ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ
РОССИЙСКОЙ ФЕДЕРАЦИИ
НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ФАКУЛЬТЕТ ЕСТЕСТВЕННЫХ НАУК
КАФЕДРА ХИМИИ ТВЕРДОГО ТЕЛА

НАУЧНО-ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР «МОЛЕКУЛЯРНЫЙ ДИЗАЙН
И ЭКОЛОГИЧЕСКИ БЕЗОПАСНЫЕ ТЕХНОЛОГИИ»

Т. Н. Дребущак

ВВЕДЕНИЕ В ХЕМОМЕТРИКУ

Учебное пособие

Новосибирск
2013

УДК 543.08(075.8)
ББК Г4я73-1
Д 730

Дребущак Т. Н. Введение в хемометрику: Учеб. пособие / Новосиб. гос. ун-т. Новосибирск, 2013. 89 с.

Рассматриваются вопросы, связанные с использованием хемометрических процедур при анализе экспериментальных данных в различных областях химии. Дается краткий теоретический материал по основам хемометрики, обзорно рассматриваются многомерные методы анализа, приводятся примеры практических заданий. Хемометрика находится на стыке химии и математики, она использует математические, статистические и другие методы для конструирования оптимальных измерительных процедур и для извлечения достоверной химической информации из экспериментальных данных. В пособии основной упор сделан на рассмотрении базовых методов анализа экспериментальных данных и на отработке навыков их практического применения. Для выполнения заданий, представленных в пособии, необходимо, чтобы студенты были знакомы с основами информатики и умели работать с электронными таблицами. Пособие написано на основе спецкурса по хемометрике, который читается на кафедре химии твердого тела на протяжении нескольких лет.

Учебное пособие предназначено для студентов и аспирантов химических специальностей.

Издание подготовлено в рамках реализации *Программы развития государственного образовательного учреждения высшего профессионального образования «Новосибирский государственный университет»* на 2009–2018 годы.

Рецензент
д-р хим. наук Ю. В. Серёткин

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Хемометрика как научная дисциплина зародилась на стыке прикладной математики и экспериментальной химии. Термин появился в 70-х годах прошлого века. В это же время появились и курсы с таким названием в университетах США, Англии, Германии и т. д.

В последние годы резко возросло количество публикаций в области хемометрики. Хемометрика развивается быстрыми темпами. Но усложнение методов анализа данных, разработка новых или различных вариаций уже известных процедур обработки многомерных данных приводит к тому, что многие химики перестают воспринимать их и использовать в своей практике, несмотря на то, что все эти сложные методы в той или иной степени реализованы в компьютерных программах. Непонимание основных, базовых методов анализа данных ведет к непониманию любых других процедур, построенных на их основе. Данный курс является вводным к хемометрике.

Наиболее часто хемометрика используется химиками-аналитиками. Большое количество примеров из литературы относится именно к аналитической химии. Хотя само понятие «хемометрика» не имеет специального предназначения для какого-либо одного из направлений экспериментальной химии. Например, в химии твердого тела, кроме химического состава вещества, важно знать кристаллическую структуру реагирующих веществ, количественно определять фазовый состав. Протекание химических реакций с участием твердых тел трудно изучать без знания кристаллической структуры исходной фазы и продукта. Если одно и то же вещество может находиться в различных полиморфных модификациях, то определить содержание того или иного полиморфа в смеси возможно только с помощью специальных физических методов, в первую очередь дифракционных. Химический анализ здесь не поможет, хотя подходы к построению градуировочных зависимостей одинаковые и в химическом анализе, и в рентгенофазовом анализе.

Точного определения, что такое хемометрика, нет. В разных учебниках предлагается немного разная трактовка этого понятия. На сайте Российского хемометрического общества дается следующее определение: хемометрика – это научная дисциплина, находящаяся на стыке химии и математики, предметом которой являются математические методы изучения химических явлений. Другое распространенное определение приводится в учебнике Массарта с соавторами: хемометрика – это химическая дисциплина, которая использует математические, статистические и другие методы, включая формальную логику, во-первых, для конструирования и выбора оптимальных измерительных процедур и экспериментов, во-вторых, для извлечения наиболее важной и достоверной химической

информации при анализе химических данных. Нетрудно заметить, что первое определение является более широким и менее понятным. Нужно ли относить к хемометрике различные численные модели, широко используемые в теоретической химии? Во втором определении напрямую сказано, что речь идет об экспериментах. На наш взгляд, это более правильно.

Собственно, суть любого эксперимента – сконструировать оптимальную измерительную процедуру и попытаться получить как можно больше информации из результатов. Можно выделить три основных этапа измерительной процедуры: 1) разработка метода (выбор метода и оптимизация); 2) проведение измерения (подготовка образца и сбор данных); 3) интерпретация данных (первичная обработка данных, перевод «сырых» данных в химическую информацию и переход от химической информации к потребительской). Хемометрика касается 1-го и 3-го этапа.

Трудно представить себе развитие науки без проведения экспериментов, получения достоверной информации и правильной интерпретации данных. Анализ данных необходим в любой области науки. Можно выделить некие общие принципы анализа данных, общие методы обработки информации. Часто эти методы основаны на статистических процедурах, так как экспериментальные данные, как правило, включают случайную компоненту. При проведении измерений практически невозможно учесть и зафиксировать все факторы, влияющие на измеряемую величину, что приводит к случайному разбросу данных вокруг истинного значения.

Практическая хемометрика – это компьютеризация, каждый шаг предполагает использование программного обеспечения. Удобным средством работы с большими массивами численных данных являются электронные таблицы. Сложные методы многомерной статистики реализованы в различных специализированных программах, например таких как Statistica.

Цель настоящего курса – знакомство с наиболее распространенными математическими методами, в первую очередь со статистическими, в приложении к экспериментальным данным в различных областях химии, овладение некоторыми практическими навыками при обработке экспериментальных данных. Надеемся, что знания, полученные в рамках этого курса, помогут в дальнейшем будущим научным сотрудникам без проблем разбираться в сложных многомерных хемометрических процедурах и выбирать оптимальные процедуры извлечения химической информации из экспериментальных данных.

Предполагается, что студенты уже знакомы с теорией вероятности и математической статистикой. Поэтому в пособии дана только краткая сводка некоторых важных для хемометрики понятий и определений. Предполагается также, что студенты знакомы с основами информатики. Приведенные в пособии примеры заданий рассчитаны на работу в электронных таблицах.

Немного о терминах… В хемометрике встретились два общепринятых термина, которые звучат одинаково, а обозначают совершенно разные вещи. С одной стороны, слово «анализ» употребляется в смысле «химического анализа», а с другой стороны, может подразумевать и «прикладной анализ случайных данных». В дальнейшем мы будем использовать слово «анализ» во втором смысле, если не указано прямо, что это химический анализ.

В конце пособия приведен список рекомендуемой литературы. Основным является ставший уже классическим учебник Массарта с соавторами (Massart et al., 1988). Большая библиография по хемометрике содержится в обзорах Родионовой и Померанцева (2006). Со времени издания первого переводного учебника по хемометрике (Шараф, 1989) другого, более современного учебника на русском языке пока что нет. Этот учебник рекомендуется как дополнительный, так как логика построения курса в нем отличается от той, что представлена в нашем пособии. Под редакцией Массарта с другой группой соавторов в 1998 г. был выпущен более полный учебник в двух томах. Остальные книги, приведенные в списке, рекомендованы в качестве дополнительной литературы. Часть материала из них вошла в переработанном и адаптированном виде в настоящее учебное пособие.

Глава 1
Основные понятия и определения

1.1. Измерения в химии

Измерения бывают прямыми и косвенными. Химия, как правило, имеет дело с косвенными измерениями. Хемометрика зародилась в недрах аналитической химии. Количество вещества или количество молекул того или иного сорта не поддается прямому измерению. Любой химик-аналитик при определении концентаций компонентов сталкивается с проблемой интерпретации экспериментально полученных результатов, извлечения химической информации. Эталона моля нет. В качестве меры используется образец сравнения (состав достаточно надежно известен) или стандартный образец (состав юридически удостоверен). Полученный экспериментально результат всегда имеет некоторую неопределенность. Эта неопределенность может быть объяснена разными причинами. Ошибку может вносить и измерительный прибор, и образец сравнения, и методика обработки и интерпретации данных, и отбор проб для анализа. Учесть и зафиксировать все факторы в реальном эксперименте практически невозможно. Если мы хотим получить достоверный результат, то экспериментально полученные данные необходимо обрабатывать с учетом вероятностного характера измеренных величин, т. е. требуется статистическая обработка. И такое положение дел не только в аналитической химии, но и в любой другой области химии при анализе экспериментальных данных.

В химии твердого тела широко используются различные физические методы: дифракционные, спектроскопические, микроскопические, термоаналитические. Достоверность полученной в экспериментах информации напрямую зависит от правильного выбора способа обработки данных и правильной интерпретации косвенных измерений.

Экспериментальные данные являются основным объектом, с которым работает хемометрика. Простейший случай – одно значение какого-либо свойства, относящееся к одному объекту. Если мы рассматриваем множество объектов и характеризуем их некоторым набором свойств, то данные можно представить в виде матрицы, каждая строка которой относится к одному объекту, а столбец – к одному свойству объекта (табл. 1). Следующая ступень усложнения данных – набор значений, относящихся к одному объекту (например, хроматограмма, дифрактограмма, спектр). В этом случае на первой ступени обработки данные можно также представить в виде матрицы, только в качестве различных признаков объектов мы рассматриваем свойство, измеренное в различных условиях (угол дифракции на дифрактограммах или волновое число в спектрах и т. д.). Ситуация еще больше усложняется, если изучается развитие процесса во времени.

В некоторых случаях для удобства обработки образец, меняющийся во времени, можно рассматривать как серию образцов, и экспериментальные данные опять же можно представить в виде матрицы, где каждая строка характеризует один и тот же образец, но в разные периоды времени. Вместо фактора времени можно рассматривать и другие внешние факторы (температура, давление и т. д.).

1.2. Признаки и шкалы

Любой объект имеет бесконечное множество свойств или признаков. Если мы выбрали определенный набор измеряемых свойств или признаков, то исходные данные можно представить в виде таблицы (см. табл. 1), или в виде матрицы $\mathbf{X} = \{x_{ij}\}$, где i – номер объекта, j – номер признака, $i = 1, \ldots, n$; $j = 1, \ldots, m$.

Таблица 1

Исходные «сырые» данные

	Признак 1	Признак 2	…….	Признак m
Объект 1	x_{11}	x_{12}	…	x_{1m}
Объект 2	x_{21}	x_{22}	…	x_{2m}
…….	…	…	…	…
Объект n	x_{n1}	x_{n2}	…	x_{nm}

Какие именно признаки нас интересуют, зависит от целей и задач конкретного исследования. Выбор конечного набора признаков – не такая уж простая задача. При многомерной постановке задачи часть признаков может оказаться неинформативными, их отбрасывание не повлияет на количество получаемой информации. Но может оказаться и наоборот, мы не включим в рассмотрение важное свойство объекта, которое влияет на изменение остальных признаков. Вопрос об информативности признака можно решить и после первичного сбора данных. При разработке конкретной методики измерений проверка признаков на информативность является одним из первых этапов оптимизации.

Разработка измерительной процедуры неразрывно связана с выбором шкалы для измерения того или иного свойства. Допустим, набор интересующих нас признаков определен, дальше нужно решить, как их измерить, т. е. как выбрать шкалы.

Шкала – инструмент измерения. Если определить измерение как процедуру приписывания чисел измеряемым объектам в соответствии с тем или иным правилом, то в таком определении обнаруживается ряд недо-

статков. С одной стороны, приписывать можно не только числа (узость определения), а с другой стороны, какие именно правила имеются в виду – это самое интересное (излишняя широта определения). Строгое математическое определение шкалы можно дать в терминах теории множеств через изоморфное отображение эмпирического множества с заданными на нем отношениями на числовое (вместо чисел можно задать и другое удобное для измерений множество, например, множество геометрических объектов). Мы не будем здесь подробно знакомиться с языком теории множеств, отметим только, что в самом общем случае измерение может быть не только количественным. Роль количественных измерений в науке безусловно велика. Но существуют и ситуации, когда строгие количественные измерения невозможны. Например, шкала твердости минералов не является количественной, так как не определена пока еще масштабная единица твердости.

В теории измерений *допустимыми преобразованиями шкал* называют такие функции, которые отображают шкалу на саму себя. По допустимым преобразованиям шкал принято выделять пять разных типов: номинальная или классификационная шкала; порядковая или ранговая шкала; шкала интервалов; шкала отношений; абсолютная шкала. Типы шкал перечислены в порядке возрастания информативности. Рассмотрим основные свойства этих шкал.

Два первых типа шкал не являются количественными.

1. *Номинальная* шкала, или шкала *наименований*. Иногда ее называют *классификационной* шкалой. Наименее информативная. Каждому объекту приписывается ярлык (принадлежность к определенной группе или классу). Примеры: структурный тип кристаллического состояния вещества, тип химической связи. В этой шкале объекты различаются по проявлению свойства, но не различаются по уровню проявления свойства. Допустимым преобразованием шкалы является любое взаимооднозначное соответствие. В результате такого рода преобразований объект не должен переходить в другой класс, а вот название ярлыка можно менять (обозначение числами, символами, словами и т. д.). Любая классификация – это измерение в шкале наименований. Эта шкала не относится к количественным, соответственно для нее не применимы никакие арифметические действия.

