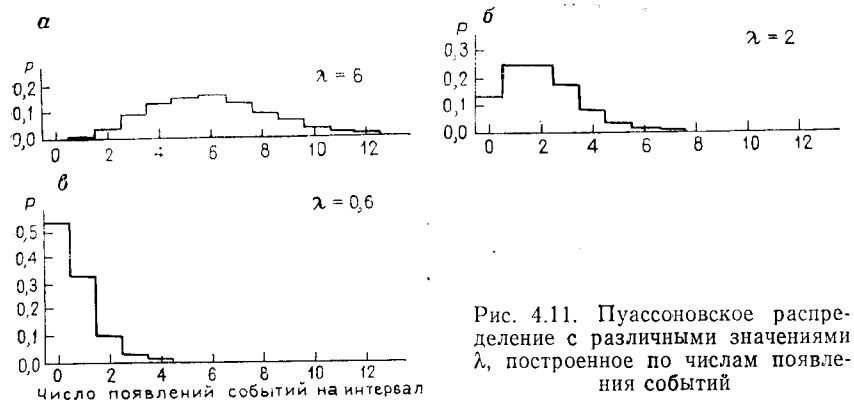


Рис. 4.11. Пуассоновское распределение с различными значениями λ, построенное по числам появления событий

кумулятивную форму с помощью преобразования

$$Y_i = t_i/T,$$

где t_i — время от начала последовательности до i-го события; T — общая длина ряда. Затем вычисляются три оценки:

$$1.\quad KC^+ = \sqrt{n}\max\left\{\frac{i}{n} - Y_i\right\},$$

$$2.\quad KC^+ = \sqrt{n}\max\left\{Y_i - \frac{i-1}{n}\right\}, \tag{4.7}$$

$$3.\quad KC^+ = \max|KC^+, KC^-|.$$

Первый критерий — это просто максимум положительных разностей между наблюденным рядом и ожидаемым, исходя из пуассоновского распределения; второй — это максимум отрицательной разности и третий — это большее из абсолютных значений двух предыдущих. Проверяемая статистика KC затем сравнивается с двухсторонними критическими значениями, приведенными в табл. 2.26. Если статистика превышает критическое значение, то максимальное отклонение больше, чем ожидаемое в выборке, полученной случайным образом из пуассоновского распределения.

Маури Шейл — это черные, содержащие кремний, глины раннего мелового возраста, встречающиеся на территории штатов Колорадо, Вайоминг и Монтана. Интервал характеризуется многочисленными бентонитовыми слоями, которые залегают в нескольких местах в Вайоминге и Монтане и обнаруживаются при бурении на ил и литейную глину. Бентониты состоят из монтмориллонита, возникающего как продукт разрушения вулканического пепла риолитового и андезитового состава. В табл. 4.7 приводятся значения мощности интервалов между последовательными бентонитовыми слоями, измеренными в обнажении

Таблица 4.7

Мощность (в футах) интервалов между последовательными бентонитовыми слоями в меловых отложениях Маури Шейл во Фремонт Кантри (штат Вайоминг)

(Верх)	47	6
29	7	3
11	8	17
6	23	4
5	10	5
10	15	4
5	2	26
17	35	4
14	4	(Низ)

Маури Шейл во Фремонт Кантри (штат Вайоминг). Эти слои представляют скопления пепла в результате бурных извержений вулканов в Идахо. Предполагается, что содержащаяся там черная глина откладывалась с постоянной скоростью, и эту последовательность значений мощности можно анализировать как ряд событий, аналогичный историческому ряду, образованному извержениями Аса.

Проверьте эти данные на тренд в значениях скорости появления. Если тренд не будет обнаружен, проверьте на автокорреляцию последовательные интервалы между событиями. Объясните: а) возможные эффекты от неодинаковых скоростей осадконакопления черной глины и б) возможность проявления активности более одного вулкана.

КРИТЕРИЙ СКАЧКОВ

Простейшая последовательность — это последовательность наблюдений, расположенных в порядке их появления, причем такая, что каждый ее элемент принадлежит одному из двух взаимоисключающих друг друга состояний. Рассмотрим трещиноватую породу с конкрециями с целью поиска в них ископаемых остатков. Дробление конкреций является испытаниями, причем каждое из них имеет два взаимоисключающих исхода: конкреция либо содержит ископаемые остатки, либо нет. Последовательность таких исходов при изучении данной породы в течение дня составляет временной ряд специального типа. Мы можем построить аналогичную последовательность экспериментально, бросая монету и отмечая выпадение герба или решки. Полученная последовательность будет напоминать следующий ряд 20 событий:

ГРГГРГРРРГРГРГГРРГГГ.

Конечно, интуиция подсказывает нам, что в этой последовательности должно появиться около десяти гербов, и мы можем определить вероятность выпадения этого (или любого другого) числа гербов. В нашем примере мы получили 11 гербов; считая, что монета правильная, мы получаем вероятность выпадения этого числа гербов в последовательности 20 испытаний, равную 0,16, или приблизительно 1/6. В эксперименте, аналогичном рассмотренному, мы можем ожидать 9, 10 и 11 выпадений гербов, т. е. немногим больше одной трети от числа испытаний. Результаты этого эксперимента подчиняются биномиальному распределению, рассмотренному в гл. 2.

Однако при этом мы не учли порядка, в котором появляются гербы. Вероятно, если бы последовательность выглядела таким образом:

ГГГГГГГГГГРРРРРРРРРР,

то это показалось бы нам очень странным, хотя вероятность получения такого же количества гербов в двадцати испытаниях такая же, как и в предыдущем примере. Другой крайний случай — попеременное появление гербов и решек:

ГРГРГРГРГРГРГРГРГРГГ,

тоже выглядит очень необычно, хотя вероятность выпадения данного числа гербов осталась неизменной. В этих примерах наше подозрение вызывает не пропорция, в которой выпадают гербы, а порядок их появления. Мы предполагаем, что выпадение герба или решки случайно, а в двух последних примерах это предположение кажется весьма неправдоподобным.

Мы можем проверить гипотезу о случайности этой последовательности путем исследования числа скачков. Скачки определяются как непрерывающиеся последовательности одних и тех же состояний. Первая последовательность содержит 13 скачков, вторая — только 2, и третья — 19. Скачки в первой последовательности подчеркнуты:

$$\underset{1}{\underline{\text{Г}}}\ \underset{2}{\underline{\text{Р}}}\ \underset{3}{\underline{\text{ГГ}}}\ \underset{4}{\underline{\text{Р}}}\ \underset{5}{\underline{\text{Г}}}\ \underset{6}{\underline{\text{РРР}}}\ \underset{7}{\underline{\text{Г}}}\ \underset{8}{\underline{\text{Р}}}\ \underset{9}{\underline{\text{Г}}}\ \underset{10}{\underline{\text{Р}}}\ \underset{11}{\underline{\text{ГГ}}}\ \underset{12}{\underline{\text{РР}}}\ \underset{13}{\underline{\text{ГГГ}}}.$$

Мы можем вычислить вероятность того, что данную последовательность скачков в эксперименте с двумя исходами (герб или решка в этом примере) можно считать случайной. С этой целью вычисляется число возможных размещений n_1 состояний 1 и n_2 состояний 2. Общее число скачков в последовательности обозначается через U; имеются таблицы, которые содержат критические значения числа скачков U для фиксированных n_1, n_2 и заданного уровня значимости α. Однако если каждое из значений n_1 и n_2 превышает десять, то распределение величины U довольно хорошо аппроксимируется нормальным распределением, и мы при использовании этого статистического критерия

можем использовать таблицы стандартного нормального распределения. Среднее число скачков в случайной последовательности с n_1 успехами и n_2 неудачами равно

$$\bar{U} = \frac{2n_1n_2}{n_1+n_2} + 1. \tag{4.8}$$

Дисперсия среднего числа скачков вычисляется по формуле

$$\sigma^2_{\bar{U}} = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}. \tag{4.9}$$

Указанные формулы позволяют определить среднее число скачков и стандартную ошибку среднего числа скачков при всевозможных размещениях n_1 и n_2 взаимоисключающих друг друга исходов. Вычислив указанные характеристики, мы можем по формуле (4.10) получить критерий Z:

$$Z = (U - \bar{U})/\sigma_{\bar{U}}, \tag{4.10}$$

где U — наблюдаемое число скачков.

Легко убедиться, что это просто формула (2.28), переписанная для величины U, относительно которой можно сформулировать и проверить ряд статистических гипотез. Например, при случайном размещении нам может потребоваться проверка гипотезы о том, что последовательность содержит более чем среднее число скачков; в этом случае нулевая гипотеза и альтернатива могут быть записаны так:

$$H_0 : U \leqslant \bar{U},$$
$$H_1 : U > \bar{U},$$

т. е. слишком большое число скачков приводит к отклонению гипотезы. Этот критерий является односторонним. Наоборот, мы можем пожелать определить, не содержит ли последовательность невероятно малое число скачков. В этом случае соответствующая гипотеза и альтернатива записываются в виде

$$H_0 : U \geqslant \bar{U},$$
$$H_1 : U < \bar{U}$$

и слишком низкое число скачков приведет к отклонению нулевой гипотезы. Этот критерий также является односторонним. Если мы хотим отклонить гипотезу о любом нарушении случайности в последовательности, то для этой цели подходит двусторонний критерий для проверки гипотезы:

$$H_0 : U = \bar{U}$$

при альтернативе

$$H_1 : U \neq \bar{U}.$$

Мы можем применить этот критерий для проверки гипотезы о случайном расположении элементов первой последовательности, содержащей результаты испытаний при 20 бросаниях монеты. Нулевая гипотеза утверждает, что нет существенного различия между наблюдаемым числом скачков и средним числом скачков для случайной последовательности того же объема. Мы используем для ее проверки двусторонний критерий. Нулевая гипотеза отвергается, если в этой последовательности имеется либо слишком мало, либо слишком много скачков и принимается альтернатива

$$H_1 : U \neq \bar{U}.$$

Выбрав 5%-ный уровень значимости ($\alpha = 0{,}05$), мы получим границы критической области $-1{,}96$ и $+1{,}96$. Сначала вычислим среднее значение и стандартное отклонение для числа скачков в случайной последовательности, имеющей n_1 гербов ($n_1 = 11$) и n_2 решек ($n_2 = 9$):

$$\bar{U} = \frac{2 \cdot 11 \cdot 9}{11+9} + 1 = 10{,}9,$$

$$\sigma^2_{\bar{U}} = \frac{(2 \cdot 11 \cdot 9)(2 \cdot 11 \cdot 9 - 11 - 9)}{(9+11)^2(9+11-1)} = 4{,}6.$$

Статистика Z равна

$$Z = \frac{U - \bar{U}}{\sigma_{\bar{U}}} = \frac{13 - 10{,}9}{2{,}1} = 1{,}0.$$

Таким образом, число скачков в последовательности меньше стандартного отклонения от среднего значения всех возможных скачков в такой последовательности и не попадает в критическую область. Следовательно, указанное число скачков не дает оснований для отклонения нулевой гипотезы и принятия предположения, что последовательность не является случайной. Другие последовательности, наоборот, дают совершенно отличные значения критерия. Так как n_1 и n_2 одинаковы для всех трех последовательностей, то $\bar{U}$ и $\sigma_{\bar{U}}$ также одинаковы. Для второй последовательности значения критерия

$$Z = \frac{2 - 10{,}9}{2{,}1} = -4{,}2,$$

для третьей

$$Z = \frac{19 - 10{,}9}{2{,}1} = 3{,}9.$$

Оба эти значения расположены за критическими пределами, и мы должны отклонить гипотезу о случайном расположении элементов последовательностей.

Геологические применения этого критерия не вполне очевидны, так как обычно приходится рассматривать последовательности с числом состояний, большим двух. Стратиграфические разрезы или, например, пересечения шлифа обыкновенно содержат не менее трех состояний, которые нельзя ранжировать никаким осмысленным образом. Мы рассмотрим способы, с помощью которых часть последовательностей можно привести к последовательности дихотомических состояний, но прежде мы остановимся на геологическом примере применения критерия скачков при изучении системы с двумя состояниями.

Обычные пегматиты образуются при кристаллизации остаточного расплава, обогащенного летучими веществами при отвердении гранитной магмы. Их структура обусловлена одновременной кристаллизацией кварца и полевого шпата в эвтектической точке. Если кристаллизация пегматита происходит без помех, то можно допустить, что зерна кварца и полевого шпата возникают в случайных точках внутри охлаждающего расплава. Эта ситуация (случайное образование зерен) остается неизменной до тех пор, пока расплав затвердеет. Однако присутствие одного кристалла, например полевого шпата, может стимулировать дополнительное образование зерен полевого шпата и привести к возникновению пестрой структуры. Наоборот, рост одного кристалла может локально исчерпать из магмы нужные составляющие и приостановить кристаллизацию, в результате чего возникает пестрая мозаика из кварца и полевого шпата. Большую плиту полированного пегматита можно рассматривать как окно в геологическую кухню, в которой студентам дана возможность изучения этих альтернативных процессов. Полированная поверхность породы позволяет легко установить контакты между слагающими ее зернами, поэтому линия, проведенная на ней, приводит к построению последовательности зерен кварца и полевого шпата. Линия на полированной плите может рассматриваться как случайная выборка из возможных последовательностей в теле пегматита, из которого была извлечена эта плита. Последовательность зерен кварца и полевого шпата вдоль линии указана в табл. 4.8. Наша задача — изучить скачки от кварца к полевому шпату, проверить случайность последовательности и определить, нет ли тенден-

Таблица 4.8

Последовательность 100 зерен полевого шпата (П) и кварца (К), полученная при пересечении пегматита

(Начало)

ПККПКККППКПКПППППППППКККПКПППКПППППКПППК
КПКПККККППППППКПППППППКККККППКККПППППППК
ПКПППППКПКПКППКПППППППКПППКККПКПКППК

(Конец)

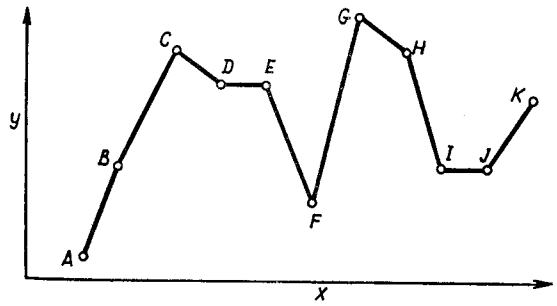


Рис. 4.12. Последовательность наблюдений, которая анализируется методом скачков вверх и вниз

ции к систематическому следованию одного состояния за самим собой или, наоборот, тенденции одного состояния следовать непосредственно за другим. Выполните исследование этих данных с помощью критерия скачков и оцените три альтернативы.

Теперь рассмотрим статистическую процедуру исследования скачков вверх и вниз. Она используется в тех случаях, когда мы имеем дело не с двумя различными состояниями, а когда последующее наблюдение больше или меньше, чем предыдущее. На рис. 4.12 изображена типичная последовательность, которую можно проанализировать методом скачков вверх и вниз.

Отрезок *ABC* изображает скачок вверх, так как каждое наблюдение превосходит предыдущее; аналогично отрезок *GHI* изображает скачок вниз. Отрезок *CDEF* изображает скачок вниз, несмотря на то, что разность между *D* и *E* равна нулю. Действительно, интервал *DE* расположен между двумя отрезками *CD* и *EF*, каждый из которых изображает скачок вниз, поэтому и весь участок *CDEF* можно рассматривать как единый скачок вниз. Интервал *IJ* можно рассматривать либо как часть отрезка *GHIJ*, изображающего скачок вниз, либо как часть отрезка *IJK*, изображающего скачок вверх, причем общее число скачков остается в любом случае неизменным. Если каждое наблюдение выражено некоторой величиной, то обычно она имеет дробную часть, и одинаковые значения (две последовательные точки с одинаковыми характеристиками) практически невозможны.

При рассмотрении только разностей между значениями в соседних точках мы преобразуем данную последовательность в последовательность, имеющую только два состояния (или три, если имеются равные значения). Последовательность, изображенную на рис. 4.12, мы можем переписать в виде

$$+\,+\,+\,-\,0\,-\,+\,-\,-\,0\,+\,.$$

Считая первый нуль минусом, мы получаем пять скачков: три

Таблица 4.9

Число радиолярий на 1 см² шлифа кремнистого сланца

(Основание разреза)	1	2	10
	2	2	12
	3	1	14
	2	0	22
	3	2	17
	5	3	19
	7	2	14
	9	0	4
	9	3	2
	11	3	1
	10	4	0
	12	9	0
	7	10	8
	4	10	14
	3	8	16
	2	9	27
	3	12	(Верх разреза)

от плюса к минусу и два от минуса к плюсу (число скачков не зависит от того, назовем ли мы второй нуль плюсом или минусом). Теперь мы можем применить процедуры, рассмотренные выше для случая последовательностей с двумя взаимоисключающими друг друга состояниями — см. формулы (4.8)—(4.10). Для того чтобы аппроксимация нормальным распределением была оправдана, необходимо иметь большую выборку, однако в большинстве геологических задач такие объемы выборок вполне доступны.

При изучении кремнистых сланцев в Скалистых горах было отмечено, что эта порода содержит необычно много хорошо сохранившихся остатков радиолярий. Их присутствие в сланцах скорее всего не является случайным, так как последовательность образцов была собрана на приблизительно равных расстояниях по разрезу. Из образцов были сделаны шлифы, в которых на площадке 10×10 мм² было подсчитано число радиолярий. Данные для 50 образцов приведены в табл. 4.9. Можно ли считать, что распространенность радиолярий изменяется в шлифе случайно? Вполне реально составить программу, которая выполнила бы все необходимые вычисления, однако усилия, которые требуется при этом затратить на программирование, по-видимому, превысят трудности, которые придется преодолевать при расчетах вручную.

В этом случае дихотомизация наблюдений достигается с помощью сравнения их величин с предыдущими наблюдениями. В действительности критерий скачков может быть применен к данным, дихотомизация которых осуществляется по произвольной схеме при условии, что проверяемая гипотеза может быть представлена дихотомически. Например, известная процедура

Таблица 4.10

Значения удельного веса образцов, собранных при пересечении магнетитового тела (хребет Ларами, штат Вайоминг)

(Западная часть)	3,57	4,58	4,22	
	3,63	5,02	3,52	
	2,86	4,68	2,91	
	2,94	4,37	3,87	
	3,42	4,88	3,52	
	2,85	4,52	3,77	
	3,67	4,80	3,84	
	3,78	4,55	3,92	
	3,86	4,61	4,09	
	4,02	4,93	3,86	
	4,56	4,60	4,13	
	4,62	4,51	3,92	
	4,31	3,98	3,54	(Восточная часть)

дихотомизации ряда наблюдений состоит в вычитании каждого наблюдения из медианы, вычисленной по всем наблюдениям, после чего проводится проверка гипотезы о случайности последовательности скачков относительно медианы с помощью критерия знаков.

На большой площади распространения докембрийских анортозитов в Ларамийском хребте в штате Вайоминг наблюдается несколько магнетитовых тел. Одно из них было вскрыто, и порода дробилась для использования в качестве добавки к буровому раствору. В карьере были отобраны образцы в точках, равномерно расположенных на пересечении магнетитового тела. Собранные образцы различались между собой; одни содержали преимущественно плагиоклаз, другие — оливин, некоторые почти целиком состояли из магнетита, а другие представляли собой смесь их трех минералов. Несоответствие ряда членов заставляет предположить, что в магнетитовом теле имеют место систематические изменения. Чтобы проверить это предположение, были проведены измерения удельного веса образцов, результаты которых приведены в табл. 4.10. Можно ли считать, что изменение удельного веса вдоль пересечения соответствует тому изменению, которого можно было бы ожидать в предположении, что состав образцов изменялся случайно относительно центрального значения?

Мы можем также провести проверку гипотезы о случайности скачков по отношению к среднему значению. Результаты этой проверки будут использованы в этой главе в разделе, посвященном тренд-анализу. Критерии скачков принадлежат к широкому классу непараметрических процедур, рассмотренных в гл. 2.

Имеются многочисленные разновидности критериев скачков, рассмотренных выше. Информацию о них читатель может по-

черпнуть в руководствах по непараметрическим статистическим методам Брэдли [8], Коновера [10], Зигеля [50]. Примеры использования критерия скачков в геологии имеются в книге Миллера и Кана [36]. Некоторые из этих авторов считают длину самого большого скачка показателем неслучайности, другие используют число точек инверсии, т. е. точек, в которых знаки последовательных наблюдений меняются. В некоторых случаях эти критерии могут оказаться более подходящими, чем процедуры, описанные выше. Вообще говоря, процедура исследования скачков вверх и вниз считается наиболее мощным приемом из критериев скачков, так как она использует изменение величины в каждой точке по отношению к прилегающим точкам. Другие дихотомические схемы отражают только изменения по отношению к одному значению, например к медиане или среднему значению.

Критерии скачков целесообразно применять в тех случаях, когда требуется выяснить причину нарушения случайности. Наличие слишком большого или слишком малого числа скачков позволяет выявить места нарушения случайности и не отождествлять их с трендом. Необходимо отметить, что сам по себе факт случайности не может быть доказан, так как условие случайности содержится в нулевой гипотезе. Мы можем только утверждать при некотором заданном уровне значимости, что нулевая гипотеза неверна и что по этой причине последовательность не является случайной. Иными словами, если наши попытки проверить неслучайность окончились провалом, то нам ничего не остается больше, кроме принятия нулевой гипотезы. В дальнейшем мы рассмотрим процедуры обнаружения тренда или систематических изменений среднего значения. Мы будем иметь возможность убедиться в том, что критерии скачков в сочетании с этими методами оказываются весьма полезными.

МЕТОД НАИМЕНЬШИХ КВАДРАТОВ И РЕГРЕССИОННЫЙ АНАЛИЗ

Во многих задачах нас могут интересовать не только имеющиеся в последовательности изменения, но также те точки, в которых эти изменения происходят. Для решения этих задач нужно иметь набор измерений изучаемой переменной, а также знать расположение точек этих измерений. Как измеряемая переменная, так и шкала, в которой в соответствующих единицах выражены элементы последовательности, должны иметь определенный размах. Оказывается, нам недостаточно простой информации о порядке следования точек. В большинстве примеров, которые мы сейчас рассмотрим, нас будет интересовать общий характер изменения данных. Информация об этом будет использована при интерполяции между данными точками для экстраполяции значений, расположенных за пределами данной

Таблица 4.11

Влажность современных илов в пробах керна, взятых на побережье Мексиканского залива, Луизиана

Глубина, футы	Влажность (граммы воды/100 г сухого осадка)	Глубина, футы	Влажность (граммы воды/100 г сухого осадка)
0	124	20	30
5	78	25	21
10	54	30	22
15	35	35	18

последовательности, для получения выводов о влиянии тренда или для получения оценок характеристик, которые могут быть интересны геологам. Если относительно распределения совокупности, из которой взяты выборки, можно сделать некоторые обоснованные предположения, то к ним можно применить статистический метод, называемый регрессионным анализом.

Данные табл. 4.11 представляют значения влажности в пробах керна современных морских илов побережья Мексиканского залива в Восточной Луизиане. Измерения получены в результате сравнения массы проб немедленно после взятия их из пробоотборника и после тщательного высушивания. Если мы сопоставим сделанные измерения и соответствующие им глубины, как это сделано на рис. 4.13, то увидим, что содержание влаги быстро падает с глубиной в верхних частях слоя ила и медленно убывает, почти стабилизируясь, в осадке вблизи основания слоя. Рассмотрим теперь различные способы исследования и записи неявных соотношений между этими наблюдениями.

Значение 47,75, указанное на рис. 4.13, — среднее содержание влаги в пробах — представляет собой точку, относительно которой дисперсия минимальна, т. е. минимальна сумма квадратов отклонений содержаний влаги относительно этой точки. Читатель должен помнить (см. гл. 2), что если некоторые пробы вызывают сомнение, то их можно заменить несмещенной и эффективной оценкой выборочного среднего, являющегося наилучшим предсказанием для дополнительных проб, которые могут быть извлечены из той же совокупности. Однако ясно, что среднее значение не может адекватно представлять данные рис. 4.13. Пробы отбирались последовательно, и потому они не являются независимыми. Еще более, чем точечная оценка, нам подошла бы прямая линия, которая выразила бы связь между содержанием влаги и глубиной на всем множестве изменения значений переменных. Интуитивные соображения подсказывают нам, что в качестве такой линии можно было бы выбрать прямую, отклонения которой от данных значений можно свести

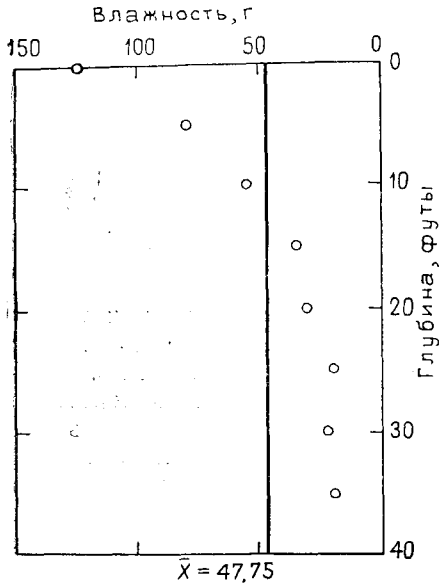


Рис. 4.13. Зависимость влажности осадка от глубины (в граммах воды на 100 г сухого осадка).

Данные собраны в скважине, пробуренной в современных илах на побережье Мексиканского залива. Отметим, что ориентация графика не соответствует ориентации, обычно используемой в математике

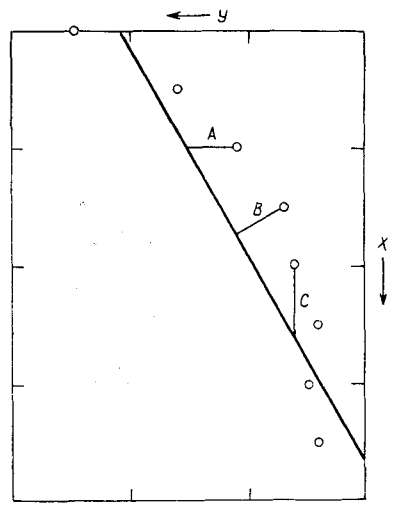


Рис. 4.14. Различные варианты критериев минимизации отклонений от линии аппроксимации:

A — минимизация отклонения влажности;
B — минимизация обобщенных отклонений;
C — минимизация отклонений глубины

до минимума. Если рассуждать по аналогии со средним, то один из способов состоит в минимизации суммы квадратов отклонений от прямой. (Среднее — это значение, относительно которого дисперсия и, следовательно, сумма квадратов отклонений, является наименьшей). Мы можем построить единственную прямую, относительно которой дисперсия минимальна. Если значения этой линейной функции в данных точках вычесть из соответствующих наблюдаемых значений, то полученное в результате множество чисел будет иметь среднее значение, равное нулю, и меньшую дисперсию, чем набор отклонений от любой другой прямой, построенной по данным точкам.

Имеется, однако, несколько способов определения и измерения отклонений от подбираемой линии. Например, мы можем рассмотреть отклонения значений влажности, отклонения глубин или некоторую их комбинацию. На рис. 4.14 отрезок *A* изображает отклонение содержаний влажности от подобранной прямой, а отрезок *C* — отклонение значения глубины от той же прямой. Отклонение *B* измерено по перпендикуляру к ней. Можно было бы построить прямые, используя любой из этих

способов измерения отклонений, но мы ограничимся лишь замечаниями по поводу каждого из этих способов. Если наша задача будет заключаться в минимизации отклонений содержаний влаги, то мы получим прямую, представляющую наилучшую оценку влажности при заданных глубинах. Наоборот, если задача будет состоять в минимизации отклонений глубин, то мы получим наилучшую оценку зависимости глубины от содержаний влаги. Третья альтернатива позволяет выразить связь между двумя переменными. В специальном наборе задач, рассматриваемых в этой главе, временны́е или пространственные интервалы считаются известными, а вторая переменная имеет непрерывное распределение. Поэтому первая альтернатива кажется наиболее подходящей для наших целей. Иными словами, содержание влаги Y рассматривается как случайная переменная, а глубина X фиксируется. Поэтому задача состоит в предсказании значений Y по значениям X. Другие случаи будут рассмотрены в следующих главах этой книги.

После того как мы условились о характеристиках прямой тренда, которую мы хотим построить, определим некоторые термины. Изучаемая переменная является зависимой (т. е. функцией) или регрессионной и обозначается Y_i. Отклонения Y_i от прямой линии должны быть минимальными. Другая переменная является независимой (или аргументом) и обозначается X_i. Пусть аппроксимирующая прямая пересекает ось Y в точке b_0 и имеет угловой коэффициент b_1. Тогда ее уравнение имеет вид

$$\widehat{Y}_i = b_0 + b_1 X_i, \tag{4.11}$$

где $\widehat{Y}_i$ — оценка для Y_i при данном значении X_i. Рассматриваемое отклонение равно $\widehat{Y}_i - Y_i$, и наша задача сводится к нахождению такой прямой, для которой сумма квадратов отклонений

$$\sum_{i=1}^{n} (\widehat{Y}_i - Y_i)^2 = \text{минимум}. \tag{4.12}$$

Получение окончательного результата требует применения дифференциального исчисления, поэтому мы не будем рассматривать доказательство, а ограничимся тем, что приведем так называемые нормальные уравнения, позволяющие найти значения b_0 и b_1 для аппроксимирующей прямой. Они имеют вид

$$\sum_{i=1}^{n} Y_i = b_0 n + b_1 \sum_{i=1}^{n} X_i; \tag{4.13}$$

$$\sum_{i=1}^{n} X_i Y_i = b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2. \tag{4.14}$$

Решая систему уравнений, получим

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \left(\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i\right) \Big/ n}{\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2 \Big/ n} = \frac{SP_{xy}}{SS_x} \qquad (4.15)$$

и

$$b_0 = \frac{\sum_{i=1}^{n} Y_i}{n} - b_1 \frac{\sum_{i=1}^{n} X_i}{n} = \bar{Y} - b_1 \bar{X}. \qquad (4.16)$$

Мы могли бы использовать эти формулы для получения коэффициентов прямой, однако легко заметить, что уравнения (4.13) и (4.14) представляют собой систему уравнений, которую можно решить, используя методы, описанные в гл. 3.

Оба эти уравнения можно записать в матричной форме:

$$\begin{pmatrix} n & \Sigma X \\ \Sigma X & \Sigma X^2 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} \Sigma Y \\ \Sigma XY \end{pmatrix}. \qquad (4.17)$$

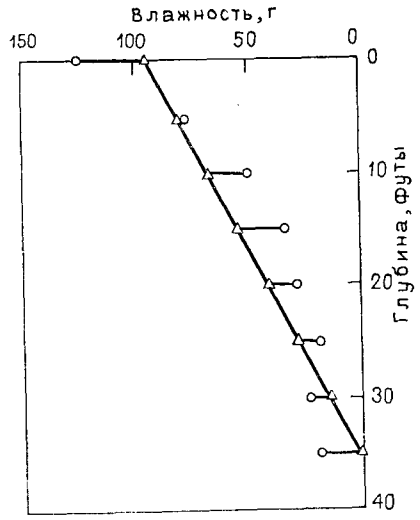
Хотя в этом простом случае использование матричного метода едва ли дает какие-либо преимущества, в более сложных ситуациях его применение оправданно. Поэтому мы приведем решение задачи о зависимости содержания влаги от глубины методами матричной алгебры и будем использовать этот метод и далее в настоящей главе. Элементы матриц таковы: $n=8$, $\Sigma X=140$, $\Sigma Y=382$, $\Sigma XY=3870$ и $\Sigma X^2=3500$. Система в матричной форме имеет вид

$$\begin{pmatrix} 8 & 140 \\ 140 & 3500 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} 382 \\ 3870 \end{pmatrix}.$$

Решив ее, получаем $b_0=94{,}67$ и $b_1=-2{,}68$. Мы можем использовать полученные значения для вычисления оценок содержаний влаги в осадке на различных глубинах. Полученные оценки опробования в точках позволяют измерить, насколько прямая, построенная по методу наименьших квадратов, соответствует исходным выборочным данным. Если бы построенная прямая проходила в точности через каждую выборочную точку, то $\hat{Y}_i$ и Y_i совпадали бы и сумма квадратов отклонений от прямой была бы равна нулю. Конечно, в приведенном примере это не так. Значения $\hat{Y}_i$ и Y_i изображены на рис. 4.15.

Мы можем определить три характеристики, которые описывают изменение зависимой переменной. Первая из них — это

Рис. 4.15. Наблюдаемые значения влажности и их оценки, полученные из линейного уравнения регрессии, построенного по методу наименьших квадратов



общая сумма квадратов (SS_T) переменной Y:

$$SS_T = \sum_{i=1}^{n} Y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right)^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2. \qquad (4.18)$$

Разделив это уравнение на $(n-1)$, получим дисперсию переменной Y:

$$s^2 = \frac{1}{n-1} SS_T = \frac{1}{n(n-1)}\left[n \sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2\right]. \qquad (4.19)$$

Вторая характеристика изменчивости зависимой переменной — это сумма квадратов отклонений оцененных значений $\hat{Y}_i$ от среднего значения $\bar{Y}$:

$$SS_R = \sum_{i=1}^{n} \hat{Y}_i^2 = \frac{1}{n}\left(\sum_{i=1}^{n} \hat{Y}_i\right)^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2. \qquad (4.20)$$

Как следует из правой части этого равенства, оценки имеют то же среднее значение, что и исходные данные. Сумма квадратов этих оценок $\hat{Y}_i$ характеризует меру изменчивости линии регрессии относительно среднего значения. Если $\hat{Y}_i$ и Y_i совпадают для всех наблюдений, то суммы квадратов, вычисленные по формулам (4.18) и (4.20), будут одинаковыми. Наоборот, если сумма квадратов (4.20) будет меньше, то разность

$$SS_D = SS_T - SS_R, \qquad (4.21)$$

называемая остаточной суммой квадратов, будет отличаться от нуля. Как легко убедиться, величину SS_D можно также вычислить по формуле

$$SS_D = \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2, \tag{4.22}$$

где SS_D является мерой отклонения прямой, построенной по методу наименьших квадратов, от результатов наблюдений. Качество приближения прямой характеризуется отношением

$$R^2 = \frac{SS_R}{SS_T}. \tag{4.23}$$

Если для имеющихся данных прямая хорошо подобрана, то это отношение будет близко к единице; ниже мы рассмотрим критерии, позволяющие судить о том, насколько хорошо это отношение характеризует качество оценки. Величину R^2 нередко выражают в процентах. Та же терминология принята в тренд-анализе, который, как мы увидим, является прямым обобщением этого метода. Необходимо отметить, что квадратный корень из R^2 равен множественному коэффициенту корреляции R:

$$R = \sqrt{R^2} = \sqrt{SS_R/SS_T}. \tag{4.24}$$

Алгебраический эквивалент этого соотношения определен в гл. 2 как коэффициент корреляции

$$r = \frac{SS_{xy}}{\sqrt{SS_x \cdot SS_y}}. \tag{4.25}$$

Таким образом, при нахождении уравнения прямой, характеризующей зависимость влажности осадка от глубины, по методу наименьших квадратов мы вычислили различные величины, необходимые при определении сумм квадратов, качества приближения и коэффициента корреляции. Вычислите величины SS_T, SS_R, SS_D, R^2 и R для данных табл. 4.11.

Совершенно очевидно, что прямая линия не всегда хорошо аппроксимирует данные даже в случаях высокой корреляции. Плохое приближение возникает как следствие ряда причин, среди которых следует отметить высокую дисперсию зависимой переменной (чрезмерный разброс данных), а также выбор неподходящей модели. В этом примере мы склонны подозревать последнее, так как расположение исходных данных наводит на мысль, что для аппроксимации более пригодна кривая, а не прямая линия. Ниже мы рассмотрим нелинейную аппроксимацию. Однако прежде нам придется изучить статистические критерии, применяемые для проверки предположения, что данные подчиняются некоторым заданным требованиям.

Если Y_i — случайная переменная, которой соответствует некоторый интервал изменения переменной X_i, то мы можем предположить, что имеющиеся данные подчиняются следующей теоретической модели:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \tag{4.26}$$

где i — номера последовательных наблюдений. Величина ε является случайной нормально распределенной величиной с нулевым средним и неизвестной дисперсией σ^2, не зависящей от величин Y_i. Иными словами, предполагается, что наблюдаемые значения Y_i являются суммами постоянной величины, связанной со средним значением (если X_i и Y_i отсчитываются от своих средних значений, то β_0 равно нулю), линейной функции от X_i и случайной компоненты ε. Это соотношение изображено на рис. 4.16. Предполагается, что для каждой точки линии регрессии существует нормальное распределение частот возможных значений переменной Y_i. Применяя метод наименьших квадратов и используя выборочные коэффициенты регрессии, вычисленные исходя из модели (4.17), мы можем оценить параметры регрессии [т. е. параметры β в формуле (4.26)] по выборочным коэффициентам регрессии [параметры b в модели (4.17)]. Если сделанные нами ограничения выполнены, то метод наименьших квадратов даст нам оценки максимального правдоподобия параметров регрессии b_1 и b_0, и построенная нами линия регрессии будет ближе к истинной прямой регрессии, чем любая другая прямая. Если построенное линейное уравнение является удачной регрессионной моделью, то дисперсия случайной компоненты равна дисперсии относительно линии регрессии.

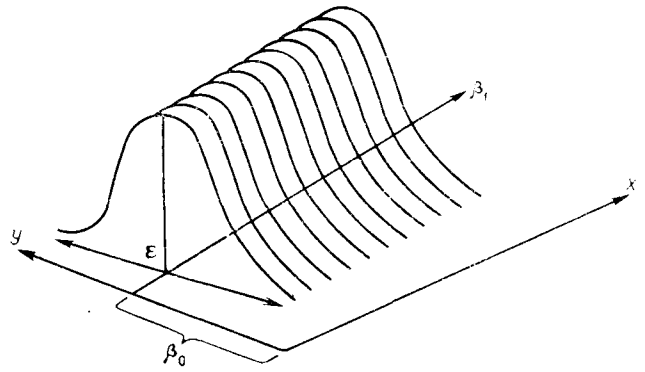


Рис. 4.16. Компоненты регрессионной модели $Y_i = \beta_0 + \beta_1 X + \varepsilon_i$.
Предполагается, что случайная компонента ε_i нормально распределена относительно линии регрессии

Таблица 4.12

Дисперсионный анализ для случая простой линейной регрессии

Источник изменчивости	Сумма квадратов	Число степеней свободы	Средние квадраты	Значение F-критерия
Линейная регрессия	SS_R	1	MS_R	MS_R/MS_D
Отклонение	SS_D	$n-2$	MS_D	
Общая дисперсия	SS_T	$n-1$		

Наоборот, если модель выбрана неудачно, то дисперсия относительной прямой регрессии будет больше, чем дисперсия величины ε.