2. *Порядковая* шкала, или *ранговая*. В основе измерения лежит процедура ранжирования. Примеры порядковых шкал: школьная оценка, шкала твердости минералов, шкала Рихтера (для оценки силы землетрясения), сортность, ряд напряженности металлов. Все объекты можно сравнить по уровню проявления свойства, но нельзя определить величину различия. Эта шкала не имеет эталона (масштабной единицы) и нет абсолютного нуля. Допустимыми являются строго возрастающие преобразования, т. е. функции преобразования f такие, что $x > y \Leftrightarrow f(x) > f(y)$. Эта шкала

также не является количественной, следовательно, арифметические действия в ней недопустимы. Распространенная ошибка – нахождение среднего арифметического в порядковой шкале.

Все остальные типы шкал относятся к количественным.

3. *Интервальная* шкала, или шкала интервалов. В этой шкале можно определить величину проявления свойства, «на сколько» один объект отличается от другого, но нельзя определить уровень исчезновения свойства. Есть масштабная единица, но нет абсолютного нуля. Допустимыми являются линейные преобразования вида $f(x) = kx + b, \quad (k > 0)$. Выбор нуля произволен. Примеры: измерение температуры в шкале Цельсия или Фаренгейта, измерение потенциальной энергии, время по календарю.

4. Шкала *отношений*. Можно определить отношение между уровнями проявления свойства, «во сколько раз» один объект отличается по свойству от другого. Есть масштабная единица и есть абсолютный нуль или полное исчезновение свойства. Например, масса, длина, объем, заряд, скорость и т. д. Допустимыми являются преобразования подобия $f(x) = kx, \quad (k > 0)$. Еще один пример – шкала температур Кельвина.

5. *Абсолютная* шкала. Это в некотором смысле вырожденный вариант шкалы отношений. Определяется накопление свойства. Существует естественная масштабная единица. Например, результат счета, порядковый номер в таблице Менделеева. Допустимым является только тождественное преобразование $f(x) \equiv x$.

Развитие наших знаний о признаке ведет к увеличению информативности шкалы. Вопрос о том, какую шкалу выбрать для измерения, не всегда связан со свойством. Часто выбор шкалы зависит от уровня наших знаний об измеряемом свойстве. Хорошим примером в этом смысле является температура. Первоначально человек научился определять температуру одного тела по сравнению с другим (холоднее, горячее). Это измерение в порядковой шкале. С появлением знания о том, что тела при нагревании расширяются, были сконструированы интервальные шкалы: Цельсия, Фаренгейта. Если одно тело имеет температуру 4°С, а другое – 2°С, абсолютно бессмысленным будет высказывание, что первое тело в два раза горячее второго. С развитием наших представлениий о температуре появились понятие асолютного нуля температуры и шкала Кельвина, которая является шкалой отношений. И если для какого-либо свойства еще не сконструированы количественные шкалы, то это может означать, что наши знания об этом свойстве могут быть не достаточно глубоки. Заметим, что всегда можно перейти от более информативной к менее информативной шкале, но не наоборот.

1.3. Погрешности измерений и причины возникновения ошибок

Любое измерение осуществляется с ошибкой или погрешностью. Невозможно учесть все источники вариации измеряемой величины. Не вдаваясь в природу возникающих погрешностей, можно провести классификацию погрешностей по источникам возникновения и по специфическому воздействию на измерительную систему.

По определению, *погрешность* E – это разность показываемого значения x_a и истинного x (т. е. измеренного идеальным прибором в идеальном эксперименте): $E = x_a - x$. Коррекцией или поправкой B называют величину, равную погрешности, но с обратным знаком: $B = x - x_a$. Действительное значение равно сумме измеренного значения и поправки.

Измерение должно быть представительным (репрезентативным). Например, при измерении температуры помещения измеренная в одной точке локальная температура объявляется температурой помещения. Подобная ситуация возникает почти всегда, когда с помощью малого числа датчиков необходимо измерить среднее значение поля величин (температура, концентрация и т. д.). Другой пример – ошибки при измерении с отбором пробы. В этих случаях при отборе пробы необходимо выполнение некоторых статистических условий, таких как однородность состава, независимость отбора (случайный выбор последовательности) и т. д. На практике ошибки представительности возникают часто, выявляются с трудом, могут появиться даже при использовании высококачественных измерительных приборов, при этом нередко ошибки представительности имеют значительную величину и могут многократно превышать остальные погрешности.

Погрешности можно классифицировать по источникам возникновения и по специфическому воздействию на измерительную систему. Для наглядности измерительную систему можно представить в виде обобщенной блок-схемы (рис. 1).

Обратное воздействие сильно проявляется, например, при зондовых измерениях параметров потока. Другой пример сильного воздействия – измерение температуры жидкости с помощью термометра, если объем жидкости слишком мал, а разница температур жидкости и термометра велика.

Действие *аддитивных* внешних помех накладывается на сигнал. При этом погрешность не зависит от значения измеряемой величины. Например, смещение нуля прибора.

Мультипликативной называется помеха (например, статическое давление, температура окружающей среды, поле тяготения), которая влияет на передаточную характеристику или изменяет ее. Результирующая погрешность зависит от измеренной величины. Например, односторонний нагрев

рычажных весов солнечными лучами, соотношение плеч рычагов изменяется. Чем больше сам вес образца, тем больше погрешность измерения на таких весах. Если температура движущейся жидкости измеряется термометром, то скорость потока является мультипликативной внешней помехой.

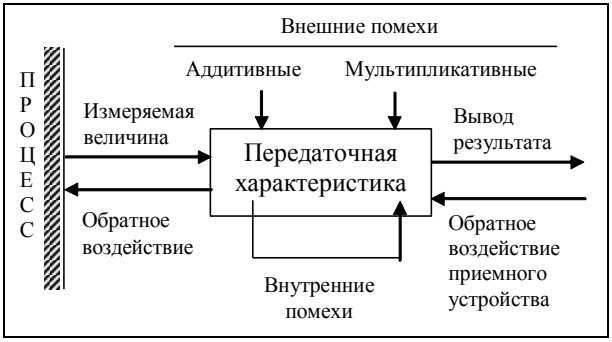

Рис. 1. Блок-схема измерительной системы

Внутренние помехи возникают из-за внутриприборных эффектов. Например, люфт при механическом преобразовании, трение и т. д. Как правило, внутренние помехи приводят к нелинейностям и, как следствие, к погрешностям.

Погрешности, связанные с процессом измерения, можно разделить на систематические и случайные, статические и динамические. Особо следует выделить влияние внешних условий применения измерительного устройства. Часто требуется соблюдение специальных предписаний при применении приборов. Ограничения устанавливаются исходя из допустимой погрешности.

Условия применения измерительного устройства считаются известными, если наряду с процессом известны и наиболее существенные влияющие внешние факторы. Однако всегда остается большое число менее значительных факторов. Погрешность, вызванная этими неучтенными факторами, является *случайной*. Заранее неизвестны ни ее абсолютная величина, ни ее знак. Точность, с которой может быть определено (с заданной вероятностью) указанное ожидаемое значение, можно оценить статистически. Эту точность не следует путать с правильностью измерения (при правильном измерении мы должны получить истинное значение).

Если основные влияющие величины известны, их стремятся поддерживать постоянными, однако они могут отклоняться от тех значений, которые были приняты при градуировке прибора. Кроме того, воспроизведение

образцовых мер никогда не бывает точным. Эти погрешности нельзя исключить повторением измерения. Они являются *систематическими* и отличаются своей воспроизводимостью. Их абсолютная величина и знак остаются неизменными при заданных условиях.

Вопрос о различии случайной и систематической погрешностей решается в зависимости от требуемой точности и от выбранной измерительной процедуры.

Передаточную характеристику можно представить в виде математического выражения, описывающего взаимосвязь входной и выходной величин. Передаточная характеристика линейного прибора для определения величин, не меняющихся во времени, есть константа. Существуют приборы и с нелинейными характеристиками, которые можно описать алгебраическими уравнениями. В этих случаях погрешность зависит только от размера измеряемого значения и не является функцией времени. Это *статические* погрешности измерения.

При измерении изменяющейся во времени величины связь между входной и выходной величинами описывается дифференциальными уравнениями. При этом погрешность зависит не только от размера измеряемой величины, но и от характера изменения ее во времени. Это *динамические* погрешности. Вообще говоря, статическая погрешность – частный случай динамической.

Перечислим погрешности, связанные с обработкой измереннных значений: погрешность отсчета и квантования; временная дискретизация; погрешность, обусловленная неадекватностью принятой гипотезы; погрешность результата измерения при расчете по нескольким измеренным значениям.

Ошибка *отсчета* в большей степени определяется видом устройства вывода. Ошибка *квантования* связана с цифровой записью величин в виде чисел. Временную дискретизацию особо следует учитывать при анализе сигналов, связанных с исследованиями динамических погрешностей.

В основе статистических методов обработки в общем случае лежат некоторые гипотезы, например, предположение о нормальном распределении случайной погрешности. В идеале при разработке новых методик измерения всегда надо проверять принятые гипотезы и предположения.

Например, для определения статической характеристики прибора построили график зависимости измеренных значений от истинных $x_a(x)$. Предположим, что связь между истинным значением и измеренным линейная. Метод обработки состоит в расчете прямой по методу наименьших квадратов (МНК). Этот метод основан на гипотезе нормального распределения относительно истинного значения и независимости распределения от величины измеряемого значения. Такое допущение оправдано в тех случаях, когда случайная погрешность обусловлена аддитивными, а не

мультипликативными помехами. Предложенный метод обработки может обусловить внесение двух дополнительных погрешностей: а) если статическая характеристика отличается от линейной, то добавляется систематическая погрешность; б) если рассеяние зависит от измеряемого значения (мультипликативная помеха), то рассчитанный угол градуировочной прямой является сомнительным. Такого рода погрешности обусловлены *неадекватностью принятой гипотезы*. Для улучшения результатов следует вводить весовые коэффициенты в МНК.

Общая блок-схема классификации погрешностей представлена на рис. 2. При расчете по нескольким измеренным значениям следует уделять особое внимание *распространению* погрешностей исходных данных на конечный результат. Систематическим и случайным погрешностям соответствуют разные законы распространения (см. п. 3.2).

Рис. 2. Классификация погрешностей

Случайная ошибка не может быть предсказана заранее, неизвестны ни ее абсолютная величина, ни ее знак. Однако можно высказать суждение о ее статистических свойствах. Систематическую ошибку нельзя исключить повторением измерения, она отличается своей воспроизводимостью. Ее абсолютная величина и знак остаются неизменными при заданных условиях эксперимента.

1.4. Случайные величины. Параметры случайных величин

В реальном эксперименте невозможно учесть все источники вариации признаков, влияющие на конечный результат. Невозможно провести два измерения одинаково, строго в одних и тех же условиях. Измеряемая величина всегда имеет некоторую неопределенность, т. е. является случайной величиной.

В самом общем случае растянутое по времени наблюдение какого-либо явления можно назвать процессом. Если каждое наблюдение дает невоспроизводимый результат, то мы имеем дело со случайным процессом. Любое наблюдение дает только один вариант из множества возможных. Случайные процессы делятся на стационарные и нестационарные. В свою очередь стационарные процессы делятся на эргодические и неэргодические. Чтобы разобраться с этой классификацией, необходимо ввести некоторые определения. Рассмотрим совокупность (ансамбль) наблюдений или реализаций, характеризующих случайный процесс (рис. 3).

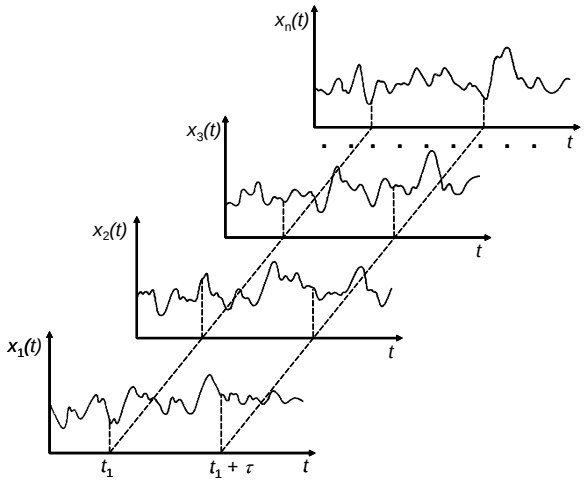

Рис. 3. Ансамбль реализаций, задающих случайный процесс

Среднее значение или математическое ожидание (первый момент) этого случайного процесса в момент времени t_1 можно вычислить, усреднив все мгновенные значения реализаций ансамбля в этот момент времени:

$$\mu_x(t_1) = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} x_k(t_1).$$

Аналогичным образом *ковариацию* (смешанный момент) значений случайного процесса в два различных момента времени вычисляют путем усреднения по ансамблю произведений мгновенных значений в моменты времени t_1 и $t_1 + \tau$:

$$R_{xx}(t_1, t_1 + \tau) = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} x_k(t_1) x_k(t_1 + \tau).$$

В общем случае, когда $\mu_x(t_1)$ и $R_{xx}(t_1, t_1 + \tau)$ зависят от момента времени, случайный процесс называется *нестационарным*. В том частном случае, когда $\mu_x(t_1)$ и $R_{xx}(t_1, t_1 + \tau)$ не зависят от момента времени t_1, инвариантны по времени и все остальные моменты, случайный процесс называется *стационарным*.

Можно вычислить характеристики случайного процесса, усреднив не по ансамблю, а по времени. Среднее значение и ковариация, вычисленные по k-й реализации, равны

$$\mu_x(k) = \lim_{T\to\infty} \frac{1}{T} \int_0^T x_k(t)dt, \quad R_{xx}(\tau, k) = \lim_{T\to\infty} \frac{1}{T} \int_0^T x_k(t) x_k(t + \tau)dt.$$

Если случайный процесс стационарен, а $\mu_x(k)$ и $R_{xx}(\tau,k)$, вычисленные по различным реализациям, совпадают, то такой процесс называется *эргодическим*.

Мы, в основном, будем иметь дело со стационарными, эргодическими процессами. В этом случае можно перейти от рассмотрения случайных процессов к рассмотрению случайных величин, убрав зависимость от времени. Оценку параметров такой случайной величины можно делать по выборке, в которой каждое измерение можно проводить в разные моменты времени и брать разные части объекта или образца. Некоторые методы разрушают образец в процессе измерения, для каждого следующего измерения берется уже другой образец. В этом случае необходимо дополнительно убедиться в однородности или гомогенности исходного объекта.

Вернемся к основным понятиям математической статистики. Если переменная может принимать конечное число значений, она называется *дискретной*. Дискретной также называется переменная, принимающая

бесконечное число значений, но это множество является счетной последовательностью. В других случаях переменная *непрерывная*. Множество значений, которое может принимать случайная переменная, называют *генеральной совокупностью*. Основными статистическими характеристиками, имеющими важное значение для описания свойств отдельных случайных величин, являются: среднее значение; дисперсия; плотность вероятности. Среднее значение характеризует центр рассеяния случайной величины, дисперсия – величину рассеяния данных или разброс. Плотность вероятности $p(x)$ задает скорость изменения вероятности $P(x)$ в зависимости от значения переменной. Основные формулы для одномерных случайных величин приведены в табл. 2.

Таблица 2

Одномерные случайные величины и моменты

Дискретные	Непрерывные
$\sum p(x_i) = 1;$ $p(x) \geq 0$	$\int_{-\infty}^{+\infty} p(x)dx = 1;\ p(x) \geq 0$ $\text{Prob}(x_a < x \leq x_b) = \int_{x_a}^{x_b} p(x)dx;$ $P(x) = \int_{-\infty}^{x} p(\xi)d\xi;\quad \frac{dP(x)}{dx} = p(x)$
Первый момент (математическое ожидание)	
$\mu = E(x) = \sum_i p_i x_i$	$\mu = E(x) = \int_{-\infty}^{+\infty} xp(x)dx$
Второй момент	
$E(x^2) = \sum_i p_i x_i^2$	$E(x^2) = \int_{-\infty}^{+\infty} x^2 p(x)dx$
Дисперсия	
$\sigma^2 = E\left((x-\mu)^2\right) = \sum_i p_i (x_i - \mu)^2$	$\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 p(x)dx$

В случае двумерных случайных величин также должно выполняться условие нормировки:

$$\int_{-\infty}^{+\infty}\int p(x,y)dxdy = 1; \quad p(x,y) \geq 0.$$

Интегральная и дифференциальная функции распределений для двумерной величины имеют вид

$$P(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} p(\xi,\eta)d\xi d\eta; \quad \frac{\partial}{\partial y}\left[\frac{\partial P(x,y)}{\partial x}\right] = p(x,y).$$

Если обе переменные *x* и *y* статистически независимы, то

$$p(x,y) = p(x)p(y).$$

В двумерном случае ковариация определяется как

$$c_{xy} = \int_{-\infty}^{+\infty}\int (x-\mu)(y-\mu)p(x,y)dxdy.$$

Тогда корреляция

$$\rho_{xy} = \frac{c_{xy}}{\sigma_x \sigma_y}; \quad -1 \leq \rho \leq 1.$$

Если какая-нибудь мера вычислена для всей генеральной совокупности, то она называется *параметром*. Мера, вычисленная по выборке, является *оценкой параметра*.

1.5. Определение основных статистик по выборке, точечные оценки параметров случайных величин

Термин «выборка» определяется как некоторое количество независимых измерений одного и того же свойства. Нужно понимать, что измеряемый по выборке параметр является оценкой, основанной на сравнительно небольшом числе наблюдений. Если бы мы задались целью провести «идеальный» эксперимент, то необходимо было бы в одних и тех же условиях на одном и том же образце провести одновременно бесконечно много измерений. Это физически невозможно. В общем случае любую функцию от всех значений в выборке $Q = f(x_1, x_2, ..., x_n)$ называют *статистикой* (не путать с общепринятым значением этого термина как названия дисциплины). Количество значений в выборке называют *объемом*.

Центр рассеяния случайной величины или оценка математического ожидания характеризуется мерами центральной тенденции. Чаще всего используется *среднее арифметическое*:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

где n – объем выборки. Кроме среднего арифметического, можно определить и другие меры центральной тенденции. *Мода* – значение, которое встречается наиболее часто в выборке. Если величина является непрерывной, то моду можно определить по частотному распределению (см. 3.1). *Медиана* – значение, которое делит группу наблюдений пополам, половина всех измерений лежит выше медианы, половина – ниже. Среднее арифметическое допустимо находить только для величин, измеренных в количественных шкалах, так как только на количественных шкалах определена операция сложения. Медиана допустима еще и в порядковых шкалах, а мода – во всех типах шкал, включая и шкалу наименований. В некоторых особых случаях для количественных шкал в качестве меры центральной тенденции выбирают среднее геометрическое, среднее гармоническое.

Величину рассеяния данных также можно характеризовать разными мерами изменчивости. Например, *размах* – разность максимального и минимального значений в выборке; полумеждуквартильный размах – половина расстояния между третьим и первым квартилями (квартили делят группу наблюдений на 4 части, второй квартиль – это медиана). Редко, но используется *среднее отклонение*, в этом случае усредняется модуль отклонения от среднего:

$$MD = \frac{1}{n}\sum_{i=1}^{n} \left| x_i - \bar{x} \right|.$$

Важной характеристикой рассеяния данных является среднеквадратичное отклонение или стандартное отклонение, определенное по выборке. *Стандартным отклонением* называется корень квадратный из дисперсии. Сводка некоторых параметров случайных величин и их оценок с формулами расчета приведена в табл. 3. Оценка коэффициента корреляции по формуле из таблицы называется *коэффициентом корреляции Пирсона*.

Выбор той или иной меры для оценки параметров генеральной совокупности связан со свойствами оценок. В табл. 3 для каждой оценки указаны кратко их свойства.

Рассмотрим некоторые свойства оценок.

Несмещенность. Если среднее выборочного распределения оценки равно величине оцениваемого параметра, то оценка называется несмещен-

ной. $\bar{x}$ – несмещенная оценка математического ожидания для любого распределения. Мода и медиана для асимметричных распределений являются смещенными оценками. s_x^2 – несмещенная оценка дисперсии σ_x^2, а вот $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ является смещенной оценкой дисперсии. Среднеквадратичное отклонение по выборке s_x – смещенная отрицательно оценка σ. Коэффициент корреляции Пирсона r_{xy} также является смещенной отрицательно оценкой ρ_{xy}.

Таблица 3

Параметры случайных величин и оценки параметров

Параметр	*Оценка*	*Свойства оценки*
Среднее μ (математическое ожидание)	Мода	Смещенная, состоятельная;
	Медиана	Смещенная, состоятельная;
	$\bar{x}$	Несмещенная, наиболее эффективная
Дисперсия σ_x^2	$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$	Несмещенная
Стандартное отклонение σ_x	$s_x = \sqrt{s_x^2}$	Смещенная отрицательно, состоятельная
Ковариация c_{xy}	$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$	Несмещенная
Корреляция ρ_{xy}	$r_{xy} = \frac{s_{xy}}{s_x s_y}$ (коэффициент корреляции Пирсона)	Смещенная отрицательно, состоятельная

Состоятельность. Состоятельная оценка, даже если она смещенная, при постоянном увеличении объема выборки приближается к значению параметра, который она оценивает. Мода, медиана, стандартное отклонение, коэффициент корреляции Пирсона являются состоятельными оценками.

Относительная эффективность оценок. Характеризует, как сильно изменяется оценка от выборки к выборке. Чем меньше дисперсия ошибки оценки, тем эффективнее оценка. Из трех мер центральной тенденции –

мода, медиана, среднее арифметическое – наиболее эффективной является среднее арифметическое.

Оценка параметра сама по себе является случайной величиной. Стандартное отклонение оценки характеризует случайную ошибку оценки. Каждая оценка имеет свое вероятностное распределение.

В табл. 4 даны формулы преобразований оценок среднего и дисперсии при преобразовании исходных переменных:

Таблица 4

Преобразования оценок

Преобразование	Среднее	Дисперсия
$y = x + b$	$y = x + b$	$s_y^2 = s_x^2$
$y = ax$	$y = ax$	$s_y^2 = a^2 s_x^2$
$y = x - x$	$y = 0$	$s_y^2 = s_x^2$
$y = \frac{x - x}{s_x}$	$y = 0$	$s_y^2 = 1$

Последняя строка в табл. 4 соответствует преобразованию, которое носит специальное название – *стандартизация*.

Для нахождения конкретного значения оценки используются все объекты (наблюдения) в выборке. Все вышеперечисленные оценки параметров относятся к описательным статистикам и являются *точечными*.

1.6. Некоторые виды распределений случайных величин

Распределения вероятности случайных величин бывают дискретными и непрерывными. Наиболее часто в прикладной статистике применяются нормальное (гауссово) распределение и основанные на нем распределения. Все эти распределения являются непрерывными. Но для начала рассмотрим два важных примера дискретных распределений: биномиальное и распределение Пуассона.

Биномиальное распределение. Наблюдения, которые могут выражаться в одной из двух возможностей (удача, неудача) с постоянной вероятностью, называют биномиальными. Пусть p – вероятность удачи при одном наблюдении, а q – вероятность неудачи (q =1 – p). Вероятность получения n_1 удач из n наблюдений определяется биномиальным распределением:

$$P(n_1) = \binom{n_1}{n} p^{n_1} q^{n-n_1}, \text{ где } \binom{n_1}{n} = \frac{n!}{n_1!(n-n_1)!}.$$

Математическое ожидание и дисперсия биномиального распределения вычисляются по формулам

$$\mu = np, \quad \sigma^2 = npq.$$

Распределение Пуассона описывает дискретную переменную, относящуюся к дискретным событиям в непрерывном интервале (например, в интервале времени). Распределение Пуассона является предельным случаем биномиального распределения, в котором n стремится к бесконечности, p стремится к 0, а np равно конечному числу λ. Вероятность получить результат n_1, $P(n_1)$, может быть вычислена по формуле

$$P(n_1) = \frac{\exp(-\lambda)\lambda^{n_1}}{n_1!}.$$

Среднее и дисперсия такого распределения равны λ. Распределение Пуассона используется, например, при описании результатов, полученных сцинтилляционными счетчиками.

Перейдем теперь к рассмотрению основного распределения в теории ошибок.

Нормальное распределение. Плотность распределения нормальной случайной величины задается соотношением

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right),$$

где μ – математическое ожидание или среднее, σ – стандартное отклонение. Обозначается оно как $N(\mu, \sigma^2)$. Нормальная интегральная функция распределения определяется соотношением

$$P(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx.$$

Стандартизованное или стандартное нормальное распределение $N(0, 1)$:

$$p(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}; \quad z = \frac{x-\mu_x}{\sigma}; \quad \mu_z = 0; \quad \sigma_z^2 = 1.$$

Нормальное распределение симметрично. Плотность вероятности нормального распределения унимодальна (имеет один максимум), монотонно изменяется по обе стороны моды. На рис. 4 изображена функция плотности вероятности стандартизованного нормального распределения.

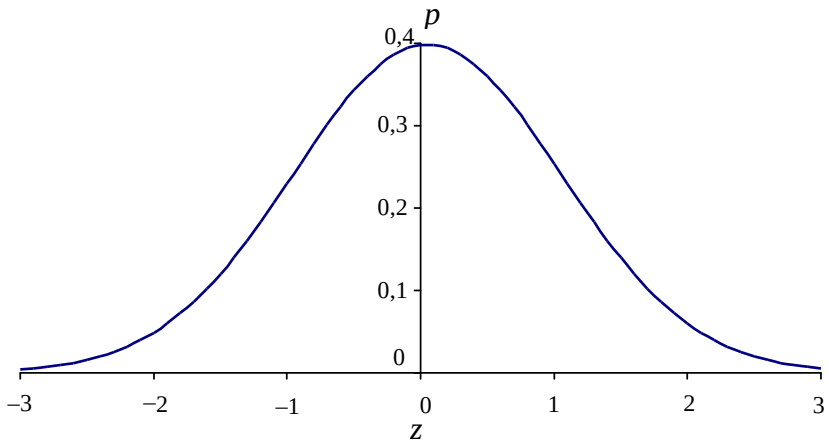

Рис. 4. Функция плотности вероятности стандартизованного нормального распределения ($\mu = 0$, $\sigma = 1$)

В электронных таблицах существуют функции прямого и обратного нормального распределения. В случае прямого нормального распределения в качестве аргументов задаются x, μ и σ^2, в результате выводится значение интегральной вероятности $P(x)$. В случае обратного нормального распределения в качестве аргументов задается вероятность α, μ и σ^2, а выводится значение x_α. Практически во всех книгах и справочниках по статистике приводятся таблицы стандартного нормального распределения. В электронных таблицах для стандартного нормального распределения существуют отдельные функции. В качестве дополнительного аргумента в электронных таблицах вводится логическая величина, позволяющая выводить либо площадь под кривой плотности вероятности до соответствующего предела, либо значение ординаты.