Можно использовать полученные суммы квадратов для вычисления оценок дисперсий, которые в свою очередь необходимы при проверке двух альтернатив. В частности, SS_D используется как оценка дисперсии относительно линии регрессии. Мы можем получить адекватную оценку для σ^2 только в том случае, если проведем измерения Y_i в каждой точке X_i, так как это единственный путь, который позволяет оценить значение дисперсии Y независимо от дисперсии X. Однако значение SS_R дает оценку дисперсии σ^2 в том случае, если наша модель правильна; если же наша модель неправильна, это значение превосходит σ^2 на некоторое положительное число, называемое смещением. Используя SS_R, можно провести дисперсионный анализ, приводящий к отклонению нулевой гипотезы в любом из двух случаев, либо когда изменчивость наблюдений слишком велика для того, чтобы сделать надежные выводы, либо если постулированная нами модель неверна. В табл. 4.12 приведена схема дисперсионного анализа.

Как указано в гл. 2, средние квадратов дают дисперсии, оценки которых получаются в результате деления соответствующих сумм квадратов на отвечающие им числа степеней свободы. Величине MS_R отвечает одна степень свободы, так как ее значение получено на основе двух «наблюдений» значений коэффициентов b_0 и b_1. Общая дисперсия имеет $n-1$ степеней свободы. Поэтому величина MS_D должна иметь число степеней свободы, равное разности между двумя указанными, т. е. $(n-1)-1=n-2$. Мы можем применить ANOVA к рассмотренной выше задаче, как это сделано в табл. 4.13. При этом проверяется следующая гипотеза:

$$H_0 : \beta_1 = 0$$

Таблица 4.13

Результаты дисперсионного анализа, проведенного для определения значимости регрессии, характеризующей зависимость содержания воды в осадке от глубины

Источник изменчивости	Сумма квадратов	Число степеней свободы	Средние квадраты	Значение F-критерия
Линейная регрессия	7546,88	1	7546,88	23,071*
Отклонение	1962,62	6	327,10	
Суммарная дисперсия	9509,50	7		

* Гипотеза о равенстве дисперсий отклоняется при 5%-ном уровне значимости ($\alpha=0{,}05$).

при альтернативе

$$H_1 : \beta_1 \neq 0.$$

Линия регрессии подчинена условию: она проходит через средние значения X и Y. Если угловой коэффициент β_1 незначимо отличается от нуля, то это эквивалентно следующему утверждению: рассеяние значений Y относительно линии регрессии не меньше, чем их рассеяние относительно $\bar{Y}$. Выберем 5%-ный уровень значимости ($\alpha=0{,}05$). Если H_0 верна, то проверяемая статистика подчиняется F-распределению с $\nu_1=1$ и $\nu_2=6$ степенями свободы, и поэтому критическая область состоит из значений, превышающих $F=5{,}99$. Вычисленное значение критерия попадает в критическую область, поэтому мы должны отклонить гипотезу о том, что дисперсия относительно линии регрессии не отличается от дисперсии, полученной по наблюдениям. Однако даже несмотря на то, что существует значительный линейный тренд, графическое представление данных позволяет предположить, что мы в состоянии провести анализ точнее.

В 15 м от первой скважины в илистых отложениях устья реки была пробурена вторая скважина. Содержания воды в грабах из этой скважины образуют последовательность измерений Y_i, позволяющих оценить σ^2. В результате мы можем определить, является ли слабая корреляция между содержанием воды в осадке и глубиной следствием сильного разброса данных или результатом непригодности уравнения, выбранного в качестве модели. Данные по второй скважине приведены в табл. 4.14. Нанесите эти точки на график и сравните полученное распределение с распределением, соответствующим данным табл. 4.11.

Таблица 4.14

Значения влажности осадков из второй скважины

Глубина, футы	Влажность (граммы воды/100 г сухого осадка)	Глубина, футы	Влажность (граммы воды/100 г сухого осадка)
0	137	20	28
5	84	25	24
10	50	30	23
15	32	35	20

Данные табл. 4.11 можно объединить вместе с данными табл. 4.14, после чего построить уравнение регрессии по всем наблюдениям. Вычисления величин SS_T, SS_R и SS_D проводятся так же, как и раньше, только теперь число наблюдений удвоилось. Так как сейчас у нас в распоряжении имеются новые наблюдения, мы можем подсчитать сумму квадратов, возникающую из-за недостаточной точности аппроксимации (SS_{LF}), и сумму квадратов, соответствующую «чистой» случайной компоненте (SS_{PE}), которые разбивают сумму квадратов отклонений на две части. В случае пар повторных наблюдений мы можем найти величину по формуле

$$SS_{PE} = 1/2 \sum_{i=1}^{n} (Y_{i1} - Y_{i2})^2. \qquad (4.27)$$

Эта величина имеет одну степень свободы для каждой точки, а остаточная сумма квадратов SS_{LF} находится путем вычитания, так же как и ее число степеней свободы

$$SS_{LF} = SS_D - SS_{PE}. \qquad (4.28)$$

Совсем не обязательно, чтобы мы проводили дублирующие измерения в каждой точке, но если это сделать, то анализ удастся осуществить более точно. Можно также использовать более двух повторений Y_i для каждого значения X_i, при этом вычисление величины SS_{PE} становится несколько более сложным. Эти и другие усовершенствования описаны в книгах по регрессионному анализу Ли [35], Дрейпера и Смита [17], и мы не будем останавливаться на них более подробно.

Схема модифицированного дисперсионного анализа приведена в табл. 4.15. Используя объединенные данные по двум скважинам, выполните дисперсионный анализ и вычислите SS_{PE} и SS_{LF}. Среднее значение суммы квадратов SS_{PE} является оценкой для $\sigma^2_{Y \cdot X}$, т. е. дисперсии относительно линии регрессии. Оно находится по формуле

$$MS_{PE} = SS_{PE}/k, \qquad (4.29)$$

Таблица 4.15

Дисперсионный анализ в случае простой линейной регрессии с повторением. Число наблюдений Y_i равно n; число точек, в которых сделаны повторные измерения, равно k

Источник изменчивости	Сумма квадратов	Число степеней свободы	Средние квадраты	F-критерий
Линейная регрессия	SS_R	1	MS_R	MS_R/MS_D[а]
Отклонение	SS_D	$n-2$	MS_D	
Недостаток точности	SS_{LF}	$(n-2)-k$	MS_{LF}	MS_{LF}/MS_{PE}[б]
Чистая случайная компонента	SS_{PE}	k	MS_{PE}	
Общая дисперсия	SS_T	$n-1$		

[а] Критерий качества приближения.

[б] Критерий соответствия модели.

где k — число точек, в которых проведены повторные измерения. В нашем случае мы сделали это для всех точек, и поэтому k равно $n/2$, так как половина наблюдений Y_i дублирована. Мы отмечали, что величина SS_D является мерой дисперсии вокруг регрессии плюс некоторое смещение, которое может возникнуть из-за выбора неподходящей модели, и поэтому среднее значение квадратов SS_{LF} является оценкой только этого смещения. Мы можем провести проверку пригодности модели, вычисляя значение

$$F = MS_{LF}/(MS_{PE}). \qquad (4.30)$$

Если вычисленное значение критерия попадает в критическую область, то мы должны сделать вывод, что построенная модель не отвечает действительности. Если проверка не приводит к отклонению модели, то обе оценки дисперсий можно сложить ($MS_{LF}+MS_{PE}=MS_D$) и оценить качество аппроксимации, как мы делали это раньше. Вычислите F-отношение и дополните табл. 4.15, а затем определите, является ли эта простая линейная модель достаточно хорошей. На рис. 4.17 изображены четыре возможные ситуации для двух характеристик, одна из которых — соответствие модели, а вторая — качество аппроксимации.

Нелинейная регрессия

После вычисления F-критерия для проверки соответствия модели выборочным данным вы можете прийти к выводу, что прямая линия неадекватно представляет выборку. Что делать после этого, зависит от предмета исследования и от ваших знаний или догадок о соотношении между переменными X и Y.

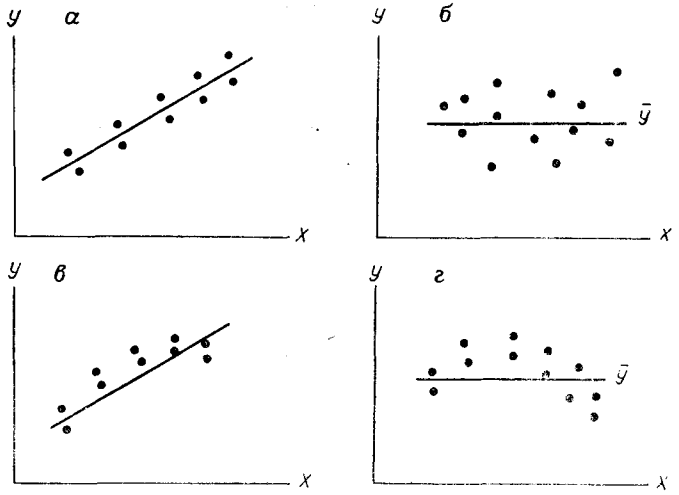


Рис. 4.17. Возможные случаи линейной регрессии [17]:

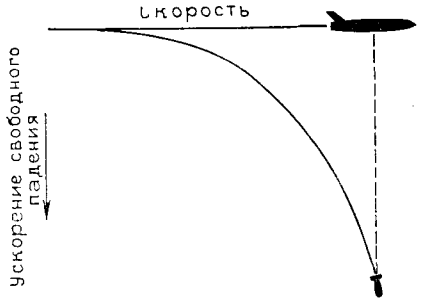
а — существенная линейная регрессия и точность аппроксимации удовлетворительная; *б* — отсутствие линейной регрессии при удовлетворительной точности аппроксимации; *в* — существенная линейная регрессия и значимый недостаток точности аппроксимации; *г* — отсутствие линейной регрессии и значимый недостаток точности аппроксимации

Иногда можно иметь вполне определенное мнение о связи переменных. Например, если самолет сбрасывает бомбу, то, пренебрегая сопротивлением ветра, мы можем предсказать ее теоретическую траекторию, которая определяется скоростью самолета и направленным вниз ускорением свободного падения (рис. 4.18); действительно, парабола очень хорошо описывает траекторию падения бомбы. С другой стороны, мы можем ничего не знать о зависимости между двумя переменными X и Y (ее может и не существовать) и просто хотим получить выражение одной из них через другую. Обычно наши задачи находятся между этими двумя крайними случаями: мы предполагаем наличие причинной связи, но не знаем ее формы. В последних двух случаях мы можем подобрать аппроксимирующее уравнение к имеющимся данным в надежде, что оно поможет нам прояснить существующие соотношения или же точно описать форму зависимости переменных X и Y. Такие уравнения выбираются потому, что с их помощью удается аппроксимировать многие классы функций, и используются в тех случаях, когда истинный вид функции неизвестен.

Возможны различные типы аппроксимирующих функций, но чаще всего используется полиномиальная аппроксимация, заключающаяся в том, что в качестве приближающей функции выбирается сумма целых степеней независимой переменной:

$$Y_i = b_0 + b_1X_i + b_2X_i^2 + b_3X_i^3 + \ldots + b_mX_i^m. \qquad (4.31)$$

Рис. 4.18. Теоретическая траектория падения бомбы, сброшенной с самолета



Уравнение, в котором все переменные суммируются, называется линейным, так как соотношения между всеми парами имеют своими графиками прямые линии. Расширение первоначального уравнения с помощью добавления следующих степеней приводит к тому, что график начинает искривляться. Один дополнительный член заставляет прямую изменить наклон, второй дополнительный член приводит к возникновению двух точек перегиба и т. д. Увеличивающаяся искривленность позволяет линии более точно подходить к исходным данным. Действительно, если число дополнительных членов достигнет $(n-1)$, то линия пройдет точно через каждую данную точку. Однако в построении такой линии мало смысла, так как она не является более эффективной, чем сами исходные данные. Кроме того, наиболее важную информацию о данном массиве можно сохранить с использованием лишь нескольких членов в полиномиальном уравнении. На рис. 4.19 изображены различные типы полиномиальных зависимостей, соответствующих различным степеням аргумента. Максимальная степень, использованная в полиномиальном уравнении, называется степенью уравнения, т. е.

$$Y_i = b_0 + b_1X_i + b_2X_i^2 + b_3X_i^3$$

— полиномиальное уравнение третьей степени. Такова же степень уравнения $Y_i = bX_i^3$, так как оно является частным случаем предыдущего при b_0, b_1 и b_2, равных нулю. Полиномиальное уравнение строится по наблюдениям с помощью метода наименьших квадратов, а процесс этого построения называется подбором кривой.

При выполнении некоторых статистических условий качество аппроксимации и ее значимость могут быть проверены с помощью регрессионных методов, аналогичных уже рассмотренным. Эти статистические процедуры являются составной частью так называемого нелинейного регрессионного анализа.

Чтобы аппроксимировать данные кривой второго порядка (или квадратичной кривой), нужно составить нормальные урав-

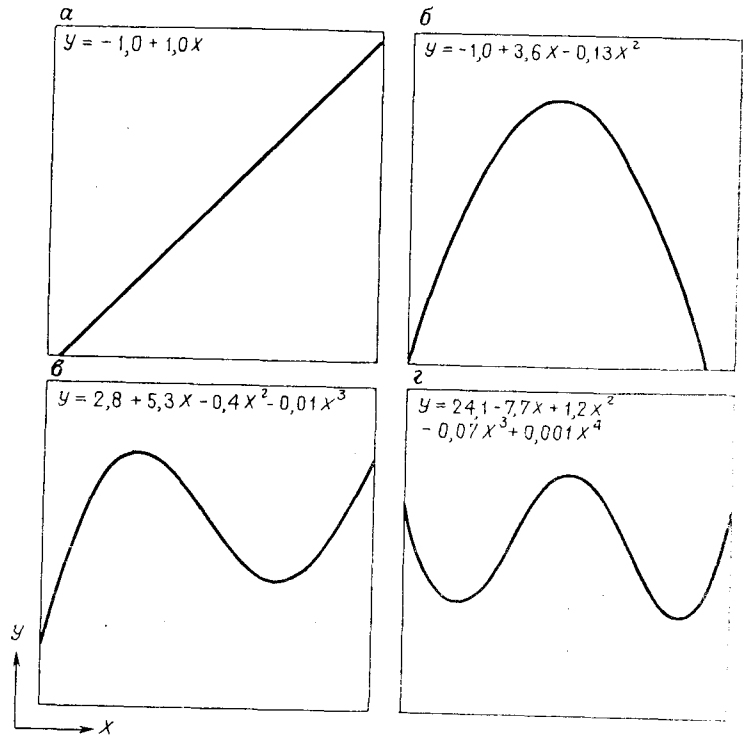


Рис. 4.19. Кривые полиномиальной регрессии для возрастающих степеней переменной X:

а — прямая линия, соответствующая многочлену первой степени; *б* — квадратическая кривая или кривая второй степени; *в* — кубическая кривая или кривая третьей степени; *г* — кривая четвертой степени

нения с включением дополнительных членов. Два нормальных уравнения (4.13) и (4.14) превращаются в совокупность трех уравнений:

$$\begin{aligned} \Sigma Y &= b_0 n + b_1 \Sigma X + b_2 \Sigma X^2, \\ \Sigma XY &= b_0 \Sigma X + b_1 \Sigma X^2 + b_2 \Sigma X^3, \\ \Sigma X^2 Y &= b_0 \Sigma X^2 + b_1 \Sigma X^3 + b_2 \Sigma X^4. \end{aligned} \tag{4.32}$$

Подразумевается, что суммирование выполняется по всем наблюдениям от 1 до n. Переписав их в матричной форме, получаем

$$\begin{bmatrix} n & \Sigma X & \Sigma X^2 \\ \Sigma X & \Sigma X^2 & \Sigma X^3 \\ \Sigma X^2 & \Sigma X^3 & \Sigma X^4 \end{bmatrix} \cdot \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} \Sigma Y \\ \Sigma XY \\ \Sigma X^2 Y \end{bmatrix}. \tag{4.33}$$

Это матричное уравнение можно решить, используя процедуру матричной алгебры, приведенную в гл. 3. Заметим, что в эту систему входят высокие степени независимой переменной. Самая высокая степень, используемая в матрице, равна удвоенной степени полинома, который мы хотим подобрать к изучаемым данным. Это обстоятельство является главным источником ошибок в вычислительных программах полиномиальной аппроксимации, так как элементы правого нижнего угла матрицы коэффициентов могут на много порядков превышать величину элементов левого верхнего угла матрицы. Это может привести к большим ошибкам округления и потере значимости в существенных цифрах, результатом чего будут неустойчивые или ненадежные решения системы уравнений. Подобное рассмотрение этих задач содержится в книге Уэстлейка [59].

Структура матрицы коэффициентов станет очевидной, если мы используем переменную X^0, которая равна 1 для всех наблюдений X_i. Мы можем занумеровать все строки и столбцы матричного уравнения следующим образом:

$$\begin{array}{c} \\ X^0 \\ X^1 \\ X^2 \\ X^3 \\ \vdots \\ X^m \end{array} \begin{array}{c} \begin{array}{ccccccc} X^0 & X^1 & X^2 & X^3 & \dots & X^m \end{array} \\ \begin{bmatrix} & & & & & \end{bmatrix} \end{array} \begin{array}{c} \begin{array}{c} b \end{array} \\ \begin{bmatrix} \\ \\ \\ \\ \\ \end{bmatrix} \end{array} = \begin{array}{c} \begin{array}{c} Y \end{array} \\ \begin{bmatrix} \\ \\ \\ \\ \\ \end{bmatrix} \end{array}. \tag{4.34}$$

Элементы матрицы коэффициентов, а также столбцов коэффициентов b и правых частей Y являются суммами смешанных произведений элементов строк и столбцов с заданными номерами. Имея в виду значение X^0, мы определяем элемент A_{11} как $\sum_{i=1}^{n} 1 \cdot 1 = n$, другие элементы верхней строки получаются умножением 1 на соответствующий столбец. Например, элемент A_{13} матрицы равен $\Sigma X^3 \cdot X^2 = \Sigma X^5$. Напомним, что при умножении показатели степени складываются, т. е. $X^a \cdot X^b = X^{a+b}$.

Для иллюстрации вычислений, используемых при построении уравнения нелинейной регрессии, проанализируем совместно данные таблиц 4.11 и 4.14. Построим квадратичную аппроксимацию, что позволит нам убедиться в том, что повышение степени приводит к значительному улучшению качества аппроксимации. Полиномиальная кривая второй степени, подобранная к этим данным, изображена на рис. 4.20. В данном случае уравнение регрессии имеет следующий вид:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 = 122{,}9 - 7{,}9 X_i + 0{,}1 X_i^2.$$

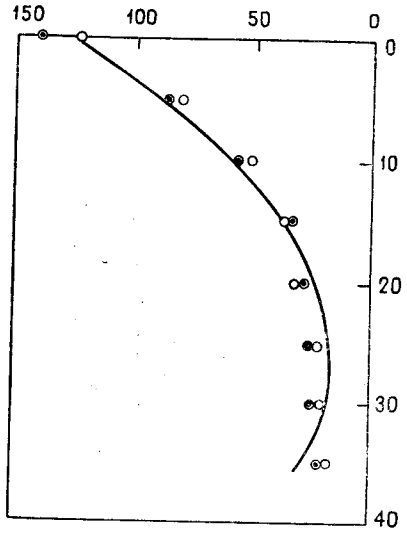


Рис. 4.20. Кривая полиномиальной регрессии второй степени, полученная по замерам влажности, взятым из таблиц 4.11 и 4.14

В этом примере необходимые для выполнения дисперсионного анализа статистики принимают следующие значения:

$$SS_T = 21363{,}0; \qquad SS_R = 20673{,}2;$$
$$SS_D = 689{,}8; \qquad SS_{PE} = 126{,}0;$$
$$R^2 = 0{,}97; \qquad R = 0{,}98.$$

Легко убедиться, что значения SS_T и SS_{PE} такие же, как и в случае линейной аппроксимации, так как они не содержат оценок величин $\hat{Y}$. Как можно было ожидать, более гибкая квадратичная кривая ближе подходит к наблюдаемым данным, чем прямая линия. Сумма квадратов отклонений относительно линии регрессии уменьшилась с 5177,8 до 689,8. Это большое уменьшение, но эта разница не всегда бывает столь значительна. Таблицу дисперсионного анализа можно снова расширить для проверки этой характеристики (табл. 4.16).

Как можно увидеть из этой таблицы, сумма квадратов получается в результате вычитания суммы квадратов для линейной функции SS_{R1} из аналогичной суммы квадратов для квадратичной функции SS_{R2}. Эта новая сумма квадратов является мерой улучшения качества аппроксимации в результате введения дополнительного члена в уравнение регрессии. В критерии «б» (см. табл. 4.16) эта величина, которая обозначена через SS_{2-1}, используется для оценки дополнительной дисперсии в регрессии. Ее значимость проверяется точно таким же образом, как и для самого уравнения регрессии. Если окончательное значение F-критерия попадает в критическую область, то

Таблица 4.16

Дисперсионный анализ для определения значимости дополнительных членов в нелинейной регрессии

Источник изменчивости	Сумма квадратов	Число степеней свободы	Средние квадраты	F-критерий
Линейная регрессия	SS_{R1}	1	MS_{R1}	
Квадратичная регрессия	SS_{R2}	2	MS_{R2}	MS_{R2}/MS_{DK} [а]
Квадратичное дополнение	SS_{2-1}	1	MS_{2-1}	MS_{2-1}/MS_{D2} [б]
Квадратичное отклонение	SS_{D2}	n—3	MS_{D2}	
Общая дисперсия	SS_T	n—1		

[а] Критерий значимости для квадратичного приближения.

[б] Критерий значимости для проверки увеличения качества квадратичной аппроксимации по сравнению с линейной.

добавочный член дает существенный вклад в регрессию и его следует сохранить. Если же полученное значение несущественно, то дополнительный член не дает значимого вклада в регрессию. Необходимо отметить, что критерий «а» может быть существенным, в то время как критерий «б» таким не является. Это происходит потому, что критерий «а» предназначен для проверки гипотезы о равенстве нулю комбинации линейных и квадратичных членов. При этом линейная часть может оказаться высоко значимой, в то время как вклад квадратичного члена будет очень низким. Тогда значение критерия «а» будет существенным благодаря значительному влиянию одного линейного члена. Иногда может оказаться, что значимыми является либо один из двух, либо оба члена, либо ни один из двух членов не является значимым.

Читатель уже мог заметить, что корреляционная зависимость всегда увеличивается при добавлении новых членов полинома. Если число членов полинома достигает (n—1), то коэффициент корреляции становится равным 1,00 независимо от степени разброса данных точек. Однако приведенные выше критерии показывают, что увеличение коэффициента корреляции не имеет статистических оснований. При этом если средние значения квадратов отклонений увеличиваются, то F-отношение, ха-

рактеризующее значимость аппроксимации, уменьшается. Оценка дисперсии частично зависит от числа наблюдений, использованных для ее вычисления, или от числа степеней свободы. Последнее постоянно уменьшается по мере увеличения числа коэффициентов в уравнении регрессии. Напомним (см. гл. 2), что мы теряем в точности на каждом оцениваемом параметре одну степень свободы и что коэффициенты b полиномиального уравнения являются оценками коэффициентов регрессии β.

При выполнении определенных статистических требований процедуру проверки гипотезы о незначимости добавляемых членов можно распространить на члены более высоких степеней в уравнении полиномиальной регрессии. Если получение дополнительных наблюдений реально, то эти критерии можно использовать в комбинации с критерием проверки отсутствия приближения или критерием чисто случайной компоненты. В табл. 4.17 приведена дополнительная схема дисперсионного анализа для квадратичной регрессии и объединенной выборки значений влажности осадка.

В некоторых случаях нас может интересовать не только получение оценок в заданных точках или отклонения от линии регрессии, но и ее наклон и значение, при котором этот наклон изменяется. В качестве примера задачи такого типа мы рас-

Таблица 4.17

Результаты дисперсионного анализа для определения значимости квадратичной регрессии содержания воды в осадке и глубины

Источник изменчивости	Сумма квадратов	Число степеней свободы	Средние квадраты	F-критерий
Линейная регрессия	16 185,19	1	16 185,19	
Квадратичная регрессия	20 673,24	2	10 336,62	$\frac{MS_{R2}}{MS_{D2}} = 277,83$[а]
Квадратичное дополнение	4 488,05	1	4 488,05	$\frac{MS_{2-1}}{MS_{D2}} = 98,96$[б]
Квадратичное отклонение	689,76	13	45,37	
Общая дисперсия	21 363,00	15		

[а] Квадратичная регрессия высоко значима.

[б] Квадратичная регрессия приводит к значительному улучшению качества аппроксимации по сравнению с линейной регрессией.

смотрим стратиграфическую последовательность по скважине, пробуренной в породах нижнего палеозоя в восточной части Оклахомы. Целью геологического исследования являлось восстановление условий, существовавших во время формирования отложений. Изученная толща состоит из слоев алевролита и песчаника, относительно которых допускалось, что они имеют морское происхождение. Геологами было сделано предположение, что осадочный бассейн постепенно наполнялся, и, по мере того как береговая линия продвигалась по направлению к местоположению скважины, мощность последовательно образовавшихся слоев песчаника увеличивалась. Толща насчитывала тысячи слоев и было бы крайне обременительно измерять каждый из них. Вместо этого была измерена мощность каждого слоя песчаника через интервал в 10 футов (3 м). Эти измерения приведены в табл. 4.18. Геологу интересно знать, существует ли зависимость между мощностью отдельного слоя и общей мощностью накопленного осадка. Отметим, что суммарная мощность X измеряется в фиксированных точках, а мощность индивидуальных слоев рассматривается как случайная величи-

Таблица 4.18

Мощность слоев песчаника в кластических отложениях нижнего палеозоя в Оклахоме

Интервал X, футы	Мощность Y, дюймы	Интервал X, футы	Мощность Y, дюймы
10	9,2	260	8,5
20	7,1	270	8,9
30	5,9	280	10,7
40	3,7	290	14,4
50	6,2	300	15,2
60	4,1	310	12,1
70	3,9	320	15,3
80	5,0	330	9,0
90	4,4	340	11,2
100	6,8	350	8,9
110	5,9	360	9,0
120	6,1	370	6,5
130	7,7	380	11,0
140	7,0	390	13,9
150	5,5	400	9,1
160	9,8	410	11,2
170	6,9	420	17,3
180	5,2	430	15,8
190	6,8	440	11,1
200	8,5	450	11,8
210	7,1	460	18,9
220	10,4	470	9,6
230	6,7	480	17,9
240	8,6	490	12,8
250	6,4	500	15,0

на, что позволяет использовать регрессионную модель. Иными словами, геолог должен проверить гипотезу, заключающуюся в том, что коэффициент регрессии b_1 значимо отличается от нуля.

После того как описанным выше методом получено уравнение регрессии $Y_i = b_0 + b_1 X_i$, можно оценить дисперсию относительно линии регрессии, используя величину MS_D. Последнюю в свою очередь можно использовать для вычисления t-статистики:

$$t = \frac{b_1}{\sqrt{MS_D / SS_X}}. \tag{4.35}$$

Среднее значение квадратов, связанное с отклонением (MS_D), равно SS_D, деленному на n—2 степеней свободы, как это указано в табл. 4.12. Исправленная сумма квадратов SS_X находится по формуле

$$SS_X = \sum_{i=1}^{n} X_i^2 - \frac{1}{n}(\Sigma X_i)^2. \tag{4.36}$$

Эта величина используется для проверки одной из гипотез:

1) $H_0 : \beta_1 = 0$,
2) $H_0 : \beta_1 \leqslant 0$,
3) $H_0 : \beta_1 \geqslant 0$

против соответствующих альтернатив:

$H_1 : \beta_1 \neq 0$,
$H_1 : \beta_1 > 0$,
$H_1 : \beta_1 < 0$.

Проверка первой нулевой гипотезы требует двустороннего критерия, так как и положительный, и отрицательный наклоны приводят к ее отклонению. Два других критерия являются односторонними. Геолог заинтересован в получении ответа на следующий вопрос: увеличивается ли мощность слоя, т. е. положителен ли наклон линии регрессии между мощностью отдельного слоя и суммарной мощностью? Поэтому для наших целей подходит первая гипотеза. Критерий будет односторонним с критической областью, расположенной справа.

Мы можем быстро вычислить необходимые для критерия величины. Некоторые значения, используемые в этом критерии, приведены ниже.

Уравнение регрессии: $Y = 4{,}25 + 0{,}020\,X$

$SS_T = 730{,}94$	$SS_R = 425{,}18$
$SS_D = 305{,}76$	$SS_X = 1\,041\,250{,}00$
$R^2 = 0{,}58$	$R = 0{,}76$

Исходные данные содержат 50 наблюдений, поэтому величина SS_D соответствует 48 степеням свободы. При 5%-ном уровне значимости ($\alpha = 0{,}05$) одностороннее критическое значение t-статистики при $\nu = 46$ степенях свободы равно 1,68. Значение t=статистики равно

$$t = \frac{0{,}02}{\sqrt{6{,}37 / 1041250{,}00}} = 8{,}09.$$

Полученное число лежит в критической области, и поэтому мы должны отклонить гипотезу о том, что наклон прямой регрессии отрицателен или равен нулю. Последовательности присуща небольшая, но вполне определенная тенденция увеличения мощности отдельных слоев.

Приведенный только что критерий является частным случаем критерия

$$t = \frac{b_1 - \beta_1}{\sqrt{MS_D / SS_X}} \tag{4.37}$$

для проверки гипотезы, заключающейся в том, что наклон линии регрессии имеет некоторое заранее заданное значение β_1.

В рассмотренном случае (критерий 4.35) значение β_1 равно нулю. Этот критерий, точнее, некоторая его разновидность, имеет важное применение в анализе временных рядов. Методы анализа временных рядов основаны на предположении об отсутствии тренда в изучаемых данных, т. е. что наклон линии регрессии по отношению к временной оси (или оси расстояний) равен нулю. Если тренд существует, то его нужно устранить, иначе анализ временных рядов теряет свою силу. Ряды, не имеющие значительного линейного тренда, называются стационарными. Если в данных имеется устойчивый или направленный тренд, то ряд называется эволюционным, или нестационарным.

Одним из предположений в теории линейной регрессии является предположение о том, что дисперсия относительно линии регрессии постоянна. Это можно проверить с помощью исследования остатков между данными и их оценками. Если дисперсия постоянна, то остатки образуют более или менее равномерную полосу около линии регрессии. Если же имеется постепенное изменение ширины полосы отклонений, то дисперсия не может быть постоянной. Эти два условия даже получили свои несколько устрашающие названия: гомоседастичности для постоянной дисперсии и гетероседастичности для изменяющейся дисперсии. Быстрый и не вполне точный способ определения типа изменения дисперсии относительно линии регрессии состоит в построении линейной регрессии для абсолютных значений отклонений. Изменение дисперсии в последовательности будет проявляться как значительный наклон.

Другое допущение регрессионного анализа состоит в том, что отклонения от линии регрессии не коррелируются между собой. Под автокорреляцией в данном случае подразумевается стремление остатков к образованию групп близких отклонений в одну и ту же сторону по отношению к линии регрессии. Присутствие ряда последовательностей автокоррелированных остатков может указывать на то, что регрессионная модель не соответствует исходным данным. Может также случиться, что автокоррелированные отклонения свидетельствуют о существовании явлений, представляющих геологический интерес. Эти вопросы будут рассмотрены подробнее в главе о тренд-анализе, где автокоррелированные положительные остатки выбираются в качестве показателей экономического потенциала запасов нефти и других полезных ископаемых. Проверка наличия автокорреляции проводится с помощью критерия скачков, применяемого к последовательности знаков отклонений от линии регрессии, или одного из методов, рассмотренных в разделе об автокорреляции.

Интересна зависимость дисперсии и автокорреляции остатков, выявленная при изучении горных разработок в северном Квебеке. На месторождении золота бульдозером была проделана длинная траншея. Вдоль нее с некоторым интервалом были

Таблица 4.19

Содержания золота в пробах из Проспект Тренс, Северный Квебек

Расстояние, футы	Значение содержаний (n·1.555 г/т)	Расстояние, футы	Значение содержаний (n·1.555 г/т)
(Северный конец траншеи)			
3,0	0,9	66,0	9,0
9,2	1,2	67,0	12,0
13,0	0,5	68,1	10,4
18,9	1,7	71,1	5,2
22,3	1,4	73,0	1,4
23,1	1,3	74,1	1,2
22,5	1,0	76,0	1,1
28,6	1,1	76,1	1,0
30,1	12,0	80,4	6,5
30,9	9,1	82,2	11,9
33,0	4,9	84,0	15,6
36,4	1,9	86,6	6,9
39,8	1,1	87,6	1,1
42,9	1,9	90,5	1,1
46,0	1,4	92,5	15,9
50,1	1,7	93,9	9,9
53,9	2,2	94,4	3,8
55,8	0,9	96,3	1,6
60,0	1,3	98,7	2,7
64,9	1,3	100,1	0,8
		(Южный конец траншеи)	

отобраны пробы, в которых определялось содержание золота. Тренд полученных значений был очевидным, а на одном конце траншеи были отмечены значительные отклонения от линии регрессии. Обычно это предвещает богатое месторождение золота. Именно такие месторождения часто обладают крайне низкими значениями содержаний золота в большей части минерализованной зоны, но наряду с этим попадаются и богатые жилы. Кроме того, в них группами или скачками встречаются большие положительные отклонения, которые также указывают на то, что траншея пересекает зону минерализованных жил. По данным табл. 4.19 проверьте наличие тренда в значениях содержаний золота и исследуйте поведение дисперсии вдоль профиля.

Используя критерий знаков, ответьте на вопросы, можно ли считать, что отклонения распределены случайно относительно линии регрессии? Вытекает ли из результатов анализа, что траншея пересекла участки минерализации? Можно ли полученные сведения об отклонениях от линии регрессии использовать для разумной экстраполяции содержаний золота вне пределов траншеи?

Ортогональная полиномиальная регрессия

Подгонка полиномиальной кривой высокой степени к данным методом наименьших квадратов требует решения большого количества совместных уравнений, что до появления ЭВМ представляло собой обременительное занятие. Как следствие этого, на более ранних этапах исследователи избегали пользоваться общими методами регрессионного анализа и, когда это только было возможно, пользовались более простым в вычислительном плане методом, называемым ортогональной полиномиальной регрессией. Для применения этого метода данные должны быть собраны с равными интервалами приращений по X. Необходимо отметить, что к упрощению вычислений приводит то, что коэффициенты ортогональных многочленов являются независимыми. Это означает, что добавление нового члена к строящемуся уравнению не изменяет уже вычисленных членов.

Ортогональные полиномы впервые появились в работах П. Л. Чебышева в девятнадцатом столетии, хотя современные процедуры их вычисления принадлежат Р. Фишеру (1925 г.). Подробное изложение теории ортогональных полиномов дано в книгах Р. Фишера [18], Дрейпера и Смита [17] и Моррисона [37], а также в других изданиях. Хотя появление ЭВМ сделало использование ортогональных многочленов не столь обязательным, они могут оказать существенную помощь в анализе данных, собираемых с регулярным интервалом. Мы рас-

Таблица 4.20

Ортогональные полиномиальные члены степеней от 1 до 4 и для числа наблюдений от 3 до 12

Члены первой степени (линейные)

3	4	5	6	7	8	9	10	11	12
−1	−3	−2	−5	−3	−7	−4	−9	−5	−11
0	−1	−1	−3	−2	−5	−3	−7	−4	−9
1	1	0	−1	−1	−3	−2	−5	−3	−7
	3	1	1	0	−1	−1	−3	−2	−5
		2	3	1	1	0	−1	−1	−3
			5	2	3	1	1	0	−1
				3	5	2	3	1	1
					7	3	5	2	3
						4	7	3	5
							9	4	7
								5	9
									11

Члены второй степени (квадратичные)

3	4	5	6	7	8	9	10	11	12
1	1	2	5	5	7	28	6	15	55
−2	−1	−1	−1	0	1	7	2	6	25
1	−1	−2	−4	−3	−3	−8	−1	−1	1
	1	−1	−4	−4	−5	−17	−3	−6	−17
		2	−1	−3	−5	−20	−4	−9	−29
			5	0	−3	−17	−4	−10	−35
				5	1	−8	−3	−9	−35
					7	7	−1	−6	−29
						28	2	−1	−17
							6	6	1
								15	25
									55

Члены третьей степени (кубичные)

3	4	5	6	7	8	9	10	11	12
	−1	−1	−5	−1	−7	−14	−42	−30	−33
	3	2	7	1	5	7	14	6	3
	−3	0	4	1	7	13	35	22	21
	1	−2	−4	0	3	9	31	23	25
		1	−7	−1	−3	0	12	14	19
			5	−1	−7	9	−12	0	7
				1	−5	−13	−31	−14	−7
					7	−7	−35	−23	−19
					14	−14	−22	−25	
							42	−6	−21
								30	−3
									33

Продолжение табл. 4.20

Члены четвертой степени (квартики)

3	4	5	6	7	8	9	10	11	12
		1	1	3	7	14	18	6	33
		−4	−3	−7	−13	−31	−22	−6	−27
		6	2	1	−3	−11	−17	−6	−33
		−4	2	6	9	9	3	−1	−13
		1	−3	1	9	18	18	4	12
			1	−7	−3	9	18	6	28
				3	13	−11	3	4	28
					7	−21	−17	−1	12
						14	−22	−6	−13
							18	−6	−33
								6	−27
									33

смотрим некоторые из этих приложений в разделе, посвященном фильтрации временных рядов.

Обычное уравнение полиномиальной регрессии

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \ldots + \beta_n X_i^n + \varepsilon \qquad (4.38)$$

может быть представлено в виде

$$Y_i = \alpha_0 + \alpha_1 \xi_{1i} + \alpha_2 \xi_{2i} + \ldots + \alpha_m \xi_{mi} + \varepsilon, \qquad (4.39)$$

где ξ_m — члены ортогональных многочленов, а α — коэффициенты метода наименьших квадратов. Численные значения переменных ξ_m могут быть определены из последовательности наблюдений X и из степени искомого уравнения регрессии. Обычно, однако, члены ортогональных многочленов находятся попросту из таблиц, аналогичных табл. 4.20. Полиномиальные члены всегда целые и для каждого наблюдения в подгоняемой последовательности требуется один член. Это наводит на мысль о необходимости решения важной задачи: если последовательность состоит из многих наблюдений, должна быть построена очень большая таблица полиномиальных членов. (Таблица для 75 наблюдений приводится Фишером и Ейтсом [19]).

Коэффициенты α находятся из следующих уравнений:

$$\alpha_0 = \frac{1}{n} \Sigma Y_i = \overline{Y}, \qquad (4.40)$$

$$\alpha_m = \frac{\Sigma Y_i \xi_m}{\Sigma \xi_m^2}. \qquad (4.41)$$

Измерения влажности в некоторых пробах, приведенные в табл. 4.11, очень подходят для исследования с применением ортогональных многочленов, так как наблюдения расположены с равным интервалом в пространстве ниже керна. Измерения, вместе с членами ξ_1 ортогональных многочленов для линейно-

го приближения по восьми наблюдениям, взятым из табл. 4.20, приведены ниже. Указаны также произведения Y и ξ_1.