Для дальнейшего рассмотрения удобно ввести понятие процентной точки распределения. Значение z_α, удовлетворяющее уравнению

$$P(z_\alpha) = \int_{-\infty}^{z_\alpha} p(z)dz = \Pr ob[z \le z_\alpha] = 1 - \alpha,$$

называется *$100 \times \alpha$-процентной точкой* нормального распределения.

Нормальное распределение играет важную роль как в описательной статистике, так и в теории статистического вывода. Оно является отличной аппроксимацией распределений частот большого числа наблюдений при влиянии множества независимых факторов. Точного нормального распре-

деления в эксперименте получить нельзя, так как идеального, бесконечного эксперимента не может быть.

Широкое распространение нормального распределения связано не в последнюю очередь с тем, что существует так называемая центральная предельная теорема.

Рассмотрим сумму $y = x_1 + x_2 + \ldots + x_n$, где n случайных независимых переменных x_i со средними μ_i и дисперсией σ_i^2. *Центральная предельная теорема* утверждает, что при $n \to \infty$ распределение величины y приближается к нормальному распределению со средним и дисперсией

$$\mu_y = \sum_{i=1}^{n} \mu_i, \quad \sigma_y^2 = \sum_{i=1}^{n} \sigma_i^2.$$

Во многих практически зачимых случаях распределений случайной величины x, не являющихся нормальными, выводы центральной предельной теоремы остаются справедливыми.

Из этой теоремы следует, в частности, что распределение выборочного среднего $\bar{x}$ стремится к нормальному с математическим ожиданием μ_x и дисперсией σ_x^2/n при увеличении объема выборки. *Стандартной ошибкой среднего* называется стандартное отклонение оценки $\bar{x}$ или

$$\sigma_{\bar{x}} = \sigma_x / \sqrt{n}.$$

Во многих случаях выборочное распределение среднего можно считать нормальным уже при $N > 4$, а при $N > 10$ приближение будет очень хорошим. При этом сама величина x необязательно должна быть распределена нормально.

При значениях $\lambda > 10$ распределение Пуассона приближается к нормальному. Биномиальное распределение также имеет тенденцию приближаться к нормальному при $np > 5$ и $nq > 5$.

Теоретически выборочные распределения можно вывести и для других оценок или выборочных параметров.

Хи-квадрат-распределение. Пусть $z_1, z_2, \ldots, z_k$ есть k независимых случайных величин, каждая из которых имеет нормальное распределение с нулевым средним и единичной дисперсией. Определим новую случайную величину вида:

$$\chi^2 = z_1^2 + z_2^2 + z_3^2 + \ldots + z_k^2$$

Эта случайная величина подчиняется так называемому хи-квадрат-распределению с k степенями свободы. Функция плотности вероятности

хи-квадрат-распределения лежит в положительной области и асимметрична (рис. 5). При $k \to \infty$ распределение стремится к нормальному.

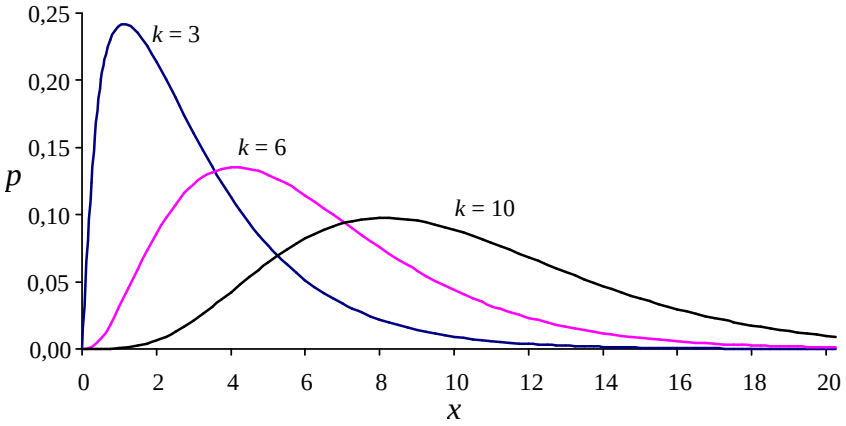

Рис. 5. Функция плотности вероятности хи-квадрат-распределения

Это распределение описывает, например, такую случайную величину, как оценка дисперсии s^2. Выборочное распределение выборочной дисперсии определяется из соотношения

$$\frac{ns^2}{\sigma_x^2} = \chi_k^2, \quad k = N - 1,$$

где N – объем выборки.

Распределение Стьюдента или t-распределение. Это распределение вывел английский статистик В. Госсет и опубликовал его под псевдонимом «Стьюдент». В то время он работал на пивоварне Гиннеса и занимался статистическим исследованием качества пива.

Случайная величина вида

$$t_k = \frac{z}{\sqrt{\chi_k^2 / k}}$$

имеет распределение Стьюдента с k степенями свободы. z – нормально распределенная стандартизованная случайная величина, χ_k^2 подчиняется хи-квадрат-распределению с k степенями свободы. Распределение Стьюдента унимодально, симметрично, среднее равно 0, дисперсия равна $k/(k-2)$. Графики распределения Стьюдента для разных степеней свобо-

ды приведены на рис. 6. При $k \to \infty$ t-распределение приближается к нормальному. На рис. 6 видно, насколько быстро распределение Стьюдента приближается к нормальному при росте числа степеней свободы k.

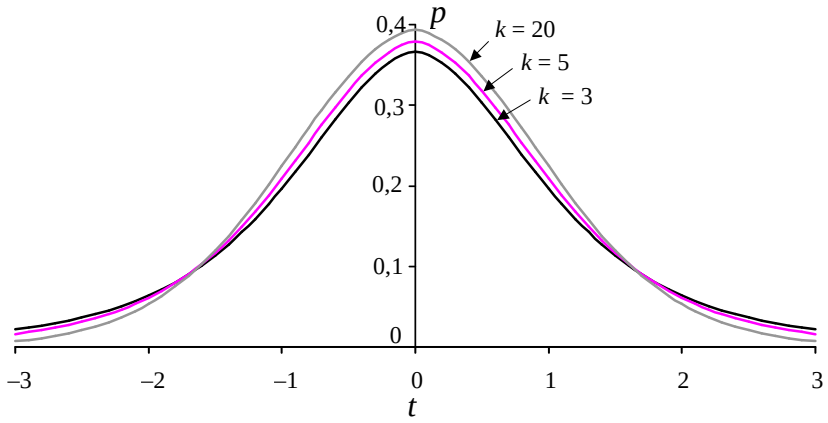

Рис. 6. Распределение Стьюдента

Выборочное распределение оценки математического ожидания при неизвестной дисперсии является распределением Стьюдента и задается соотношением

$$\frac{(\bar{x} - \mu_x)\sqrt{N}}{s} = t_k, \quad k = N - 1.$$

Распределение Стьюдента в основном используется для малых объемов выборок. Считается, что при $N > 30$ вместо t-распределения можно использовать нормальное распределение.

F-распределение. Определим случайную величину в виде соотношения

$$F_{k_1 k_2} = \frac{\chi^2_{k_1} / k_1}{\chi^2_{k_2} / k_2},$$

где в числителе и знаменателе стоят случайные величины, распределенные по хи-квадрат распределению со своими степенями свободы. Эта случайная величина подчиняется F-распределению с k_1 и k_2 степенями свободы. F-распределение унимодально и асимметрично, определено только при $x \geq 0$ (рис. 7). Отметим, что статистика t_n^2 имеет F-распределение с $k_1 = 1$ и $k_2 = n$ степенями свободы.

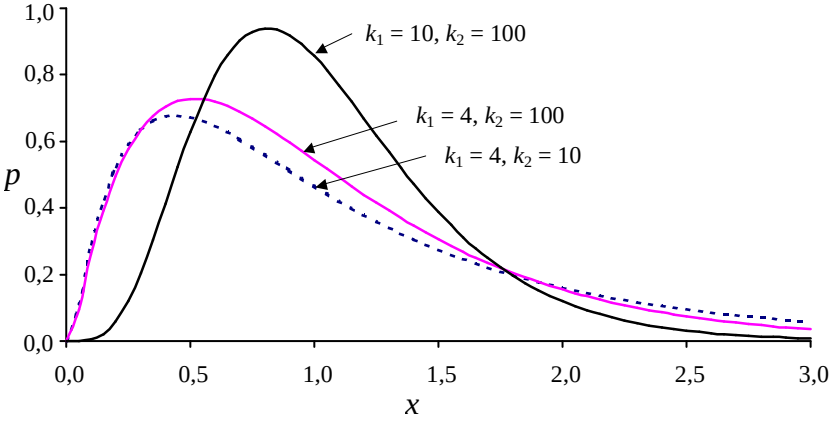

Рис. 7. *F*-распределение

F-распределение в теории статистического вывода используется для сравнения дисперсий. Отношение двух выборочных дисперсий подчиняется *F*-распределению и задается соотношением

$$\frac{s_x^2/\sigma_x^2}{s_y^2/\sigma_y^2} = F_{k_x k_y}, \quad k_x = N_x - 1, k_y = N_y - 1.$$

Заметим, что если выборки сделаны из одной и той же случайной величины $x = y$, то вышеприведенное соотношение превращается в следующее:

$$\frac{s_1^2}{s_2^2} = F_{k_1 k_2}, \quad k_1 = N_1 - 1, k_2 = N_2 - 1.$$

1.7. Интервальное оценивание, доверительные интервалы

Оценки параметров случайных величин бывают точечные и интервальные. Среднее, дисперсия, стандартное отклонение, коэффициент корреляции, рассчитанные по выборке, являются точечными оценками. Они не позволяют судить о степени близости выборочных значений к оцениваемому параметру. Более информативно строить интервал, который накрывает оцениваемый параметр с известной степенью достоверности, *доверительный интервал*. Степень достоверности или уровень доверия мы задаем сами. Если вероятность риска (ошибки) обозначить через α, то степень достоверности будет выражаться, как $1 - \alpha$. Смысл уровня риска или, как еще говорят, *уровня значимости* в том, что в $\alpha \times 100$ % выборок

наш построенный доверительный интервал не будет содержать истинного значения параметра. Доверительный интервал можно построить, если известно выборочное распределение рассматриваемой оценки.

Допустим, нам надо построить доверительный интервал для x, вычисленного по N независимым наблюдениям. Можно сделать следующее вероятностное утверждение:

$$\text{Prob}\left[z_{1-\alpha/2} < \frac{(x-\mu_x)\sqrt{N}}{\sigma_x} \le z_{\alpha/2} \right] = 1-\alpha,$$

где $z_{1-\alpha/2}$ и $z_{\alpha/2}$ – процентные точки стандартного нормального распределение $N(0, 1)$.

Уже сделанная оценка среднего либо попадет, либо не попадет в этот интервал. Если производится много выборок и для каждой из них вычисляется значение x, то можно ожидать, что участвующая в формуле величина будет попадать в указанный интервал с относительной частотой, примерно равной $1 - \alpha$. Доверительный интервал для математического ожидания μ_x можно построить по выборочному значению x:

$$\left[x - \frac{\sigma_x z_{\alpha/2}}{\sqrt{N}} \le \mu_x < x + \frac{\sigma_x z_{\alpha/2}}{\sqrt{N}} \right].$$

Если σ_x неизвестна, то доверительный интервал строится по распределению Стьюдента:

$$\left[x - \frac{s_x t_{n;\alpha/2}}{\sqrt{N}} \le \mu_x < x + \frac{s_x t_{n;\alpha/2}}{\sqrt{N}} \right], \quad n = N-1.$$

В формулах использованы свойства симметричности распределений:

$$z_{1-\alpha/2} = -z_{\alpha/2}; \quad t_{1-\alpha/2} = -t_{\alpha/2}.$$

Истинное значение попадает в указанный интервал с доверительной вероятностью $100 \times (1 - \alpha)$ %. Подобные утверждения можно сделать относительно любых оценок параметров, лишь бы были известны соответствующие выборочные распределения.

Доверительный интервал для оценки дисперсии можно построить, используя процентные точки хи-квадрат-распределения:

$$\left[\frac{ns^2}{\chi^2_{n,\alpha/2}} \le \sigma_x^2 < \frac{ns^2}{\chi^2_{n,1-\alpha/2}} \right], \quad n = N-1.$$

В этом случае необходимо найти две процентные точки, так как хи-квадрат-распределение асимметрично.

Чтобы построить доверительный интервал для оценки коэффициента корреляции r, используют дополнительное преобразование Фишера z_r, которое есть в статистических таблицах

$$z_r = \ln \sqrt{\frac{1+r}{1-r}}.$$

Величина z_r приближенно распределена нормально со средним z_ρ и дисперсией $1/(N-3)$. Тогда строят доверительный интервал для z_r по нормальному распределению, а потом делают обратное преобразование. Если доверительный интервал для r накрывает нулевое значение, то корреляция считается статистически не доказанной.