Y_i	=	124	78	54	35	30	21	22	18
ξ_1	=	—7	—5	—3	—1	1	3	5	7
$Y_i\xi_1$	=	—868	—390	—162	—35	30	63	110	126

Коэффициент α_0 аппроксимирующей прямой есть попросту среднее значение Y_i: $\alpha_0 = 382/8 = 47{,}75$.

Коэффициент α_1 находится умножением каждого наблюдения Y_i на соответствующий полиномиальный член, суммированием затем делением на сумму квадратов членов:

$$\alpha_1 = \Sigma Y_i\xi_1/\Sigma\xi_1^2 = -1126/168 = -6{,}70.$$

Уравнение линейной регрессии содержания влаги на глубине поэтому есть

$$\hat{Y}_i = 47{,}75 - 6{,}70\xi_{1i}.$$

Используя обычную регрессию, мы получим соотношение

$$\hat{Y}_i = 94{,}67 - 2{,}68X_i.$$

Эти два уравнения отличаются потому, что ортогональные многочлены выражены через ξ_{1i}, а не через X_i. Однако если мы воспользуемся этими двумя уравнениями, то они оба дадут одну и ту же оценку для Y_i. Например, предположим, что мы вычислили гипотетическое содержание влаги в керне на глубине 9 м, соответствующей семнадцатому измерению в последовательности. Член ортогонального многочлена, соответствующий семнадцатому измерению, есть $+5$, так что два альтернативных уравнения будут иметь вид

$$\hat{Y} = 94{,}67 - 2{,}68\,(30) = 14{,}27$$

и

$$\hat{Y} = 47{,}75 - 6{,}70\,(5) = 14{,}25.$$

С точностью до ошибок округления эти уравнения эквивалентны.

Другое преимущество ортогональных многочленов станет очевидным, если мы захотим построить аппроксимацию с помощью уравнения более высокой степени. Для этого требуется повторить все вычислительные процедуры, только подставляя соответствующие полиномиальные члены более высокой степени. Уже найденные коэффициенты остаются неизменными. Например, для получения уравнения регрессии для содержаний влаги на глубине с точностью до членов второго порядка, мы

выберем члены второй степени для восьми наблюдений из табл. 4.20:

Y_i	=	124	78	54	35	30	21	22	18
ξ_2	=	7	1	—3	—5	—5	—3	1	7
$Y_i\xi_2$	=	868	78	—162	—175	—150	—63	22	126

Сумма произведений равна $\Sigma Y_i\xi_2 = 544$, так что коэффициент при квадратичном члене равен $\alpha_2 = 544 : 168 = 3{,}24$. Регрессия второго порядка влажности на глубине поэтому есть

$$\hat{Y}_i = 47{,}75 - 6{,}70\xi_{1i} + 3{,}24\xi_{2i}.$$

Аналогично можно аппроксимировать данные последовательно полиномами более высоких степеней, вплоть до кривой семнадцатой степени, которая должна пройти в точности через каждое значение. Коэффициенты ортогональной регрессии α могут быть преобразованы в коэффициенты обычной регрессии β простой подстановкой в уравнение, которое используется для определения членов ортогональной регрессии. Детали см. в книгах Дрейпера и Смита [17] и Остла и Менсинга [39], которые также приводят уравнения для прямого определения различных сумм квадратов, необходимых при дисперсионном анализе и при проверке значимости полиномиальной регрессии высших степеней.

Приведенная главная ось

Рассмотренные выше регрессионные методы позволяют построить линейную аппроксимацию для совокупности двумерных наблюдений, так что квадрат отклонений одной из переменных от прямой минимален. Если отклонения в направлении Y минимизированы, то получается одно множество коэффициентов линейной регрессии, но если отклонения минимизируются в направлении X, то получается другое множество коэффициентов. Если эти две прямые изобразить на графике (рис. 4.21), то они пересекутся в точке $\overline{X}$, $\overline{Y}$. Косинус угла между этими прямыми прямо связан с коэффициентом корреляции между X и Y.

Бывает так, что физические условия диктуют нам необходимость считать одну переменную функцией другой, или же цель выполняемого исследования указывает, какая из двух переменных должна быть зависимой в уравнении регрессии. Однако иногда оказывается невозможно из разумных соображений решить, какая переменная должна быть X, а какая Y. Это случается, например, в биометрии, где бывает полезно знать соотношение между двумя множествами измерений, таких, как длина и ширина раковин, однако совершенно неясно, какое множество измерений должно быть функцией от других. Аналогичные об-

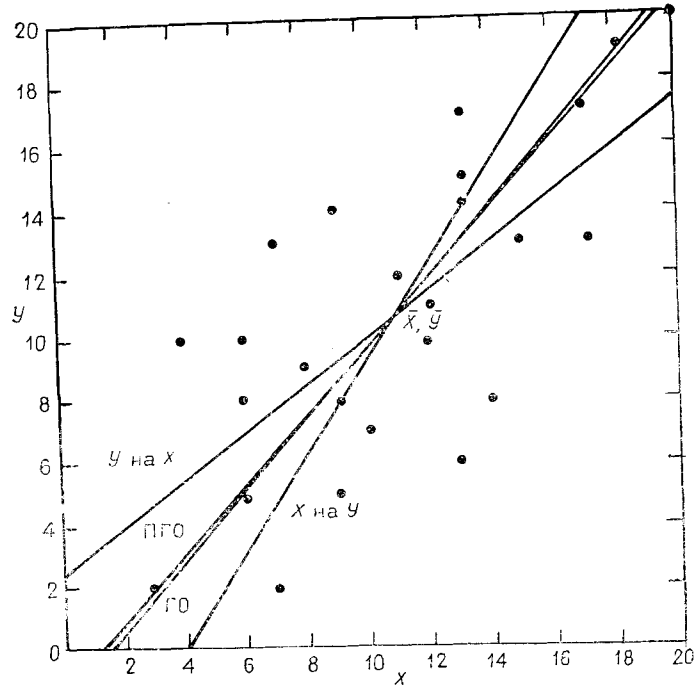


Рис. 4.21. Диаграмма рассеяния двумерных данных, взятых из таблиц 6—19. Изображены также линии регрессии Y на X и X на Y, приведенная главная ось (ПГО) и главная ось (ГО)

стоятельства возникают в петрофизике, где общая проблема состоит в том, чтобы связать два (сделанных различными методами) ряда измерений, такие, как времена звуковых переходов и измерения плотности нейтронов. Оба типа измерений подвержены ошибкам, и никакое из них не может рассматриваться как функция другого; в этих случаях как раз очень полезно представить графически оба ряда переменных, выразив некоторым образом их взаимную связь.

Первое решение, которое приходит в голову, — это подбор прямой, которая минимизирует отклонения наблюдений от этой прямой как в направлении оси X, так и в направлении оси Y одновременно. Такая линия должна расщеплять разность между линиями регрессии X по Y и Y по X, что соответствует визуальному впечатлению от тренда в наблюдениях. Поэтому было бы целесообразно приписать это расщепление рассеянию данных точек по обеим переменным, а не отклонениям единственной переменной от подбираемой прямой.

Имеется два метода определения такой прямой. Один метод состоит в минимизации квадратов отклонений от прямой

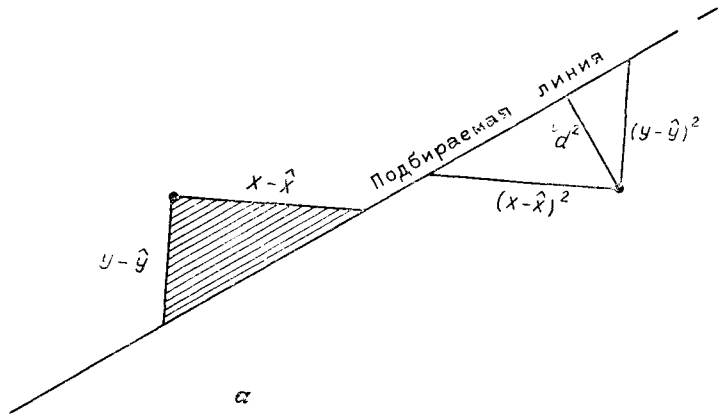


Рис. 4.22. Критерии аппроксимации ПГО и ГО:

а — приведенная главная ось минимизирует произведение отклонений $X—X^1$ и $Y—Y^1$ от подбираемой прямой; эта процедура эквивалентна минимизации площадей заштрихованных треугольников; *б* — главная ось минимизирует сумму квадратов отклонений $(X—X^1)^2$ и $(Y—Y^1)^2$, в результате минимизируются квадраты отклонений по перпендикуляру d^2

как в направлении X, так и в направлении Y одновременно. В силу теоремы Пифагора это эквивалентно минимизации квадратов перпендикуляров отклонений от подбираемой прямой (рис. 4.22). Такая прямая называется главной осью и может быть найдена как главный собственный вектор матрицы дисперсий и ковариаций X и Y. Процедуру вычисления главной оси мы обсудим в гл. 6 в разделе, посвященном методу главных компонент.

Другой метод состоит в минимизации произведения отклонений в направлениях X и Y. В действительности этот метод приводит к минимизации суммы площадей треугольников, образованных наблюдениями и подбираемой прямой (см. рис. 4.22), что приводит к прямой, называемой обычно приведенной главной осью (прямой ПГО). Большинство статей на эту тему было опубликовано в журналах «Biometrics» и «Biometrika», что отражает популярность этого метода среди ученых, занимающихся вопросами роста организмов. Хотя свойства приведенной главной оси мало интересовали статистиков, все же они были исследованы Кермаком и Халденом [31] и Красклом [33].

Резюме этих исследований для геологов дается Тиллом [52], а более подробное изложение имеется у Миллера и Кана [36].

Приведенная главная ось определяется с помощью обычного линейного уравнения, имеющего два коэффициента: один, представляющий начальную точку, другой — наклон:

$$Y = b_0 + b_1 X.$$

Таблица 4.21

Суммы квадратов и другие статистики для данных табл. 6.19 (см. кн. 2)

$n_X=25$	$n_Y=25$
$\Sigma X=272$	$\Sigma Y=267$
$\overline{X}=10{,}88$	$\overline{Y}=10{,}68$
$s_X^2=20{,}3$	$S_Y^2=24{,}1$
$s_X=4{,}51$	$S_Y=4{,}91$
$SS_X=487{,}2$	$SS_Y=578{,}4$

$\text{cov}_{XY}=15{,}6$

$SP_{XY}=374{,}4$

$r=0{,}71$

Наклон определяется как отношение стандартных отклонений двух переменных X и Y, или

$$b_1 = S_Y/S_X. \tag{4.42}$$

Так как n одно и то же для обоих стандартных отклонений, b_1 может быть найдено с помощью эквивалентного уравнения

$$b_1 = \sqrt{SS_Y/SS_X}. \tag{4.43}$$

Коэффициент b_0 приведенной главной оси дается формулой

$$b_0 = \overline{Y} - b_1\overline{X}. \tag{4.44}$$

Вычисление приведенной главной оси продемонстрируем на данных табл. 6.9 (см. книгу 2), представленных на рис. 4.21. С помощью данных табл. 6.9 будут также проиллюстрированы вычисления метода главных компонент (или, используя терминологию настоящего параграфа, метода нахождения главных осей). Суммы, суммы квадратов и попарные произведения, средние, дисперсии и ковариации приведены в табл. 4.21. Используя их, мы можем сначала вычислить обычную регрессию Y на X и X на Y. Для регрессии Y на X

$$b_1 = SP_{XY}/SS_X = 374{,}4/487{,}2 = 0{,}77,$$

$$b_0 = \overline{Y} - b_1\overline{X} = 10{,}68 - 0{,}77\,(10{,}88) = 2{,}43.$$

Таким образом, уравнение регрессии есть $Y=2{,}46+0{,}77\,X$. Для регрессии X на Y

$$b_1 = \frac{SP_{XY}}{SS_Y} = \frac{374{,}4}{578{,}4} = 0{,}65,$$

$$b_0 = \overline{X} - b_1\overline{Y} = 10{,}88 - 0{,}65\,(10{,}68) = 3{,}97,$$

что дает уравнение регрессии $X=3{,}97+0{,}65\,Y$. Для приведенной главной оси

$$b_1 = \sqrt{\frac{SS_Y}{SS_X}} = \sqrt{\frac{578{,}4}{487{,}2}} = 1{,}09,$$

$$b_0 = \overline{Y} - b_1\overline{X} = 10{,}68 - 1{,}09\,(10{,}88) = -1{,}18$$

Уравнение прямой ПГО имеет вид $Y=-1{,}18+1{,}09\,X$. Сравнение дает следующее: первый собственный вектор ковариационной матрицы X и Y есть

$$I = \begin{pmatrix} 0{,}66 \\ 0{,}75 \end{pmatrix}.$$

Это означает, что собственный вектор имеет наклон, характеризуемый длиной 0,75 единиц по оси Y и 0,66 единиц по оси X, что эквивалентно равенству 1,14 коэффициента b_1. Коэффициент b_0 равен $\overline{Y}-b_1\overline{X}=10{,}68-1{,}14\,(10{,}88)=-1{,}72$. Уравнение главной оси может быть записано в виде $Y=-1{,}72+1{,}14\,X$. На рис. 4.2 изображены две линии регрессии, главная ось и приведенная главная ось. Заметим, что приведенная главная ось и главная ось очень похожи друг на друга. Приведенная главная ось делит пополам угол между линией регрессии Y на X и линией регрессии X на Y; главная ось соответствует несколько большей дисперсии Y, что соответствует повороту на чуть более крутой угол.

Стандартные ошибки коэффициентов обеих приведенных главных осей могут быть легко вычислены, затем можно сформулировать приближенные критерии значимости. Однако не существует эквивалентов хорошо обоснованного дисперсионного анализа тому анализу, который следует выполнить в условной регрессии. Стандартная ошибка наклона ПГО равна

$$se_{b_1} = b_1\sqrt{\left(\frac{1-r^2}{n}\right)}. \tag{4.45}$$

Эквивалентность угловых коэффициентов b_1 и b_2 двух приведенных главных осей можно проверить с помощью критерия

$$Z = \frac{b_1 - b_2}{\sqrt{se_{b_1}^2 - se_{b_2}^2}}, \tag{4.46}$$

с которым легко узнать вариант одного из элементарных критериев, обсужденных в гл. 2. Проверяемая статистика Z распределена приблизительно нормально и ее значимость может быть определена из таблицы стандартного нормального распределения.

Стандартная ошибка определения коэффициента b_0 равна

$$se_{b_0} = s_Y \sqrt{\frac{1-r^2}{n}\left(1 + \frac{\overline{X}^2}{s_X^2}\right)}. \qquad (4.47)$$

Равенство (4.47) можно использовать для построения приближенных доверительных интервалов для вычисленного значения b_0. Аналогично стандартная ошибка в определении углового коэффициента может быть использована для определения приблизительного доверительного интервала вокруг b_1. В сущности, вообще было неправомерно использовать критерии проверки значимости коэффициентов приведенной главной оси. Из-за отсутствия теоретического обоснования этих процедур приведенная главная ось может быть использована для целей описания, а не для проверки статистической значимости.

СПЛАЙНЫ

Некоторые данные удобно представлять себе как струны в пространстве пар координат, т. е. наблюдения состоят из измерений двух свойств, совокупность которых может рассматриваться как последовательность точек в двумерном пространстве. Для целей наглядного представления желательно связать эти точки гладкой непрерывной линией. Мы сделаем это с помощью сплайн-функции.

Сплайны есть один из широких классов кусочно-определенных функций, которые могут быть использованы для представления кривых в двумерных или в трехмерных пространствах. Математический сплайн получил свое имя благодаря физическому двойнику, гибкому чертежному инструменту, сделанному из узкой полоски дерева или пластика, который может деформироваться, принимая любую форму в соответствии с каким-либо объектом неправильной формы. Чертежный инструмент закреплен свинцовыми гирями, называемыми «утками», которые фиксируют положение инструмента в точках привязки. Между «утками» инструмент изгибается так, чтобы получилась гладкая непрерывная кривая. Аналогично математический сплайн ограничен определенными точками, но между ними он изгибается так, чтобы в результате получилась гладко изменяющаяся линия.

Сплайны не являются ни аналитическими функциями, ни статистическими моделями, такими, например, как полиномиальная регрессия, описанная ранее. Скорее они являются совершенно произвольными объектами, лишенными какого-либо теоретического обоснования, исключая то, что они определяют

характеристики самой линии. Однако они очень полезны для интерполяции и важны в обеспечении мягкой структуры дисплеев ЭВМ. Интерактивные вычислительно-графические системы получили более широкое распространение при построении геологических и геофизических моделей. Подбор кривых с помощью сплайнов играет важную роль в этих системах.

Сплайны кусочно являются многочленами, подчиненными условию непрерывности производной в общих точках соседних кусков или сегментов. Наиболее общий сплайн состоит из кубических многочленов, которые являются функциями вида

$$Y = \beta_1 + \beta_2 X + \beta_3 X^2 + \beta_4 X^3.$$

Кривая, определенная кубическим многочленом, должна проходить ровно через четыре точки, но для аппроксимации более длинной последовательности необходимо использовать последовательность полиномиальных сегментов. Чтобы убедиться в том, что нет разрывов при изменении наклона или кривизны между соседними сегментами, полиномиальная функция подгоняется не по четырем точкам, а только по двум. Это позволяет нам использовать дополнительные ограничения, которые обеспечат непрерывность первых производных результирующего сплайна в точках сочленения (наклон линии одинаков по обе стороны сочленения). Сплайн степени m будет иметь непрерывные производные в точках сочленения вплоть до порядка $m-1$.

Изложение теории построения уравнений сплайнов потребовало бы использовать дифференциальное исчисление, владение которым не считается обязательным для читателя этой книги. Поэтому мы просто укажем необходимые уравнения в форме, удобной для вычислений, и остановимся на их приложениях, интересующихся теорией сплайнов мы отсылаем к отличному вводному курсу Роджерса и Адамса [45], посвященному проблемам графического изображения с помощью ЭВМ, и к монографии Типпера [53], касающейся геологических приложений методов построения аппроксимирующих поверхностей.

Математические обозначения, используемые в теории сплайнов, несколько неожиданны, они поясняются с помощью рис. 4.23, на котором представлено множество четырех наблюдений, связанных кусочно-определенной сплайн-функцией. Наблюдения представлены точками, обозначенными P_i, причем подразумевается, что P_i в действительности является вектором в декартовой системе координат, т. е. $P_i = [X_i, Y_i]$. Интервалы между последовательными точками можно измерить хордой (или прямолинейным отрезком, соединяющим две точки), которой можно приписать число t_i, где i — номер второй точки. Кубическая сплайн-функция строится по паре точек; на рисунке указаны три последовательных сплайна, один из точки P_1 в точку P_2, другой из P_2 в P_3 и третий из P_3 в P_4.

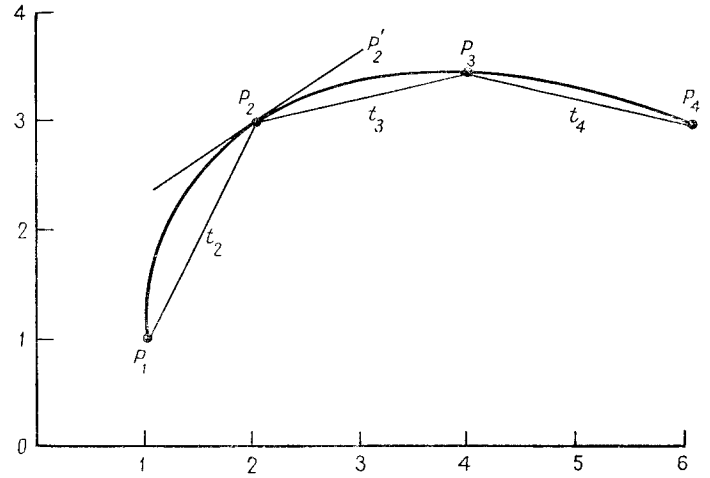


Рис. 4.23. Четыре точки, связанные функцией кубического сплайна.
Исходные наблюдения обозначены через P_i. Расстояния по хордам между точками равны t_i. Касательная к сплайну во внутренней точке P_2 обозначена через P_2'

В общем виде уравнение сплайна может быть записано в форме многочлена третьей степени от параметра t:

$$\hat{P}_t = \beta_1 + \beta_2 t + \beta_3 t^2 + \beta_4 t^3 \qquad (4.48)$$

— это представление в виде кубического многочлена координат сплайна на некотором расстоянии t вдоль интервала между двумя точками. Для определения коэффициентов этого уравнения требуется знать координаты точек, определяющих концы сплайна, и наклоны касательных прямых в этих точках. В добавление к этому мы можем указать граничные условия, определяющие поведение аппроксимирующей линии на первом и последнем участках. Конечно, заданы координаты точек. По этим данным требуется определить наклоны касательных векторов. Граничные условия могут быть выбранными по-разному в зависимости от вида линии в ее узловых точках. Мы рассмотрим только граничные условия, называемые релаксационными, или натуральными; они не требуют задания касательных векторов в концевых точках.

Для нахождения касательных векторов во внутренних точках (P_2 и P_3 на рис. 4.23) мы должны решить ряд совместных уравнений вида

$$[M]\,[P'] = [B], \qquad (4.49)$$

где неизвестный вектор коэффициентов P' определяет искомые касательные. Матрица в левой части уравнения является тридиагональной, т. е. в ней все элементы являются нулевыми, ис-

ключая диагональные элементы и элементы, стоящие непосредственно выше и ниже диагонали. Для обращения таких матриц известен специальный метод. При релаксационных граничных условиях матрица $[M]$ имеет размер $n \times n$ и выглядит так

$$[M] = \begin{bmatrix} 1{,}0 & 0{,}5 & 0 & 0 & 0 & \dots & 0 \\ t_3 & 2(t_2+t_3) & t_2 & 0 & 0 & \dots & 0 \\ 0 & t_4 & 2(t_3+t_4) & t_3 & 0 & \dots & 0 \\ 0 & 0 & t_5 & 2(t_4+t_5) & t_4 & \dots & 0 \\ 0 & 0 & 0 & t_6 & 2(t_5+t_6) & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \dots & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \dots & \\ 0 & 0 & 0 & 0 & 0 & 2 & 4 \end{bmatrix} \qquad (4.50)$$

Вектор правой части $[B]$ имеет вид

$$[B] = \begin{bmatrix} \frac{3}{2t_2}(P_2 - P_1) \\ \frac{3}{t_2 t_3} t_2^2 (P_3 - P_2) + t_3^2 (P_2 - P_1) \\ \frac{3}{t_3 t_4} t_3^2 (P_4 - P_3) + t_4^2 (P_3 - P_2) \\ \frac{3}{t_4 t_5} t_4^2 (P_5 - P_4) + t_3^2 (P_4 - P_3) \\ \frac{3}{t_5 t_6} t_5^2 (P_6 - P_5) + t_6^2 (P_5 - P_4) \\ \vdots \\ \frac{6}{t_n}(P_n - P_{n-1}) \end{bmatrix} \qquad (4.51)$$

Матричное уравнение решается обращением матрицы $[M]$ и затем умножением этой обратной матрицы на матрицу $[B]$. Заметим, что так как координаты точек P_i определяются значениями X и Y, то матрица $[B]$ имеет порядок $n \times 2$, где n — число точек, по которым строится сплайн-аппроксимация. В (4.51) указан вид членов матрицы $[B]$. Первый столбец матрицы $[B]$ находится подстановкой вместо длин хорд значений t_k и значений координат X наблюдений. Второй столбец строится аналогично, только подставляются координаты Y.

Матрица решений P' также имеет размеры $n \times 2$. Каждая строка P' характеризует наклон касательной к сплайну в точке наблюдения, заданной координатами X и Y.

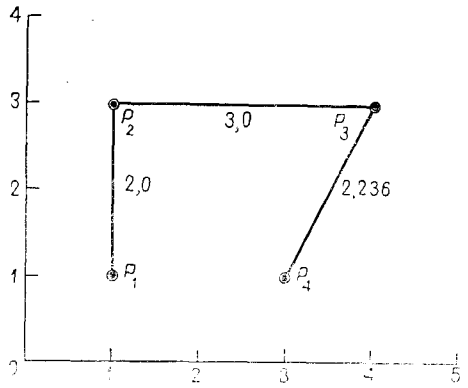


Рис. 4.24. Четыре точки, по которым строится кубический сплайн. Указаны длины хорд между точками

Для нахождения четырех коэффициентов β, определяющих k-ый сплайн (т. е. линию, связывающую точки P_k и P_{k+1}), мы имеем

$$\beta_1 = P_k;$$
$$\beta_2 = P'_k;$$
$$\beta_3 = \frac{3(P_{k+1} - P_k)}{t_{k+1}^2} - \frac{2P'_k}{t_{k+1}} - \frac{P'_{k+1}}{t_{k+1}};$$
$$\beta_4 = \frac{2(P_k - P_{k+1})}{t_{k+1}^2} + \frac{P'_k}{t_{k+1}^2} + \frac{P'_{k+1}}{t_{k+1}^2}. \qquad (4.52)$$

Наконец, если четыре коэффициента найдены для k-го звена, можно определить точки вдоль звена кривой в пределах этого интервала. Длина хорды между точками k и $k+1$ должна быть разделена на соответствующее число частей и эти последовательные расстояния должны быть подставлены вместо t в уравнение (4.48). Это обеспечит нам множество регулярно расположенных в пространстве точек, связанных между собой так, что получается кривая сплайн-аппроксимации. Этот процесс повторяется для каждого сегмента кусочно определенного сплайна, причем используются как угловые коэффициенты, так и длины хорд, связывающих внутренние точки, а также координаты точек с целью нахождения нового множества коэффициентов для каждого участка сплайна.

Проиллюстрируем этот метод на примере кубического сплайна, при этом будем использовать четыре точки, указанные на рис. 4.24, координаты которых имеют вид

$$P = \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 4 & 3 \\ 5 & 1 \end{bmatrix}$$

Длины хорд равны $t_2=2{,}0$; $t_3=3{,}0$ и $t_4=2{,}236$. Это все, что требуется для построения матрицы $[M]$ определенной по формуле (4.50).

$$M = \begin{bmatrix} 1{,}0 & 0{,}5 & 0 & 0 \\ 3{,}0 & 10{,}0 & 2{,}0 & 0 \\ 0 & 2{,}236 & 10{,}472 & 3{,}0 \\ 0 & 0 & 2{,}0 & 4{,}0 \end{bmatrix}.$$

Матрица, обратная к ней, равна

$$[M]^{-1} = \begin{bmatrix} 1{,}1375 & -0{,}0625 & 0{,}0139 & -0{,}0104 \\ -0{,}3749 & 0{,}1250 & -0{,}0279 & 0{,}0209 \\ 0{,}0934 & -0{,}0311 & 0{,}1184 & -0{,}0888 \\ -0{,}0467 & 0{,}0156 & -0{,}0592 & 0{,}2944 \end{bmatrix}.$$

Мы должны также определить матрицу правой части $[B]$. Необходимая информация для нахождения элементов $[B]$ состоит из длин хорд и координат точек. Так как каждая точка имеет две координаты, то вектор $[B]$ имеет два столбца, первый для X, второй — для Y

$$\begin{bmatrix} \frac{3}{2\cdot 2}(1-1) & \frac{3}{2\cdot 2}(3-1) \\ \frac{3}{2\cdot 3}[2^2(4-1)+3^2(1-1)] & \frac{3}{2\cdot 3}2^2(3-3)+3^2(3-1) \\ \frac{3}{3\cdot 2{,}236}[3^2(3-4)+2{,}236^2(4-1)] & \frac{3}{3\cdot 2{,}236}[3^2(1-3)+2{,}236^2(3-3)] \\ \frac{6}{2{,}236}(3-4) & \frac{6}{2{,}236}(1-3) \end{bmatrix} =$$

$$= \begin{bmatrix} 0 & 1{,}5 \\ 6 & 9 \\ 2{,}683 & -8{,}050 \\ -2{,}683 & -5{,}367 \end{bmatrix}.$$

Умножая $[B]$ на $[M]^{-1}$, получаем

$$[P]' = \begin{bmatrix} -0{,}3097 & 1{,}2026 \\ 0{,}6187 & 0{,}6750 \\ 0{,}3723 & -0{,}6160 \\ -0{,}3552 & -1{,}0328 \end{bmatrix}$$

Теперь мы имеем все, что необходимо для вычисления коэффициентов сплайна для каждого звена в нашем примере. Для по-

лучения уравнения первого из них надо подставить соответствующие значения t, P и P' в уравнение (4.52). Получим для координаты X

$$\beta_1 = 1;$$

$$\beta_2 = -0{,}3097.$$

$$\beta_3 = \frac{3(1-1)}{2^2} - \frac{2(-0{,}3097)}{2} - \frac{0{,}6187}{2} = 0{,}0004;$$

$$\beta_4 = \frac{2(1-1)}{2^3} + \frac{(-0{,}3097)}{2^2} + \frac{0{,}6187}{2^2} = 0{,}0773;$$

для координаты Y

$$\beta_1 = 1;$$

$$\beta_2 = 1{,}2026;$$

$$\beta_3 = \frac{3(3-1)}{2^2} - \frac{2(1{,}2026)}{2} - \frac{0{,}6750}{2} = -0{,}0401;$$

$$\beta_4 = \frac{2(1-3)}{2^3} + \frac{1{,}2026}{2^2} + \frac{0{,}6750}{2^2} = -0{,}0306,$$

или

$$[B] = \begin{bmatrix} 1 & 1 \\ -0{,}3097 & 1{,}2026 \\ 0{,}0004 & -0{,}0401 \\ 0{,}0773 & -0{,}0306 \end{bmatrix}.$$

Аналогичным образом мы можем определить коэффициенты сплайна для второго и третьего звена. Они равны

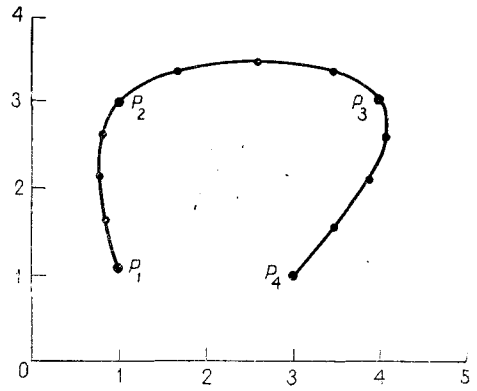
$$\begin{bmatrix} 1 & 3 \\ 0{,}6187 & 0{,}6750 \\ 0{,}4634 & -0{,}2447 \\ -0{,}1121 & 0{,}0066 \end{bmatrix} \begin{bmatrix} 4 & 3 \\ 0{,}3723 & -0{,}6160 \\ -0{,}5506 & -0{,}1872 \\ 0{,}0823 & 0{,}0280 \end{bmatrix}$$

Наконец, коэффициенты звеньев сплайна можно использовать для определения координат промежуточных точек на сплайне между каждой парой наблюдений. Если мы вычислим координаты большого числа таких точек и свяжем их затем прямыми линиями, визуально мы получим непрерывную гладкую кривую. Это как раз тот метод, на основе которого графическая система ЭВМ вычисляет и вычерчивает гладкие искривленные линии. Для иллюстрации мы ограничимся рассмотрением трех промежуточных точек каждого сплайна.

Для нахождения промежуточных точек мы сначала разделим каждую хорду на четыре части; расстояния в $t_k/4$, $2t_k/4$ и

Рис. 4.25. Гладкая сплайн-функция, состоящая из 30 сегментов между каждой точкой рис. 4.24.

Три промежуточных точки каждого сплайна, вычисленные как указано в тексте, изображены маленькими точками



$3t_k/4$ определяют значения t, которые должны быть подставлены в уравнение сплайна. Для первого сплайна эти расстояния равны 0,5; 1,0 и 1,5.

Вставляя их в уравнение (4.48), сначала для X, а потом для Y, получаем

$$\hat{P}_{5X} = 1 - 0{,}3097\,(0{,}5) + 0{,}0004\,(0{,}5^2) + 0{,}0773\,(0{,}5^3) = 0{,}8549;$$

$$\hat{P}_{5Y} = 1 + 1{,}202\,(0{,}5) - 0{,}0401\,(0{,}5^2) - 0{,}0306\,(0{,}5^3) = 1{,}5874.$$

Аналогично мы можем вычислить координаты первого сплайна на расстоянии $t = 1{,}0$ и $t = 1{,}5$. Они равны

для $t_{1,0}$ [0,7679 2,1319],

для $t_{1,5}$ [0,7969 2,6104].

Этот процесс повторяется для второго и третьего сплайнов, а в результате получаем следующее множество координат

для сплайна 2	[1,6774	3,3714]
	]2,5924	3,4841]
	[3,4612	3,3548]
для сплайна 3	[4,050	2,6020]
	[3,8431	2,1163]
	[3,3442	1,5722]

Эти результаты представлены на рис. 4.25. Также показан гладкий сплайн, порожденный вычислением 30 промежуточных точек между каждой парой узлов. Хотя процедура вычисления коэффициентов сплайна запутанная, овладение ею позволяет относительно просто получать столько точек на кривой, сколько потребуется.

СЕГМЕНТИРОВАННЫЕ ПОСЛЕДОВАТЕЛЬНОСТИ

Зонирование

Зонирование — это разделение последовательности в относительно однородные сегменты, каждый из которых отличен от прилегающих сегментов. Палеонтологи, например, прибегают к зонированию стратиграфической последовательности на основе соответствующей избыточности ископаемых микроэлементов. Данные опробования могут быть разделены в относительно однородные интервалы и будут представлять зоны постоянной литологии, соответствующей стратиграфическим единицам. Данные траверсов, полученные воздушной радиометрической разведкой, подразделяются на зоны, которые могут быть интерпретированы как пояса однородных скальных структур или минерализаций.

Имеется два основных подхода к зонированию. Простейший из них называется «локальным поиском границ». Он основан на поиске внезапного изменения средних значений, или, что эквивалентно, на поиске наиболее сильного изменения градиента последовательности. Для определения границ между почвенными зонами вдоль траверса Уэбстер [56] развил метод скользящего среднего. Ряд значений изучается последовательным движением короткого интервала вдоль последовательности. Скользящий интервал называется окном. Он распадается на две части: сегмент от точки $i+h$ последовательности до точки i и другой сегмент от точки i до точки $i-h$.

Мера, называемая обобщенным расстоянием, вычисляется для разностей между сегментами внутри двух половинок окна. Обобщенное расстояние есть отношение, образованное делением квадрата разности между средними значениями двух сегментов на объединенную дисперсию последовательностей в сегментах. Обозначим это среднее сегмента от x_i до x_{i+h} через $\bar{X}_1$ и его дисперсию через s_1^2; среднее сегмента от x_i до x_{i-h} через $\bar{X}_2$ и дисперсию через s_2^2. Тогда обобщенная разность есть

$$D^2 = \frac{(\bar{X}_1 - \bar{X}_2)^2}{s_1^2 + s_2^2}. \qquad (4.53)$$

Заметим, что объединенная дисперсия есть просто сумма дисперсий двух сегментов, так как оба сегмента содержат одно и то же число наблюдений. Также заметим, что первые и последние h точек последовательности не могут попасть в различные зоны.

Представление на графике h как функции D^2 приводит к преобразованию исходного траверса в новую последовательность, в которой границы зон имеют вид острых пиков. На рис. 4.26, *а* представлено первоначальное шестикилометровое сечение вдоль

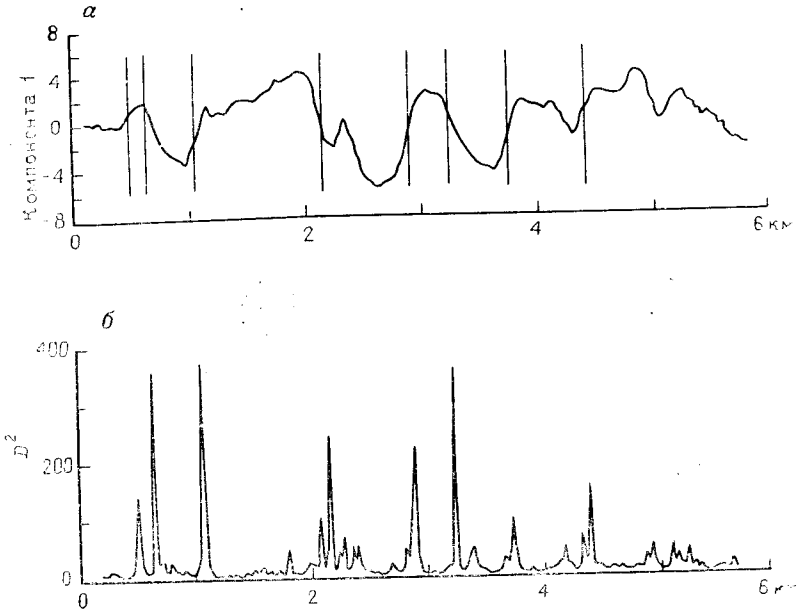


Рис. 4.26. Сечение, на котором указаны изменения свойств почвы вдоль 6-километровой линии в верхней части Тейме Валли (Англия):
а — изменение 27 свойств по первой главной компоненте; *б* — значение D^2 вдоль траверса. Максимумы определенных границ указаны на графике *а*

верхней части Тейм Вэлли в Англии [56]. Образцы почв были собраны с 20-метровым интервалом и проанализированы по 27 свойств. Это огромное множество измерений было сжато с помощью метода главных компонент, который будет рассмотрен в гл. 6. Здесь мы только заметим, что указанное пересечение представляет наиболее эффективную линейную комбинацию исходных переменных. На рис. 4.26, *б* представлена зависимость D^2 от расстояния вдоль пересечения, вычисленная с помощью расщепляющегося скользящего окна, которое охватывает 18 точек.

Уэбстер замечает, что выполнение этой процедуры зависит от изменчивости исходной последовательности и длины скользящего окна. Длинное окно должно усреднять вдоль малых зон и может пропустить короткие интервалы. Однако оно будет чувствительно к ошибкам изменчивости, возникающим вследствие шумовых помех в записи исходных данных. Короткое окно более чувствительно и будет отождествлять малые зоны, но может привести к иррегулярному неинтерпретируемому графику D^2. Уэбстер опубликовал программу на языке ФОРТРАН, которая реализует этот метод нахождения границ зон [57].

Одно из возражений против методов локального поиска границ состоит в том, что они могут привести к получению необычного числа границ, в частности, в наиболее изменчивой части последовательности. Глобальное зонирование основано на другой идее, а именно на разбиении последовательности в заранее заданное число сегментов, которые внутренне настолько однородны, насколько это возможно, и отличны от прилегающих сегментов настолько, насколько это возможно.