1.8. Теория статистического вывода, проверка гипотез

Теория статистического вывода основана на предположениях о случайном выборе из генеральной совокупности. Статистическая гипотеза – это утверждение относительно неизвестного параметра. Оценка параметра по выборке никогда не будет в точности равна истинному значению из-за выборочной изменчивости. Возникает вопрос, при каком отклонении выборочного значения от истинного это отличие можно приписать естественной статистической изменчивости. Ответ можно дать в статистических терминах, вычислив вероятность любого значимого отклонения.

При проверке любой статистической гипотезы решение никогда не принимается со стопроцентной уверенностью, всегда есть риск принятия неправильного решения.

Выделим основные этапы проверки статистической гипотезы.

Этап 1. Формулируется проверяемая гипотеза. Обычно стараются выдвигать так называемую *нуль-гипотезу*, которую обозначают H_0. Термин пришел из области философии. Доказательства собираются для аннулирования гипотезы. Считается, что гипотезу нельзя доказать конечным числом фактов, а вот опровергнуть можно одним-единственным фактом. Кроме того, выдвигают альтернативную гипотезу, которую обозначают H_1. При выдвижении альтернативной гипотезы возможны два варианта: односторонняя или двусторонняя гипотезы. Поясним на примере. Пусть относительно математического ожидания некоторой случайной величины выдвинута нуль-гипотеза

$$H_0: \mu = 0.$$

Тогда альтернативную гипотезу можно сформулировать двумя разными способами:

$H_1: \mu \neq 0$	$H_1: \mu > 0$
(*двусторонняя гипотеза*)	(*односторонняя гипотеза*)

Этап 2. Высказываются предположения, необходимые для определения выборочного распределения статистики, оценивающей параметр гипотезы, т. е. выбирается *критерий* проверки гипотезы. Берется выборочное распределение для случая, когда гипотеза верна. Для приведенного выше примера нуль-гипотезы и альтернативных гипотез можно выбрать *z*-критерий, имеющий нормальное распределение $N(0, 1)$, если объем выборки достаточно большой или известна дисперсия:

$$z = \frac{\bar{x} - \mu}{\sqrt{\sigma^2 / N}}.$$

Если дисперсия неизвестна и $N < 30$, то необходимо выбирать *t*-критерий, подчиняющийся распределению Стьюдента:

$$t = \frac{\bar{x} - \mu}{\sqrt{s^2 / N}}.$$

Этап 3. Принимается степень риска для неправильного вывода. Риск, представленный как вероятность, обозначается α и называется *уровнем значимости*. Исходя из принятого риска строится критическая область, т. е. определяется группа (интервал) значений критерия, позволяющих принять решение об ошибочности нуль-гипотезы. Таким образом, вся область значений делится на области принятия и отвержения гипотезы. Для приведенного выше примера выбор соответствующих областей изображен на рис. 8.

Этап 4. Из генеральной совокупности извлекается одна выборка. По ней рассчитывается значение статистики (критерия проверки) и принимается решение относительно истинности. Если значение критерия попадает в критическую область, то гипотеза отклоняется.

При принятии какого-либо решения вероятны два типа ошибок. Выделяют ошибки первого и второго рода (табл. 5). Вероятность ошибки первого рода равна уровню значимости α. Для определения вероятности ошибки второго рода следует уточнить отклонение истинного значения от постулируемого гипотезой. Например, выдвинута гипотеза, что параметр равен φ_0, а на самом деле $\varphi = \varphi_0 \pm d$. Вероятность того, что оценка попадет в область принятия гипотезы равна β (см. рис. 9). Следовательно, вероятность ошибки второго рода равна β при выявлении отклонений величины на $\pm d$ от гипотетического значения.

Таблица 5

Типы ошибок в статистическом выводе

		Наше предположение (принятое решение)	
		H_0 принята	H_0 отвергнута в пользу H_1
Фактически	Верна H_0	Решение правильное, вероятность $(1 - \alpha)$	Ошибка I рода, вероятность α
	Верна H_1	Ошибка II рода, вероятность β	Решение правильное, вероятность $(1 - \beta)$

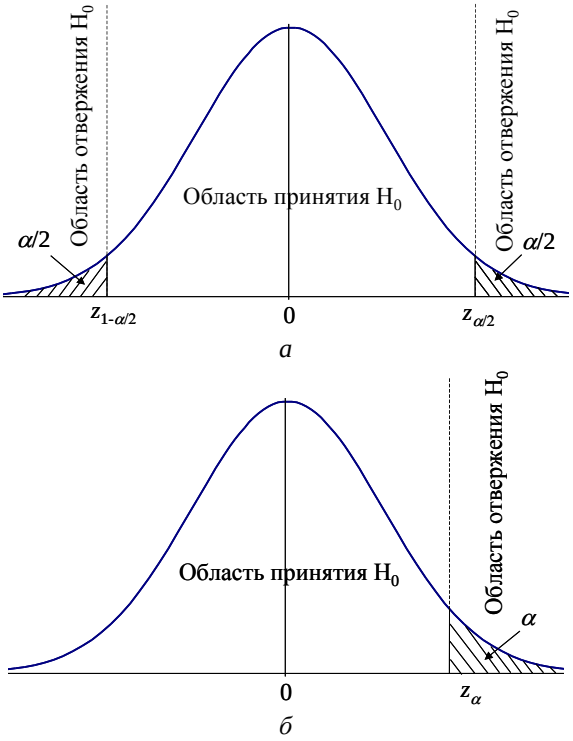

Рис. 8. Критические области при проверке статистической гипотезы:
а – двусторонняя гипотеза, *б* – односторонняя гипотеза

Вероятность $(1 - \beta)$ называется *мощностью критерия*. Единственный способ одновременно уменьшить и α, и β состоит в увеличении объема

выборки N (рис. 10). При увеличении объема выборки дисперсия оценки среднего уменьшается и распределение становится более узким. Следовательно, меньше становится площадь перекрытия.

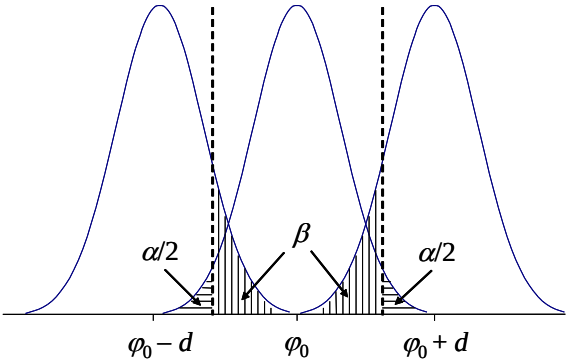

Рис. 9. Определение ошибки второго рода

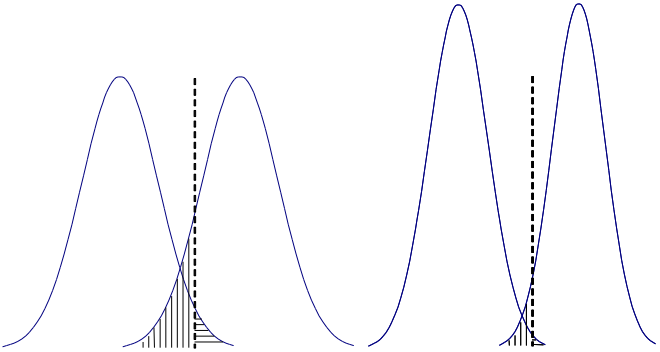

Рис. 10. Уменьшение ошибки второго рода при увеличении объема выборки

Мощность критерия для данного значения проверяемого параметра увеличивается с ростом объема выборки. Мощность критерия также увеличивается с увеличением ошибки первого рода α. Для фиксированных значений α и N мощность критерия увеличивается, когда истинное значение проверяемого параметра сильнее отклоняется от значения, принятого в нуль-гипотезе.

1.9. Корреляция и регрессия. Метод наименьших квадратов

Одна из наиболее распространенных задач анализа экспериментальных данных – поиск взаимосвязи и взаимозависимости двух и более случайных переменных. Рассмотрим сначала двумерный случай. Выше мы уже давали определение коэффициентов ковариации и корреляции. Ковариация (соизменение) характеризует взаимозависимость переменных. Ковариация положительного знака означает прямую связь, отрицательного – обратную связь. Нулевая ковариация характеризует отсутствие связи или независимость переменных. Но абсолютное значение коэффициентов ковариации зависит от самой измеряемой величины, что создает трудности при сравнении степени связи разных пар переменных. Коэффициент корреляции лишен этих недостатков, так как он нормирован на стандартные отклонения. Таким образом, независимо от того, для каких пар признаков мы находим коэффициент корреляции, его значение лежит в пределах от –1 (строгая обратная связь) до +1 (строгая прямая связь). При линейных преобразованиях значений переменной величина коэффициента корреляции не меняется.

Наличие корреляции двух переменных отнюдь не означает, что между ними существует причинная связь. Нельзя сказать определенно, что от чего зависит, что является причиной, а что следствием. Возможно, существует некий скрытый (латентный) фактор, который воздействует на обе эти переменные. Коэффициент корреляции Пирсона r_{xy} симметричен и не зависит от перестановки x и y.

Из того, что коэффициент корреляции равен 0, нельзя делать вывод, что x и y не связаны друг с другом. Возможно, что существует сильная нелинейная связь. Коэффициент корреляции Пирсона является мерой степени *линейности* связи. Поэтому случаи нелинейной связи надо рассматривать отдельно. Часто используют различные преобразования, чтобы перейти к линейному случаю и дальше уже использовать методы корреляционного и регрессионного анализа.

Если мы имеем дело с N случайными величинами, то можно определить совместное N-мерное нормальное (гауссово) распределение. Замечательная особенность этого распределения состоит в том, что все его свойства определяются исключительно средними значениями каждой переменной и ковариациями. Например, двумерное нормальное распределение случайных величин x и y обладает рядом важных свойств.

1. Распределение значений x без учета y, которому они соответствуют, есть нормальное распределение.

2. Распределение y без учета x, которому они соответствуют, есть нормальное распределение.

3. Для каждого фиксированного значения x значения y подчиняются нормальному распределению с дисперсией $\sigma^2_{y,x}$, одинаковой для всех x.

4. Для каждого фиксированного значения y значения x подчиняются нормальному распределению с дисперсией $\sigma^2_{x,y}$, одинаковой для всех y.

5. Средние значения y для каждого отдельного значения x ложатся на прямую.

Из 5-го свойства следует, что использование прямой для прогнозирования y по x по отдельной выборке разумно и никакая другая кривая не может дать лучших результатов. Свойства 3 и 4 используются в дисперсионном анализе.

Коэффициент корреляции позволяет установить степень взаимосвязи. Однако наряду с этим желательно иметь модель этой связи, по которой можно было бы предсказывать значения одной случайной величины по значениям другой. В этом случае используется *регрессионный анализ*, основанный на предположении о нормальном распределении случайных величин, т. е. на предположении о линейной связи.

Рассмотрим одномерную линейную модель. Предположим, что оценка величины y по x определяется как $\hat{y} = a + bx$. Если данные связаны идеальной линейной зависимостью, то предсказанное значение точно равняется измеренному значению. На практике обычно наблюдается разброс. Это означает, что $\hat{y}$ должно быть равно среднему значению всех наблюдений, сделанных при фиксированном значении x.

Как и любая выборочная оценка, коэффициенты a и b, найденные по выборке, могут отличаться от истинных. Обозначим истинные значения коэффициентов α и β. Тогда измеренное значение имеет вид

$$y_i = \alpha + \beta x_i + e_i .$$

Отдельное измерение содержит случайную компоненту e_i, которую называют *остатком* (рис. 11).

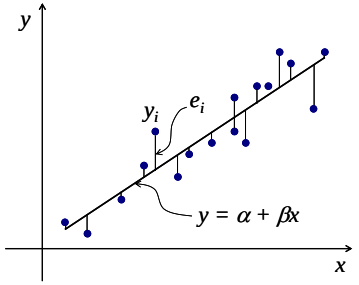

Рис. 11. Линейная регрессия

Константы a и b можно найти с помощью *метода наименьших квадратов* (МНК). Минимизируется сумма квадратов остатков или сумма квадратов отклонений измеренного и предсказанного по модели значения:

$$Q = \sum_{i=1}^{n} (y_i - a - bx_i)^2, \qquad \begin{cases} \dfrac{dQ}{da} = 0; \\ \dfrac{dQ}{db} = 0. \end{cases}$$

Здесь n – объем выборки. Решая систему линейных уравнений относительно a и b, получим:

$$b = \frac{n\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n\sum_i x_i^2 - \left(\sum_i x_i\right)^2};$$

$$a = y - bx\,.$$

Для каждой выборочной оценки можно определить дисперсии и построить соответствующие доверительные интервалы. Приведем формулы для расчета дисперсий величины y по x ($s_{y/x}^2$), коэффициентов линейного уравнения (s_a^2 и s_b^2) и конкретной оценки $\hat{y}$:

$$s_{y/x}^2 = \frac{1}{n-2}\sum_i (y_i - \hat{y}_i)^2;$$

$$s_b^2 = \frac{s_{y/x}^2}{\sum_i (x_i - x)^2};$$

$$s_a^2 = s_{y/x}^2 \left(\frac{1}{n} + \frac{x^2}{\sum_i (x_i - x)^2} \right);$$

$$s_{\hat{y}}^2 = s_{y/x}^2 \left(\frac{1}{n} + \frac{(x_i - x)^2}{\sum_i (x_i - x)^2} \right).$$

Распределение остатков e_i является нормальным $N(0, \sigma^2)$, причем все остатки имеют одинаковую дисперсию, не зависящую от значения x. Это свойство называют *гомоскедактичностью*.