Одна из первых и наиболее практичных процедур такого рода была рекомендована Джиллом [21], который использовал итерационную версию дисперсионного анализа. Сначала последовательность делится на два сегмента, очень короткую начальную часть и остаток последовательности. Сумма квадратов внутри сегментов SS_w вычисляется по формуле

$$SS_w = \sum_{j=1}^{m} \sum_{i=1}^{n_j} (x_{ij} - \bar{X}_{.j})^2 \Big/ \sum_{j=1}^{m} n_j - m, \qquad (4.54)$$

где x_{ij} — i-я точка внутри j-го сегмента, $\bar{X}_j$ — среднее j-го сегмента, n_j — число точек в j-м сегменте, m — число сегментов. Сумма квадратов между сегментами SS_b является мерой изменчивости средних сегментов относительно $\bar{X}_{..}$, общего среднего объединенной последовательности или

$$SS_b = \frac{1}{m-1} \sum_{j=1}^{m} (\bar{X}_{.j} - \bar{X}_{..})^2. \qquad (4.55)$$

Разбиение на два сегмента движется вдоль последовательности, и в каждом положении вычисляются две величины SS_w и SS_b. Для каждого возможного положения границы вычисляется отношение

$$R = \frac{SS_b - SS_w}{SS_b}. \qquad (4.56)$$

Положение, соответствующее максимальному значению R, выбирается в качестве первой зональной границы.

Далее, повторением процесса вставки дополнительной границы, которая снова должна давать максимум R, эти две зоны снова подразделяются. Зонирование повторяется до тех пор, пока вся последовательность не разделится в заранее заданное число зон или пока величина R не перестанет увеличиваться при добавлении новых границ.

Процедура Джилла была использована для осуществления автоматического цифрового зонирования данных скважин. Недавно очень похожая процедура была опубликована Хоукингом и Мерриэмом [27], [28]. Они использовали глобальную оптимизацию, основанную на методах динамического программирования. Как и алгоритм Джилла, этот алгоритм является итеративным, но в силу его рекурсивности он имеет преимущество перед принципом оптимальности Беллмана в том, что в конце процедуры выбора границ зон дает самое лучшее разбиение из всех возможных. При использовании не рекурсивных процедур всегда возможно такое явление: позиция, выбранная как наилучшая граница между двумя зонами, не будет наилучшей, если в одну из зон будет вставлена дополнительная граница.

Хоукинг и Мерриэм вычисляют величину, которая есть сумма внутризонных дисперсий, эквивалентная величине SS_w в алгоритме Джилла. Если эта величина вычисляется для всех возможных разбиений последовательности в два сегмента, то результат будет некоторым табличным SS_w для $n-1$ возможных положений первой границы. Для каждого возможного первого разбиения затем вычисляется новое значение SS_w для всех возможных позиций второй границы, которая делит последовательность на три зоны. Выбирая наименьшее значение SS_w для второго разбиения, получаем оптимальное соответствующее положение первой границы и оно остается наилучшим, сколько бы дополнительных границ мы ни вставляли.

Процесс итерируется по третьему циклу, и для каждой комбинации оптимальной первой границы со всевозможными вторыми границами находятся всевозможные третьи границы и вычисляются соответствующие им значения SS_w. Выбор наименьшего значения SS_w затем определяет оптимальное положение второй границы. Процесс повторяется снова и снова до тех пор, пока не будет найдено заданное число границ.

В силу рекурсивной природы алгоритма окончательное множество зон обладает наименьшей возможной дисперсией среди любого возможного множества с тем же числом зон, покрывающих весь интервал. К сожалению, этот метод непрактичен при очень длинных последовательностях из-за высокой цены вычислительных работ, выполняемых для достижения оптимума.

Классификация по сходству

В этом разделе мы остановимся на методах упорядочения наблюдений на основе их относительного сходства по некоторому признаку. Если наблюдения характеризуются некоторым множеством переменных, то это в сущности требует их проектирования некоторым образом на единственную прямую, на которой их положение логически соответствует их расположению в исходном множестве данных. Это может быть осуществлено одним из многих методов, таких, например, как метод главных компонент или факторный анализ, которые рассматриваются в

последней главе. Такой классификацией может быть хронологическое упорядочение, широко используемое в археологии. К сожалению, нет гарантий того, что последовательность наблюдений, упорядоченная по сходству, будет хронологически упорядочена некоторым осмысленным образом. Понятия упорядочения по сходству и простого упорядочения ранее широко не использовались в геологии, исключая применения численной таксономии в палеонтологии [51]. Однако имеется область, в которой понятие упорядочения по сходству оказывается полезным, а именно, исследование геологической корреляции двух стратиграфических последовательностей.

Два петрографических ряда данных можно исследовать совместно на основе сходства их записей в буровых скважинах. При этом их перетасовывают, подобно колоде карт, с помощью процедур динамического программирования [23]. Каждая точка одной последовательности сравнивается с наиболее близкой точкой другой последовательности, причем соблюдается лишь условие, что стратиграфический порядок должен быть сохранен в обеих последовательностях. Таким путем достигается истинное упорядочение по сходству, так как окончательное размещение имеет смысл как литологический, так и хронологический. Можно ввести дополнительные ограничения, которые приведут к объединению специфических точек двух последовательностей, либо обусловят объединение какого-либо заданного сегмента одной последовательности с некоторой точкой другой. Таким образом, если основания маркеров в пределах двух сравниваемых последовательностей отождествляются, то эти основания должны соответствовать друг другу. Остальные члены коррелированы на основе наибольшего сходства, подчиняясь лишь следующим ограничениям: линии корреляции не могут пересекаться и основания маркеров должны быть коррелированы.

Алгоритм, опубликованный Гордоном и Реймент [23], напоминает алгоритм зонирования Хоукинга и Мерриэма [28]. Сначала каждая точка последовательности, соответствующей первой скважине, сравнивается с каждой точкой последовательности, соответствующей второй скважине. Имеется n наблюдений из первой скважины и m — из второй, результат сравнения — таблица размера $n \times m$. При этом можно использовать множество различных сравнительных мер, но Гордон и Реймент используют простую меру несходства:

$$D_{jk} = \sum_{l=1}^{p} \omega_l (u_{lj} - v_{lk}), \qquad (4.57)$$

где u_{lj} — отклик каротажной переменной l на глубине j в первой скважине, а v_{lk} — отклик той же переменной на глубине k

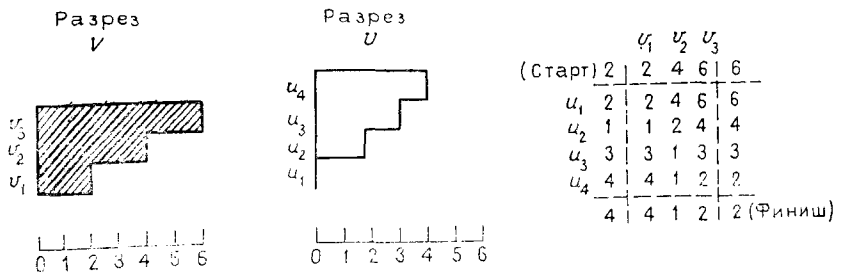


Рис. 4.27. Искусственные стратиграфические разрезы, изображенные вместе. Разрез V содержит три интервала, а разрез U — четыре. Характеристика, измеренная на каждом интервале, изменяется от 1 до 6. Матрица содержит простые меры расхождения между всеми возможными парами интервалов в разрезах U и V

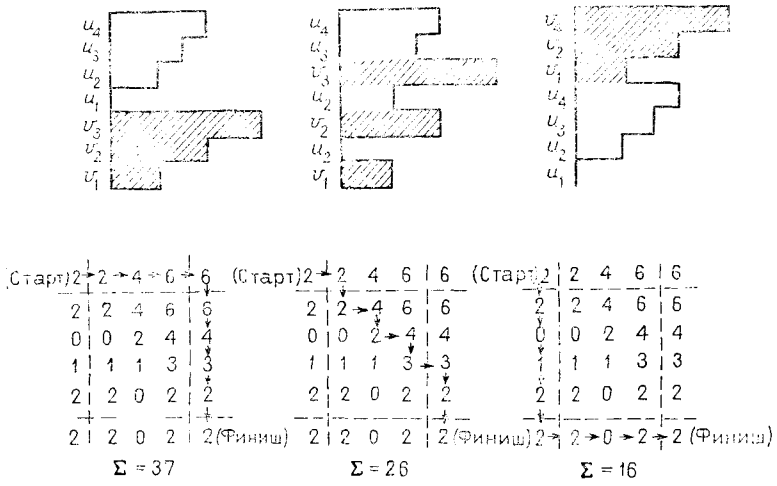


Рис. 4.28. Результаты, полученные объединением последовательностей U и V. Стрелки в матрицах показывают последовательность путей от основания до вершины. Общее расхождение обозначено через Σ

во второй скважине. Если угодно, различным переменным можно приписать веса ω_l.

Алгоритм динамического программирования позволяет найти единственный путь в этой таблице из левого верхнего угла в правый нижний угол таким образом, чтобы сумма мер несходства была минимальной. Стратиграфический порядок сохраняется благодаря тому, что исследователь может двигаться только вниз и направо. На рис. 4.27 изображен несложный пример, в котором проводится сравнение трех интервалов каротажа одной скважины с четырьмя интервалами другой. Заметим, что бо́льшая часть строк и столбцов матрицы повторяется; это

позволяет алгоритму спустить вниз ту или иную каротажную диаграмму к началу и закончить процесс. Некоторые примеры путей в матрице вместе с соответствующими им результатами стратиграфической корреляции приведены на рис. 4.28.

Для нахождения оптимального пути используется рекурсивная процедура. Начиная с левого верхнего угла (начало) выбираем первый интервал равным u_1 или v_1; мера несходства вдоль любого пути одинакова и равна 2+2. Если выбрать u_1, следующий интервал может быть либо u_2 с общей мерой несходства 2+2+0=4, или v_1 с общей мерой несходства 2+2+ +2=6. Другой вариант может быть таким: если v_1 выбрали на первом шаге, на втором шаге можно выбрать либо v_2, либо u_1. Пусть v_2 дает вклад в общую меру несходства, равную 2+ +2+2=6. Выбирая из этих двух случаев тот, в котором путь (начало)$\rightarrow u_1 \rightarrow u_2$ имеет минимальную меру несходства, получаем, что собственно первый шаг осуществляется в u_1.

Выбрав в качестве начальной точки u_1 с двумя возможными вариантами второго шага либо в u_2, либо в v_1, исследуем возможные варианты третьего шага. Снова имеется четыре возможности: из u_2 в u_3 с общей мерой несходства 2+2+0+1=5; из u_2 в v_1 (2+2+0+0=4); из v_1 в v_2 (2+2+2+4=10); и из v_1 в u_2 (2+2+2+0=6). Наименьшая сумма соответствует пути (начало)$\rightarrow u_1 \rightarrow u_2 \rightarrow v_1$, так что оптимальный второй шаг — переход в точку u_2. Далее проводится следующая итерация, в которой исследуется результат осуществления четырех возможных переходов из u_2. Путь с минимальным значением меры несходства определяет оптимальный шаг после u_2. Процесс повторяется до тех пор, пока нижняя (левая) точка (конец) будет достигнута. В приведенном примере оптимальный путь есть

$$(\text{начало}) \rightarrow u_1 \rightarrow u_2 \rightarrow v_1 \rightarrow u_3 \rightarrow v_2 \rightarrow u_4 \rightarrow v_3 \rightarrow (\text{конец})$$

с общей мерой несходства 10. Матрица и полученные стратиграфические последовательности изображены на рис. 4.29.

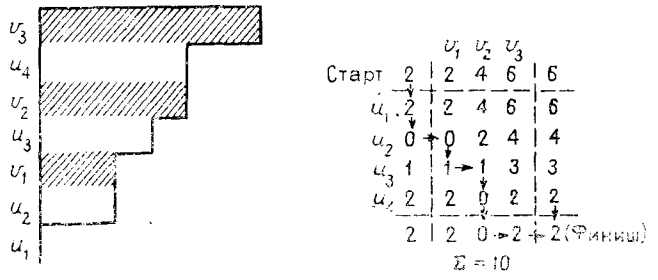


Рис. 4.29. Оптимальное объединение последовательностей U и V.

Стрелки в матрицах указывают порядок путей в сравниваемых последовательностях. Общее расхождение равно 16, т. е. наименьшему возможному значению для любой последовательности

В силу своей гибкости этот метод слежения кажется потенциально очень эффективным средством исследования корреляции. К сожалению, даже при использовании мощного аппарата динамического программирования он требует большого машинного времени, и совместное отслеживание длинных последовательностей можно рекомендовать лишь при выполнении исследований исключительной важности.

АВТОКОРРЕЛЯЦИЯ

На рис. 4.30 представлены результаты гамма-каротажа части скважины в разрезе Пенсильванских отложений в западном Канзасе. Разрез состоит из измененных известняков и сланцев. Из-за радиации калия-40 в глинистых минералах сланцы характеризуются относительно высоким фоном, в то время как известнякам свойственна низкая радиоактивность. В этом частном разрезе было замечено наличие циклотем, т.е. более или менее регулярных повторений литологических разновидностей. Беглого взгляда на результаты каротажа достаточно, чтобы убедиться в том, что известняк переслаивается со сланцами, которые имеют приблизительно ту же мощность.

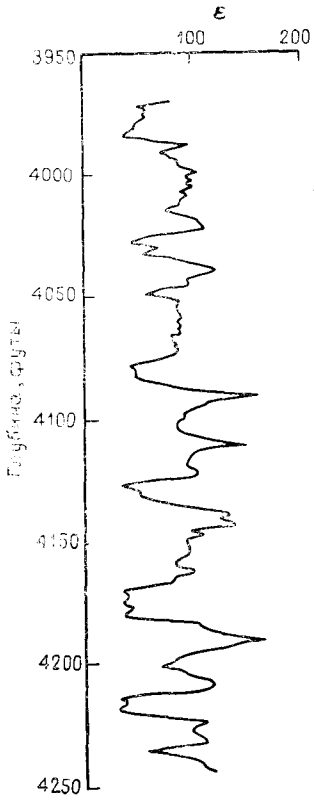


Рис. 4.30. Результаты гамма-каротажа части последовательности пенсильванских отложений в нефтяной скважине в Западном Канзасе

Повторения, так же как и другие свойства последовательности, устанавливаются с помощью вычисления меры сходства между членами этой последовательности, т. е. последовательность сравнивается с самою собой в последовательных положениях и вычисляется степень сходства между соответствующими интервалами. Если каждая точка сравнивается последовательно со всякой другой точкой, то обнаруживаются все позиции хорошего соответствия и также определяется степень несходства в других позициях. Для осуществления этой операции временной ряд должен иметь некоторые характеристики. Он должен состоять из последовательности наблюдений переменной Y, измеренной в последовательные моменты времени или в точках пространства. Каждое

наблюдение должно быть отделено от предшествующего наблюдения некоторым интервалом по времени или некоторым расстоянием, которые являются постоянными для данного ряда. Положение наблюдения в ряде мы будем обозначать нижним индексом, например Y_t. Поэтому необязательно явно указывать время или расстояние для переменной X, так как оно вполне характеризуется индексом и может быть в случае необходимости определено благодаря тому, что $X=\Delta t$, где Δ — расстояние между соседними точками. Весь временной ряд содержит n точек и имеет общую длину $T=\Delta(n-1)$.

Расстояние между двумя любыми точками Y_t и $Y_{t+\tau}$ называется лагом длины τ, где τ — число интервалов между двумя точками. Это — смещение временно́го ряда относительно себя самого в предшествующий момент времени или в предшествующем положении. Между временны́ми рядами и цепями можно провести аналогию. Каждой связи в цепи соответствует наблюдение в ряде. Если мы приложим два одинаковых сегмента цепи друг к другу и сравним их между собой, то получим попарное сравнение с лагом 0. Если мы сдвинем одну цепь на одно звено так, чтобы первое звено первой последовательности сравнивалось со вторым звеном второй, то все другие звенья также сдвинутся, и такое положение сравниваемых последовательностей называется имеющим лаг 1. Цепи могут быть сдвинуты более чем на одно звено, и попарное сравнение тогда будет иметь лаг 2 и так далее.

Автоковариация с лагом τ — это ковариация между всеми наблюдениями Y_t и наблюдениями $Y_{t+\tau}$, т. е. ковариация вычисляется между членами самого ряда и того же ряда, смещенного на лаг длины τ. Определяющее уравнение автоковариации есть

$$\mathrm{cov}_\tau = \frac{1}{n-\tau}\sum_{t=1+\tau}^{n} Y_t Y_{t-\tau} - \bar{Y}_t \bar{Y}_{t-\tau}. \qquad (4.58)$$

Автоковариация с лагом 0 — это просто дисперсия временно́го ряда. Если ряд очень длинный, а лаг τ короткий, то среднее данного ряда и сдвинутого рядов в сущности тождественны и уравнение (4.58) может быть упрощено. Однако если τ является значимой дробной частью длины временно́го ряда, то различие между средними становится существенным. Вычислительный эквивалент уравнения (4.58) есть

$$\mathrm{cov}_\tau = \frac{(n-\tau)\sum_{t=1+\tau}^{n} Y_t Y_{t-\tau} - \sum_{t=1+\tau}^{n} Y_t \sum_{t=1+\tau}^{n} Y_{t-\tau}}{(n-\tau)(n-\tau-1)}. \qquad (4.59)$$

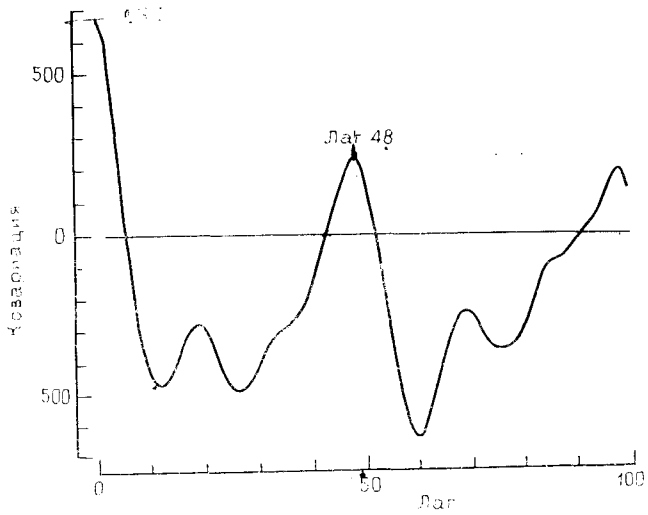


Рис. 4.31. Автокорреляционная функция гамма-каротажа, изображенного на рис. 4.30.
Лаг 48 соответствует сдвигу по глубине на 48 футов

Имеется соглашение о том, что автоковариация вычисляется для лагов от 0 до примерно $n/4$. Полученные значения можно представить как автоковариограмму или автоковариационную функцию, которая представляет зависимость автоковариации от лага. Рис. 4.31 представляет автоковариационную функцию для данных каротажа скважин, изображенных на рис. 4.30. Кривая начинается с максимального значения 672 для лага $\tau=0$, затем убывает и поднимается снова до лага 48, который соответствует расстоянию в пространстве, приблизительно равному 48 футам, так как данные гамма-каротажа были оцифрованы с интервалом в 1 фут. Это есть приблизительно величина вертикального расстояния между последовательными положениями известняков в последовательности, представленной на рис. 4.30.

Единицы, в которых измеряется автоковариация — это квадраты единиц измерений временного ряда; в нашем примере — это квадраты единиц электродвижущей силы E. Это означает, что автоковариация чувствительна к изменениям масштаба временно́го ряда, что доставляет затруднения при сравнении двух автоковариограмм. Однако если временной ряд стандартизован вычитанием из каждого наблюдения среднего и делением на стандартное отклонение, ряд будет представлен в единицах стандартного отклонения и автоковариация будет иметь стандартизованный вид. Как отмечалось в гл. 2, ковариация стандарти-

зованной переменной есть корреляция, то же верно и для автоковариаций.

Прежде чем стандартизировать наш временной ряд, мы можем вычислить сразу автокорреляцию, разделив автоковариацию на дисперсию ряда, т. е.

$$r_\tau = \frac{\operatorname{cov} v_\tau}{\operatorname{var} Y} = \frac{\sum_{t=1+\tau}^{n} Y_t Y_{t-\tau} - \bar{Y}_t \bar{Y}_{t-\tau}}{\sum_{t=1}^{n} (Y_t - \bar{Y})^2}. \qquad (4.60)$$

Автокорреляция временно́го ряда конечной длины n с лагом τ, который является значимой дробной частью n, может быть вычислена как отношение двух величин: числитель — выражение (4.59), знаменатель — оценка $\sqrt{\operatorname{var} Y_1 \operatorname{var} Y_{t+\tau}}$. Поскольку пределы суммирования в числителе и знаменателе одинаковы, то внешний вид результата можно упростить, сократив число наблюдений n—τ. Как и в формуле (4.59) пределы суммирования будут от $t=1+\tau$ до n.

$$r_\tau = \frac{\Sigma Y_t Y_{t-\tau} - \Sigma Y_t \Sigma Y_{t-\tau}}{\sqrt{\left[\Sigma Y_t^2 - (\Sigma Y_t)^2\right]\left[\Sigma Y_{t-\tau}^2 - (\Sigma Y_{t-\tau})^2\right]}}. \qquad (4.61)$$

На рис. 4.32 представлена автокоррелограмма данных гамма-каротажа, представленных на рис. 4.30. Заметим, что она идентична по форме автоковариационной функции, исключая лишь то, что она принимает значения между $+1{,}0$ и $-1{,}0$.

Необходимо отметить, что в терминологии теории временны́х рядов имеются некоторые разногласия. Некоторые авторы считают автокорреляцию параметром совокупности и используют термин сериальная корреляция как эквивалент статистики, вычисляемой по выборке. Другие используют термин сериальная корреляция в значении корреляции между двумя временны́ми рядами. Некоторые авторы последнюю корреляцию называют кросс-корреляцией. Мы будем использовать как термин автокорреляции, так и термин взаимная корреляция и не будем делать различий между статистиками и параметрами.

Заметим, что коррелограмма в действительности является двусторонней, с отрицательной частью — зеркальным отражением положительной части. Мы поясним это явление, вернувшись к нашей аналогии с двумя цепями. Если одну из цепей обозначить через A, а другую — через B и последовательно передвигать цепь A вперед по цепи B, то получающиеся коэффициенты автокорреляции составят положительную часть коррелограммы. Но если двигать цепь B вперед по цепи A и вычислять автокорреляцию, то лаги будут отрицательными, так как относительное движение двух последовательностей поменяло

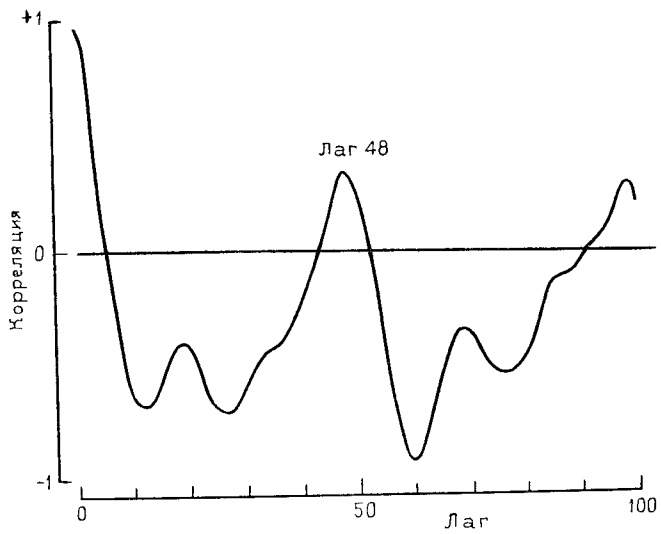


Рис. 4.32. Автокоррелограмма гамма-каротажа, изображенного на рис. 4.30

направление на обратное. Цепи A и B идентичны, поэтому безразлично, в каком направлении одна из них сдвигается относительно другой, так как $r_{-\tau}=r_\tau$. На практике мы не рассматриваем отрицательную часть автокоррелограммы, но она играет некоторую роль при разложении в ряд Фурье.

Коррелограммы напоминают характеристики временны́х рядов. Сравнивая коррелограммы временно́го ряда с коррелограммами идеализированных моделей, легко убедиться в том, что они хорошо согласуются. Простейшая модель временно́го ряда — это последовательность независимых и нормально распределенных наблюдений. Это значит, что между наблюдениями в некоторый момент времени t и наблюдениями в другой момент времени $t+\tau$ нет никакой связи. Такое условие естественно в том случае, когда временной ряд порожден случайным процессом. Ожидаемая автокорреляция для этой модели есть $\rho_\tau=0$ для всех лагов τ, больших нуля. Теоретическая коррелограмма в этом случае — прямая линия, проходящая через точку $r=0$.

Другие модели предполагают наличие некоторой зависимости между последовательными наблюдениями. Например, предположим, что некоторый процесс порождается случайными нормально распределенными наблюдениями, которые усредняются тотчас же, как были порождены, т. е.

$$Y_t = \sum_{t-w}^{t} \frac{Z_t}{w}.$$

Это модель скользящего среднего и она имеет коррелограмму вида

$$\rho_\tau = 1 - \tau/w. \qquad (4.62)$$

Имеется широкий спектр моделей возрастающей сложности, которые можно было бы предложить. Хорошее введение в эту тему приводят Юл и Кендел [63], в применении к гидрологии — Йевьевич [64]. Рис. 4.33 иллюстрирует построение сложного временного ряда, являющегося линейной комбинацией более простых рядов. На рис. 4.33, *а* представлена регулярная правильная синусоидальная волна и ее коррелограмма. По мере того как ряд сдвигается по фазе относительно самого себя или в случае, если пики спариваются с впадинами, коррелограмма изменяется от +1 до 0 и затем до —1. Затем коэффициент корреляции снова растет, пока не достигнет значения +1 в тот момент, когда сигнал сдвинется ровно на одну длину волны. На рис. 4.33, *б* изображен сигнал, полученный с помощью первой из рассмотренных выше моделей, — последовательность случайных чисел. Коррелограмма быстро убывает от 1 с нулевым лагом, потом слабо колеблется около нуля. Оба эти ряда являются стационарными, т. е. в наблюдениях нет значительного тренда. Нестационарный сигнал изображен на рис. 4.33, *в*, где наблюдения, образующие последовательность, неуклонно растут по величине. Соответствующая коррелограмма показывает, что корреляция неуклонно убывает. На рис. 4.33, *г* изображена комбинация случаев 4.33, *а* и 4.33, *б*, т. е. синусоидальной волны и наложенного на нее шума. Полная автокорреляция возможна только при нулевом лаге, однако о периодической компоненте временно́го ряда напоминает пик коррелограммы, имеющей два пересечения с нулевым уровнем. График 4.33, *д* представляет собой комбинацию графиков 4.33, *а*, 4.33, *б* и 4.33, *в*, т. е. синусоидальная волна складывается с графиком линейного тренда и с наложенным на них шумом. Заметим, что тренд значительно снижает наши возможности выделить периодическую компоненту в сигнале.

Как отмечалось выше, математическое ожидание или среднее значение автокорреляции для последовательности случайных чисел равно нулю. Ожидаемая теоретическая дисперсия автокорреляции случайной последовательности при любом лаге τ равна

$$\sigma_\tau^2 = \frac{1}{n - \tau - 3}. \qquad (4.63)$$

Эти два параметра определяют совокупность элементов случайного временного ряда данной длины n [63]. Как уже отмечалось в гл. 2, можно определить вероятность появления некоторого случайного события для нормального распределения с

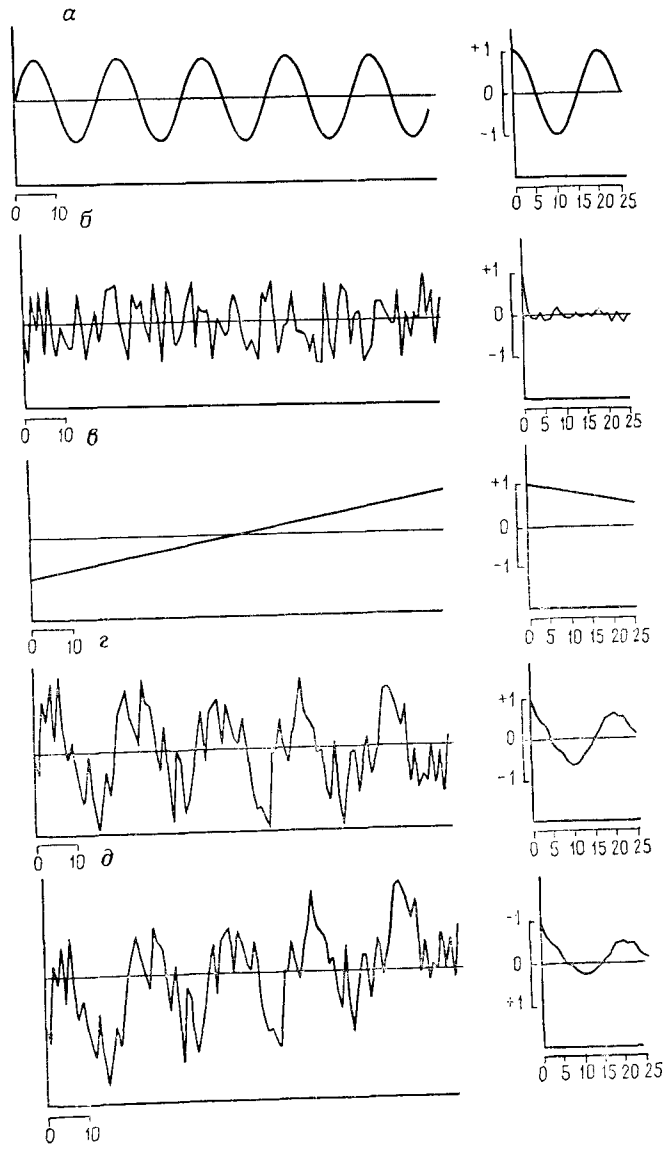


Рис. 4.33. Примеры идеализированных временных рядов и их автокорреляционные функции:

а — синусоидальная волна с длиной 20 единиц; *б* — последовательность случайных чисел или шум; *в* — последовательность, возрастающая по линейному закону, или «тренд»; *г* — сумма синусоидальных волн и случайного шума (последовательность *а* плюс последовательность *б*); *д* — сумма синусоидальной волны со случайным шумом и линейным трендом (последовательность *а* плюс последовательность *б* плюс последовательность *в*)

известным средним значением и заданной дисперсией, если ввести стандартизованную величину

$$Z = \frac{x - \mu}{\sigma}. \qquad (4.64)$$

Извлекая квадратный корень из выражения (4.63), мы получим стандартное отклонение автокорреляции. Подставив среднее значение и стандартное отклонение в формулу (4.64), получим

$$Z_\tau = \frac{r_\tau - 0}{1/\sqrt{n - \tau + 3}} = r_\tau \sqrt{n - \tau + 3}. \qquad (4.65)$$

Эту величину можно использовать в качестве статистики для проверки гипотезы, заключающейся в том, что автокорреляция r_τ равна нулю при условии, когда длина последовательности n велика, а лаг τ мал. «Много» и «мало» — термины, которые трудно определить точно. Обычно n выбирают не более 50, а τ — не превосходящим $n/4$. (Некоторые авторы устанавливают еще более низкие пределы для τ, например $n/10$ или еще меньше). Эти ограничения основываются на том, что если лаг увеличивается, то значение r_τ приходится вычислять по все меньшему и меньшему числу наблюдений. Это приводит не только к увеличению дисперсии r_τ, но также к все более значительному нарушению предположения о том, что автокорреляция вычисляется по выборке из временного ряда бесконечной длины. По указанным причинам высокие значения автокорреляционной функции при большом лаге не играют существенной роли, пока сам временной ряд во много раз длиннее.

В табл. 4.22 представлены необычные для геолога данные, а именно наблюдаемые значения истинного временного ряда из геологического прошлого. Только при очень редком стечении обстоятельств удается датировать прошедшие события по результатам изучения пород, что в свою очередь позволяет определить временную шкалу для последовательности геологических данных. Эоценовые озерные отложения в Скалистых горах состоят из тонкослоистых доломитизированных нефтяных сланцев мощностью в сотни футов. Было установлено, что слоистость претерпевает изменения, т. е. является результатом сезонных климатических изменений в бассейне озера. Измерения мощности этих слоев позволили определить годичные изменения скорости осадконакопления за время существования озера. Табл. 4.22 содержит значения мощностей слоев, замеренных в разрезе этих отложений вблизи западного побережья одного из самых больших озер. Мы попытаемся ответить на несколько вопросов, связанных с этими данными. Например, существовал ли тренд в скорости отложения доломитов с течением времени, причиной

Таблица 4.22

Мощность последовательности слоев в разрезе нефтеносных сланцев Грин Ривер, мм

(Верх разреза)	6,0	8,6	10,8	4,2	
	7,2	9,0	9,5	4,5	
	7,1	12,0	8,1	5,9	
	7,1	13,7	7,2	7,3	
	7,2	14,0	7,1	7,3	
	7,4	13,6	6,8	6,7	
	8,0	12,1	7,0	6,0	
	8,6	12,9	7,1	5,8	
	10,0	12,8	5,6	5,7	
	11,4	11,1	3,8	6,5	
	12,0	9,0	3,4	8,2	
	11,0	7,5	4,2	10,2	
	9,6	7,5	4,8	12,3	
	8,7	8,4	4,5	13,2	
	7,6	8,4	3,6	13,2	
	7,2	7,9	3,0	12,4	
	7,2	7,0	2,8	9,7	
	7,8	6,7	4,1	9,2	
	8,1	6,8	6,8	9,3	
	7,8	7,3	8,1	8,3	
	7,1	7,3	7,8	6,0	
	7,2	7,2	6,4	5,7	
	7,1	8,1	4,6	6,1	
	7,0	9,8	3,7	6,3	
	7,0	11,0	4,0	6,3	(Основание разреза)
	7,7				

которого могло явиться постепенное изменение климата? Имеется ли очевидная цикличность мощности слоев, которая, возможно, имеет связь с астрономическими явлениями? При наличии цикличности мы можем определить многолетние периоды (связанные, например, с солнечными пятнами за последние 11 лет). Было проведено 101 наблюдение.

Выполните анализ данных и определите, имеются ли значимый тренд или периодичность в данных, представленных в таблице, характеризующей изменения мощности. Напомним, что данные должны быть стационарными, так что до выполнения автокорреляционного анализа любой значимый тренд нужно устранить.

В геологии имеется одна большая область, в которой исследования проводятся постоянно, — это предсказание землетрясений. Эти исследования хорошо субсидируются правительством, первые доклады указывают на некоторый успех в области краткосрочного прогнозирования больших землетрясений. Это делается с помощью изучения некоторых сейсмических волн, предшествующих основному событию.

Долгосрочный прогноз землетрясений — это совсем другая проблема. Ее решение требует обнаружения периодичностей или

Таблица 4.23

Показатель активности известных землетрясений за период 1770—1869 гг. [42]

1770	66	1795	78	1820	90	1845	86
1771	62	1796	110	1821	86	1846	127
1772	66	1797	79	1822	119	1847	201
1773	197	1798	85	1823	82	1848	76
1774	63	1799	113	1824	79	1849	64
1775	0	1800	59	1825	111	1850	31
1776	121	1801	86	1826	60	1851	138
1777	0	1802	199	1827	118	1852	163
1778	113	1803	53	1828	206	1853	98
1779	27	1804	81	1829	122	1854	70
1780	107	1805	81	1830	134	1855	155
1781	50	1806	156	1831	131	1856	97
1782	122	1807	27	1832	84	1857	82
1783	127	1808	81	1833	100	1858	90
1784	152	1809	107	1834	99	1859	122
1785	216	1810	152	1835	99	1860	70
1786	171	1811	99	1836	69	1861	96
1787	70	1812	177	1837	67	1862	111
1788	141	1813	48	1838	26	1863	42
1789	69	1814	70	1839	106	1864	97
1790	160	1815	158	1840	108	1865	91
1791	92	1816	22	1841	155	1866	64
1792	70	1817	43	1842	40	1867	81
1793	46	1818	102	1843	75	1868	162
1794	96	1819	111	1844	99	1869	137

тренда в исторических регистрациях землетрясений, которые могут быть экстраполированы на будущее. Цель таких исследований — не предсказание конкретного землетрясения, а предсказание тех периодов, когда сейсмическая активность будет необыкновенно высокой.

В табл. 4.23 приведены показатели сейсмической активности во всем мире за 100 лет, основанные на ежегодных данных регистрации сильных землетрясений [42]. Исследовать эти записи — это значит определить, имеется ли в них значимый тренд, периодичности или другие специфические особенности. Если ответ положителен, то как эта информация может быть использована для построения прогнозной модели?

ВЗАИМНАЯ КОРРЕЛЯЦИЯ

Если можно провести сравнение временно́го ряда с самим собой для последовательных лагов с целью обнаружения зависимости во времени, то, вероятно, можно также сравнить два временны́х ряда друг с другом для того, чтобы определить положения наибольшего соответствия. Два типа информации выходят на первый план при таком сравнении: сила связи между

двумя рядами и лаг или сдвиг во времени или расстояние между ними в их положении максимальной эквивалентности. Процесс сравнения двух временных рядов при последовательных лагах называется взаимной корреляцией (кросс-корреляцией). Во многих примерах невозможно обозначить положение нулевого лага, так как любой из двух рядов может быть ведущим для другого. Поскольку два ряда не идентичны, то кросс-взаимная коррелограмма не симметрична относительно середины; лаги, с которыми ряд A ведет ряд B, отличаются от лагов, с которыми B ведет A (рис. 4.34). Осложнения возникают в тех случаях, когда ряд A не обязательно имеет ту же длину, что и ряд B. Действительно, один из подходов к «автоматической корреляции» (под этим понимается установление равенства двух стратиграфических последовательностей в геологическом смысле) состоит в скольжении короткой дизъюнктивной части одного стратиграфического интервала после другого целого стратиграфического разреза с целью определения положения наибольшего согласования.

Определение коэффициента взаимной корреляции такое же, как и определение обычного коэффициента корреляции, и лишь немного отличается от коэффициента автокорреляции. Если мы введем обозначения Y_{1i} и Y_{2i} для двух рядов и определим n^* как число перекрывающихся позиций в двух цепях, то взаимная корреляционная функция для m сходных позиций вычисляется по формуле

$$r_m = \frac{n^*\Sigma Y_1 Y_2 - \Sigma Y_1 \Sigma Y_2}{\sqrt{[n^*\Sigma Y_1^2 - (\Sigma Y_1)^2]\,[n^*\Sigma Y_2^2 - (\Sigma Y_2)^2]}} \qquad (4.66)$$

или

$$r_m = \frac{\mathrm{cov}_{1,2}}{S_1 S_2}, \qquad (4.67)$$

где $\mathrm{cov}_{1,2}$ — ковариация перекрывающихся участков двух последовательностей 1 и 2, S_1 и S_2 — соответствующие стандартные отклонения. Заметим, что суммирования происходят только по таким отрезкам двух последовательностей, которые перекрываются в спаренной позиции. Спаренные позиции нумеруются последовательно, как указано на рис. 4.34, и кросс-коррелограмма есть график спаренной позиции в зависимости от кросс-корреляции.

Так как суммирования распространяются только на перекрытый отрезок, то вычисление коэффициента взаимной корреляции основывается на дисперсии, вычисленной по общим точкам. Наоборот, коэффициент автокорреляции вычисляется с помощью дисперсии, вычисленной по всей цепи. Дисперсия используется в знаменателе определяющего автокорреляционного уравнения и

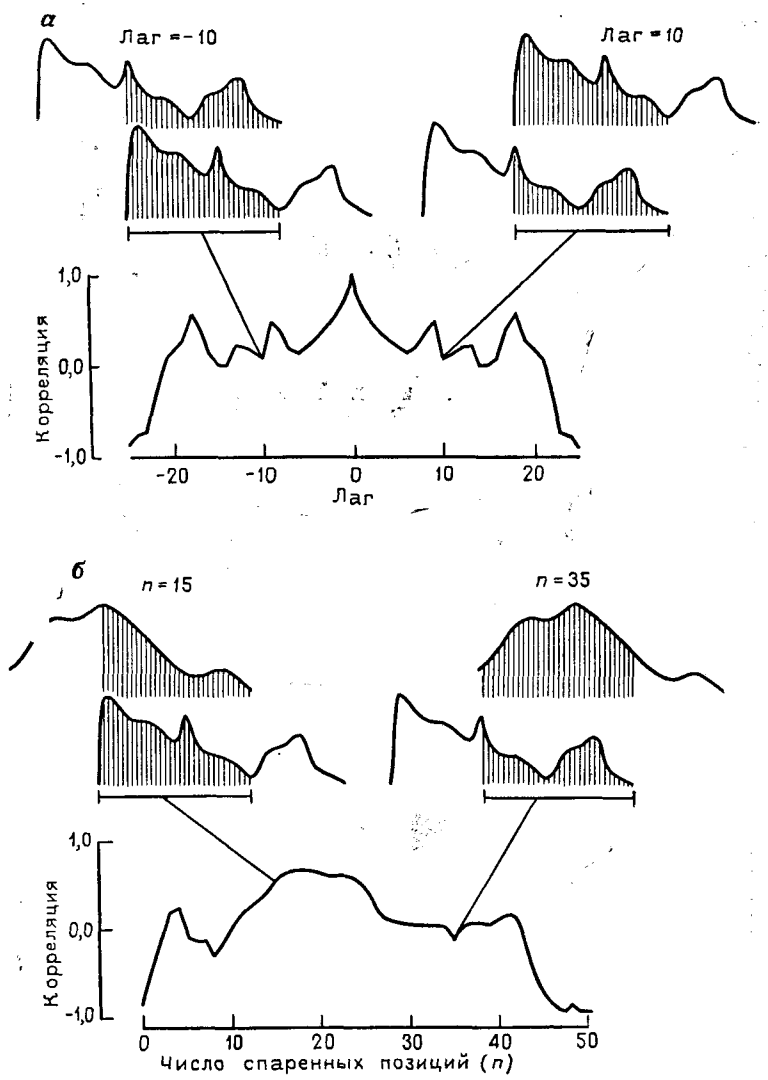


Рис. 4.34. Коррелограммы:

а — полученная при вычислении автокорреляционной функции для положительных и отрицательных значений лага (коррелограмма симметрична относительно нулевого значения лага); *б* — полученная для двух различных последовательностей (коррелограмма асимметрична, если обе последовательности не совпадают)

считается постоянной для всех значений лага. Дисперсия, вычисленная для целой последовательности, более устойчива, чем оценка, полученная для более короткого отрезка этого ряда, и поэтому более предпочтительна. Однако при вычислении коэффициентов взаимной корреляции не следует ожидать, что дисперсии будут постоянными на протяжении обеих цепей, в особенности в том случае, когда одна из цепей короче другой.

Значимость коэффициента взаимной корреляции можно установить с помощью приближенного критерия

$$t = r_m \sqrt{\frac{n^* - 2}{1 - r_m^2}}, \qquad (4.68)$$

имеющего $n^* - 2$ степеней свободы. Этот критерий выводится из критерия значимости коэффициента корреляции между двумя выборками, взятыми из нормальной совокупности. При этом нулевая гипотеза заключается в том, что истинный коэффициент корреляции между двумя последовательностями в заданной спаренной позиции равен нулю, т. е. мы имеем случайный ряд, что можно ожидать в случае, если две последовательности независимы.

Взаимная корреляция — наиболее подходящий прием для сравнения двух рядов, которые имеют временную зависимость между собой. В качестве примера можно проанализировать данные, приведенные в табл. 4.24. «Арсенал» в Скалистых горах — это завод по производству различных ядовитых веществ военного назначения; он расположен в Денвере (штат Колорадо), вблизи от первой гряды Скалистых гор. Огромные количества загрязненной воды возникают как побочный продукт военной промышленности. Пытаясь избавиться от этой грязной воды, в 1961 г. пробурили для ее закачки скважину в основании гор. К сожалению, скважина проникла в вертикальный срез большой складки вдоль фронта Скалистых гор, и очевидно, что закачка под большим давлением грязной жидкости послужила увеличению подвижности складки в результате ее смазки. Один ряд данных, приведенных в табл. 4.24, — это помесячные записи объемов жидкости, введенных в скважину «Арсенала» в Скалистых горах за 4 года использования скважины. Другой ряд представляет собой число землетрясений, зарегистрированных в Денвере в каждом месяце. При изучении статистической связи между этими двумя временными рядами Бадуэлл [4] представил их в кумулятивной форме и сделал вывод о том, что имеется резко выраженный 3-месячный лаг между закачкой и инцидентным землетрясением. К сожалению, этот метод визуального сравнения двух кривых скорее всего ошибочный, так как при его использовании устанавливается соответствие между двумя рядами записей, имеющих разный масштаб.

Таблица 4.24

Объем жидкости, закачанной в скважину, пробуренную в Скалистых Горах, и число зарегистрированных землетрясений в Денвере (штат Колорадо) [4]

Время года	Объем закачанной жидкости (миллионы галлонов)	Число землетрясений
1962 г. Март	4,2	—
Май	7,2	2
Апрель	8,4	12
Июнь	8,0	35
Июль	5,2	23
Август	6,0	29
Сентябрь	5,0	24
Октябрь	5,6	8
Ноябрь	4,0	6
Декабрь	3,6	20
1963 г. Январь	6,0	25
Февраль	7,6	22
Март	7,8	21
Апрель	6,4	42
Май	3,6	21
Июнь	4,0	8
Июль	3,4	6
Август	2,4	10
Сентябрь	3,9	11
Октябрь	0	12
Ноябрь	0	4
Декабрь	0	2
1964 г. Январь	0	5
Февраль	0	2
Март	0	9
Апрель	0	9
Май	0	2
Июнь	0	4
Июль	0	4
Август	0	5
Сентябрь	0,6	2
Октябрь	1,8	14
Ноябрь	2,4	2
Декабрь	2,0	7
1965 г. Январь	2,0	1
Февраль	1,7	30
Март	1,7	9
Апрель	3,6	19
Май	4,0	11
Июнь	6,4	38
Июль	8,9	62
Август	5,4	48
Сентябрь	6,4	87
Октябрь	3,8	5

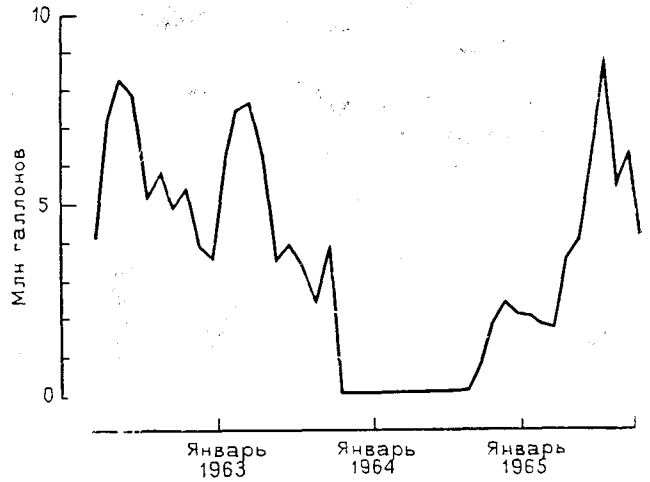


Рис. 4.35. Количество отравленной жидкости, закачанной в скважину в Скалистых горах

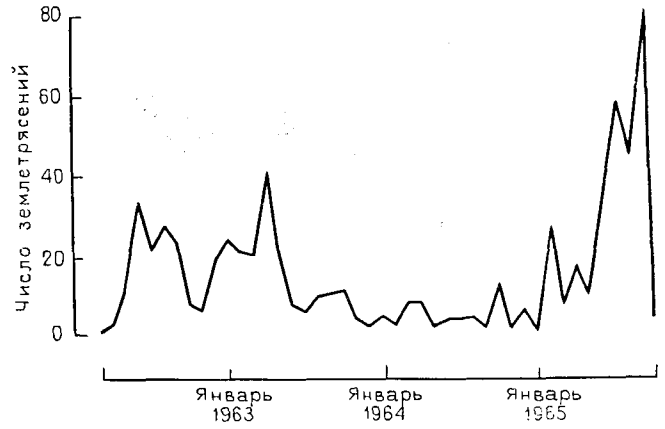


Рис. 4.36. Число зарегистрированных в течение месяца землетрясений, эпицентр которых был расположен вблизи Денвера (штат Колорадо)

Эти два ряда записей представлены в графической форме на рисунках 4.35 и 4.36. Взаимная коррелограмма этих двух временных рядов показана на рис. 4.37. Так как в этом примере оба ряда имеют общее начало и временную шкалу, то позиции спаривания можно охарактеризовать как «положительные» и «отрицательные» лаги, начиная с положения начального совпадения. Положение наибольшего соответствия между двумя ря-

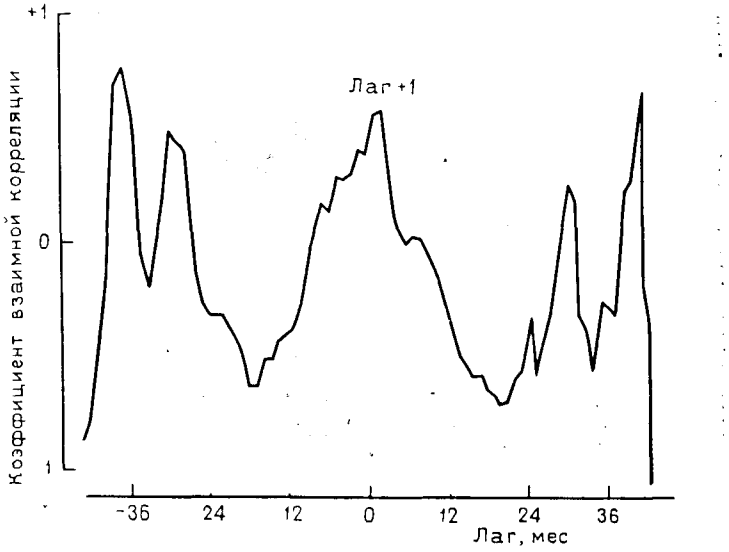


Рис. 4.37. Взаимная коррелограмма, характеризующая связь числа землетрясений в течение месяца и объемов отравленной жидкости, введенной в течение месяца в скважину.

Максимальная корреляция имеет место при лаге +1 (явно высокие корреляции при очень больших лагах статистически незначимы из-за низкого числа наблюдений)

дами случается при лаге +1, когда число землетрясений в данном месяце сравнимо с объемом грязной воды, закачанной месяцем раньше. Корреляция в этой позиции спаривания есть $r = 0{,}60$. Вторая наивысшая кросс-корреляция $r = 0{,}57$ получается при лаге 0. Это указывает на то, что тектонический отклик на операцию закачки возникает очень быстро и распространяется на период около месяца.

Иногда бывает необходимо сравнить два периодических временных ряда с одинаковыми периодами. Использовав преимущество периодичности, вычислим взаимную коррелограмму в «круговой» форме. Действительно, каждый временной ряд вкладывается в окружность и начало ряда таким образом связывается с его концом. Вместо того чтобы представлять себе два ряда цепями, движущимися друг относительно друга, их можно представить себе в форме колес. Каждое колесо имеет одно и то же число делений вокруг его обода, каждое из которых соответствует одному из последовательных наблюдений. Если одно колесо вращается по отношению к другому, то можно провести кросс-сравнение между ними в n различных положениях до начала повторений.

В Шизапик Бей, как и во многих устьях рек, в течение дневного приливно-отливного цикла наблюдается комплексное изме-

Таблица 4.25

Соленость воды в Шизапик Бей. Место отбора проб — Аннаполис (штат Мэриленд) вдали от побережья, 1927 г. [58]

Время	Прилив или отлив	Соленость, мкг/г	
		поверхностная	придонная
3 июля 14.30	1/4 отлива	6,97	11,10
16.00	1/2 »	6,20	11,54
17.30	Отлив	5,93	12,12
19.00	»	6,32	13,52
20.30	1/4 прилива	6,36	13,35
22.00	1/2 »	6,72	12,83
23.30	3/4 »	6,80	13,31
4 июля 1.00	Прилив	6,90	13,02
2.30	1/4 отлива	7,14	12,14
5.30	1/2 »	6,91	12,44
4.00	Отлив	6,76	12,60
7.00	Начало прилива	6,74	12,79
8.30	3/8 »	6,20	13,46
10.00	1/2 »	7,45	12,33
11.30	3/4 »	7,47	12,40
13.00	Прилив	7,47	12,14

нение солености из-за перемешивания пресной воды с морской. Чистая вода Шизапик Ривер течет сквозь плотную морскую воду в бассейн; во время отлива эта вода движется дальше вниз по устью. Однако имеется встречный поток вдоль дна, который во время малой воды переносит плотную морскую воду вверх по бассейну. В табл. 4.25 приведены результаты измерения солености проб воды, взятых с интервалом в 1,5 ч в течение суток (средняя глубина 11 м). В графической форме данные представлены на рис. 4.38.

Корреляционная функция представлена на рис. 4.39 и, очевидно, указывает на лаг 2, представляющий разницу в 3 ч между пиком появления соленой воды вблизи дна и максимумом содержания соли в поверхностных водах. Так как записи представляют 24-часовой дневной повторяющийся цикл, то взаимная корреляция может быть представлена в круговой форме. Поэтому коэффициенты взаимной корреляции можно найти для любых лагов вплоть до 24 ч.

Взаимная корреляция и геологическая корреляция

Геологи постоянно делают попытки применить методы взаимной корреляции для сравнения стратиграфических последовательностей, однако, сталкиваясь с трудностями, обращаются к автоматической геологической корреляции. Несмотря на то что литература по этим вопросам очень обширна, их усилия дости-

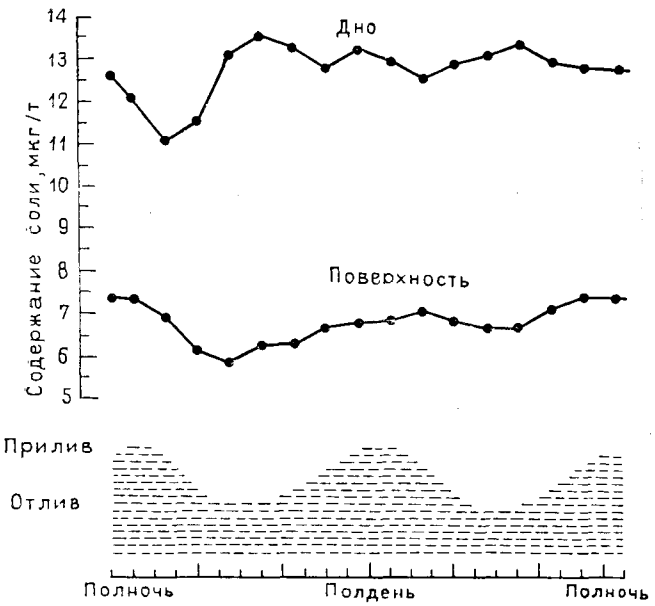


Рис. 4.38. Содержания соли в придонных водах и поверхностных водах в Шизапик Бей вблизи Аннаполиса (штат Мэриленд)

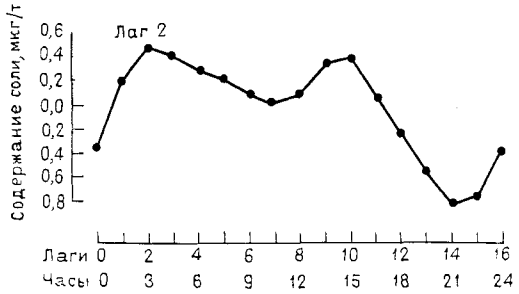


Рис. 4.39. Взаимная коррелограмма между содержаниями соли в придонных и поверхностных водах в Шизапик Бей.

Максимум соответствия наблюдается при лаге 2, представляющем 3-часовое расхождение

гают успеха лишь при исключительных обстоятельствах. Причины этого легко понять. Метод взаимной корреляции основан на предположении, что две сравниваемые последовательности наблюдались в дискретном множестве точек, равномерно распределенных в пространстве, и что интервалы сбора данных в обеих последовательностях были одинаковы. К сожалению, трудно собрать стратиграфические данные так, чтобы удовлетворить

Таблица 4.26

Мощность слоев в разрезе нефтеносных сланцев свиты Грин Ривер. (Разрез расположен в 10 милях севернее разреза, приведенного в табл. 4.22)

Разрез В	Мощность, мм			
(Верх разреза)	10,8	15,6	9,0	
	11,7	15,0	9,2	
	11,0	13,4	10,7	
	9,9	14,6	10,3	
	9,8	13,0	8,9	
	9,9	10,3	9,4	
	10,0	9,4	10,6	
	10,0	9,0	12,6	
	10,2	10,1	14,2	
	10,8	10,3	12,3	
	11,3	9,2	11,1	
	12,0	9,1	11,0	(Основание разреза)
	13,5			

этим требованиям. Вообще невозможно разместить точки опробования так, чтобы они располагались равномерно в геологическом времени, и если точки расположены на равных расстояниях, то мы можем допустить, что скорость осадконакопления была постоянной в течение всего времени образования двух последовательностей. Если вообразить себе, что стратиграфические разрезы аналогичны записям, производимым регистрирующими устройствами, то они будут выглядеть так, как если бы эти устройства работали с различными и неизвестными скоростями в различные времена и даже большую часть времени работали вспять.

Теперь обратимся к проблеме геологической корреляции, где специальные условия не только благоприятствуют использованию взаимной корреляции, но и вселяют надежду на успех. В табл. 4.26 представлены измерения мощности слоев, полученные из некоторой последовательности эоценовых озерных отложений в Скалистых горах. Эти отложения, являющиеся частью свиты Грин-Ривер, состоят из тонких слоистых доломитовых нефтяных сланцев мощностью в десятки метров. Аналогичный разрез представлен на рис. 4.22. Этот разрез удален от первого лишь на 16 км, и, по-видимому, короткая последовательность эквивалентна некоторой части более длинной. Какие-либо маркирующие слои или отличительные свойства в этой монотонной части отсутствуют, и поэтому мы должны проводить вычисления корреляций при попарном сравнении мощности индивидуальных пластов в двух разрезах. Вычислите взаимную коррелограмму и определите положение наилучшей корреляции между двумя разрезами. Как вы объясните другие позиции с меньшим, но значимым попарным сравнением?

Взаимная ассоциация

Рассмотренные выше методы предназначены для сравнения двух последовательностей, образованных значениями некоторых измеряемых переменных. К сожалению, многие стратиграфические исследования приводят к последовательностям, которые содержат только номинальные данные, и не позволяют применить методы взаимной корреляции. Задача состоит в том, чтобы найти меру сходства, которая, в отличие от коэффициента корреляции, не основывалась бы на результатах измерения изучаемых переменных. Одним из наиболее часто используемых геологами методов является метод взаимной связи. Исходные данные представляют собой ряд состояний (например, таких, как литологический тип породы, т. е. известняк, песчаник, сланец и т. д.), наблюдаемых в стратиграфическом разрезе. Эти состояния взаимно исключают друг друга и не могут быть ранжированы никаким осмысленным образом. В методе взаимосвязей две такие последовательности передвигаются одна относительно другой, и для перекрывающихся отрезков оценивается степень соответствия. Для каждого положения подсчитывается общее число сравнений и число совпадающих состояний, а затем вычисляется отношение числа совпадений к общему числу сравнений. Это отношение можно использовать в качестве коэффициента сходства двух цепей в положении перекрытия.

Предположим, что мы при изучении разреза закодировали литологические типы пород по схеме: песчаник — 1, сланец — 2, известняк — 3, уголь — 4, алевролит — 5. Ясно, что способ кодирования совершенно произволен. Таким образом, два стратиграфических разреза можно представить следующим образом:

(а) 5 1 2 1 2 5 2 3 2 4 2 5

(б) 3 2 1 5 2 5 1 2 1 2 3 2

Сдвинем второй разрез относительно первого и в каждом случае подсчитаем число совпадений:

Первое положение:

5 1 2 1 2 5 2 3 2 4 2 5

3 2 1 5 2 5 1 2 1 2 3 2

Число сравнений равно 1, число совпадений — 0, отношение — 0.

Второе положение:

5 1 2 1 2 5 2 3 2 4 2 5

3 2 1 5 2 5 1 2 1 2 3 2

Число сравнений равно 2, число совпадений — 0, отношение — 0.

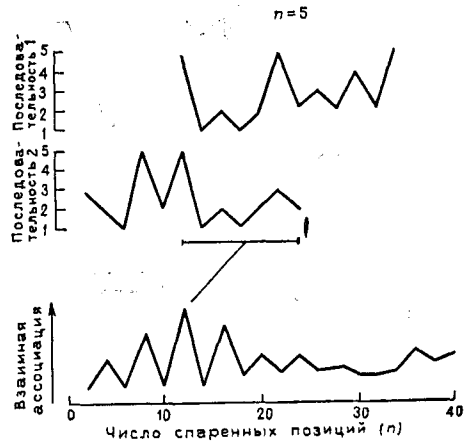


Рис. 4.40. Сравнение двух нечисловых последовательностей с помощью метода взаимосвязей.
Две последовательности изображены в положении максимума сходства

Третье положение:

5 1 2 1 2 5 2 3 2 4 2 5

3 2 1 5 2 5 1 2 1 2 3 2

Число сравнений равно 3, число совпадений — 1, отношение — 0,33.

Если на графике каждому положению сравнения поставить соответствующее отношение, то мы получим линию, аналогичную коррелограмме. Такой график для двух сечений изображен на рис. 4.40.

Значимость отношения сходства можно установить с помощью χ^2-критерия. Чтобы выполнить эту проверку, мы должны определить число совпадений и число несовпадений в перекрывающихся участках. Для этого достаточно вычислить вероятность появления заданного числа совпадений в двух совершенно случайных последовательностях. Эти случайные последовательности должны содержать такое же число наблюдений каждого типа, как и две данные выборки.

Эти две последовательности будем называть соответственно цепью 1 и цепью 2. Предположим, что имеется m категорий, на которые классифицируются наблюдения. Если мы обозначим число наблюдений, попадающих в класс с номером (a), для цепи 1 через X_{1k}, то общая длина цепи 1 будет равна $n_1 = \sum_{k=1}^{m} X_{1k}$. В цепи 2 имеется также m категорий, каждая с X_{2k} наблюдениями, и поэтому общая длина цепи 2 равна $n_2 = \sum_{k=1}^{m} X_{2k}$. Чтобы определить вероятность совпадения данного числа наблюдений в двух последовательностях, мы должны разделить сумму про-

изведений $\sum_{k=1}^{m} X_{1k}X_{2k}$ по всем категориям на произведение длин цепей, т. е.

$$P = \frac{1}{n_1 n_2} \sum_{k=1}^{m} X_{1k}X_{2k}. \qquad (4.69)$$

В нашем примере число категорий равно 5. Первая цепь содержит два песчаника, пять сланцев и т. д. Числа наблюдений для каждой категории в двух цепях приведены в табл. 4.27. Последний столбец таблицы содержит произведение числа наблюдений в каждой категории, а последняя строка — соответствующие суммы по столбцам. Вероятность появления хотя бы одного совпадения в любом положении двух случайных последовательностей при их сравнении применительно к нашим разрезам равна $39/(12 \cdot 12) = 0{,}27$. Вероятность отсутствия совпадений $1{,}00 - 0{,}27 = 0{,}73$.

Мы вычислили вероятность получения хотя бы одного совпадения в двух цепях со случайно расположенными элементами, соответствующими одному и тому же распределению наблюдений, как и в рассмотренных нами двух последовательностях. Среднее значение числа совпадений для интервала перекрытия определяется как произведение величины P на число сравнений. Аналогично среднее число несовпадений равно произведению величины $(1-P)$ на число сравнений. Предположим, что две последовательности в нашем примере полностью перекрываются, в результате чего получается 12 позиций для сравнения. Среднее число совпадений в последовательности равно $12 \times 0{,}27 = 3{,}2$, а среднее число несовпадений равно $12 \times 0{,}73 = 8{,}8$. Используя

Таблица 4.27

Число наблюдений каждого из литологических состояний в двух стратиграфических последовательностях

Категория	Цепь 1	Цепь 2	$X_{1k}X_{2k}$
1. Песчаник	2	3	6
2. Сланец	5	5	25
3. Известняк	1	2	2
4. Уголь	1	0	0
5. Алевролит	3	2	6
Суммы	$n_1 = 12$	$n_2 = 12$	$\sum_{k=1}^{m} X_{1k}X_{2k} = 39$

эти величины, построим χ^2-статистику:

$$\chi^2 = \frac{(O-E)^2}{E} + \frac{(O'-E')^2}{E'}, \qquad (4.70)$$

где O — наблюдаемое число совпадений; O' — наблюдаемое число несовпадений; E — среднее число совпадений; E' — среднее число несовпадений.

Если гипотеза о случайности совпадений верна, то эта статистика имеет χ^2-распределение с одной степенью свободы. Если среднее число совпадений мало, как в случае совпадений вблизи концов цепей, то критерий можно улучшить, используя поправку Йейтса см. [18]. Она сводится к вычитанию $^1/_2$ из абсолютного значения разности наблюдаемого и среднего значений:

$$\chi^2 = \frac{(O-E-1/2)^2}{E} + \frac{(O'-E'-1/2)^2}{E'}. \qquad (4.71)$$

Число степеней свободы при этом остается неизменным, $\nu = 1$.

В нашем примере $O = 2$, $O' = 10$, $E = 3{,}2$, $E' = 8{,}8$, поэтому критерий χ^2 равен

$$\chi^2 = \frac{(2-3{,}2)^2}{3{,}2} + \frac{(10-8{,}8)^2}{8{,}8} = 0{,}61$$

и является незначимым. Можно сделать вывод, что наблюдаемое число совпадений между двумя последовательностями в данной позиции сравнения не больше, чем среднее значение для двух случайных последовательностей с аналогичной структурой. Поправка Йейтса для малых выборок не изменяет этого вывода.

Так как мы имели дело с дискретным рядом событий (совпадение или несовпадение), то распределение частот совпадений в условиях нулевой гипотезы подчиняется биномиальному закону. Среднее число возможных совпадений для двух случайных последовательностей, которые сравниваются при n^* спаренных позициях, равно

$$\overline{E} = P(n^*). \qquad (4.72)$$

Расстояние наблюдаемого числа совпадений от этого среднего в единицах стандартного отклонения выражается по формуле

$$S = \frac{O-E}{E(1-P)}. \qquad (4.73)$$

К сожалению, мы не можем вычислить вероятность получения наблюдения, имеющего столь большое отклонение от среднего значения биномиального распределения, если не будем иметь в распоряжении его подробных таблиц. Последние не вполне пригодны при больших значениях n, которые как раз и нужны в задачах изучения взаимосвязей. Однако мы можем ис-

пользовать нормальное приближение к биномиальному распределению [40], выражаемое формулой (4.74), и найти вероятность появления заданного значения отклонения от среднего. С этой целью можно использовать таблицы стандартного нормального распределения.

$$Z = \sqrt{n^*}\left(2\arcsin\sqrt{O/n^*} - 2\arcsin\sqrt{P}\right). \qquad (4.74)$$

Подставляя требуемые значения для двух последовательностей в формулу (4.74), получим

$$Z = \sqrt{12}\left(\arcsin\sqrt{2/12} - 2\arcsin\sqrt{0{,}27}\right),$$

или

$$Z = 3{,}46\,(2\arcsin 0{,}41 - 2\arcsin 0{,}53) = -0{,}97.$$

Сравнивая две случайные последовательности, мы можем ожидать три совпадения, а наблюдаем два, что на одно стандартное отклонение отличается от ожидаемого числа совпадений. Наблюдаемое число совпадений в этой позиции вполне может быть случайным.

В качестве примера применения этого корреляционного метода рассмотрим задачу сопоставления множеств наблюдений в разрезах угольного бассейна центральной Англии. Хорошие обнажения пород здесь редки, а электрокаротаж не дает какой-либо информации, поэтому большая часть данных о стратиграфической последовательности получена из глубоких выработок, таких, как карьеры и шахты. Закодируем литологические разновидности пород таким образом: 1 — песчаник; 2 — алевролит; 3 — сланец, не содержащий фауны; 4 — подстилающая глина; 5 — уголь; 6 — сланец, содержащий фауну; 7 — известняк. Первый разрез изучался в затопленной угольной шахте. Второй, менее мощный разрез обнажен в стене открытого угольного карьера в 10 км от первого. Найдите положения наилучшего совпадения короткой и длинной последовательностей. Данные приведены в табл. 4.28.

Из предыдущего примера ясно, что метод изучения взаимосвязей не обязательно эквивалентен методу взаимной корреля-

Таблица 4.28

Два закодированных стратиграфических разреза в центральной Англии

	(Основание)														
Разрез шахты	2	4	5	6	3	4	5	3	1	4	5	3	4	5	
	3	4	5	4	5	3	2	4	5	3	4	5	3	1	
	4	5	4	5	6	3	4	5	6	3	4	5	2	1	
	3	4	5	3	5	3	2	4	5	3	5	2			
	(Основание)										(Верх)				
Разрез карьера	4	5	3	4	5	4	5	3	2	1	2	4	5	3	

Обозначения: 1 — песчаник, 2 — алевролит, 3 — сланец, не содержащий фауны, 4 — подстилающая глина, 5 — уголь, 6 — сланец, содержащий фауну, 7 — известняк.

ции. Имея дело с временны́ми рядами, мы должны предполагать, что наблюдения располагаются в точках вдоль некоторой прямой: это ограничение отсутствует в анализе взаимосвязей. Наши данные могут просто состоять из последовательности состояний, перечисленных в том порядке, в котором они встречаются. Как и в только что приведенных стратиграфических разрезах, расстояния между последовательными точками в этом случае несущественны.

Аналогичным образом нечисловые последовательности можно сравнивать с самими собой; этот процесс называется автоассоциацией. Автоассоциация полезна при исследовании периодичностей в порядке следования состояний и очень широко использовалась при исследовании циклотем [47].

В этом случае сравниваются не две последовательности, а одна последовательность сама с собой. Укажем вероятность хотя бы одного совпадения в этом случае. Биномиальная вероятность получения данного числа совпадений в случайной последовательности при сравнении ее самой с собой составит

$$P = \left(\sum_{k=1}^{m} X_k^2 - n\right) \Big/ (n^2 - n). \qquad (4.75)$$

Мы предполагаем, что последовательность представляет собой случайное размещение m состояний или классов, причем каждое состояние встречается X_k раз. Общее число наблюдений равно $\sum_{k=1}^{m} X_k = n$. Эту вероятность надо прямо подставлять в (4.72) и использовать при вычислении χ^2-распределения и стандартного отклонения. Критерий предназначен для проверки нулевой гипотезы, заключающейся в том, что число совпадений не отличается существенно от ожидаемого числа совпадений для случайной последовательности при сравнении ее с самой собой.

Для иллюстрации применения метода автоассоциаций можно использовать данные из разреза шахты (см. табл. 4.27). Если в разрезе шахты будет содержаться много повторяющихся элементов, то это приведет к необыкновенно высоким значениям отношений, характеризующих совпадения, и к значительным отклонениям от ожидаемого среднего числа совпадений. Интерпретация графиков, характеризующих взаимосвязи, проводится аналогично интерпретации коррелограмм. Однако коэффициент взаимосвязи вычисляется на основании номинальных данных, и в силу этого информация, содержащаяся в последовательности, значительно беднее, чем в эквивалентном временно́м ряду. Так как мы используем качественные данные, то не можем ожидать того же результата, который можно было бы получить при анализе настоящих временны́х рядов. Этот фактор необходимо учитывать при интерпретации результатов по взаимосвязям и автоассоциациям.

ПОЛУВАРИОГРАММЫ

Термин геостатистика сейчас широко применяется к специальным ветвям прикладной статистики, начало развития которой было положено Г. Матероном в Центре математической морфологии в Фонтенбло, Франция. Цель геостатистики состояла в исследовании проблем, которые возникают тогда, когда обычная статистическая теория используется в оценке изменений содержания руды в рудном теле. Однако так как геостатистика есть абстрактная теория статистического поведения, она применима ко многим обстоятельствам в различных областях геологии и других естественных наук.

Ключевое понятие геостатистики — понятие регионализованной переменной, которая имеет свойства, промежуточные между свойствами полностью случайных величин и полностью детерминированных переменных. Типичные регионализованные переменные являются функциями, описывающими естественные явления и имеющими географическое распределение, такие, как высота над уровнем поверхности, изменения содержаний в рудном теле или спонтанный электрический потенциал, измеренный в скважине методом каротажа. В отличие от случайных, регионализованные переменные непрерывны от точки к точке, но изменения их настолько сложны, что они не могут быть описаны какой-либо регулярной детерминированной функцией.

Даже несмотря на то что регионализированная переменная непрерывна в пространстве, обычно невозможно знать ее значение в любой точке. Вместо этого ее значения известны только благодаря пробам, которые берутся в определенных местах. Размер, вид, ориентация и пространственное размещение этих проб составляют базу регионализованной переменной. Эта переменная при изменении хотя бы одного из этих параметров будет иметь различные характеристики. Например, предположим, что мы хотим определить изменчивость сорта вкрапленных руд в молибденовом месторождении. Вероятно, результат анализа 5-сантиметрового керна, полученного при алмазном бурении, будет отличаться от результата измельчения проб из рудных отвалов. В обоих случаях мы могли бы взять в точности то же самое число проб, и они могли бы быть получены из идентичных местоположений в руднике. Однако тот факт, что объемы проб в одном множестве измеряются в кубических сантиметрах, а в другом — в кубических метрах, неизбежно должен влиять на схему изменчивости сортности руды, которую мы картируем в руднике. Главная задача геостатистики — связать результаты, полученные по одной базе (например, образцы керна), с результатами, полученными для другой базы (например, эксплуатационные блоки).

Геостатистика позволяет дать оценки формы регионализированных переменных в одном, двух и трех измерениях. В следующей главе мы более подробно рассмотрим процедуру оценки, называемую крайгингом. Сейчас мы коснемся лишь одной из важнейших статистических характеристик геостатистики, а именно понятия полудисперсии, которое используется для выражения скорости изменения регионализованных переменных вдоль заданного направления. Оценка полудисперсии содержит процедуры, аналогичные процедурам анализа временных рядов, следовательно, приводит к необходимости использования геостатистики.

Полудисперсия есть мера степени пространственной зависимости между пробами взоль заданной базы. Для простоты мы предположим, что пробы являются точечными измерениями некоторого свойства, такого, как глубина подповерхностного горизонта. Для облегчения вычислений мы будем далее предполагать, что база регулярная, т. е. пробы равномерно расположены в пространстве вдоль прямых линий. Если расстояние между пробами по прямой линии равно некоторой величине Δ, то полудисперсия может быть вычислена для расстояний, кратных Δ:

$$\gamma_h = \frac{1}{2n} \sum_{i=1}^{n-h} (X_i - X_{i+h})^2. \qquad (4.76)$$

В этих обозначениях X_i — значение регионализованной переменной, взятой в точке i; X_{i+h} — другое значение, взятое через h интервалов. Мы поэтому нашли сумму квадратов разностей между значениями регионализованной переменной в паре точек, разделенных расстоянием Δh. Число точек равно n, так что число сравнений между парами точек есть $n-h$.

Если мы вычислим полудисперсии для различных значений h, то мы можем нанести результаты на график в виде полувариограммы, являющейся аналогом коррелограммы. На рис. 4.41 представлена полувариограмма, соответствующая глубине сейсмически отражающего горизонта и построенная по измерениям вдоль сейсмического профиля, приведенного на рис. 4.42. Заметим, что когда расстояние между точками опробования равно нулю, то значение в каждой точке сравнивается с самим собой. Следовательно, все разности равны нулю, и полудисперсия для γ_0 есть нуль. Если Δh — малое расстояние, точки при сравнении оказываются очень похожими, и полудисперсия будет мала. По мере увеличения расстояния Δh сравниваемые точки становятся слабее связанными друг с другом и расстояния между ними увеличиваются, что приводит к большим значениям γ_h. Предположим, что на некотором расстоянии сравниваемые точки находятся так далеко, что они не связаны друг с другом, и их квадраты разностей будут равны по величине дисперсии относительного среднего значения. Полудисперсия более не растет и полувариограмма переходит в плоскую область, называемую поро-

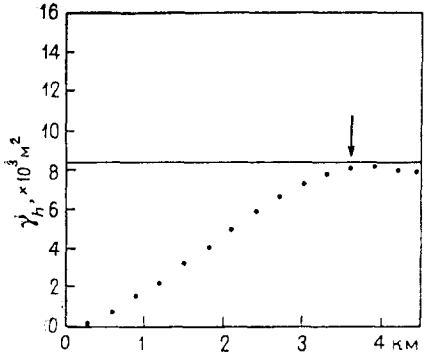


Рис. 4.41. Полувариограмма абсолютных отметок кровли меловой формации, измеренной вдоль морского сейсмического разреза в Магеллановом проливе, Чили [38].

Линия, изображенная точками, представляет порог, или дисперсию, возвышений и равна 8380 м². Ранг, указанный стрелкой, — расстояние, ниже которого разность между дисперсиями и порогом считается пренебрежимо малой (3,5 км)

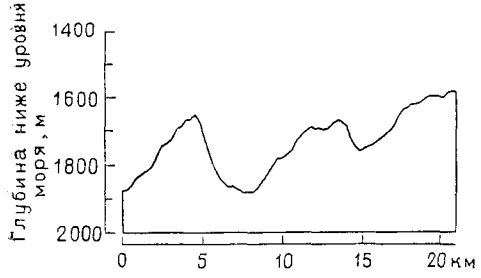


Рис. 4.42. Подпочвенные структурные абсолютные отметки кровли меловой формации, оцененные по отражению сейсмических волн вдоль 21-километрового морского траверса в Магеллановом проливе. Сейсмические измерения взяты в 300-метровом интервале [38]

том. Расстояние, на котором полудисперсия приближается к дисперсии, называется рангом, или размахом регионализованной переменной, оно определяет окрестность, в пределах которой все положения связаны друг с другом.