Анализ остатков во многих случаях позволяет определить, адекватно ли линейная модель описывает экспериментальные данные, выявить гетероскедактичность. При гетероскедактичном распределении остатков возможны варианты перехода к гомоскедактичному случаю.

Долю общего разброса данных относительно выборочного среднего, которую можно объяснить выбранной моделью, называют *коэффициентом детерминации*. В линейном случае двух переменных этот коэффициент равен коэффициенту корреляции Пирсона в квадрате r_{xy}^2 и принимает значения от 0 до 1. Дисперсия остатков равна

$$s_e^2 = s_y^2(1 - r_{xy}^2).$$

В уравнении $\hat{y} = a + bx$ наклон прямой b вычисляется из предположения, что x – независимая переменная, y – зависимая (линейная регрессия y на x). Если поменять ролями переменные, получим регрессию x на y с другим наклоном b', тогда

$$\hat{x} = x + b'(y - y).$$

Можно показать, что эти два наклона связаны с выборочным коэффициентом корреляции соотношением $r_{xy}^2 = bb'$ (рис. 12).

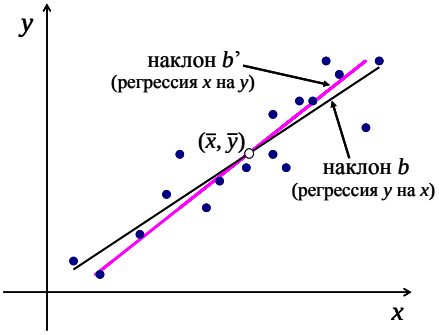

Рис. 12. Две линии регрессии на одних и тех же данных. Точка пересечения этих прямых соответствует выборочным средним

Чем сильнее корреляция, тем ближе наклоны этих прямых, в предельном случае при $r_{xy} = 1$ прямые совпадают, а все экспериментальные точки

лежат на линии регрессии. Если корреляция отсутствует ($r_{xy} = 0$), регрессионные прямые перпендикулярны друг другу. Тогда говорят, что признаки *x* и *y* независимы, или *ортогональны*.

В неколичественных шкалах (шкала наименований или порядковая шкала) коэффициент корреляции Пирсона вычислить нельзя. Тогда вводят другие меры связи. Например, для переменных в порядковых шкалах можно использовать коэффициент ранговой корреляции Спирмена (основан на коэффициенте Пирсона, но оперирует рангами) или коэффициент τ Кендалла (основан на вычислении доли совпадений порядка и его инверсии). Конкретные формулы для этих коэффициентов зависят от того, являются ли ранги связанными.

Для начала определим процедуру *ранжирования* выборки. Выборка сортируется и каждому значению приписывается ранг начиная с 1, одинаковым значениям присваиваются обязательно одинаковые ранги. Последний ранг должен совпадать с объемом выборки, т. е. с количеством объектов или измерений.

Для несвязанных рангов коэффициент корреляции Спирмена можно вычислить по приближенной формуле

$$r_s = 1 - \frac{6\sum (x_i - y_i)^2}{N(N^2 - 1)},$$

где x_i, y_i – ранги, N – объем выборки. Интерпретация коэффициента корреляции Спирмена такая же, как и коэффициента корреляции Пирсона.

При вычислении коэффициента τ Кендалла сначала определяют общее число «совпадений» и «инверсий». Пусть объектам присвоены ранги по *x* и по *y*. Для некоторой пары объектов констатируется «совпадение», если их порядок по *x* и *y* одинаков. Всего существует $N(N - 1)/2$ пар объектов. Тогда для несвязанных рангов

$$\tau = \frac{(\textit{общее число совпадений}) - (\textit{общее число инверсий})}{N(N-1)/2}.$$

Если два объекта выбираются случайно из выборки объемом *N*, то разность между вероятностью того, что они будут иметь одинаковый порядок как по *x*, так и по *y*, и вероятностью того, что у них будет наблюдаться различие в порядках, равна величине τ.

Примеры этих коэффициентов приведены для того, чтобы показать, что меру связи или корреляцию можно оценить не только по коэффициенту Пирсона. Для классификационных шкал не определена даже процедура ранжирования. Тогда нужно использовать другие коэффициенты, например, коэффициент связи φ для дихотомических данных.

Глава 2
Краткий обзор многомерных методов анализа

2.1. Стратегия обработки многомерных данных

Методы многомерной статистики изучают взаимосвязи большого количества признаков и большого количества объектов. Все методы можно условно разделить на несколько групп, среди которых важнейшими являются методы факторного, кластерного и дискриминантного анализа. На практике исследователь полагается в основном на компьютерные программы, которые часто предусматривают разные варианты вычислений. В идеале применение различных методов должно приводить к практически эквивалентным результатам. Тем не менее нужно осознавать, что большинство многомерных задач не имеет единственного или наилучшего решения. Всегда остается неоднозначность, разрешить которую методами математической статистики невозможно. Необходимо принятие внестатистических решений.

Вернемся к рассмотрению таблицы исходных данных (см. табл. 1). Можно выделить два разных типа задач. Первый тип – анализ свойств или признаков объектов, поиск факторов, влияющих на изменчивость признаков (анализ столбцов). Второй тип – сравнительный анализ объектов по набору признаков, поиск факторов, позволяющих выделить группы или классы схожих объектов (анализ строк).

По матрице $\mathbf{X} = \{x_{ij}\}$, где i – номер объекта, j – номер признака, $i = 1, \ldots, n$; $j = 1, \ldots, m$, можно построить матрицу связи или *матрицу корреляции* признаков

$$\mathbf{R} = \begin{pmatrix} r_{11} & \cdots & r_{1m} \\ \vdots & & \vdots \\ r_{m1} & \cdots & r_{mm} \end{pmatrix}.$$

Так как коэффициент корреляции симметричен ($r_{xy} = r_{yx}$), матрица симметрична относительно диагонали, а все $r_{ii} = 1$. Вместо коэффициентов корреляции иногда используют другие виды коэффициентов связи. Если вместо коэффициентов корреляции стоят коэффициенты ковариации, то матрицу называют *ковариационной* (обозначают **C**). В идеале, если мы предложили адекватную модель и по ней предсказали весь набор свойств объектов, то корреляционные матрицы, вычисленные по модели и по экспериментальным данным, должны совпадать. На таком сравнении строятся методы *корреляционного анализа*. Матрица корреляции используется также во многих методах факторного анализа.

Аналогично матрице корреляции можно построить *матрицу близости* объектов:

$$\mathbf{D} = \begin{pmatrix} d_{11} & \cdots & d_{1n} \\ \vdots & & \vdots \\ d_{n1} & \cdots & d_{nn} \end{pmatrix}.$$

Элементами матрицы близости являются расстояния между парами объектов в многомерном пространстве признаков. Как определить расстояние (метрику) – отдельный вопрос. Наиболее привычной является евклидова метрика

$$d(x, y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}, \quad x = (x_1, \ldots x_m),\ y = (y_1, \ldots y_m).$$

Евклидово расстояние не всегда удобно применять в компьютерных алгоритмах из-за корня квадратного. Можно использовать, например, так называемую «городскую» или «манхэттенскую» метрику:

$$d(x, y) = \sum_{i=1}^{m} |x_i - y_i|.$$

Расстояние между двумя точками в этом случае измеряется не по кратчайшей прямой, соединяющей точки, а так, как если бы мы в городе шли от одной точки до другой по улицам, которые идут строго параллельно или перпендикулярно друг другу. Отсюда и название метрики.

Сложнее определить расстояние от точки до множества или расстояние между двумя множествами. Существует множество вариантов разных метрик, но все они должны подчиняться основному определению и удовлетворять трем условиям. Функция $d(\cdot, \cdot)$ называется *расстоянием* (метрикой), если для любых трех объектов a, b и c в заданном пространстве признаков выполняются свойства:

1) неотрицательность – $d(a, b) \geq 0$, $a = b \Leftrightarrow d(a, b) = 0$;
2) симметричность – $d(a, b) = d(b, a)$;
3) неравенство треугольников – $d(a, b) \leq d(a, c) + d(c, b)$.

По матрице близости можно проверять модели, объясняющие не изменчивость признаков, а сходство объектов, что позволяет находить алгоритмы классификации объектов по признакам.

При построении различного рода моделей обычно начинают с линейных зависимостей (регрессионный анализ). Выше уже рассматривалась двумерная линейная модель регрессионного анализа. В многомерном

случае $y = f(x_1, \ldots, x_m)$ уравнение прямой выглядит следующим образом (уравнение множественной регрессии):

$$y = a + b_1 x_1 + b_2 x_2 + \cdots + b_m x_m.$$

Используя, как и в двумерном случае, метод наименьших квадратов, можно определить все коэффициенты линейной модели однозначно. Кроме того, для линейной модели можно найти множественный коэффициент детерминации (r^2), который определяется как отношение дисперсии предсказанных значений к дисперсии фактических значений и показывает, какая доля ошибки была объяснена построением линейной модели. Чем выше коэффициент детерминации, тем лучше модель описывает данные. Естественно, что это справедливо только в случае линейной зависимости. При отклонении от линейности необходимо выбирать другие модели или делать преобразования первичных данных.

2.2. Определение источника вариации данных. Дисперсионный анализ

Допустим, мы измерили определенный набор признаков, т. е. получили таблицу исходных данных. Реализация измерения каждого конкретного свойства или признака каждого отдельного объекта в реальном (не идеальном) эксперименте является случайной величиной. Возникает вопрос, какая часть дисперсии в наших данных происходит вследствие систематических причин (так называемых факторов), а какая из-за случайного разброса. В этом случае часто используют методы *дисперсионного анализа*. Устоявшееся название в англоязычной литературе, а соответственно и в статистических пакетах программ – ANOVA (ANalysis Of VAriance) или MANOVA (для многофакторных задач). Можно определить ANOVA как «статистический метод анализа измерений, зависящих от разного рода факторов, действующих одновременно, который позволяет решить, какой вид воздействия важен, и оценить это влияние».

Начнем с проверки влияния какого-либо одного фактора. Мы фиксируем фактор и проводим измерения при различных уровнях проявления этого фактора (контролируемый фактор). Уровень проявления фактора не обязательно измеряется в количественных шкалах, он может быть задан в любой шкале, начиная со шкалы наименований (например, адрес лаборатории, в которой проводили измерения). Проверяем, влияет ли этот фактор на разброс данных. В этом случае используется *однофакторный дисперсионный анализ*.

Проводится эксперимент с одним исследуемым фактором, который имеет J уровней. На каждом уровне берется n_j наблюдений. Основные допущения относительно наблюдений:

– независимы;
– взяты из нормальной генеральной совокупности с дисперсией σ^2;
– дисперсии на всех J уровнях одинаковы (гомоскедактичность).

Экспериментальные данные можно представить в виде таблицы (табл. 6), где $i = 1,\ldots,n_j$; $j = 1,\ldots,J$. Количество измерений n_j в общем случае может быть разным на разных уровнях j.

Тогда внутригрупповое среднее определяется как

$$x_{\bullet j} = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij}.$$

Принято ставить точку на месте того подстрочного индекса, по которому идет усреднение. Общее среднее имеет вид

$$x_{\bullet\bullet} = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j} x_{ij}, \text{ где } N = \sum_{j=1}^{J} n_j.$$

Таблица 6

Таблица данных для дисперсионного анализа

j / n	1	2	.....	J
1	x_{11}	x_{12}	.....	x_{1J}
2	x_{21}	x_{22}	.....	x_{2J}
.	.			.
.	.			.
.	.		x_{ij}	.
.	.			.
n_j	$x_{n_1 1}$		.....	$x_{n_J 1}$
	$\bar{x}_{\bullet 1}$	$\bar{x}_{\bullet 2}$	.....	$\bar{x}_{\bullet J}$

Общую сумму квадратов отклонений от среднего можно представить в виде суммы двух слагаемых:

$$SS_t = \sum_i\sum_j \left(x_{ij} - x_{\bullet\bullet}\right)^2 = \sum_i\sum_j \left(x_{ij} - x_{\bullet j}\right)^2 + \sum_j n_j \left(x_{\bullet j} - x_{\bullet\bullet}\right)^2 = SS_w + SS_b.$$

Первое слагаемое (SS_w) – внутригрупповая сумма квадратов, а второе (SS_b) – межгрупповая сумма квадратов.

Этапы однофакторного дисперсионного анализа.

1. Для объяснения данных постулируется линейная модель

$$x_{ij} = \mu + \mu_j + e_{ij},$$

где μ – общее среднее; μ_j – константа для всех n_j значений в группе j; e_{ij} – ошибка линейной модели (остаток). Считаем, что x_{ij} распределено нормально. Остатки тоже распределены нормально с той же дисперсией и с нулевым средним.

2. Формулируется нуль-гипотеза H_0 и альтернативная ей H_1:

H_0: $\mu_1 = \mu_2 = \ldots = \mu_j$ (как следствие, =0);
H_1: по крайне мере два μ_j различны.