Для некоторой произвольной точки в пространстве мы можем представить себе окрестность как симметричный интервал (или площадь, или объем, в зависимости от размерности) вокруг точки. Если регионализованная переменная стационарна или всюду имеет одно и то же среднее значение, то любое положение вне этого интервала совершенно независимо от центральной точки и не может давать информацию вокруг значения регионализированной переменной в этой точке. В пределах этой окрестности, однако, регионализированная переменная во всех наблюдаемых точках связана с регионализованной переменной в центральной точке и, следовательно, может быть использована для оценки ее значения. Если мы используем множество измерений, сделанных в точках внутри этой окрестности для оценки значения регионализованной переменной в центральной точке, то полувариограмма обеспечит собственные веса, которые должны быть приписаны каждому измерению.

Остановимся коротко на факте, который будет нам полезен позже. Полудисперсия равна не только среднему значению квадратов разностей для пар точек, расположенных на расстоянии

Δh друг от друга, но и дисперсии этих разностей, т. е. полудисперсия может быть определена по формуле

$$\gamma_h = \frac{1}{2n}\sum\left\{(X_i - X_{i+h}) - \frac{1}{n}\Sigma(X_i - X_{i+h})\right\}^2. \qquad (4.77)$$

Заметим, что среднее значение регионализованной переменной X_i есть также среднее регионализованной переменной X_{i+h}, так как это — те же самые наблюдения, только взятые в другом порядке, т. е.

$$\frac{\Sigma X_i}{n} = \frac{\Sigma X_{i+h}}{n}.$$

Поэтому их разность должна быть равна нулю

$$\frac{\Sigma X_i}{n} - \frac{\Sigma X_{i+h}}{n} = 0.$$

Комбинируя суммы, получаем

$$(\Sigma X_i - \Sigma X_{i+h})/n = [\Sigma(X_i - X_{i+h})]/n = 0.$$

Подставляя в (4.77), мы видим, что числитель второго члена равен нулю, так что это уравнение совпадает с уравнением (4.76). Заметим, что это соотношение строго справедливо только тогда, когда регионализованная переменная стационарна. Если данные не стационарны, то среднее значение последовательности изменяется вместе с h, и (4.77) должно быть модифицировано.

Как и следовало ожидать, имеются математические соотношения между полудисперсией и другими статистиками, такими, как автоковариация и автокорреляция. Если регионализованная переменная стационарна, то полудисперсия для расстояния Δh равна разности между дисперсией и пространственной автоковариацией для того же расстояния (рис. 4.43). Если регионализованная переменная не только стационарна, но и стандартизована так, чтобы среднее равнялось нулю, а дисперсия единице, то полувариограмма будет зеркальным отражением автокорреляционной функции (рис. 4.44).

К сожалению, часто регионализованные переменные не стационарны, скорее они отражают изменения их средних значений от точки к точке. Если мы попытаемся построить полувариограмму для такой переменной, то обнаружим, что она может не иметь описанных выше свойств. Однако если пересмотреть определение полудисперсии, приведенное в формуле (4.77), то мы заметим, что оно состоит из двух частей, первая соответствует разностям переменной в паре точек, а вторая — среднему этих разностей. Если регионализованная переменная стационарна, то, как мы видели, вторая часть равна нулю, а если она нестационарна, то это среднее будет иметь некоторое не равное нулю значение. Действительно, регионализованная переменная может быть рассмотрена как состоящая из двух частей, называемых остатком и дрифтом. Дрифт — это математическое ожидание регионализованной переменной в точке i, или с точки зрения вы-

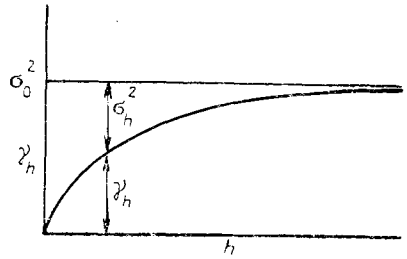


Рис. 4.43. Соотношение между полудисперсией γ_h и автоковариацией для стационарной регионализованной переменной.

σ_0^2 — дисперсия наблюдений или автоковариация для лага 0. Для значений вне этого множества $\gamma_h = \sigma_0^2$

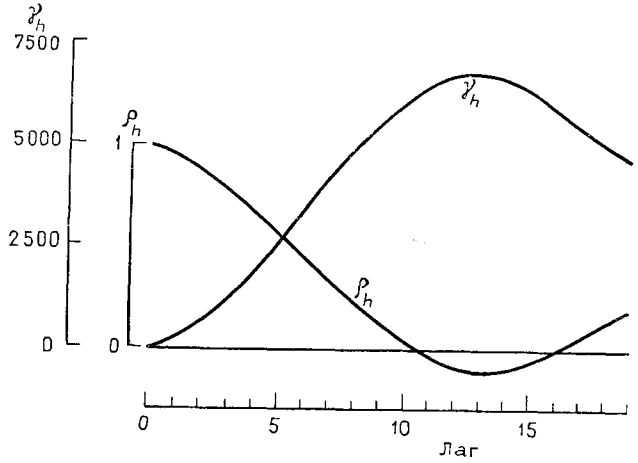


Рис. 4.44. Соотношение между полудисперсией γ_h и автокорреляцией ρ_h для стационарной регионализованной переменной

числений, взвешенное среднее всех точек внутри окрестности вокруг точки i. Дрифт будет иметь вид кривой, аппроксимирующей регионализованную переменную. Если дрифт вычесть из регионализованной переменной, то остатки $R_i = X_i - \overline{X}_i$ сами дадут регионализованную переменную, имеющую средние значения, равные нулю. Другими словами, остатки будут стационарными и можно построить их полувариограмму.

Здесь мы неожиданно приходим к проблеме цикличности. Дрифт может быть оценен, если мы знаем размер окрестности и веса, приписанные точкам внутри этой окрестности. Однако веса могут быть вычислены только в том случае, если мы знаем полудисперсии, соответствующие расстояниям между точкой i, центром окрестности и различными другими точками Вычисленный однажды дрифт вычитаем из значений, полученных при наблюдении. Полученную стационарную переменную в свою очередь можно использовать для оценки размера окрестности в виде полувариограммы.

Теперь ослабим строгость наших определений и перейдем к методу проб и ошибок. Сначала нужно признать, что нельзя определить окрестность в том смысле, в котором мы использовали этот термин. Вместо этого окрестность определяется как

удобный, но все же произвольный интервал, в пределах которого мы уверенно можем утверждать, что все позиции связаны друг с другом. Предположим, что в пределах этой произвольной окрестности дрифт можно аппроксимировать простым выражением, например

$$\overline{X}_0 = \Sigma b_1 X_i$$

(линейный дрифт) или же

$$\overline{X}_0 = \Sigma (b_1 X_i + b_2 X_i^2)$$

(квадратичный дрифт). В эти формулы входят координаты всех заданных точек внутри произвольной окрестности, так что существуют взаимосвязи между размером окрестности, дрифтом и полувариограммой остатков. Если окрестность велика, вычисления дрифта будут основаны на большом числе точек, и дрифт можно будет представить очень гладкой кривой. В этом случае остатки будут характеризоваться большой изменчивостью, а полувариограмма окажется сложной по форме. Следовательно, специфика размера малой окрестности будет влиять на большую изменчивость оценки дрифта, на уменьшение остатков и на простоту вариограммы.

Определение коэффициентов b дрифта требует решения некоторого числа совместных уравнений повышенной сложности, описание которых откладывается до раздела, посвященного крайгингу. Единственные переменные в этих уравнениях — это полудисперсии, соответствующие различным расстояниям между точкой с номером i и другими точками в рассматриваемой окрестности. Однако они еще не дают полувариограммы, из которой следует получить необходимые полудисперсии. Можно допустить, что полувариограмма имеет какой-либо естественный для нее вид, и использовать его в качестве первого приближения. К счастью, легко предвидеть вид простой полувариограммы, и это позволяет использовать окрестность настолько малого размера, насколько это возможно.

Экспериментальные оценки дрифта вычитаются из соответствующих наблюдений, в результате чего получается множество экспериментальных остатков. По этим остаткам можно вычислить полувариограмму и затем сравнить ее с той полувариограммой, которая была выбрана в качестве первого приближения. Если сделанные предположения были правильными, то обе они совпадут, и можно считать, что форма дрифта и полувариограммы определены успешно. Однако более вероятно, что они отличаются, и следует проделать вычисления еще раз.

Процесс совместного построения удовлетворительных выражений для полувариограммы и дрифта является важной составной частью «структурного» анализа. В некотором смысле это — искусство, требующее опыта, терпения и иногда удачи. Этот процесс не всегда приводит к приемлемым решениям, так как

они неоднозначны; много комбинаций дрифта, окрестностей и моделей полувариограмм могут дать примерно одинаковые результаты. Он пригоден особенно в том случае, когда регионализованные переменные неустойчивы или же мы располагаем лишь короткой последовательностью. В таких обстоятельствах трудно сказать, когда мы достигнем эффекта от комбинации оценок.

Полувариограмма отражает пространственное поведение регионализованных переменных или их остатков. Некоторые идеализированные формы полувариограмм даны на рис. 4.45.

На рис. 4.45, *а* приведена полувариограмма параболического типа, касающаяся оси X в начале координат. Она иллюстрирует очень гладкое изменение регионализованной переменной. На рис. 4.45, *б* представлена полувариограмма, имеющая вид прямой линии; она указывает на умеренное и непрерывное изменение регионализованной переменной. Истинная случайная переменная не будет непрерывной, и ее полувариограмма будет горизонтальной линией, ордината которой равна дисперсии (рис. 4.45, *в*). В некоторых случаях полувариограмма не проходит через начало координат, а имеет при абсциссе, равной нулю, ненулевое значение. Этот случай соответствует «эффекту самородков». Он изображен на рис. 4.45, *г*. В теории величины γ_0 должна быть равна нулю. Эффект самородков возникает тогда, когда регионализованная переменная настолько ошибочно определена на коротком расстоянии, что полувариограмма выходит из нуля на уровень эффекта самородков на расстоянии, меньшем, чем интервал опробования.

Моделирование полувариограмм

В принципе экспериментальная полувариограмма может быть прямо использована для получения оценок, которые мы рассмотрим в следующей главе. Однако полувариограмма известна только в дискретном наборе точек, расположенных на расстояниях Δh; на практике, однако, полувариограммы могут по-

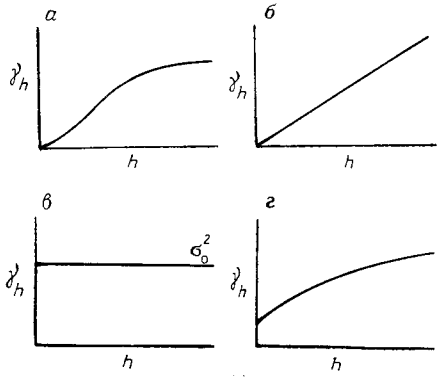


Рис. 4.45. Идеализированные полувариограммы:

а — параболическая форма, показывающая отличную непрерывность регионализованной переменной; *б* — линейная форма, показывающая умеренную непрерывность; *в* — горизонтальная форма уровня σ_0^2, соответствующая случайной переменной, не имеющей пространственной автокорреляции; *г* — эффект самородков, или явное отклонение полувариограммы от начала координат, показывающее, что регионализованная переменная сильно изменчива при расстояниях, меньших чем интервал опробования

требоваться для любых расстояний независимо от того, является ли оно кратным Δ или нет. По этой причине дискретная экспериментальная полувариограмма должна быть представлена некоторой непрерывной функцией, которая может быть вычислена для любого желаемого расстояния.

Подбор модельного уравнения к экспериментальной полувариограмме проводится обычно на глазок, методом проб и ошибок. Кларк [9] описывает и дает примеры ручных вычислений, в то время как Олеа [38] приводит программу вычисления линейной полувариограммы, имеющей тот же угловой коэффициент в начале координат, как и экспериментальная полувариограмма.

В идеале модель, выбранная для представления полувариограммы, начинается в начале координат, гладко возрастает до некоторого верхнего предела, затем остается на одном постоянном уровне. Сферическая модель, представленная на рис. 4.46, обладает этими свойствами. Она определена по формуле

$$\gamma_h = \sigma_0^2 \left(\frac{3h}{2a} - \frac{h^3}{2a^3} \right) \tag{4.78}$$

для всех расстояний вплоть до области влияния полувариограммы a. За пределами этой границы $\gamma_h = \sigma_0^2$. Сферическая модель обычно характеризуется как идеальная форма полувариограммы. Иногда используется другая модель — экспоненциальная

$$\gamma_h = \sigma_0^2 \left(1 - e^{\frac{-h}{a}} \right). \tag{4.79}$$

На рис. 4.47 сравниваются сферическая и экспоненциальная модели. Экспоненциальная кривая никогда не достигает своего предельного значения, а приближается к нему асимптотически. Значит, полудисперсия экспоненциальной модели ниже, чем сферическая, для всех значений h, меньших, чем размер области влияния. Линейная модель проще, чем сферическая или экспоненциальная, так как она имеет только один параметр, наклон. Модель имеет вид

$$\gamma_h = \alpha h \tag{4.80}$$

и представляет собой прямую, проходящую через начало координат. Очевидно, эта модель не может иметь пика, так как она растет неограниченно. Иногда линейная модель произвольно модифицируется с помощью вставки внезапного излома в точке пика, как, например,

$$\begin{aligned} \gamma_h &= \alpha h \quad \text{для} \quad h < a, \\ \gamma_h &= \sigma_0^2 \quad \text{для} \quad h \geqslant a. \end{aligned} \tag{4.81}$$

Армстронг и Джебин [3] подвергают критике такие модели, так как использование крайгинга для получения оценок предполагает непрерывность и гладкое изменение полувариограммы. Однако для расстояний, значительно меньших границы, линей-

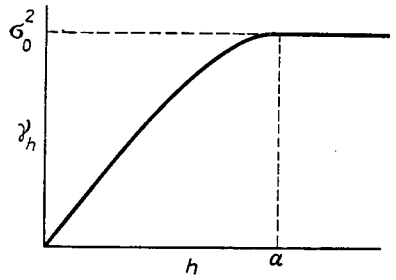


Рис. 4.46. Сферическая модель полувариограммы.
Буквой *a* обозначена абсцисса точки, в которой полувариограмма становится равной дисперсии σ_0^2

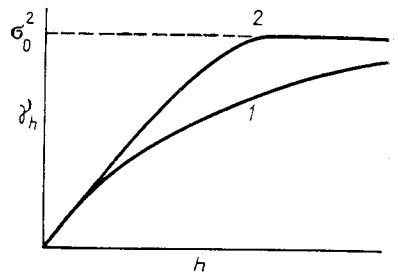


Рис. 4.47. Экспоненциальная и сферическая модели полувариограммы. Обе модели имеют один и тот же начальный наклон [9]:
1 — экспоненциальная модель; *2* — сферическая модель

ная модель является идеально хорошей аппроксимацией. Это хорошо видно на рис. 4.47, где обе модели, сферическая и экспоненциальная, почти совпадают с прямой линией вблизи начала координат. Если регионализованная переменная была опробована с достаточной плотностью, относительно области влияния, то значимых различий между оценками, полученными в условиях линейной, сферической или другой модели нет.

СПЕКТРАЛЬНЫЙ АНАЛИЗ

Спектральный анализ — это метод разложения изменчивости временного ряда на отдельные компоненты, соответствующие продолжительности или длине интервалов, в пределах которых эта изменчивость проявляется. Для его осуществления временной ряд рассматривается как сумма большого числа более простых временны́х рядов, имеющих вид регулярных синусоид с различными амплитудами, длинами волн и начальными точками. В вычислениях можно предполагать, что синусоиды независимы друг от друга. Так как сумма всех этих синусоид дает исходный временной ряд, то сумма изменений во всех синусоидах должна быть равна общей изменчивости ряда.

В различных руководствах спектральный анализ имеет различные названия, например гармонический анализ, анализ Фурье или частотный анализ. Название этого метода восходит к теории музыки, изучающей закономерности, связанные с колебаниями. В XVII веке Кеплер, применяя соотношения гармонического анализа, найденные им в арифметике, геометрии и музыке, открыл законы планетарного движения. Однако Жан Батист Фурье (1768—1830 гг.) доказал общую теорему о том, что любая однозначная непрерывная функция может быть представлена в виде ряда из синусоид. Эта теорема теперь носит его имя. Прежде чем подробно излагать теорию рядов Фурье, мы должны определить ряд терминов. Большинство этих терминов

взято из электротехники и используется при анализе электрических сигналов. Хотя электрический сигнал представляет собой изменение энергии волны с течением времени, инженеры любят представлять себе его «замороженным» в осциллографе. При изложении теории рядов Фурье инженеры всегда предполагают, что сигнал изменяется со временем, однако тот факт, что они исследуют сигнал как пространственное явление на экране осциллографа, указывают, что время и пространство считаются равноправными. Несомненно, математически это корректно, и та неосознанная легкость, с которой инженеры пользуются таким допущением, может убедить нас в том, что эти два понятия, как правило, взаимозаменяемы. Поэтому наше изложение терминологически несколько отличается от такового в большинстве работ по анализу сигналов.

На рис. 4.48 изображен повторяющийся сигнал, который можно описать чистой синусоидальной волной. Расстояние от одной точки волны до эквивалентной точки на следующей волне называется длиной волны λ. Частота f — характеристика, обратная длине волны, т. е. $f=\frac{1}{\lambda}$ — число волн, укладывающихся в единицу длины или времени. В большинстве инженерных расчетов сигнал характеризуется частотой. В геологических задачах интереснее иметь дело с длиной волны. Время, требуемое для того чтобы правильный сигнал повторился, называется его периодом. Термин период является эквивалентом длины волны, но период измеряется в единицах времени, например в миллисекундах, а не в единицах расстояния, т. е. в сантиметрах. При описании таких явлений, как морские волны, удобно пользоваться как длиной волны, так и периодом. Длина волны в этом случае измеряется как расстояние между гребнями соседних волн, период — как время, требуемое для того чтобы две последовательные волны прошли фиксированную точку отсчета, например конец волнореза. Это временной промежуток между моментом появления одной волны и моментом появления другой. (Возникновение термина периодический обязано слову период, которое означает сигнал, повторяющийся с правильными интервалами.) Половина расстояния от впадины волны до гребня называется амплитудой (A).

На рис. 4.49 изображены две одинаковые синусоидальные кривые, смещенные относительно друг друга. Соответствующие им амплитуды и длины волн одинаковы. Разность между значениями Y_1 и Y_2 при данном значении x_i определяется различием в фазе между двумя волнами. Это различие можно описать с помощью фазового угла Ф, а его определение уяснить себе с помощью рис. 4.50. Простое механическое приспособление для получения синусоидальной волны состоит из диска радиуса r, вращающегося с постоянной скоростью. Карандаш, закрепленный

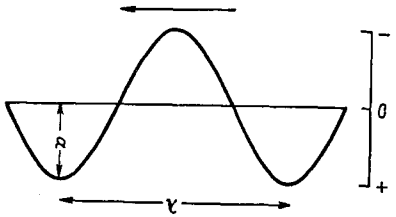


Рис. 4.48. Регулярное повторение синусоидальной волны:
λ — длина волны; A — амплитуда

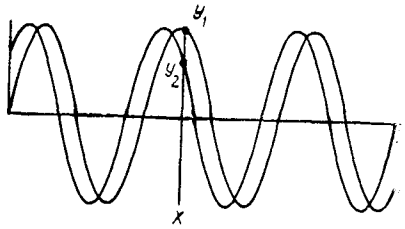


Рис. 4.49. Две синусоидальные волны, которые идентичны по форме.
Различия между Y_1 и Y_2 для специфического значения X характеризуют разность фаз двух волновых форм

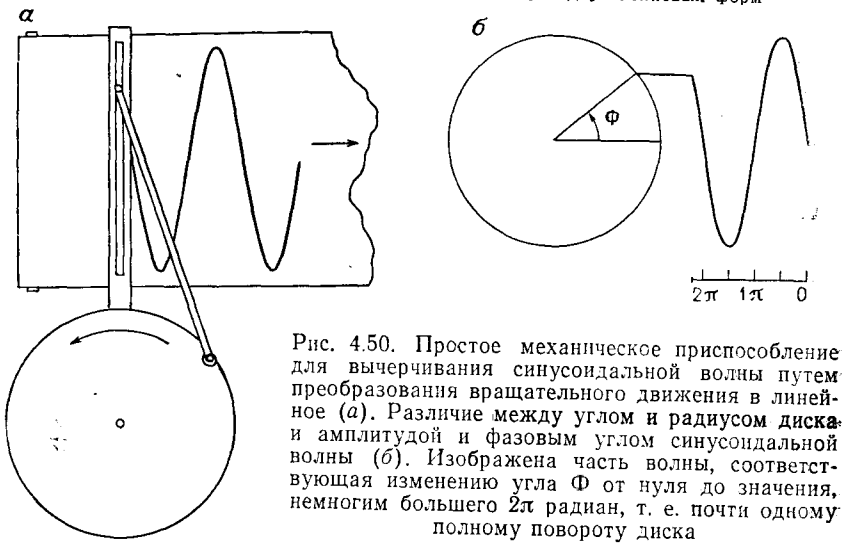


Рис. 4.50. Простое механическое приспособление для вычерчивания синусоидальной волны путем преобразования вращательного движения в линейное (*а*). Различие между углом и радиусом диска и амплитудой и фазовым углом синусоидальной волны (*б*). Изображена часть волны, соответствующая изменению угла Ф от нуля до значения, немногим большего 2π радиан, т. е. почти одному полному повороту диска

на стержне, присоединенном к краю диска, вычерчивает линию на бумаге, двигающейся с постоянной скоростью под концом стержня. Эта линия является синусоидальной кривой, имеющей амплитуду $A=r$ и длину волны, представляющую собой функцию скорости вращения диска и скорости движения бумаги. Значение Y в любом положении равно $Y=A\sin\varphi$.

Если мы хотим вычертить полную волну, то точка закрепления стержня на диске должна сделать полный оборот вокруг центра диска, т. е. повернуться на 360° или на 2π радиан. Предположим, что мы начинаем вычерчивание при произвольном начальном положении карандаша на бумаге, которое мы назовем нулевым. Угол α между карандашом, центром диска и горизонтальной линией на бумаге имеет некоторое значение φ (рис. 4.51). Если мы в процессе работы с прибором сдвинемся вниз на расстояние x_i и остановимся, то карандаш окажется в точке,

координата Y которой равна $A\sin(2\pi x_i/X+\varphi)$. Угол α_i будет $(2\pi x_i/X+\varphi)$. В этих выражениях X есть общая длина записи. Важно отметить, что начальный угол $\alpha=\varphi$ в исходном положении является постоянным для всех последующих значений амплитуды координаты Y и угла α. Константа φ называется фазой, фазовым углом, или фазовой константой. Три параметра — амплитуда, длина волны и фазовый угол — полностью описывают форму волны. Фазовый угол синусоидальной волны измеряется от горизонтальной линии, проходящей через центр.

Если мы хотим построить ряд волн различной формы, то для этого нужно изменить амплитуду, длину волны или фазовый угол. Можно также описать косинусоидальную кривую, по форме идентичную синусоидальной, но характеризующуюся фазовым углом β, измеряемым от вертикальной прямой, проходящей через центр устройства. Иными словами, косинусоидальная волна отличается от синусоидальной по фазе на 90°, или на $\pi/2$ радиан. Если к нашему устройству присоединить другой карандаш, как это указано на рис. 4.52, то мы сможем вычертить одновременно как синусоидальную, так и косинусоидальную волны. Очевидно, что они отличаются одна от другой только фазовым углом. Мы можем суммировать основные тригонометрические соотношения с помощью рис. 4.53, представляющего синусоидальную кривую. Это разложение дано в терминах функции косинуса, но эквивалентное разложение, ведущее к тому же соотношению, основано на синусе. Допустим, что волна на расстоянии X или за время X укладывается целиком. Любая точка X в пределах этого интервала может быть выражена в радианах с помощью формулы обращения

$$\theta = 2\pi x/X. \tag{4.82}$$

Теперь координаты X покрывают интервал от 0 до 2π радиан. Для удобства мы предположим, что амплитуда кривой равна единице. Тогда уравнение кривой имеет вид

$$Y = \cos\theta. \tag{4.83}$$

Эта амплитуда может быть заменена любой другой амплитудой просто умножением на коэффициент A:

$$Y = A\cos\theta. \tag{4.84}$$

Число циклов, которое встречается внутри основного интервала (частота), может быть увеличено или уменьшено умножением θ на коэффициент k:

$$Y = \cos k\theta. \tag{4.85}$$

Как нарисовано, максимум волны находится в начале координат, но он может быть сдвинут в любое положение вычитанием фазового угла φ:

$$Y = \cos(\theta - \Phi). \tag{4.86}$$

Комбинируя все эти модификации, мы найдем, что любая кри-

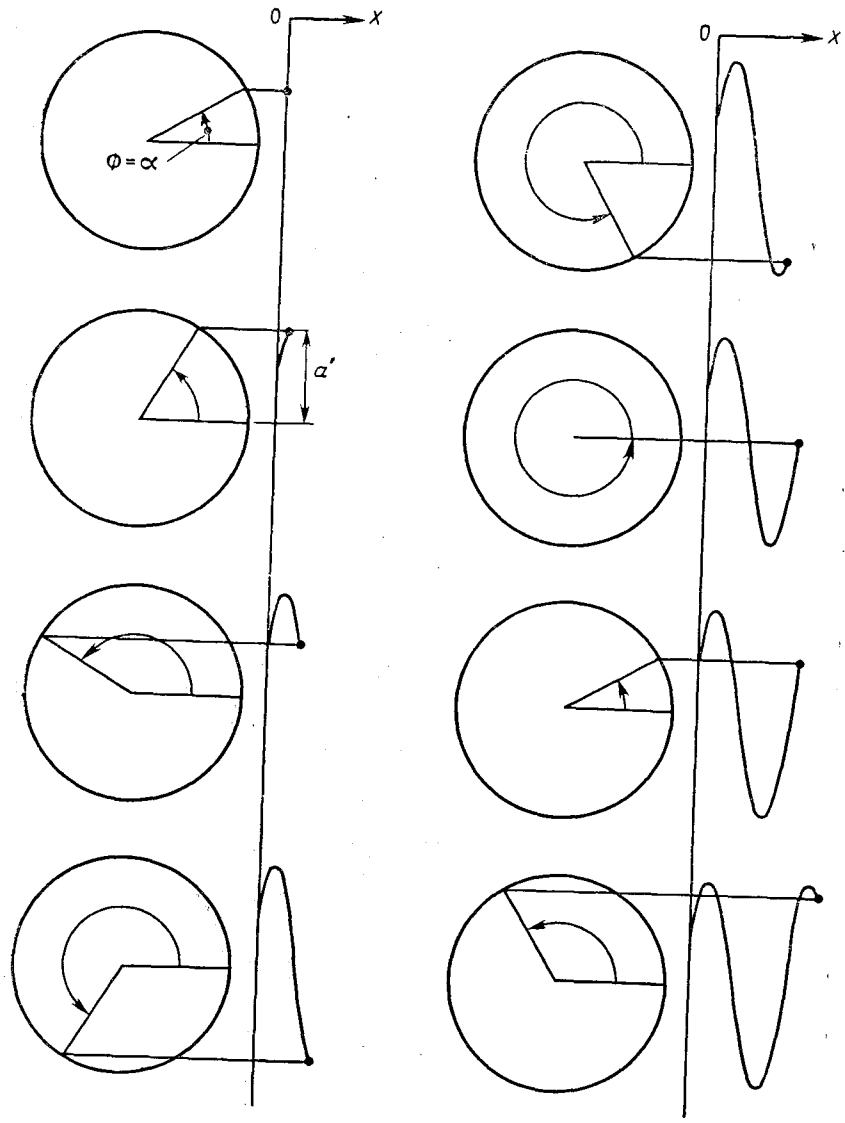


Рис. 4.51. Постепенное изменение фазового угла от начального значения Φ. Для последовательных значений x_i, $Y_i = A \sin(2\pi x_i/X + \Phi)$ и $\alpha_i = 2\pi x_i/X + \Phi$

вая регулярной синусоидальной формы может быть записана в виде

$$Y_k = A_k \cos(k\theta - \Phi_k). \qquad (4.87)$$

Число k называется числом гармоник, или числом циклов на ба-

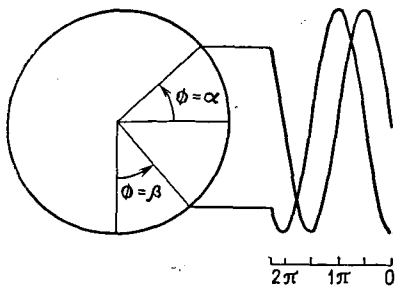


Рис. 4.52. Синусоидальная и косинусоидальная волны, соответствующие вращению диска.

α и β — соответственно фазовые углы для синусоидальной и косинусоидальной волн. Изображение соответствует изменению угла Φ от нуля до значения, немногим большего 2π радиан, т. е. одному повороту диска

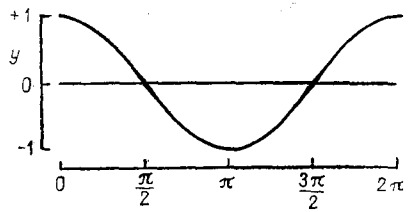


Рис. 4.53. Синусоидальная волна. Общая длина кривой равна 2π радиан

зисный интервал. Так как в спектральном анализе можно скомбинировать много различных гармоник для получения исходного временно́го ряда, то индекс k необходим для обозначения волновой формы частного вида.

Далее мы можем использовать тригонометрическое равенство для разности двух углов:

$$\cos(R - S) = \cos S \cos R + \sin S \sin R.$$

Переписав равенство (4.87), получим

$$Y_k = A_k \cos \Phi_k \cos k\theta + A_k \sin \Phi_k \sin k\theta. \qquad (4.88)$$

Это выражение может быть упрощено, если определить коэффициенты $\alpha_k = A_k \cos \varphi_k$ и $\beta_k = A_k \sin \varphi_k$. При этом получаем

$$Y_k = \alpha_k \cos k\theta + \beta_k \sin k\theta. \qquad (4.89)$$

Любой временной ряд, какой бы сложности он ни был (исключая тот случай, когда он непрерывен или без изломов и для каждого значения x может быть определено только одно значение Y), может быть представлен в виде суммы рядов косинусоидальных волн, определенных таким образом, как изображено на рис. 4.54. Это — выражение соотношения Фурье:

$$Y = \sum_{k=0}^{\infty} \alpha_k \cos k\theta + \beta_k \sin k\theta. \qquad (4.90)$$

Заметим, что (4.90) — это уравнение линейное, т. е. все члены складываются вместе. Оно напоминает по форме полиномиальную регрессию, которую мы рассмотрели ранее в этой главе. В уравнении Фурье тригонометрические члены $\cos k\theta$ и $\sin k\theta$ эквивалентны степенным членам таких многочленов, как X^2 и X^3. Коэффициенты α_k и β_k этих членов могут быть найдены ме-

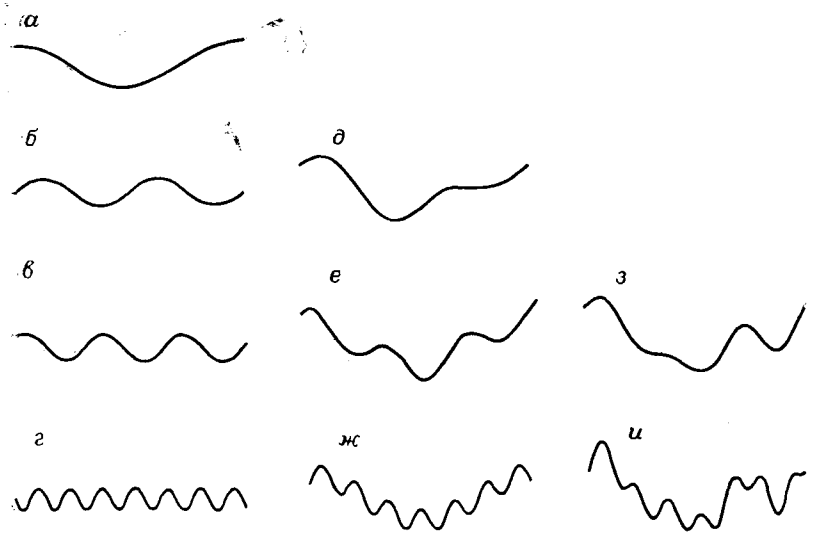


Рис. 4.54. Результат сложения последовательных гармоник косинусоидальной волны.

а — фундаментальная или первая гармоника $A=0{,}5$, $k=1$, $\Phi=0°$; *б* — вторая гармоника $A=0{,}3$, $k=2$, $\Phi=0°$; *в* — третья гармоника $A=0{,}3$, $k=3$, $\Phi=45°$; *г* — седьмая гармоника $A=0{,}25$, $k=7$, $\Phi=180°$; *д* — первая плюс вторая гармоника, $(а)+(б)$; *е* — первая плюс третья гармоника, $(а)+(в)$; *ж* — первая плюс седьмая гармоника, $(а)+(г)$; *з* — первая плюс вторая и третья гармоники, $(а)+(б)+(в)$; *и* — первая плюс вторая, третья и седьмая гармоники $(а)+(б)+(в)+(г)$

тодом наименьших квадратов. Однако если мы попытаемся оценить эти коэффициенты матричными методами и если требуется определить много гармоник, то мы очень скоро столкнемся с непреодолимыми вычислительными трудностями.

Гармонический анализ

Если временной ряд представляет действительно периодическое явление, мы можем использовать технику так называемого гармонического анализа для разложения ряда на его составные части. Временны́е ряды, обладающие свойством периодичности, встречаются в природе. Это, например, флуктуации приливов и отливов, сезонные изменения температуры, последовательности слоев. Они также включают записи, порожденные такими приборами, как дифрактометр Х-лучей (преобразование Фурье таких данных является специальной темой). Временной ряд может быть совокупностью наблюдений в дискретном множестве точек, расположенных с равным интервалом Δ в пространстве. Если имеется n этих точек, одна из которых есть j, то коэффициенты α_k и β_k ряда Фурье могут быть вычислены по следующим

формулам:

$$\beta_k = \frac{2}{k}\sum_{j=0}^{n-1} Y_j \sin\left(\frac{2\pi jk}{n}\right), \tag{4.91}$$

$$\alpha_k = \frac{2}{k}\sum_{j=0}^{n-1} Y_j \cos\left(\frac{2\pi jk}{n}\right). \tag{4.92}$$

Эти уравнения не столь сложны, как может показаться на первый взгляд; выражения в скобках есть просто представленное в радианах положение j-й точки, если общую длину временно́го ряда определить как 2π радиан. В силу тригонометрических соотношений коэффициент β_0 всегда равен нулю, а α_0 упрощается

$$\alpha_0 = \frac{1}{n}\sum_{j=0}^{n-1} Y_j, \tag{4.93}$$

что есть как раз среднее временно́го ряда. Индексы изменяются от 0 в начале временно́го ряда до n—1 в его конце, а не от 1 до n для того, чтобы правильно перевести положение первого наблюдения в радианы.

Определив коэффициенты α_k и β_k k-ой гармоники, мы можем определить амплитуду волновой формы в виде

$$A_k = \sqrt{\alpha_k^2 + \beta_k^2}. \tag{4.94}$$

Фазовый угол k-ой гармоники равен

$$\Phi_k = \mathrm{tg}^{-1}(\beta_k/\alpha_k). \tag{4.95}$$

Выражение tg^{-1} означает: «найти угол, тангенс которого задан».

Если в дискретном наборе точек собраны данные для регулярной синусоиды, то дисперсия этой выборки связана с амплитудой этой волновой формы. В пределе дисперсия есть просто половина квадрата амплитуды, или

$$s_k^2 = A_k^2/2 = (\alpha_k^2 + \beta_k^2)/2. \tag{4.96}$$

Так как теорема Фурье утверждает, что временной ряд можно рассматривать как сумму многих синусоидальных функций или гармоник, то дисперсия временно́го ряда является просто суммой дисперсий этих самых гармоник. Мы можем поэтому выразить дисперсию k-ой гармоники как пропорциональную часть от общей дисперсии временно́го ряда, из которого она была получена. Если мы вычислим длинный ряд гармоник, то дисперсии последовательных гармоник можно нанести на периодограмму, или, как иногда говорят, изобразить дискретный или линейный энергетический спектр. Это есть попросту представление зависимости частоты или дисперсии от числа гармоник k, и традиционно представляется в виде, указанном на рис. 4.55. (Термины энергия и дисперсия являются синонимами; первый из них

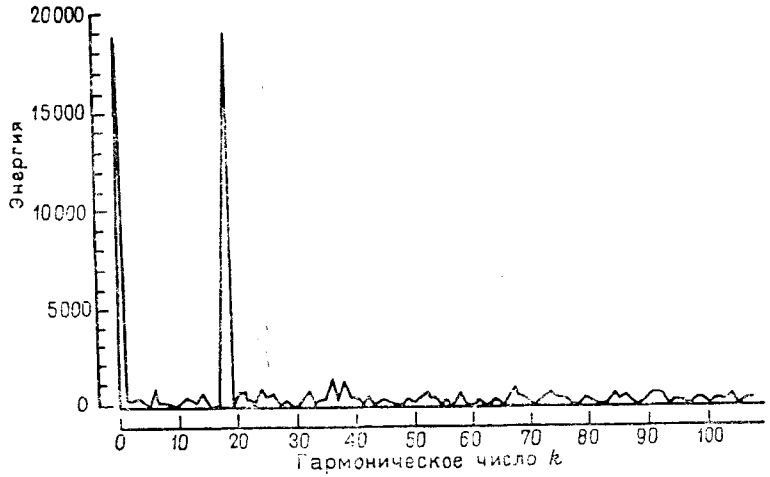


Рис. 4.55. Периодограмма месячного потока Кейв-Крик, Кентукки.

Специальные значения указывают, что поток имеет единственный периодический годичный цикл. Энергии приведены в квадратах единиц измерения, которые равны сотым частям дюйма

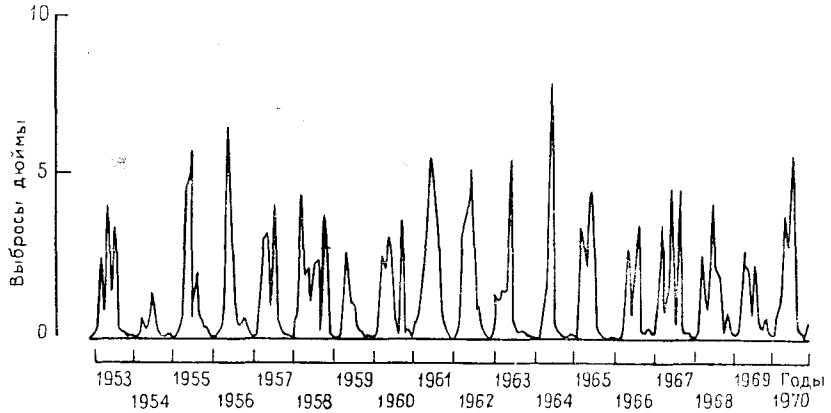


Рис. 4.56. Месячный поток Кейв-Крик, Кентукки за период с октября 1952 г. до сентября 1970 г.