Если у нас хотя бы в одной из групп наблюдается сильное отклонение среднего от среднего в остальных группах (что нельзя объяснить случайным отклонением), то межгрупповая дисперсия будет заведомо больше внутригрупповой. Поэтому для проверки нуль-гипотезы выбирается *F*-критерий сравнения дисперсий (табл. 7).

Таблица 7

Однофакторный дисперсионный анализ

Источник вариации	Сумма квадратов *SS*	Степени свободы *df*	Средний квадрат (дисперсия) *MS*	*F*	*F критическое*
Между группами	SS_b	$J-1$	$MS_b = \frac{SS_b}{J-1}$	$\frac{MS_b}{MS_w}$	$F_{1-\alpha,J-1,N-J}$
Внутри групп	SS_w	$N-J$	$MS_w = \frac{SS_w}{N-J}$		
Итого (все группы вместе)	SS_t	$N-1$	$MS_t = \frac{SS_t}{N-1}$		

3. Выбирается уровень значимости α.

4. Проводятся вычисления сумм квадратов, степеней свободы, средних квадратов (см. табл. 7).

5. Рассчитывается *F*-отношение и сравнивается с процентной точкой *F*-распределения $F_{1-\alpha,J-1,N-J}$.

Если *F*-отношение больше критической точки, то H_0 отвергается.

Несмотря на то, что при формулировании модели дисперсионного анализа накладывались довольно жесткие ограничения (нормальность, независимость, гомоскедактичность), сама процедура ANOVA является достаточно устойчивой к нарушениям этих ограничений в некоторых пределах.

Выводы статистиков говорят о том, что нарушение гипотезы о нормальном распределении в дисперсионном анализе практически не имеет значения. Вероятность ошибки первого рода остается практически той же самой. При одинаковом объеме выборок в группах влиянием неоднородности дисперсий на уровень значимости *F*-критерия можно пренебречь.

Если мы не можем при проверке влияния одного фактора избавиться от одновременного действия других факторов, то можно попытаться провести анализ одновременно для нескольких факторов, фиксируя каждый на своем уровне (MANOVA). Для двух факторов разработаны процедуры двухфакторного дисперсионного анализа. Возможны два варианта: без повторений и с повторениями.

Рассмотрим первый вариант – *двухфакторный дисперсионный анализ без повторений*. Мы остаемся в рамках представления данных в виде табл. 6, только будем считать, что действие второго фактора *B* фиксируется по строкам и разное в разных строках. Тогда число строк во всех группах будет одинаковым и равным I (вместо n_j). По каждой строке можно найти среднее, аналогично нахождению среднего по столбцу:

$$x_{i\bullet} = \frac{1}{J}\sum_{j=1}^{J} x_{ij}.$$

Линейная модель постулируется в виде

$$x_{ij} = \mu + \mu_j + \nu_i + e_{ij},$$

где ν_i – константа для всех значений в i-й строке.

Если нет влияния фактора, то дисперсии должны быть одинаковы. Общая сумма квадратов отклонений может быть представлена в виде

$$SS_t = SS_A + SS_B + SS_w,$$

$$SS_A = I\sum_{j=1}^{J}\left(x_{\bullet j} - x_{\bullet\bullet}\right)^2,$$

$$SS_B = J\sum_{i=1}^{I}\left(x_{i\bullet} - x_{\bullet\bullet}\right)^2,$$

$$SS_w = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(x_{ij} - x_{i\bullet} - x_{\bullet j} + x_{\bullet\bullet}\right)^2.$$

Проверяются две гипотезы: первая – относительно влияния фактора A, вторая – относительно влияния фактора B. Сравниваются дисперсии по каждому фактору с внутригрупповой дисперсией (табл. 8). По *F*-

распределению находятся критические точки для каждого фактора. Для каждой из гипотез находится свое значение *F*-критерия и своя процентная точка для выбранного уровня значимости (разные степени свободы).

Таблица 8

Двухфакторный дисперсионный анализ без повторений

Источник вариации	*SS*	*df*	*MS*	*F*	*F* критическое
Влияние фактора *A*	SS_A	$J-1$	$MS_A = \frac{SS_A}{J-1}$	$\frac{MS_A}{MS_w}$	$F_{1\text{-}\alpha,J\text{-}1,dfw}$
Влияние фактора *B*	SS_B	$I-1$	$MS_B = \frac{SS_B}{I-1}$	$\frac{MS_B}{MS_w}$	$F_{1\text{-}\alpha,I\text{-}1,dfw}$
Внутри групп	SS_w	$df_w = (I-1)\times(J-1)$	$MS_w = \frac{SS_w}{(I-1)(J-1)}$		
Итого (все группы вместе)	SS_t	$N-1$	$MS_t = \frac{SS_t}{N-1}$		

Рассмотрим второй вариант – *двухфакторный дисперсионный анализ с повторениями*. Организация данных несколько усложняется (табл. 9).

Таблица 9

Экспериментальные данные в двухфакторном дисперсионном анализе с повторениями

A / *B*	1	…..	*J*
1	x_{111} x_{112} … x_{11K}	…..	x_{1J1} x_{1J2} … x_{1JK}
2	x_{211} x_{212} … x_{21K}		x_{2J1} x_{2J2} … x_{2JK}
….	….	x_{ijk}	….
I	x_{I11} x_{I12} … x_{I1K}	…….	x_{IJ1} x_{IJ2} … x_{IJK}

Предполагается, что первый фактор A имеет J уровней, второй фактор B имеет I уровней, а число измерений с фиксированными значениями ij равно K (число значений в каждой клеточке таблицы экспериментальных данных). Каждое индивидуальное значение теперь имеет три индекса.

В двухфакторном дисперсионном анализе с повторениями можно проверить три гипотезы: 1) влияние фактора A; 2) влияние фактора B; 3) перекрестное влияние двух факторов $A \times B$, так как можно выделить еще и вклад в дисперсию SS_{AB} от взаимодействия двух факторов (табл. 10).

Таблица 10

Двухфакторный дисперсионный анализ с повторениями

Источник вариации	*SS*	*df*	*MS*	*F*	*F критическое*
Влияние фактора A	SS_A	$J-1$	$MS_A = \frac{SS_A}{J-1}$	$\frac{MS_A}{MS_w}$	$F_{1-\alpha,J-1,dfw}$
Влияние фактора B	SS_B	$I-1$	$MS_B = \frac{SS_B}{I-1}$	$\frac{MS_B}{MS_w}$	$F_{1-\alpha,I-1,dfw}$
Перекрестное влияние $A \times B$	SS_{AB}	$df_{AB} = (I-1)\times(J-1)$	$MS_{AB} = \frac{SS_{AB}}{df_{AB}}$	$\frac{MS_{AB}}{MS_w}$	$F_{1-\alpha,df_{AB},df_w}$
Внутри групп	SS_w	$df_w = IJ\times(K-1)$	$MS_w = \frac{SS_w}{IJ(K-1)}$		
Итого (все группы вместе)	SS_t	$N-1$	$MS_t = \frac{SS_t}{N-1}$		

Сумма квадратов внутри ячеек:

$$SS_w = \sum_i \sum_j \sum_k \left(x_{ijk} - x_{ij\bullet}\right)^2.$$

Доля перекрестного влияния:

$$SS_{AB} = K \sum_i \sum_j \left(x_{ij\bullet} - x_{i\bullet\bullet} - x_{\bullet j\bullet} + x_{\bullet\bullet\bullet}\right)^2.$$

Общая дисперсия:

$$SS_t = SS_A + SS_B + SS_{AB} + SS_w.$$

Тогда в таблице дисперсионного анализа добавится еще одна строка (см. табл. 10). Для каждого *F*-критерия строится своя критическая область по процентной точке *F*-распределения. В каждой из трех гипотез степени свободы среднего квадрата отклонения (дисперсии), стоящего в числителе, разные.

В результате делается три статистических вывода. Если значение критерия *F* лежит в критической области, то гипотеза о том, что соответствующий фактор не влияет, отвергается на выбранном уровне значимости.

2.3. Факторный анализ. Метод главных компонент

В факторном анализе предполагается, что наблюдаемые признаки являются линейной комбинацией некоторых латентных (скрытых) факторов. Группируются признаки (рис. 13). Некоторые из факторов допускаются общими для нескольких признаков, другие – характерными для каждой переменной отдельно. Характерные факторы ортогональны друг другу, следовательно, они не вносят вклад в ковариацию (корреляцию) между переменными. Только общие факторы, число которых предполагается гораздо меньше числа переменных, вносят вклад в корреляцию.

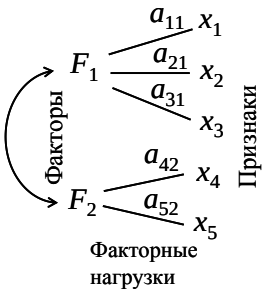

Рис. 13. Пример модели с двумя общими факторами

Анализируется матрица корреляции. Задача состоит в том, чтобы подобрать коэффициенты в факторной модели (нагрузки на факторы) так, чтобы получить наилучшее совпадение экспериментальной матрицы корреляции и вычисленной по факторной модели. Так как корреляционная матрица по выборке всегда отличается от матрицы по генеральной совокупности, точную структуру факторной модели получить нельзя. Мы можем говорить только об оценках параметров факторной структуры.

Существует множество вариантов факторного анализа. Мы их перечислять не будем. Во всех случаях можно выделить три основных этапа факторного анализа:

1) подготовка соответствующей матрицы ковариации (корреляции);

2) выделение первоначальных факторов;
3) вращение модели в целях получения окончательного решения.

Отдельно рассматривается вопрос о количестве факторов. Если пространство общих факторов найдено, то с помощью поворота осей можно получить бесконечное множество решений. Подбор подходящей системы координат – важная задача. Нужно так подобрать оси, чтобы результат можно было проинтерпретировать в терминах предметной области.

На первом этапе часто используют *анализ главных компонент* (англоязычная аббревиатура PCA). Этот метод иногда выступает как самостоятельный. Поясним его суть на примере двух признаков. Главные компоненты находятся в два этапа: перенос начала координат в центр рассеяния данных и вращение для получения оптимального решения (рис. 14). В этом методе при выборе главных компонент объясняется максимальная доля дисперсии.

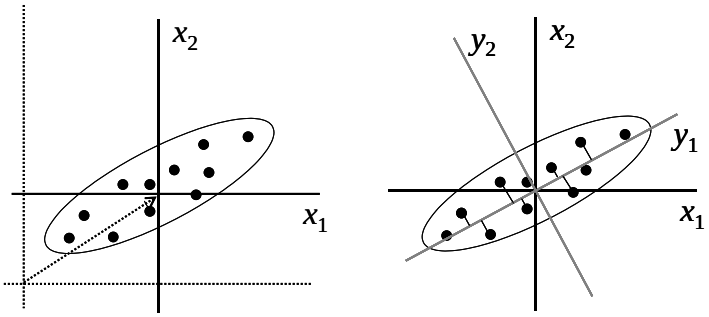

Рис. 14. Переход к главным компонентам

Пусть мы перенесли начало осчета в центр и построили диаграмму рассеяния по двум признакам x_1 и x_2. Новые компоненты y_1 и y_2 выбираются так, чтобы сумма квадратов расстояний до новой оси y_1 была минимальной (не путать с МНК, где минимизируется сумма квадратов отклонений по оси ординат, а не расстояний до прямой, см. рис. 11). Вторая ось выбирается перпендикулярно первой. Тогда

$$y_1 = a_{11}x_1 + a_{12}x_2;$$
$$y_2 = a_{21}x_1 + a_{22}x_2;$$
$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

a_{ij} – *нагрузки* на главные компоненты.

Данные вдоль оси y_1 имеют максимальный разброс, т. е. максимальную дисперсию. Потери информации при таком повороте координат не происходит. Тем не менее y_1 является наиболее информативной осью. Если зависимость x_1 и x_2 строго линейная, то y_1 содержит всю информацию и отпадает необходимость во второй оси. Можно сократить размерность данных. В многомерном случае оси располагаются по мере убывания доли дисперсии вдоль них, но перпендикулярно остальным осям. Как правило, в многомерном случае оси, вдоль которых доля дисперсии < 10–20 %, можно отбросить как мало информативные (решение зависит от конкретной задачи). Таким образом проводят *сжатие данных*.

В факторном методе при выборе главных компонет задача стоит в объяснении корреляции. Математический метод получения направлений главных осей основан на нахождении собственных чисел и векторов по матрице корреляции.

$$\det(\mathbf{R} - \mathbf{I}\lambda) = 0,$$

где **I** – единичная матрица, λ – собственное число. Решим это уравнение в двумерном случае:

$$\mathbf{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}; \quad \mathbf{R} = \begin{pmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{pmatrix}; \quad r_{11} = r_{22} = 1; \quad r_{12} = r_{21};$$

$$\det\begin{pmatrix} 1-\lambda & r_{12} \\ r_{12} & 1-\lambda \end{pmatrix} = 0; \quad \lambda^2 - 2\lambda + (1 - r_{12}{}^2) = 0;$$

$$\lambda_1 = 1 + r_{12}; \quad \lambda_2 = 1 - r_{12}.$$

Если r близок к 1, то $\lambda_1 \approx 2$, $\lambda_2 \approx 0$. Для некоррелированных данных оба собственных числа равны единице ($\lambda_1 \approx 1$, $\lambda_2 \approx 1$). Эти свойства сохраняются и для большей размерности. Сумма собственных чисел равна числу переменных (m).

$$\begin{pmatrix} \textit{Доля, соответствующая} \\ \textit{данной компоненте} \end{pmatrix} = \begin{pmatrix} \textit{Соответствующее} \\ \textit{собственное число} \end{pmatrix} \Big/ m.$$

Этот метод также помогает уменьшить размерность данных. Можно уменьшить размерность данных, но при этом латентная факторная структура остается «вещью в себе».