был введен в электротехнике и теперь широко используется во всех областях техники, анализе сигналов и распознавании изображений.) Исходный степенной ряд представлен на рис. 4.56. Временной ряд представляет месячный сток воды в Кейв Крик в Кентукки; данные приведены в табл. 4.29, взятой из [24]. Периодограмма показывает, что изменения месячных стоков имеют годичный (12-месячный) период, плюс случайная компонента.

Спектр, вычисленный указанным выше образом, иногда на-

Таблица 4.29

Месячный сток воды в Кейв Крик, Кентукки (в сотнях дюймов)

Год	Октябрь	Ноябрь	Декабрь	Январь	Февраль	Март	Апрель	Май	Июнь	Июль	Август	Сентябрь
1953	2	5	19	240	86	416	147	354	31	18	7	1
1954	0	2	4	54	22	40	139	35	8	7	6	14
1955	2	4	30	73	463	579	59	197	55	24	28	3
1956	4	6	13	59	637	469	192	28	32	64	38	8
1957	7	10	172	308	325	103	392	68	24	6	5	2
1958	3	106	432	200	221	117	235	236	19	369	170	12
1959	6	9	17	270	195	112	102	24	24	5	4	2
1960	3	36	269	219	313	291	68	19	364	138	14	30
1961	12	52	79	204	295	532	476	414	159	48	18	4
1962	2	6	76	346	401	508	330	79	96	38	8	7
1963	39	141	124	150	146	548	52	25	14	29	11	3
1964	1	4	3	87	173	788	45	21	11	8	2	16
1965	15	7	347	276	230	449	146	31	8	5	1	2
1966	4	2	2	48	281	79	202	332	25	14	41	11
1967	7	119	357	97	161	466	50	476	33	14	15	7
1968	9	38	271	135	98	425	238	199	91	29	75	16
1969	14	22	112	278	216	73	237	74	40	27	66	17
1970	7	25	91	130	389	291	568	206	38	14	6	27

Примечание. Данные собраны в течение богатого осадками года, начиная с октября предыдущего года и кончая сентябрем, для того чтобы избежать излома последовательности в середине зимы [24].

зывается необработанным спектром и является оценкой истинного спектра, или спектра совокупности. Эти оценки связаны с очень большими стандартными погрешностями, которые не могут быть уменьшены при увеличении n по причинам, которые будут объяснены в следующем разделе. Некоторые из этих ошибок могут оказаться ложными в силу смешивания недопустимо высоких частот с низкими частотами. Эти высокие частоты, длины волн которых меньше, чем двойное расстояние между точками опробования, не могут быть обнаружены. Наивысшая частота, которая может быть обнаружена, есть частота Найквиста, ее длина волны равна 2Δ, где Δ — расстояние между последовательными наблюдениями. Причину, по которой дисперсии частот вне предела Найквиста совпадают с дисперсиями более низких частот, можно легко увидеть на рис. 4.57.

Тем не менее можно статистическими методами проверить, существует ли доминантная компонента в периодограмме. Критерий, рекомендованный Фишером, основан на вычислении спект-

Рис. 4.57. Высокоэнергетическая синусоидальная волна (прерывистая линия), наблюденная в дискретном множестве точек, дает отчетливо видную волну низкой частоты (сплошная линия)

рального значения s_k^2, которое превышает значение σ_k^2 временно́го ряда, составленного из независимых случайных точек. Критерий основан на вычислении отношения

$$\hat{g} = s^2_{\max}/2s^2, \tag{4.97}$$

где $s^2_{\max}$ — наибольший пик периодограммы и s^2 — дисперсия всего временно́го ряда. Критическое значение g для спектральной вероятности p дается формулой

$$g \approx 1 - e^{\frac{\ln p - \ln m}{m-1}}, \tag{4.98}$$

где $m = n/2$, если временной ряд содержит четное число наблюдений, и $m = (n-1)/2$, если n нечетно. Если проверяемая статистика $\hat{g}$ превышает критическое значение g, то можно предположить, что периодическая компонента существует. Если проверяемая статистика не превосходит критическое значение, то наблюдаемый спектральный пик $s^2_{\max}$ может возникнуть случайным образом.

Непрерывный спектр

Гармонический анализ и построение периодограммы или линейного спектра целесообразны в тех случаях, когда исследуемый временной ряд действительно периодичен. В природе встречается немного естественных явлений, обладающих истинной периодичностью, не считая явления, связанные с астрономическими циклами, такими, как месячные приливы и отливы или сезонные изменения. Большинство геологических временны́х рядов не периодические, а случайные. Последовательность называется случайной, если она может быть охарактеризована только с помощью статистических характеристик, в противоположность детерминированной последовательности, в которой состояние может быть точно предсказано по ее коэффициентам. Даже в тех случаях, когда временной ряд не содержит действительно периодических компонент, методы спектрального анализа позволяют получить ценную информацию о том процессе, который породил данную последовательность. Большинство временны́х рядов можно считать непрерывными последовательностями, даже несмотря на то, что данные по ним собираются в дискретном множестве точек. Для таких последовательностей можно вычислить непрерывный спектр, или функцию спектральной плотности, причем дисперсия временного ряда разделена пропорционально в множестве частотных полос. Непрерывный спектр имеет вид непрерывной кривой, выражающей зависимость дисперсии от частоты, и аналогичен непрерывной функции распределения вероятностей с дисперсиями, пропорциональными площадям под спектральной кривой между граничными частотами. Общая площадь под спектром равна общей дисперсии временно́го ряда, для которого она вычислялась. Наоборот, линейный спектр периодического временного ряда показывает дисперсии, присущие определенным индивидуальным частотам.

Каждая конкретная последовательность наблюдений, которую мы хотим анализировать, может рассматриваться как случайная выборка из большого, возможно, бесконечного множества таких временны́х рядов, которые могут генерироваться изучаемыми процессами. Полное множество временны́х рядов называется ансамблем, и это есть как раз та совокупность, из которой наша конкретная выборка извлекается.

Временной ряд называется стационарным, если его свойства не изменяются со временем. Пространственный ряд с теми же самыми характеристиками называется однородным. Если временной ряд подразделить на малые сегменты и средние по всем этим сегментам одинаковы (и совпадают со средним всего временно́го ряда), то в этом случае говорят о стационарности первого порядка или о стационарности в среднем. Если в дополнение к этому автоковариация изменяется только с изменением лага, а не в зависимости от положения внутри временно́го ряда, то говорят, что ряд обладает свойством стационарности второго порядка. Это свойство иногда называют слабой стационарностью, или стационарностью в более широком смысле. Если все моменты высшего порядка зависят только от лага и не зависят от положения, то ряд называется сильно стационарным или обладает стационарностью в строгом смысле.

Если временной ряд не только сильно стационарен, но и все его статистики инвариантны от ряда к ряду внутри ансамбля, то ансамбль называется эргодическим. Многие статистические критерии теории временных рядов обладают свойством эргодичности так же, как одномерные статистические критерии обладают свойством однозначности дисперсии. Если мы имеем несколько временны́х рядов, являющихся реализациями одного и того же случайного процесса, то мы можем провести проверку их на эргодичность. Вообще же мы имеем только один временной ряд, и в этом случае никакая проверка такого рода невозможна. Мы можем проверить стационарность по средним значениям и дисперсии для единственного временно́го ряда, применяя регрессию или подразделяя ряды на сегменты и проверяя, одинаковы ли эти статистики для различных сегментов. Если да, то ряд является собственно стационарным и можно предположить, что он обладает свойством эргодичности. Если ряд не является собственно стационарным, то ансамбль, из которого он извлечен, не может быть эргодическим.

Иногда временной ряд может быть сделан стационарным применением простой процедуры уравнивания, или вычитанием линейного тренда из наблюдений, т. е. с помощью метода наименьших квадратов вычисляется линейная регрессия Y_i на j. Тогда новый временной ряд определяется как ряд отклонений:

$$Y_j' = Y_j - \hat{Y}_j, \tag{4.99}$$

где $\hat{Y}_j$ — предсказанные значения, определяемые с помощью уравнения регрессии. Если исходный ряд характеризовался малым изменением его среднего значения, то новый ряд будет иметь стационарное среднее, равное нулю.

Для стационарного случайного временно́го ряда, являющегося непрерывным и представленного дискретным набором равномерно распределенных в пространстве точек, непрерывная дисперсия (или энергетический спектр) может быть вычислена одним из двух методов. Более новый и более широко используемый метод основан на вычислении многих значений линейного спектра с помощью алгоритма быстрого преобразования Фурье. Затем эти спектральные значения усредняются вдоль частот для получения сглаженной оценки непрерывного спектра. Как следует из названия, быстрое преобразование Фурье очень эффективный метод, а именно, он приводит к желаемому результату за $n \log_2 n$ арифметических операций, в то время как другие методы требуют выполнения n^2 операций. Представим соотношение Фурье в виде

$$s_k^2 = \frac{1}{n} \sum_{j=-\frac{n}{2}}^{\frac{n}{2}} Y_j e^{-i2\pi j k/n}, \tag{4.100}$$

где $i = \sqrt{-1}$ — мнимая единица. Выполнение преобразований, диктуемых алгоритмом быстрого преобразования Фурье, требует математических и вычислительных навыков, которые находятся вне рамок излагаемого в этой книге. Поэтому мы не развиваем эту тему, хотя в настоящее время эти методы наиболее широко используются в спектральном анализе. Отличное введение в теорию быстрого преобразования Фурье вместе с алгоритмом дано Рейнером [43], который также дает примеры применения быстрого преобразования Фарье в географии. Первая основополагающая статья, посвященная алгоритму быстрого преобразования Фурье, была опубликована Джентльменом и Сендом [20], она основана на математических методах, которые впервые были введены Кули и Тьюки [11]. Современные руководства по анализу временны́х рядов (Блюмфельд [7], Бендат и Пирсол [6]) трактуют этот предмет весьма широко. Несколько более старый метод вычисления непрерывного спектра состоит в нахождении преобразования Фурье автокорреляционной функции временного ряда. Развитый первоначально Бартлеттом [5], этот метод дает те же результаты, что и более новый метод быстрого преобразования Фурье, он легче для понимания, хотя и не удобен для вычислений. Он еще широко используется, особенно тогда, когда автокорреляционная функция также представляет интерес и когда временной ряд не очень длинный. Енкинс и Уоттс [29], например, обсуждают этот метод, а Йевьевич [64] применяет его к решению гидрогеологических задач.

Преобразование Фурье автокорреляционной функции дает энергетический спектр; для временно́го ряда, заданного дискретным множеством выборочных характеристик. Енкинс и Уоттс [29] дают преобразование в следующем виде:

$$s_f^2 = \frac{1}{2} A_f^2 = \sum_{l=-(L-1)}^{L-1} r_l e^{i2\pi f l}. \tag{4.101}$$

Здесь s_f^2 — дисперсия или энергия в частотной полосе с центром в точке f; r_l — автокорреляция с лагом l; максимальный лаг в автокоррелограмме есть L, $i = \sqrt{-1}$.

Заметим, что это равенство очень напоминает комплексное преобразование Фурье, заданное уравнением (4.100). Однако равенство (4.101) может быть значительно упрощено. Хотя автокорреляционная функция может быть вычислена как для отрицательных, так и для положительных лагов, она симметрична относительно лага нуль; т. е. $r_l = r_{-l}$. Поэтому вычисления необходимо проделать только для значений l от нуля до $L-1$ и затем результат удваивается. Далее, так как наши данные состоят из конечного множества вещественных чисел, которые равномерно расположены в пространстве, мы можем использовать упрощенное представление через косинусы, а не общую форму. Более простой вид уравнения (4.101), пригодный для вычислений, таков:

$$s_f^2 = A_f^2/2 = 2\left(\sum_{l=1}^{L-1} r_l \cos 2\pi f l\right). \tag{4.102}$$

Необходимо работать с частотами, а не с гармониками, так как специфические лаги автокорреляционной функции не преобразуются в простые кратные друг другу в частотном спектре. Соотношение между лагом и частотой дается формулой

$$f_j = \frac{j}{2\Delta L}, \tag{4.103}$$

где Δ — расстояние между последовательными наблюдениями во временном ряде, и j представляет как j-й лаг, так и j-ю спектральную полосу. Частота j-й полосы есть f_j и измеряется в циклах на единицу (единицы те же, что и единицы Δ, расстояния между данными точками). Если Δ измеряется в миллиметрах, то f_j задается в циклах на миллиметр.

Используя уравнение (4.102), мы получаем так называемый необработанный спектр. Из-за стандартной ошибки в каждой из спектральных полос его нельзя считать удовлетворительным. Это препятствие нельзя преодолеть обычным образом, удлиняя временной ряд (увеличивая n — общее число наблюдений). Известно, что при увеличении n при фиксированном лаге l в фор-

мулах произведений $(x_j - \bar{X})(x_{j+1} - \bar{X})$ увеличивается количество информации и автокорреляционная функция r_l при этом будет иметь меньшую стандартную ошибку. Однако это неверно для s_f^2. Екинс и Уоттс [29] показали, что информация, содержащаяся в каждой оценке спектра s_f^2, распространяется на полосу частот, ширина которой вокруг f равна $\pm l/n$. По мере увеличения n общая информация, содержащаяся в дисперсии спектра, распределяется на увеличивающееся число полосок уменьшающейся ширины. В результате увеличение n делает возможной оценку дисперсии в более узких полосках частот, давая смещенную оценку истинного спектра. Однако оценка стандартной ошибки этих более узких полос не улучшается.

Бартлетт [5] впервые разработал метод уменьшения стандартной ошибки в спектральных оценках на основе алгоритма, называемого либо методом скользящего окна, либо фильтрацией, либо сглаживанием, либо пространственным усреднением. Автокорреляционную функцию можно определить взвешенным методом лагового окна или фильтра, и взвешенная автокоррелограмма может быть преобразована в спектральную область.

Веса придают особое значение более коротким лагам, которые основаны на большем числе наблюдений, чем более длинные лаги. С другой стороны необработанный спектр может быть вычислен и затем сглажен взвешиванием и усреднением вместе с прилегающими спектральными значениями. Множество весов называется спектральным окном или фильтром и является преобразованием Фурье лагового окна. Оба метода дают эквивалентные результаты, которые можно характеризовать как смазывание или сглаживание дисперсий примыкающих частот.

Множество окон или фильтров рассматривалось различными авторами; схема окон, имеющая наилучшие свойства, представляет собой наивысшее достижение в анализе временных рядов. Одно из наиболее широко используемых окон — это фильтр Тьюки — Ханнинга, который в спектральной области имеет вид

$$s_f^2 = \frac{1}{4} s_{f-1}^2 + \frac{1}{2} s_f^2 + \frac{1}{4} s_{f+1}^2. \qquad (4.104)$$

Преобразование Фурье фильтра Тьюки — Ханнинга в области лага есть

$$W_l = \frac{1}{2}\left(1 + \cos\frac{\pi l}{L}\right). \qquad (4.105)$$

Эта формула определяет веса автокорреляционной функции, которая гладко убывает от значения $W_0 = 1{,}0$ при нулевом лаге до $W_L = 0{,}0$ при максимальном лаге L. Каждый коэффициент автокорреляции умножается на соответствующий ему вес до выполнения преобразования Фурье. Уравнение (4.102) принимает вид

$$s_f^2 = 2\left(1 + s\sum_{l=1}^{L-1} W_l r_l \cos 2\pi f l\right). \qquad (4.106)$$

Можно предложить и другие окна, обычно либо в области лагов, либо в области частот, так как их применение бывает легче то в одной форме, то в другой. Широко используемое Парзеновское окно имеет вид

$$W_l = 1 - \frac{6l^2}{L^2}\left(1 - \frac{l}{L}\right) \qquad (4.107)$$

для лагов между нулем и $L/2$, и

$$W_l = 2(1 - l/L)^3 \qquad (4.108)$$

для лагов от $L/2$ до L.

Практическая стратегия исследования спектральных характеристик случайных временны́х рядов включает следующие шаги.

1. Если необходимо, избавить от тренда или уравнять данные, вычислив коэффициенты регрессии Y_j на j и затем найти остатки: $Y_j' = Y_j - \hat{Y}_j$.

2. Вычислить автокорреляцию исходного ряда (если он стационарен) или ряда остатков Y' вплоть до максимума лага L, который не превосходит $n/3$. Некоторые авторы рекомендуют вычислять более узкие пределы для максимума лага, например $n/5$, $n/6$ или даже $n/10$.

3. Вычислить необработанную спектральную плотность s_f^2, используя дискретное преобразование Фурье, определенное уравнением (4.102).

4. Сгладить необработанный спектр, используя окно Тьюки — Ханнинга (уравнение 4.104) в качестве скользящего среднего для получения сглаженной оценки дисперсии спектра.

Очевидно, в спектральном анализе имеется много способов принять решение, которое окажет влияние на окончательный результат. Они включают выбор n, длины анализируемой последовательности (если это находится во власти исследователя); максимального лага L; и окна, используемого для сглаживания. Связи между этими переменными и их влияние как на разрешение, так и на значимость оглаженного спектра обсуждены подробно в книге Енкинса и Уоттса [29].

Выполняя анализ Фурье, мы преобразовали наши данные из одной области в другую. Мы начнем с наблюдений за формой зависимости значений Y_j от координат точек пространства или времени x_j. Последовательность точек образует сигнал или волновую форму, определенную в плоскости координат X и Y. В этом случае говорят, что данные находятся во временно́й или пространственной области и зависят от переменной X, обозначающей координаты точек во времени или расстояние. Определив компоненты частот в сигнале, мы преобразовали данные в частотную область. Физическая аналогия может быть изображена на рисунке, где представлен эффект отражения солнечных лучей от зеркальных призм (рис. 4.58). Сноп белого света можно рассматривать как комплексную волновую форму, изменяю-

щуюся с течением времени и состоящую из света различных цветов (или длин волн). Призма действует как частотный анализатор и разделяет пучок на его составляющие части, в результате чего на дисплее получается радуга. Каждая окрашенная полоса отделяется от соседней промежуточной зоной, ширина которой пропорциональна разности их длин волн или частот, а интенсивность каждой полосы пропорциональна вкладу данной длины волны в общую интенсивность исходного пучка. Исследование спектра источника света может сказать нам многое о строении источника света, его температуре, природе вещества, сквозь которое проходит свет и так далее. Аналогично исследование частотного спектра данной последовательности может рассказать нам о его природе и источнике, дать информацию, которую нельзя получить никаким другим способом.

ФИЛЬТРЫ

Фильтром называется математический оператор, который изменяет заданный временной ряд в другой временной ряд, имеющий требуемую форму. Такие операторы называются «фильтрами», так как первоначально они представлялись электронными фильтрующими целями, состоящими из сети сопротивлений и емкостей, используемых для выборочного подавления или усиления специфических частот в электронных сигналах. Теперь те же функции по преобразованию цифровых сигналов выполняются на компьютерах математическими фильтрами. Основы теории фильтрации были заложены Норбертом Винером в работах по статистической теории связи, и эта теория находит свое практическое применение при обработке сейсмических сигналов и

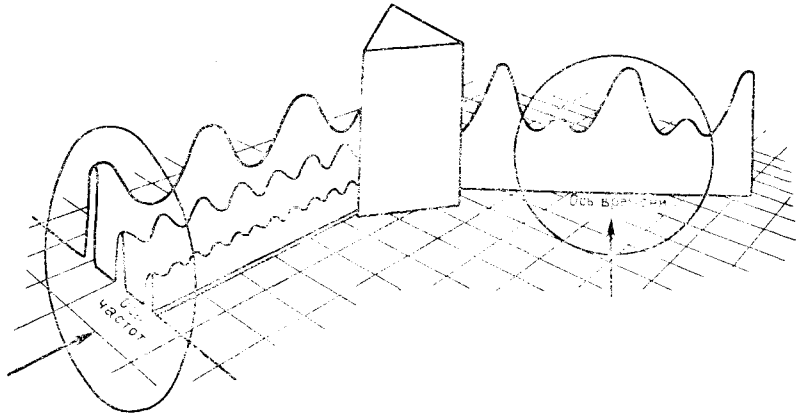


Рис. 4.58. Призма служит частотным анализатором, преобразующим белый свет (в пространственной или временной области) в его составные части разных цветов (область частот)

данных аэрофотосъемок. В силу этого большинство терминов теории фильтрации произошло не из классической статистики, а из инженерной электроники и физики.

Методы теории фильтрации получили наиболее широкое применение при исследовании отраженных сейсмических данных. Исследование сейсмических свойств отраженных сигналов основано на создании входного сигнала (сейсмическая энергия производится взрывом динамита или вибратором), который затем фильтруется в результате поглощения энергии сигнала по мере его путешествия сквозь земную толщу. Высокие частоты поглощаются более сильно, так что природа сейсмических колебаний изменяется по мере их передвижения. Добавим, что волны низких частот перемещаются через горные породы более быстро, так как начальный острый всплеск становится смазанным и ослабленным. Выходной сигнал, который обнаруживается набором поверхностных улавливателей, состоит из отражений сигналов, возвратившихся из более глубоких горизонтов. К сожалению, после их путешествия отраженные волны сильно деформированы и их трудно интерпретировать.

Задача геофизика — устранить настолько, насколько это возможно, вредные эффекты физической фильтрации, которым сейсмический сигнал был подвержен. Это делается путем обработки сейсмических записей цифровыми фильтрами, которые устраняют нежелательные частоты, обостряют отражения и подавляют шум в записи до тех пор, пока требуемый, но ослабленный сигнал не будет выделен.

В силу того что многие возмущения сейсмических сигналов частотно согласованы, геофизиков особенно интересуют фильтры, которые предпочтительно генерируют или устраняют специфические частоты. Сейсмическая фильтрация часто проводится в этой области частот с помощью преобразования Фурье сейсмических сигналов и устранения нежелательных частей спектра и последующего преобразования отфильтрованного спектра обратно в действительную область. Однако в точности тот же процесс может быть реализован преобразованием фильтра области Фурье в реальные данные и последующей сверткой его преобразования с сейсмическими записями сигналов. Так как эти фильтры предназначены для прохождения определенных частот и подавления других частот, то они называются высокочастотными, низкочастотными или полосными фильтрами.

Теоретически можно построить совершенные фильтры, которые будут пропускать сигналы специфических частот и не пропускать другие частоты, но такие фильтры обычно трудно реализовать. Как правило, эти теоретически совершенные фильтры требуют бесконечного числа членов. Если они усекаются до приемлемых размеров, то фильтры сами начинают оказывать влияние на выходной сигнал. Подобрать фильтры таким образом мо-

жет только высококвалифицированный работник, так как фильтр должен быть достаточно коротким и экономически выгодным и должен давать хорошую аппроксимацию требуемого на выходе сигнала с минимальным числом нежелательных побочных эффектов. Подробное изложение теории фильтров специфических частот содержится во многих специальных книгах. Мы же обратимся к концептуально более простым типам фильтров и к вопросам широкого использования анализа временны́х рядов во многих областях.

Временной ряд, который должен быть подвергнут фильтрации, называется входящим, а преобразованный временной ряд, выходящий из фильтра, называется выходящим. Оператор фильтрации сворачивает входящий ряд с фильтром, в результате получается выходящий ряд:

$$[C] = [B] * [f]. \tag{4.109}$$

Если на выходе мы хотим получить определенную форму, скажем $[D]$, то фильтр может быть выбран таким образом, что действительный выходящий ряд $[C]$ был настолько близок к требуемому выходящему ряду, насколько это возможно. Чтобы это сделать, вычислим элементы фильтра, который минимизирует сумму квадратов разностей между $[C]$ и $[D]$. Соответствующий математический метод очень напоминает алгоритм вычисления коэффициентов линейной регрессии. Робинсон и Трейтель [44] дают очень подробное и исключительно ясное изложение этого метода, называемого схемой наименьшего квадратичного фильтра.

Рассмотрим простой пример. Предположим, что входящий вектор B имеет два элемента $[b_0, b_1]$. Тогда фильтр также содержит два элемента $[f_0, f_1]$. Так как выход из фильтра является сверткой входа и фильтра, то он состоит из трех элементов.

$$[C] = [B] \times [f],$$

$$[C] = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} \begin{bmatrix} f_0 & f_1 \end{bmatrix} = \begin{bmatrix} f_0 b_0 & f_1 b_0 \\ f_0 b_1 & f_1 b_1 \end{bmatrix}, \tag{4.110}$$

$$[C] = [f_0 b_0 \quad f_0 b_1 + f_1 b_0 \quad f_1 b_1].$$

Если мы хотим, чтобы выход из фильтра равнялся некоторому заданному $[D]$, то можно написать следующую систему совместных уравнений

$$\begin{aligned} f_0 b_0 + 0 &= d_0, \\ f_0 b_1 + f_1 b_1 &= d_1, \\ 0 + f_1 b_1 &= d_2 \end{aligned} \tag{4.111}$$

или в матричной форме

$$\begin{bmatrix} b_0 & 0 \\ b_1 & b_0 \\ 0 & b_1 \end{bmatrix} \begin{bmatrix} f_0 \\ f_1 \end{bmatrix} = \begin{bmatrix} d_0 \\ d_1 \\ d_2 \end{bmatrix}. \tag{4.112}$$

Будем обозначать матрицу в левой части через $[\beta]$.

Матричное уравнение переопределено, т. е. имеется три уравнения с двумя неизвестными. Однако если обе части уравнения умножить на одну и ту же величину, то уравнение остается неизменным. Так, умножив обе части на транспонированную к $[\beta]$ матрицу, получим

$$[\beta]'[\beta][f] = [\beta]'[D]$$

или

$$\begin{bmatrix} b_0 & b_1 & 0 \\ 0 & b_0 & b_1 \end{bmatrix} \begin{bmatrix} b_0 & 0 \\ b_1 & b_0 \\ 0 & b_1 \end{bmatrix} \begin{bmatrix} f_0 \\ f_1 \end{bmatrix} = \begin{bmatrix} b_0 & b_1 & 0 \\ 0 & b_0 & b_1 \end{bmatrix} \begin{bmatrix} d_0 \\ d_1 \\ d_2 \end{bmatrix},$$

что можно записать в виде

$$\begin{bmatrix} b_0^2 + b_1^2 & b_0 b_1 \\ b_0 b_1 & b_0^2 + b_1^2 \end{bmatrix} \begin{bmatrix} f_0 \\ f_1 \end{bmatrix} = \begin{bmatrix} b_0 d_0 + b_1 d_1 \\ b_0 d_1 + b_1 d_2 \end{bmatrix}. \tag{4.113}$$

Геофизики называют матрицу в левой части автокорреляционной матрицей входа. Правая часть тогда есть кросскорреляции между входом и выходом. Действительно, элементы этих матриц не являются коэффициентом корреляции, а являются неисправленными суммами квадратов и кросспроизведений. Однако знаменатели любого корреляционого члена одинаковы и потому отсутствуют; решение этого уравнения будет одним и тем же независимо от того, используются ли коэффициенты корреляции или квадраты и кросспроизведения. Решение находится обращением матрицы, стоящей в левой части, и последующим умножением на вектор правой части. Если матрицу в левой части обозначить через $[R]$, а вектор правой части через $[X]$, то получим $[f] = [R]^{-1}[X]$.

Одно из общих применений теории фильтрации состоит в исследовании встречаемости специальной волновой формы или ее цифрового представления в некотором временно́м ряде. Аналитик может попытаться превратить эту волновую форму в нечто более явное — острый пик или излом в выходном сигнале.

Пусть входящая волна B есть [21] и требуемый выходной сигнал имеет вид $[D] = [300]$. Они представлены на рис. 4.59. Подставляя эти значения в (4.113), получаем

$$\begin{bmatrix} 2 & 0 \\ 1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} f_0 \\ f_1 \end{bmatrix} = \begin{bmatrix} 3 \\ 0 \\ 0 \end{bmatrix}$$

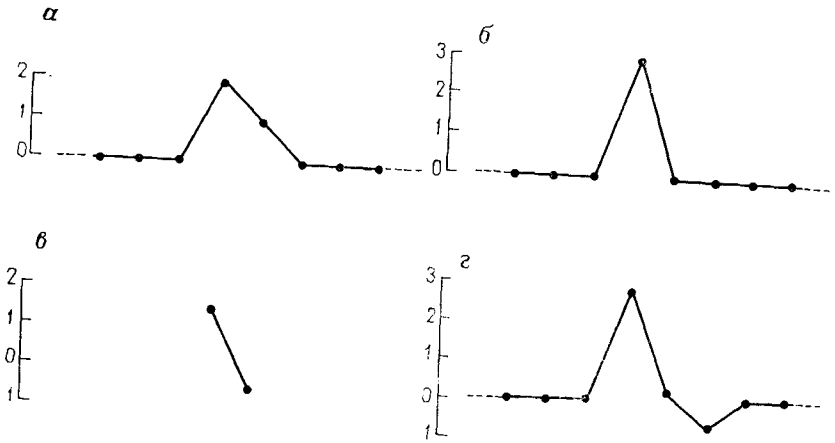


Рис. 4.59. Произвольный вход со значениями [2 1] (*а*); требуемая форма отклика [3 0 0] (*б*); фильтр наименьших квадратов [1,43—0,57] (*в*); отклик фильтра наименьших квадратов [2,9 0,3—0,6] (*г*)

Умножая слева обе части на $[\beta]'$, получаем

$$\begin{bmatrix}2 & 1 & 0\\ 0 & 2 & 1\end{bmatrix}\begin{bmatrix}2 & 0\\ 1 & 2\\ 0 & 1\end{bmatrix}=\begin{bmatrix}f_0\\ f_1\end{bmatrix}=\begin{bmatrix}2 & 1 & 0\\ 0 & 2 & 1\end{bmatrix}\begin{bmatrix}3\\ 0\\ 0\end{bmatrix}=$$

$$=\begin{bmatrix}5 & 2\\ 2 & 5\end{bmatrix}\begin{bmatrix}f_0\\ f_1\end{bmatrix}=\begin{bmatrix}6\\ 0\end{bmatrix}$$

Подстановка матрицы левой части дает

$$\begin{bmatrix}5 & 2\\ 2 & 5\end{bmatrix}^{-1}=\begin{bmatrix}0{,}2381 & -0{,}0952\\ -0{,}0952 & 0{,}2381\end{bmatrix}.$$

Искомые коэффициенты фильтра есть

$$\begin{bmatrix}f_0\\ f_1\end{bmatrix}=\begin{bmatrix}0{,}2381 & -0{,}9852\\ -0{,}0952 & 0{,}2381\end{bmatrix}\begin{bmatrix}6\\ 0\end{bmatrix}=\begin{bmatrix}1{,}4286\\ -0{,}5712\end{bmatrix}.$$

Графическое представление фильтра дано на рис. 4.59. Мы можем определить его преобразование, используя его для преобразования первоначального входного сигнала. Это делается подстановкой коэффициентов в определяющее уравнение фильтра, заданное уравнением (4.109):

$$[C]=[2\ \ 1]*[1{,}4286\ \ -0{,}5712].$$

Свертка имеет вид

$$[C]=\begin{bmatrix}2\\ 1\end{bmatrix}*\begin{bmatrix}1{,}4286 & -0{,}5712\\ 2{,}8572 & -1{,}1424\\ 1{,}4286 & -0{,}5712\end{bmatrix}$$

$$=[2{,}8572\quad(1{,}4286-1{,}1424)\quad -0{,}5712]$$

$$=[2{,}8572\quad 0{,}2862\quad -0{,}5712]$$

Выход фильтра указан на рис. 4.59; очевидно, мы не преуспели в создании чистого острия [3 0 0], но выход является очень хорошей аппроксимацией. Действительно, это есть наилучшая возможная аппроксимация для этого частного входа и фильтра, состоящего только из двух членов. Теперь мы можем измерить, насколько хорош фильтр, сравнив его выход с тем выходом, который мы хотели получить. Это делается вычислением ошибки, определенной как сумма квадратов разностей между C и D:

$$\varepsilon=(3-2{,}8572)^2+(0-0{,}2862)^2+(0+0{,}5712)^2=0{,}1428.$$

Если задать на выходе другие сигналы, не являющиеся пиком, то в некоторых случаях они дадут более эффективный фильтр. Общие типы фильтров носят названия ящик (квадратная волна), фильтр Гаусса, пила.

Сглаживание и временной тренд-анализ

Возможно, наиболее известные геологам фильтры — это те, которые предназначены для уменьшения дисперсии временного ряда. Это — произвольные фильтры, общее назначение которых — сгладить данную последовательность; выход из фильтра — это аппроксимация, тесно связанная с входом. Обоснование необходимости такого процесса фильтрации состоит в том, что временной ряд состоит из двух компонент — «длиннодействующего» сигнала или несущего информацию и наложенного на него случайного шума. В силу его природы такой шум должен быть «короткодействующей» компонентой. Поскольку сигнал мало изменяется от точки к точке, а шум имеет противоположную тенденцию, то среднее по нескольким примыкающим точкам будет стремиться к значению самого сигнала.

Один из методов сглаживания состоит в аппроксимации коротких сегментов исходной последовательности гладкими линиями или кривыми. Эти кривые могут быть подогнаны методом наименьших квадратов с помощью изложенной ранее техники регрессионного анализа. Для этой цели наиболее подходящими оказываются ортогональные многочлены, так как они требуют только перемножения ряда членов и наблюдений в аппроксимируемом сегменте, затем суммирования и деления, чтобы получить некоторый коэффициент аппроксимирующей кривой. Если коэффициенты вычислены, можно построить уравнение и исход-

ные наблюденные значения должны быть заменены на предсказанные значения $\hat{Y}_i$. До тех пор, пока порядок аппроксимирующей кривой меньше числа точек в сегменте, кривая будет аппроксимацией, подверженной меньшим изменениям, чем исходные наблюдения и, следовательно, более гладкой.

Обычно последовательность не сглаживается рядом неперекрывающихся сегментов, так как они имеют резкие изломы в своих звеньях. Вместо этого операция сглаживания состоит в построении аппроксимации к малому сегменту и определении предсказанных значений $\hat{Y}_i$, соответствующих среднему наблюдению в сегменте. Следующий сегмент выбирается так, что он перекрывает все, кроме одного, наблюдения в первом сегменте, и затем процесс повторяется. Наконец, исходное множество данных заменяется на сглаженный ряд, полученный из кривых, подогнанных к этим перекрывающимся сегментами.

Так как множество коэффициентов должно быть определено и вычислено для каждого перекрывающегося сегмента, и имеется почти так же много сегментов во временном ряде, как имеется исходных наблюдений, то это затяжной процесс даже при использовании ортогональных многочленов. Однако сами коэффициенты не являются необходимыми, все, что нам надо, — это оценки $\hat{Y}_i$ в центральных точках сегментов. С этим ограничением можно переписать уравнения ортогональных многочленов так, чтобы получить альтернативное множество ортогональных членов, которые прямо обеспечат получение оценок в центральных точках. Детали этого метода приводятся Савицким и Голеем [48].

Поскольку мы интересуемся только оценкой в центральной точке короткой последовательности, n должно быть всегда нечетным числом. Значит, как квадратичная, так и кубичная кривые будут иметь одно и то же значение в центральной точке, при этом в любом из этих случаев будет использовано одно и то же множество весов. Это верно и для кривых четвертого и пятого порядка. Конечно, если к данным подогнать линейную функцию, то значение в центральной точке будет просто средним точек в сегменте. Это эквивалентно высказыванию, что все члены линейного ортогонального многочлена равны 1.

Таблица 4.30 отражает изложенное выше; она дает ортогональные полиномиальные весовые функции ω для нечетного числа точек в последовательности для порядков кривых 2 и 3, а также для порядков 4 и 5.

Для выполнения сглаживания с использованием ортогональных многочленов мы просто свернем данную последовательность с множеством полиномиальных членов и затем разделим результат на сумму членов, т. е.

$$C = B \ast \frac{[\omega]}{\Sigma\omega}.$$

Так как множество членов симметрично относительно центрального члена, результат одинаков, если мы применим метод скользящего среднего, которое есть просто скользящее кросспроизведение между множеством членов и временны́м рядом, деленным на $\Sigma\omega$. Этот процесс, возможно, больше знаком геологам под названием временно́го тренд-анализа. Среди уравнений, используемых для указанных выше целей, пятичленный фильтр Шеппарда — Тодда, обычно записываемый как скользящее среднее в виде

$$\hat{Y}_i = \frac{1}{35}[17Y_i + 12(Y_{i+1} + Y_{i-1}) - 3(Y_{i+2} + Y_{i-2})].$$

Если сравнить веса уравнения скользящего среднего Шеппарда с ортогональными полиномиальными членами, приведенными в табл. 4.30, то мы увидим, что это просто квадратное уравнение, натянутое методом наименьших квадратов на пять точек. Другие сглаживающие уравнения, используемые во временно́м тренд-анализе, также либо идентичны, либо очень напоминают ортогональные многочлены. Эти сглаживающие функции описаны в классической работе Уиттекером и Робинсоном [60]. На рис. 4.60 представлены результаты сглаживания временно́го ряда каротажа скважин с помощью различных уравнений. Совсем, не очевидно, что эти функции скользящего среднего являются фильтрами, или что процесс скользящего среднего мате-

Таблица 4.30

Ортогональные полиномиальные весовые функции для аппроксимации (сглаживания). Функция прямо дает оценку среднего значения в последовательности из n точек, где n — нечетное число от 5 до 17 [48]

Квадратичные и кубичные члены								Члены четвертого и пятого порядков						
n	5	7	9	11	13	15	17	n	7	9	11	13	15	17
	−3	−2	−21	−36	−11	−78	−21		5	15	18	110	2145	195
	12	3	14	9	0	−13	−6		−30	−55	−45	−198	−2860	−195
	17	6	39	44	9	42	7		75	30	−10	−160	−2937	−260
	12	7	54	69	16	87	18		131	135	60	110	165	117
	−3	6	59	84	21	122	27		75	179	120	390	3755	135
		3	54	89	24	147	34		−30	135	143	600	7500	415
		−2	39	84	25	162	39		5	30	120	677	10125	660
			14	69	24	167	42			−55	60	600	11053	825
			−21	44	21	162	43			15	−10	390	10125	883
				9	16	147	42				−45	110	7500	825
				−36	9	122	39				18	−160	3755	660
					0	87	34					−198	−135	415
					−11	42	27					110	−2937	135
						−13	18						−2860	−117
						−78	7						2145	−260
							−6							−195
							−21							195
$\Sigma\omega$	35	21	231	429	143	1105	323	$\Sigma\omega$	231	429	429	2431	46189	4199

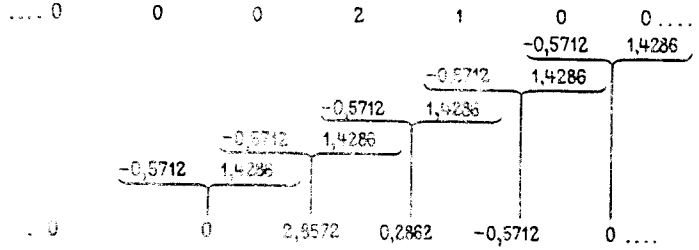
матически эквивалентен свертке двух векторов. Однако они полностью эквивалентны, что легко доказать, используя численный пример, рассмотренный выше. Необходимо предостеречь читателя: для того чтобы использовать фильтр, найденный методом наименьших квадратов, как взвешенное скользящее среднее, а не как некоторый вектор в свертке, необходимо обратить порядок элементов фильтра. Это значит, что первый элемент фильтра становится последним членом скользящего среднего, второй — предпоследним членом и так до тех пор, пока последний элемент не станет первым членом в выражении скользящего среднего.

Так как фильтры симметричны, то совершенно не очевидно, что порядок элементов в большинстве скользящих средних был изменен на противоположный, т. е. $f_{i+j}=f_{i-j}$. В геофизических применениях большинство фильтров являются асимметричными и обращение собственного порядка элементов для использования в скользящем среднем оказывается критическим.

Мы сейчас применим двухэлементный асимметрический фильтр, вычисленный первоначально для временно́го ряда

$$\ldots 0\ 0\ 0\ 2\ 1\ 0\ 0\ldots,$$

в котором читатель легко может узнать волновую форму рис. 4.59, вложенную в последовательность нулей. Если порядок коэффициентов фильтра обратить и использовать в качестве весов в скользящем среднем, то результат будет такой:



Это в точности тот же результат, который мы получили применением операции свертки.

Производные

Ортогональные многочлены также могут быть преобразованы так, чтобы дать другой полезный тип выхода фильтра. Хорошо известен способ вычисления первой производной полиномиальной кривой. Первая производная может рассматриваться как мера изменения наклона подгоняемой линии или скорость изменения значений $\hat{Y}$. Если $\hat{Y}$ увеличивается или уменьшается внутри интервала с примерно постоянной скоростью, то производная

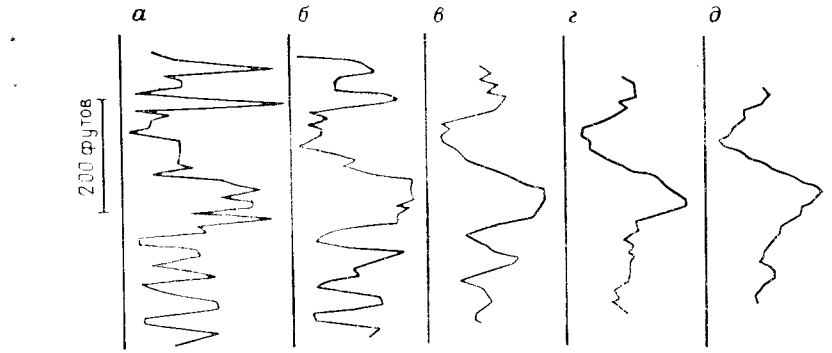


Рис. 4.60. Результаты сглаживания последовательности значений каротажной диаграммы в скважине с помощью различных уравнений [26, рис. 4.2]:
а — исходные данные; *б—д* — число членов: 5 (*б*); 9 (*в*), 13 (*г*), 17 (*д*)

этой функции на этом интервале будет близка к нулю. Большая положительная или отрицательная производная указывает на скачок изменения наклона, возможно, обязанный скачку среднего значения или наличию острого пика.

Если подгоняемая кривая определена ортогональным многочленом, то производная может быть выражена через ортогональные члены. Так как нас интересует только вычисление производной подгоняемой линии в центральной точке интервала, то полиномиальное уравнение может быть переписано так, чтобы прямо дать значение производной. Этот процесс описан Савицким и Голеем [48], которые дают также таблицы для нахождения первой и высших производных ортогональных многочленов. Табл. 4.31 содержит члены для нахождения первой производной, взятые из работы [48].

Оператор производной используется в точности как сглаживающий фильтр в двух случаях: либо при свертывании производных членов с исходной последовательностью данных, либо как набор весов в скользящем среднем. Любой из этих процессов дает отклик, который эквивалентен подгонке полиномиальных кривых к последовательным сегментам и вычислению производных подгоняемых кривых.

Мы можем увидеть результат применения оператора производной к произвольной ступенчатой функции, представленной на рис. 4.61. Крутой угол между двумя сегментами временного ряда выражается как пик в производной, в то время как остаток временно́го ряда имеет локальные производные, равные нулю.

Полиномиальные сглаживающие фильтры могут быть использованы в сочетании с фильтрами дифференцирования, если исходный временной ряд содержит так много ошибок, что производные становятся неинтерпретируемыми. В табл. 4.32 и на рис. 4.62 приведены значения пористости, измеренной в нефтя-

Таблица 4.31

Ортогональные полиномиальные функции для первой производной. Функция прямо дает оценку первой производной в центральной точке полиномиальной кривой, подгоняемой в n точках методом наименьших квадратов, где n — нечетное число от 5 до 17

	Производная квадратичного многочлена						
n	5	7	9	11	13	15	17
	−2	−3	−4	−5	−6	−7	−8
	−1	−2	−3	−4	−5	−6	−7
	0	−1	−2	−3	−4	−5	−6
	1	0	−1	−2	−3	−4	−5
	2	1	0	−1	−2	−3	−4
		2	1	0	−1	−2	−3
		3	2	1	0	−1	−2
			3	2	1	0	−1
			4	3	2	1	0
				4	3	2	1
				5	4	3	2
					5	4	3
					6	5	4
						6	5
						7	6
							7
							8
$\Sigma\omega^2$	10	28	60	110	182	280	408
	Производная кубичного многочлена и многочлена четвертой степени [48]						
n	5	7	9	11	13	15	17
	1	22	86	300	1133	12922	748
	−8	−67	−142	−294	−660	−4121	−98
	0	−58	−193	−532	−1578	−14150	−643
	8	0	−126	−503	−1796	−18332	−930
	−1	58	0	−296	−1489	−17842	−1002
		67	126	0	−832	−13843	−902
		−22	193	296	0	−7506	−673
			142	503	832	0	−358
			−86	532	1489	7506	0
				294	1796	13843	358
				−300	1578	17842	673
					660	18332	902
					−1133	14150	1002
						4121	930
						−12922	643
							98
							−748
Нормализующий член	12	252	1188	5148	24024	334152	23256

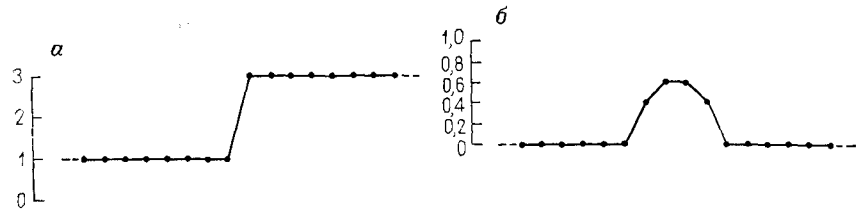


Рис. 4.61. Ступенчатая функция, состоящая из последовательностей единиц и троек (*а*). Графическое изображение первой производной ступенчатой функции, вычисленной на основе пятичленной квадратичной полиномиальной аппроксимации (*б*)

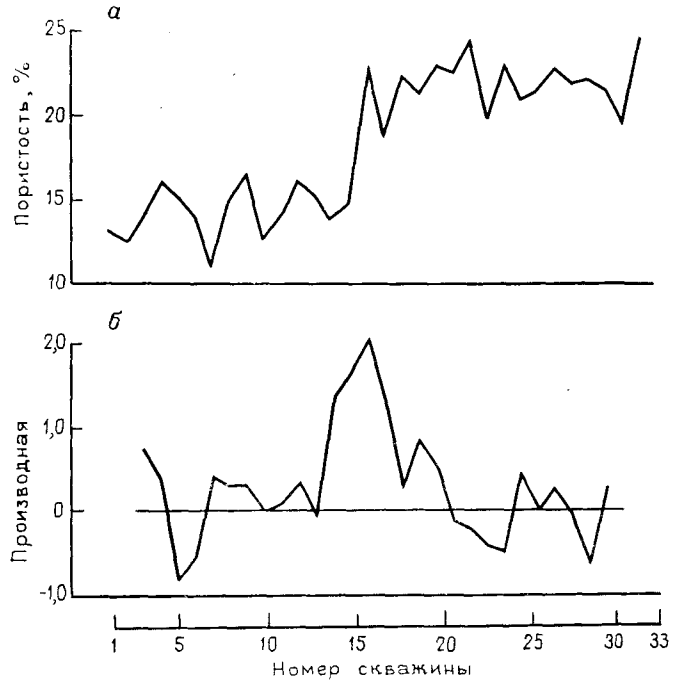


Рис. 4.62. Измерения пористости, сделанные с интервалом 10 футов (3 м) в скважине, пробуренной сквозь нефтяной резервуар на арктическом побережье Аляски:

а — необработанные данные; *б* — первая производная необработанных данных

ной скважине. На рис. 4.62 также представлены их первые производные или, скорее, первые производные квадратичных функций, подогнанных к последовательным множествам из пяти точек. Измерения пористости обладают большой изменчивостью, частично из-за изменчивости экспериментальных данных, но также в силу того, что использование очень маленьких объемов проб приводит к непредставительной выборке из резервуара. Хо-

Таблица 4.32

Значения пористости (в %), измеренные в продуктивном слое нефтяной скважины, пробуренной на арктическом побережье Аляски. Пробы взяты с интервалом в 3 м

(Верх)	1 13,2	9 6,7	17 18,9	25 21,2
	2 12,6	10 12,8	18 22,5	26 21,7
	3 14,3	11 14,2	19 21,6	27 23,0
	4 16,2	12 16,3	20 23,2	28 22,1
	5 15,2	13 15,4	21 22,8	29 22,4
	6 14,1	14 14,0	22 24,7	30 21,8
	7 11,2	15 15,0	23 20,0	31 19,8
	8 15,3	16 23,2	24 23,4	32 24,7 (Основание)

тя график первой производной ясно показывает максимальное изменение пористости в образце с номером 16, ошибки в исходных данных также оказывают влияние на остатки.

На рис. 4.63 представлены измерения пористости после сглаживания пятичленным квадратичным фильтром. Хотя различия в пористости между верхней и нижней частями резервуара сохраняются, большие флуктуации от точки к точке погашаются. Если необработанные данные действительно составлены как смесь «истинных» пористостей со случайными ошибками, то гладкая аппроксимация может давать более близкую к действительности картину изменения пористости, чем исходные данные.

Первая производная сглаженных данных также представлена на рис. 4.63. Явное изменение пористости между пробами с номерами 1 и 16 ясно показано. Как и следовало ожидать, все производные остатков сглаженной кривой близки к нулю.

Геофизики устанавливают качество фильтра, исследуя частотный спектр до и после фильтрации. Изменения в спектре указывают на то, как фильтр действует, и на то, насколько высокий эффект он дает. Мы можем выполнить простой анализ на качество фильтра, сравнивая дисперсию временного ряда до

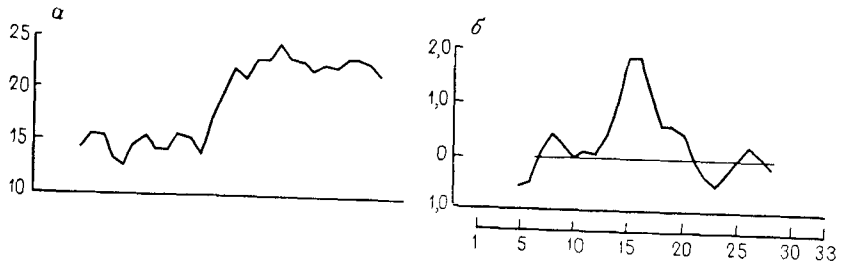


Рис. 4.63. Сглаженные значения пористости из арктической скважины:
а — данные, сглаженные с помощью пятичленного квадратичного многочлена; *б* — первая производная сглаженных данных

фильтрации с дисперсией после фильтрации. Необходимые суммы квадратов для вычисления дисперсии таковы:

для исходного временно́го ряда

$$SS_0 = \Sigma Y^2 - \frac{1}{n}(\Sigma Y)^2,$$

$$SS_0^* = \Sigma Y^2 - \frac{1}{n^*}(\Sigma Y)^2,$$

для ряда, прошедшего через фильтр

$$SS_f = \Sigma \hat{Y}^2 - \frac{1}{n^*}(\Sigma \hat{Y})^2,$$

для отклонений

$$SS_d = \frac{1}{n^*}\Sigma (Y - \hat{Y})^2.$$

Приближенное процентное отношение сумм квадратов имеет вид $(SS_f/SS_0^*)\ 100\%$. В этих равенствах все суммирования проводятся по n данным точкам исходной последовательности или по точкам сглаженной последовательности. В частности, отметим, что вычисляются два значения для суммы квадратов первоначальных данных. Первая из них учитывает все точки данной последовательности от $i=1$ до n, вторая включает только те наблюдения, для которых вычисляются оценки Y_i. Таким образом, из-за потери данных на концах сглаженной последовательности мы имеем $n^*=n-(m-1)$. Хотя в этом случае, как и в регрессионном анализе, можно было бы ожидать выполнения равенства $SS_0^*=SS_f+SS_d$, тем не менее его нет в методе скользящего среднего. Это происходит по той причине, что вычисление оценок $\hat{Y}_i$ вблизи концов последовательности данных отчасти основано на использовании значений, не входящих в вычисление SS_0^*. Поэтому значение процентного отношения сумм квадратов является приближенным, но его можно использовать как показатель эффективности процесса фильтрации.

АНАЛИЗ ВЗАИМОЗАМЕНЯЕМОСТИ

На рис. 4.64 представлен гипотетический стратиграфический разрез с закодированной последовательностью различных литологических разновидностей пород. Если вы исследуете закодированную последовательность, то заметите, что в ней часто встречаются последовательности $A \to B \to C$ и $A \to D \to C$. Это наводит на мысль о том, что состояния B и D как-то связаны и одно может быть заменено другим в предлагаемой последовательности. Это свойство двух или более состояний встречаться в одном и том же окружении называется взаимозаменяемостью,

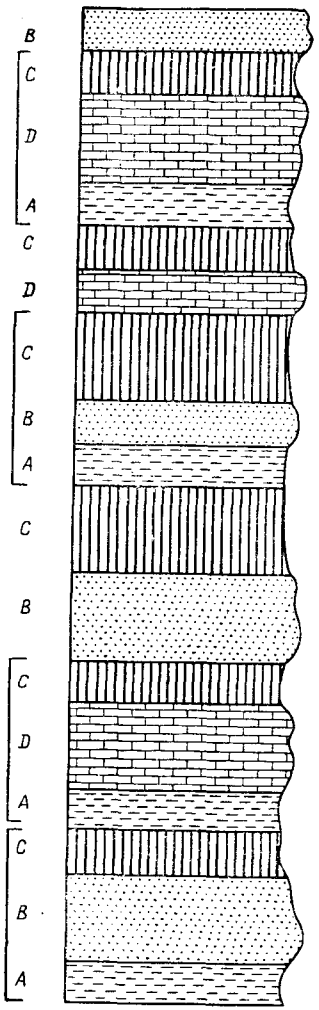


Рис. 4.64. Часть четырехчленной стратиграфической последовательности вида $A \to B \to C$ и $A \to D \to C$

и исследование этого явления близко к выявлению групп состояний, заранее не очевидных. Такие исследования применялись в различных областях, как, например, при анализе частей речи и автоматизации процесса фотографирования [46]. Эти методы представляются перспективными и для исследования геологических данных.

Анализ матрицы переходных вероятностей позволяет установить две величины: взаимозаменяемость первого порядка слева и взаимозаменяемость первого порядка справа. Эти термины возникли при исследовании написанных текстов, в которых состояния представлены словами, а чтение осуществляется слева направо. Если за какими-либо двумя отдельно взятыми словами постоянно следует одно и то же слово, то они могут быть заменены друг другом. Это явление называется левой взаимозаменяемостью, так как слово, которое может быть заменено другим в последовательности, располагается слева. Поскольку стратиграфические последовательности читаются снизу вверх, то более удобным в этом случае является термин «взаимозаменяемость снизу». Матрица левой взаимозаменяемости получается из матрицы переходных вероятностей в результате вычисления отношений произведений строк матрицы.

Предположим, что мы имеем матрицу переходных вероятностей порядка $m \times m$, где P_{ij} — вероятность получения состояния A_j после состояния A_i. Отношение произведений, определяющее левую взаимозаменяемость двух состояний, скажем A_r и A_s равно

$$L_{rs} = \frac{\sum_{j=1}^{m} P_{rj} P_{sj}}{\sqrt{\sum_{j=1}^{m} P_{rj}^2 \sum_{j=1}^{m} P_{sj}^2}}. \qquad (4.114)$$

Для нахождения знаменателя возводим в квадрат каждый элемент строк r и s, суммируем квадраты по строкам, перемножаем суммы и находим квадратный корень из этого произведения. Для нахождения числителя нужно умножить каждый элемент r-й строки на соответствующий элемент s-й строки и найти сумму этих произведений. Отношение двух полученных величин и есть отношение произведений или мера левой взаимозаменяемости L_{rs}. Заметим, что если строки с номерами r и s идентичны ($r=s$), то числитель и знаменатель совпадают, и отношение равно 1,0. Поэтому мера взаимозаменяемости изменяется в пределах от 0 до 1,0, т. е. $0 \leqslant L_{rs} \leqslant 1$. Так как значения вероятностей P_{ij} находятся с помощью деления каждого элемента строки матрицы переходных частот на сумму строки, то тот же результат можно получить прямым вычислением отношения произведений строк матрицы переходных частот.

Конечный результат вычислений, проводимых со строками матрицы переходных вероятностей, — получение симметричной матрицы отношений произведений. Диагональные элементы, конечно, будут равны 1,0. Другие элементы характеризуют степень сходства между парами различных состояний, основанную на процентном отношении чисел, характеризующих порядок следования некоторого третьего состояния за данной парой состояний.

Матрицы переходных частот и вероятностей для полной последовательности, из которой взят изображенный на рис. 4.64 участок, имеют следующий вид

Матрица переходных частот

	A	B	C	D	Сумма по строкам
A	0	11	2	10	23
B	4	0	13	3	20
C	14	6	0	9	29
D	4	3	15	0	22

Матрица переходных вероятностей (для переходов снизу вверх)

	A	B	C	D	Сумма по строкам
A	0,00	0,48	0,09	0,43	1,00
B	0,20	0,00	0,65	0,15	1,00
C	0,48	0,21	0,00	0,31	1,00
D	0,18	0,14	0,68	0,00	1,00

Мера левой взаимозаменяемости между состояниями A и B вычисляется как отношение произведений следующих строк:

$$\begin{matrix} A \\ B \end{matrix} \quad \begin{bmatrix} 0{,}00 & 0{,}48 & 0{,}09 & 0{,}43 \\ 0{,}20 & 0{,}00 & 0{,}65 & 0{,}15 \end{bmatrix}$$

Вычисляя это значение по формуле (4.114), получаем $L_{AB} = 0{,}27$, что дает нам элементы L_{12} и L_{21} матрицы мер левой взаимозаменяемости. Аналогичным образом находятся ее остальные элементы. Полная матрица имеет следующий вид:

$$\begin{bmatrix} 1{,}00 & 0{,}27 & 0{,}60 & 0{,}25 \\ 0{,}27 & 1{,}00 & 0{,}34 & 0{,}96 \\ 0{,}60 & 0{,}34 & 1{,}00 & 0{,}27 \\ 0{,}25 & 0{,}96 & 0{,}27 & 1{,}00 \end{bmatrix}$$

Высокая взаимозаменяемость сверху показывает, что существует сильная тенденция к появлению одного и того же состояния вслед за двумя другими состояниями, т. е. они появляются в одинаковом окружении. Низкая взаимозаменяемость снизу, наоборот, показывает, что вслед за двумя состояниями появляются различные состояния.

Матрица мер правой взаимозаменяемости (или взаимозаменяемости сверху) находится с помощью вычисления отношения произведений столбцов матрицы переходных вероятностей (сверху вниз), которые в свою очередь находятся с помощью деления каждого элемента матрицы переходных частот на сумму элементов соответствующего столбца. Эта матрица вероятностей содержит относительные частоты, с которыми выбранное состояние является предшествующим или последующим для другого состояния. Матрица в точности такая же, как мы могли бы получить, если бы приняли конец данной последовательности за начало, вычислили вероятности переходов сверху вниз и затем матрицу переходных вероятностей. Матрица переходных вероятностей сверху вниз указана ниже.

Матрица переходных частот

	A	*B*	*C*	*D*
A	0	11	2	10
B	4	0	13	3
C	14	6	0	9
D	4	3	15	0
Суммы по столбцам	22	20	30	22

Матрица переходных вероятностей (для переходов сверху вниз)

	A	*B*	*C*	*D*
A	0,00	0,55	0,07	0,45
B	0,18	0,00	0,43	0,14
C	0,64	0,30	0,00	0,41
D	0,18	0,15	0,50	0,00
Суммы по столбцам	1,00	1,00	1,00	1,00

В любом случае, имеем ли мы дело со строками или столбцами, все элементы строки или столбца делятся на сумму элементов этой строки или столбца, причем эти суммы сокращаются при вычислении отношения произведений между строками или столбцами. Поэтому тот же результат получается, если произвести эти операции прямо с матрицей переходных частот. Мера правой взаимозаменяемости состояний A и B вычисляется как отношение произведений двух столбцов:

$$\begin{matrix} A \\ \begin{bmatrix} 0{,}00 \\ 0{,}18 \\ 0{,}64 \\ 0{,}18 \end{bmatrix} \end{matrix} \longleftrightarrow \begin{matrix} B \\ \begin{bmatrix} 0{,}55 \\ 0{,}00 \\ 0{,}30 \\ 0{,}15 \end{bmatrix} \end{matrix}$$

Мера правой взаимозаменяемости равна $R_{AB}=0{,}49$. Другие элементы находятся аналогично. Полная матрица правой взаимозаменяемости имеет вид

$$\begin{bmatrix} 1{,}00 & 0{,}49 & 0{,}37 & 0{,}63 \\ 0{,}49 & 1{,}00 & 0{,}27 & 0{,}94 \\ 0{,}37 & 0{,}27 & 1{,}00 & 0{,}21 \\ 0{,}63 & 0{,}94 & 0{,}21 & 1{,}00 \end{bmatrix}$$

Интерпретация этой матрицы такая же, как и интерпретация матрицы мер левой взаимозаменяемости; отличие состоит лишь в том, что сходство в ней устанавливается на основании тенденции одних и тех же состояний быть предшествующими или последующими.

Наконец, мы можем определить матрицу мер взаимной заменяемости как матрицу произведений всех пар значений мер левой и правой взаимозаменяемости, т. е.

$$C_{ij} = L_{ij}R_{ij}. \tag{4.115}$$

Эта мера характеризует близость одного состояния с другим через относительные частоты, с которыми эти состояния появляются в последовательности, т. е. входят между сходными состояниями. Произведение множественных элементов двух матриц взаимозаменяемости имеет вид

$$\begin{bmatrix} 1{,}00 & 0{,}13 & 0{,}22 & 0{,}16 \\ 0{,}13 & 1{,}00 & 0{,}09 & 0{,}90 \\ 0{,}22 & 0{,}09 & 1{,}00 & 0{,}06 \\ 0{,}16 & 0{,}90 & 0{,}06 & 1{,}00 \end{bmatrix}$$

Мы можем построить «дерево» иерархического размещения, в котором указаны связи одного состояния с другим на основе их взаимной заменяемости и в котором изучаемые состояния расположены в порядке их наибольшей взаимозаменяемости. Методы построения таких «деревьев» с помощью ЭВМ будут рассмотрены в гл. 6. Теперь же мы ограничимся изучением результатов группирования «дерева» матриц взаимозаменяемости, построенного для стратиграфического разреза, представленного на рис. 4.64. Такие три «дерева» изображены на рис. 4.65. Ясно, что B и D тесно связаны как с точки зрения общего предшествующего, так и с точки зрения последующего состояния.

Появившиеся недавно публикации, посвященные анализу взаимозаменяемости в геологических задачах, не могут служить еще ярким доказательством полезности, применения этих методов в геологии. Эксперименты, использующие данные по стратиграфическому разрезу в Канзасе, оказались полезными в интерпретации циклотем. Все отложения, вошедшие в этот большой разрез, были расклассифицированы на 18 литологических со-

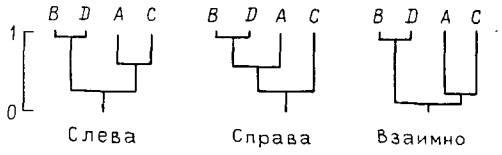


Рис. 4.65. Группы стратиграфических состояний, построенные по трем типам мер взаимозаменяемости.

Состояния *B* и *D* аналогичны по их положениям в стратиграфической последовательности

стояний. Анализ взаимозаменяемости был использован с целью нахождения минимального числа литологических разновидностей, необходимых для определения последовательности. Благодаря этому удалось избежать классификации известняков по их расположению в разрезе и объективно создать определенную циклическую модель [14]. Вероятно, этот метод можно успешно использовать при решении других стратиграфических задач, а также при исследованиях минеральных парагенезисов.

На этом мы заканчиваем рассмотрение методов исследования последовательностей данных. Нами описаны наиболее часто используемые в настоящее время в геологии процедуры и методы, которые позволили получить интересные результаты и в других областях. Однако наше изложение никаким образом нельзя считать исчерпывающим, и может случиться, что в конечном итоге полезными для наук о Земле окажутся совсем другие методы. Однако рассмотренные вопросы охватывают значительный круг задач и составляют основу для последующего изучения и продолжения исследований.

Отметим, что геологи не изучали проблему анализа последовательностей данных в той же мере, как задачи исследования распределенных данных в пространстве. Вполне вероятно, что анализ карт получил свое развитие благодаря тому, что он широко использовался для предсказания финансовых затрат при разведке месторождений как нефтяных, так и твердых полезных ископаемых. Несомненно, методы анализа двумерных данных очень важны, и мы рассмотрим их в гл. 5. Однако предсказание финансовых прибылей не может служить доказательством в пользу применения рассмотренных методов для изучения последовательностей данных. Скорее следует обратить внимание на потенциальные возможности метода автоматической корреляции каротажных диаграмм скважин или на метод изучения продолжительности времени между вулканическими извержениями или землетрясениями. Эти задачи еще не решены, и ни один из изложенных здесь методов не может претендовать на то, что он даст удовлетворительное решение. Однако по мере того как мы будем больше узнавать о геологических последовательностях, будут появляться все более мощные и усовершенствованные методы. Конечным результатом наших исследований должно быть не решение специфических задач, а расширение наших знаний о процессах, которые происходят внутри Земли.

СПИСОК ЛИТЕРАТУРЫ

1. *Agterberg F. P.* Mathematical models in ore evaluation. Canadian Research Soc. Jour., 1967, *5,* p. 144—158.

2. *Anderson R. Y.* and *Koopmans L. H.* Harmonic analysis of varve time series. Jour. Geophysical Research, 1963, *68,* p. 877—893.

3. *Armstrong M.* and *R. Jabin R.* Variogram models must be positive definite. Jour. Int'l Assoc. Mathematical Geology, 1981, *13,* N 5, p. 455—459.

4. *Bardwell G. E.* Some statistical features of the relationship between Rocky Mountain Sarsenal waste disposal and frequency of earthquakes. Geol. Soc. America, Engineering Geology Case Histories, 1970, N 8, p. 33—37.

5. *Bartlett M. S.* Smoothing periodograms from time series with spectra continuous. Nature, 1948, *161,* p. 686—687.

6. *Bendat J. S.* and *Piersol A. G.* Random data: Analysis and measurement procedures: Wiley Interscience Inc., New York, 1971, 407 p.

7. *Bloomfield P.* Fourier analysis of time series, An Introduction. Wiley Interscience Inc., New York, 1976, 258 p.

8. *Bradley J. V.* Distribution-free statistical tests. Prentice-Hall Inc., Englewood Cliffs, N. J., 1968, 388 p.

9. *Clark I.* Practical geostatistics. Applied Science Publishers, London, 1979, 129 p.

10. *Conover W. J.* Practical nonparametric statistics. John Wiley and Sons, Inc., New York, 1980, 493 p.

11. *Cooley J. W.* and *Tukey J. W.* An algorithm for machine computation of complex Fourier series. Mathematical Computing, 1965, p. 297—301.

12. *Cox D. R.* and *Lewis P. A. W.* The statistical analysis of series of events. Methuen and Co., Ltd., London, 1966, 285 p.

13. *Cox D. R.* and *Miller H. D.* The theory of stochastic processes. John Wiley and Sons, Inc., New York, 1965, 398 p.

14. *Davis J. C.* and *Cocke J. M.* Interpretation of complex lithologic successions by substituability analysis, in Merriam D. F., ed., Mathematical models of sedimentary processes. Plenum Press, New York, 1982, p. 27—52.

15. *Doveton J. H.* An application of Markov chain analysis to the Ayrshire Coal Measures succession. Scottish Jour. Geology, 1971, *7,* p. 11—27.

16. *Doveton J. H.* and *Skipper K.* Markov chain and substituability analysis of turbidite succesion. Cloridorme Formation (Middle Ordovician). Caspé, Quebec. Canadian Jour, Earth. Sciences, 1974, *11,* p. 472—488.

17. *Draper N. R.* and *Smith H.* Applied regression analysis 2nd ed. John Wiley and Sons, Inc., New York, 1981, 709 p.

18. *Fisher R. A.* Statistical methods for research workers. 14th ed. Hafner Publ. Co., New York, 1970, 362 p.

19. *Fisher R. A.* and *Yates F.* Statistical tables for biological, agricultural and medical research, 6th ed., Oliver and Boyd, London, 1963, 126 p.

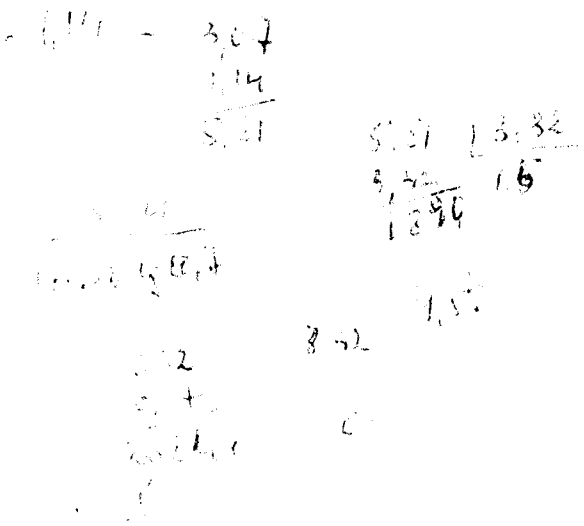
20. *Gentleman W. M.* and *Sande G.* Fast Fourier Transforms—for fun and profit. Bell Telephone Laboratories, Murray Hill, N. J., 1966, 65 p.

21. *Gill D.* Application of statistical zonation method to reservoir evaluation and digitized-log analysis. Bull. American Assoc. Petroleum Geologists, 1970, *54,* no. 5, p. 719—729.

22. *Goodman L. A.* The analysis of cross-classified data. Independence, quasi-independence and interactions in contingency tables with and without missing entries. American Statistical Assoc., Jour., 1968, *63,* p. 1091—1131.

23. *Gordon A. D.* and *Reyment R. A.* Slotting of borehole sequences. Jour. Int'l Assoc. Math. Geology, 1979, *11*, no. 3, p. 309—327.
24. *Haan C. T.* Statistical methods in hydrology. Iowa State Univ. Press., Ames Ia., 1977, 378 p.
25. *Harbaugh I. W.* and *G. Bonham-Carter.* Computer simulation in geology: John Wiley and Sons, Inc., New York, 1970, 575 p.
26. *Harbaugh J. W.* and *Meriam D. F.* Computer applications in stratigraphic analysis. John Wiley and Sons, Inc., New York, 1968, 282 p.
27. *Hawkins D. M.* and *Merriam D. F.* Optimal zonation of digitized sequential data. Jour. Int'l Assoc. Mathematical Geology, 1973, *5*, no. 4, p. 389—395.
28. *Hawking D. M.* and *Merriam D. F.* Zonation of multivariate sequences of digitized geologic data. Jour. Int'l Assoc. Mathematical Geology, 1974, *6*, no. 3, p. 263—269.
29. *Jenkins G. M.* and *Watts D. G.* Spectral analysis and the applications Holden-Day, San Francisko, 1968, 525 p.
30. *Kemeny J. G.* and *Snell J. L.* Finite Markov chains. Van Nostrand Co., Inc., Princeton, N. J., 1960, 210 p.
31. *Kermack K. A.* and *Haldane B. S.* Organic correlation and allomatry. Biometrika, 1950, *37*, p. 30—41.
32. *Krumbein W. C.* FORTRAN IV computer program for Markov chain experiments in geology. Kansas Geological Survey Computer Contribution, *13*, 1967, 38 p.
33. *Kruskal W.* On the uniqueness of the line of organic correlation. Biometrics, 1953, *9*, p. 47—58.
34. *Kuno H.* Catalogue of the active volcanoes of the world including solfatarafie 1ds part XI, Japan, Taiwan and Marianas, Inter. Volcanological Assoc., Naples, 1962, 332 p.
35. *Li J. C. R.* Statistical Inference, v. 1 and 2. Edward Bros., Inc., Ann. Arbor Mich, 1964, 658 p. (v. 1), 575 p. (v. 2).
36. *Miller R. L.* and *Kahn J. S.* Statistical analysis in the geological sciences John Wiley and Sons, Inc., New York, 1962, 483 p.
37. *Morrison D. E.* Applied linear statistical methods. Prentice-Hall, Inc., Englewood Cliffs, N. J., 1983, 562 p.
38. *Olea R. A.* Measuring spatial depedence with semivariograms. Kansas Geological Survey Series on Spatial Analysis, 1977, no. 3, Lawrence, Kans., 29 p.
39. *Ostle B.* and *Mensing R.* Statistics in research, 3rd edition, Iowa State Univ. Press, Ames, *1a*, 1975, 612 p.
40. *Owen D. B.* Handbook of statistical tables. Pergamon Press, London, 1962, 580 p.
41. *Panofsky H. A.* and *Brier G. W.* Some applications of statistics to meteorology. Pennsylvania State Univ., University Park, Pa., 1965, 224 p.
42. *Quenouile M. H.* Associated measurements, Butterworths, London, 1952, 279 p.
43. *Ragner J. N.* An introduction to spectral analysis. Pion Ltd., London, 1971, 174 p.
44. *Robinson E. A.* and *Treitel S.* Geophysical signal analysis. Prentice-Hall, Inc., Englewood Cliffs, N. J., 1980, 486 p.
45. *Rogers D. F.* and *Adams J. A.* Mathematical elements for computer graphics. McGraw-Hill, Inc., New York, 1976, 239 p.
46. *Rosenfeld A.* and *Huang H. K.* An application of cluster detection to text and picture processing. Tech. Rep. 69—68, computer Science Center, Univ of Maryland, Collegue Park, Md., 1968, 64 p.
47. *Saskin M. J.* and *Merriam D. F.* Autoassociassion, a new geological tool. Jour. Int'l Assoc. Mathematical Geology, 1969, *1*, no. 1, p. 7—16.
48. *Savitzky A.* and *Golay J. E.* Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry, 1964, *36*, no. 8, 1627—1639.
49. *Schwarzacher W.* Sedimentation models and quantitative stratigraphy. Elsevier Publ. Co., Amsterdam, 1975, 382 p.

50. *Siegel S.* Nonparametric statistics for behavioral sciences. McGraw-Hill, Inc., New York, 1956, 312 p.
51. *Sneath P. H. A.* and *Sokal R. R.* Numerical Taxonomy. The principles and practice of numerical classification. W. H. Freeman and Co., San Francisco, 1973, 573 p.
52. *Till R.* Statistical methods in Earth sciences. John Wiley and Sons, Inc., New York, 1974, 154 p.
53. *Tipper J. C.* Surface modelling techniques. Kansas Geological Survey Series on Spatial Analysis, no. 4, Lawrence, Kans., 1979, 108 p.
54. *Türk G.* Transition analysis of structural sequences. Discussion Geol. Soc. America Bulletin, 1979, Part 1, *90*, p. 989—992.
55. *Vistelius A. B.* Sedimentation time trend functions and their application for correlation of sedimentary deposits. Jour. Geology, 1961, *69*, p. 703—738.
56. *Webster R.* Automatic soil-boundary location from transect data. Jour. Int'l Assoc. Mathematical Geology, 1973, *5*, no. 1, p. 27—37.
57. *Webster R.*, DIVIDE: A FORTRAN IV program for segmenting multivariate one-dimentional spatial series. Computers and Geosciences, 1980, *6*, no. 1, p. 61—68.
58. *Wells R. C., Bailey R. K.* and *Henderson E. P.* Salinity of the water of Chessapeake Bay. U. S. Geological Survey, Prof. Paper, 1928, *154*, p. 105—152.
59. *Westlake J. R.* A handbook of numerical matrix inversion and solution of linear equations. John Wiley and Sons, Inc., New York, 1968, 171 p.
60. *Whittaker E. T.* and *Robinson G.* The calculus of observations, 4th ed., Blackie and Son, Ltd., Glasgow, 1944, 395 p.
61. *Wickman F. E.* Repose-period patterns of volcanoes. Arkiv för Mineralogi och Geologi, Bd. 4, 1966, p. 291—366.
62. *Wilkes M. V.* A short introduction to numerical analysis. Cambridge Univ. Press, Cambridge, 1966, 76 p.
63. *Yule G. U.* and *Kendall M. G.* An introduction to the theory of statistics. 14 th ed., Hafner Publ. Co., New York, 1969, 701 p.
64. *Yevjevich V.* Stachastic processes in hydrology. Water Resources Publications, Fort Collins, Colo., 1972, 276 p.



ПРОИЗВОДСТВЕННОЕ (ПРАКТИЧЕСКОЕ) ИЗДАНИЕ

Дэвис Джон С.

СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ В ГЕОЛОГИИ

Заведующий редакцией *В. А. Крыжановский*
Редактор *А. М. Антокольская*
Переплет художника *А. Н. Курьеровой*
Художественный редактор *Г. Н. Юрчевская*
Технический редактор *С. Г. Веселкина*
Корректор *Н. А. Громова*

ИБ № 8701

Сдано в набор 06.04.90. Подписано в печать 06.08.90. Формат 60×90¹/₁₆. Бумага тип. № 1. Гарнитура Литературная. Печать высокая. Усл. печ. л. 20,0. Усл. кр.-отт. 20,0. Уч.-изд. л. 21,07. Тираж 4060 экз. Заказ 201/2262—2
Цена 1 р. 80 к.

Ордена «Знак Почета» издательство «Недра». 125047, Москва, пл. Белорусского вокзала, 3

Московская типография № 11 Государственного комитета СССР по печати. 113105, Москва, Нагатинская ул., д. 1