Для минимизации остаточной корреляции после выделения определенного числа факторов используются различные методы (метод наименьших квадратов, метод максимального правдоподобия, метод минимальных остатков, методы вращений и т. д.). Все критерии, по которым выбирается наилучшее факторное решение, требуют итерационных схем.

Например, в *методах вращений* (нахождение легко интерпретируемых факторов с помощью процедуры вращения) можно выделить три подхода.

Первый – графический (проведение новых осей). Если есть скопления точек, то оси удобно проводить через эти скопления. Исходная переменная – точка в факторном пространстве, нагрузки – координаты, сами факторы – оси. Эти методы трудно формализуются. Специалисты расходятся в определении «простоты» для таких «несовершенных» структур, как факторная.

Второй подход – аналитический. Выбирается некий объективный критерий, которым надо руководствоваться при вращении. В критерии *квартимакс* вращение проводят так, чтобы максимизировать дисперсию квадратов факторных нагрузок переменной. В критерии *веримакс* вместо дисперсии квадратов факторных нагрузок переменной используют дисперсию квадратов нагрузок фактора. Алгоритмически квартимакс проще, чем веримакс, но последний дает лучшее разделение факторов. Существуют и другие критерии.

Третий подход – задание априорной целевой матрицы. Цель вращения – нахождение факторного отображения, наиболее близкого к некоторой заданной матрице.

Результаты факторного анализа можно использовать для *факторного шкалирования*, обратной задачи, которая используется для создания новых факторных шкал. Факторная шкала позволяет присваивать каждому объекту некоторые числовые оценки значений выделенных факторов, используя значения наблюдаемых переменных. Шкалирование всегда связано с некоторой неопределенностью, так как факторы точно через наблюдаемые переменные не выражаются.

Для проверки и подтверждения достоверности того или иного метода факторного анализа или шкалирования используются различные статистические критерии и теория статистического вывода.

2.4. Кластерный анализ

Кластерный анализ – это множество вычислительных процедур, используемых при создании классификации. Группируются объекты по степени сходства или близости. В результате образуются кластеры или группы очень схожих объектов. В этой области многомерного анализа существует большой разнобой в терминологии и методологии. В отличие от факторного анализа многие методы кластерного анализа являются довольно простыми процедурами, которые не имеют, как правило, достаточного статистического обоснования, т. е. являются эвристическими. Разные кластерные методы могут порождать и порождают разные кластерные решения для одних и тех же данных. Поэтому желательно иметь

проверочную методику того, насколько «естественны» выделенные группы. Цель кластерного анализа заключается в *поиске* существующих структур. В то же время его действие состоит в *привнесении* структуры в анализируемые данные.

Семейство методов кластерного анализа очень разнообразно: иерархические агломеративные, иерархические дивизимные, итеративные методы группировки, факторные методы, методы сгущений, методы, использующие теорию графов. Для определения близости объектов применяют различные меры сходства: меры расстояния, коэффициенты корреляции, коэффициенты ассоциативности, вероятностные коэффициенты сходства и т. д. В качестве расстояний тоже можно выбрать различные метрики. Существенным недостатком расстояний является то, что оценка сходства сильно зависит от различий в сдвигах данных. Чтобы избежать этого, часто вводят нормировку (стандартизацию) данных, хотя такое преобразование не всегда адекватно. В качестве меры сходства можно выбрать коэффициенты корреляции между объектами, но статистического смысла они в этом случае не имеют, поскольку среднее определяется по совокупности разнотипных переменных, а не по совокупности объектов. Смысл «среднего» в этом случае не ясен.

Рассмотрим в качестве примера метод одиночной связи, который относится к иерархическим агломеративным методам. На первом шаге находятся два наиболее схожих объекта в матрице близости **D**. Следующий кандидат на включение в кластер присоединяется к существующей группе, если он имеет наибольшее сходство с одним из объектов кластера. Алгоритм можно представить в виде дендрограммы (рис. 15). Другие агломеративные методы отличаются правилами объединения (видами связи объектов). Объекты распределяются по кластерам за один проход, и плохое начальное разбиение не может быть изменено на следующих шагах.

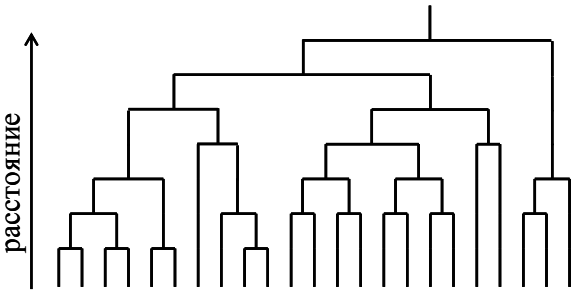

Рис. 15. Дендрограмма при кластерном анализе

В итеративных методах кластеризации можно выделить следующие шаги: 1) разбивают объекты на некоторое заданное число кластеров, вычисляют центры тяжести этих кластеров; 2) помещают каждую точку данных в кластер с ближайшим центром тяжести; 3) вычисляют новые центры тяжести, кластеры не заменяют на новые, пока не будут просмотрены все данные; 4) шаги 2 и 3 повторяют до тех пор, пока не перестанут меняться кластеры. К итеративным методам относится, например, метод *k*-средних. Значение *k* (количество кластеров) задается пользователем. Объекты приписываются к ближайшим центрам тяжести по евклидовым расстояниям. В итеративных методах не нужно хранить многомерную матрицу близости (как для иерархических методов). Кроме того, итеративные методы могут компенсировать последствия плохого исходного разбиения данных. Но большинство итеративных методов не допускает перекрытия кластеров, да и перебор всех возможных разбиений слишком сложен.

Методы сгущения уникальны тем, что могут создавать перекрывающиеся кластеры. Объектам разрешается быть членами нескольких кластеров. Оптимизируется некий статистический критерий, называемый «функцией когезии».

В каждом методе есть свои достоинства и недостатки. Четыре фактора оказывают на работу методов кластеризации большое влияние: 1) характеристики кластерной структуры (форма, размеры); 2) наличие выбросов и степень полноты классификации; 3) степень перекрытия кластеров; 4) выбор меры сходства. Например, итеративные методы группировки приводят к кластерам гиперсферической формы. Еще эти методы имеют тенденцию находить кластеры приблизительно равных размеров. А вот метод одиночной связи тяготеет к образованию больших продолговатых кластеров, но он дает хорошее восстановление кластерной структуры, если кластеры хорошо отделены и разделены. Проверить обоснованность кластерного решения чрезвычайно трудно. Пока что нет какого-то единого хорошего критерия. Допустим, мы решили проверить кластерное решение на другой, повторной выборке из одной и той же генеральной совокупности. При неудачной попытке повторить кластерное решение, оно отвергается, но успешное повторение не дает гарантии достоверности этого решения. Еще раз подчеркнем, что многие кластерные методы анализа не имеют достаточного статистического обоснования.

2.5. Дискриминантный анализ

Дискриминантный анализ помогает выявить различия между группами и дает алгоритм классификации. Основное предположение: объекты должны принадлежать одному из двух (или более) классов (групп). Строят

дискриминантные функции, по которым определяют отнесение к классу. В химии методы распознавания образов могут быть использованы для интерпретации спектроскопических данных.

Признаки, применяемые для того, чтобы отличить один класс от другого, называют *дискриминантными переменными*. Число объектов в общем случае должно превышать число дискриминантных переменных примерно в 2 раза. Чаще всего дискриминантные функции являются линейными комбинациями дискриминантных переменных. Важно, что закон распределения является многомерным нормальным. Это позволяет точно определить вероятность принадлежности к данному классу и критерий значимости. Ни одна дискриминантная переменная не может быть линейной комбинацией других. Недопустимы переменные, коэффициент корреляции которых равен 1.

В основе методов дискриминантного анализа лежат либо методы множественной регрессии, либо методы дисперсионного анализа. Если классифицирующие переменные можно считать зависимыми от дискриминантной, то задача аналогична множественной регрессии, только зависимая переменная измеряется в шкале наименований. Но когда наоборот, значение дискриминантной переменной зависит от классов, то дискриминантный анализ является обобщением дисперсионного анализа. Это типично для задач, когда принадлежность объекта к некоторому классу вызывает изменение одновременно в нескольких переменных.

Введем обозначения: g – число классов, p – число дискриминантных переменных, n_i – число объектов класса i (частота), n – общее число объектов. Должно быть: $g \geq 2$, $n_i \geq 2$, $0 < p < (n - 2)$. Дискриминантные переменные должны измеряться в количественных шкалах. Каноническая дискриминантная функция является линейной комбинацией дискриминантных переменных:

$$f_{km} = u_0 + u_1 x_{1km} + u_2 x_{2km} + \ldots + u_p x_{pkm},$$

где f_{km} – значение канонической дискриминантной функции m-го объекта в группе k, x_{ikm} – значение дискриминантной переменной x_i для m-го объекта в группе k. Коэффициенты u_i для первой функции выбираются так, чтобы ее средние значения для разных класов как можно больше отличались друг от друга. Коэффициенты второй функции выбираются также, но значения второй функции должны быть некоррелированы со значениями первой. И так далее. Максимальное число дискриминантных функций, которые можно получить, равно числу классов без 1 ($g - 1$) или числу дискриминантных переменных p, если $p < g - 1$.

Для определения положения класса можно вычислить его центроид. *Центроид* – воображаемая точка, координаты которой есть средние значе-

ния переменных в данном классе (вектор средних значений для данного класса). Чтобы различать относительное положение центроидов, не нужна слишком большая размерность, на единицу меньше числа классов. Центроиды задают пространство. Точка, в которой каждая ось имеет нулевые значения, называется *главным центроидом* (определяется по средним значениям всей совокупности объектов). Оси выбирают так же, как в методе главных компонент, но не для объектов, а для центроидов. Если расположение классов действительно различается, то степень разброса наблюдений внутри классов будет меньше общего разброса. Для измерения разброса служат матрицы ковариации (корреляции). Находят матрицу разностей межгрупповой и внутригрупповой ковариации и дальше решают систему уравнений, из которых и находят коэффициенты u_i.

По найденным коэффициентам для каждого конкретного объекта можно вычислить значения f_{km} и по ним провести сравнение. Для двух дискриминантных фунций удобно использовать графическое представление. Можно изобразить на плоскости положения центроидов групп и индивидуальных наблюдений. Пример такого представления изображен на рис. 16. Дискриминантными функциями являются оси координат. Видно, что центроиды хорошо разделяются по этим функциям. Вдоль первой функции (ось абсцисс) наблюдается наибольший разброс. Эта функция является самым мощным дискриминатором, но группы 1 и 4 по этой функции не разделяются, для их разделения нужна функция 2 (ось ординат).

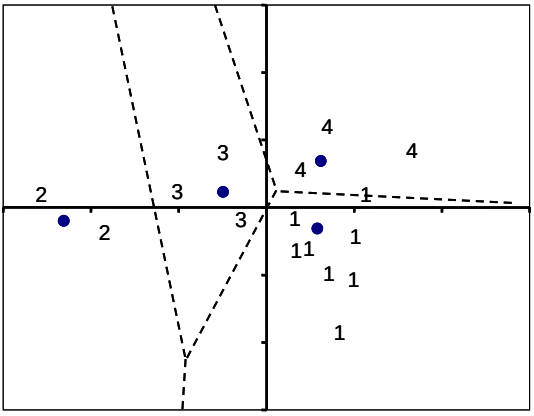

Рис. 16. Двухкоординатный график групповых центроидов и наблюдений, центроиды изображены кружками, цифры соответствуют номеру группы

Если дискриминантных функций больше двух, то графическое представление уже не является наглядным. Если мы имеем дело с большим количеством наблюдений, то более распространенным методом графического представления является построение гистограмм по каждой функции для каждой отдельной группы.

Для определения взаимозависимости отдельных переменных и дискриминантных функций находят коэффициенты корреляции. Эти коэффициенты называются полными *структурными коэффициентами*. Они показывают, насколько тесно связаны переменные с дискриминантными функциями. Когда абсолютная величина структурного коэффициента велика, почти вся информация о дискриминантной функции заключена в этой переменной. Иногда по доминантной переменной дают «имя» дискриминантной функции.

Максимальное число канонических дискриминантных функций меньше любого из чисел p и $(g - 1)$. Некоторые функции являются либо нулевыми, либо статистически мало значимыми. Полезность каждой функции оценивают по различным критериям, для чего используют теорию статистического вывода. Например, в качестве критерия значимости можно выбрать долю дисперсии дискриминантной функции, которая объясняется разбиением на классы.

Используя дискриминантные функции, мы можем провести классификацию, т. е. предсказать класс, к которому наиболее вероятно принадлежит объект. Выборку, по которой проводится разделение классов, можно назвать обучающей. Дискриминантные функции дают нам полезную информацию об отдельных объектах, о различиях между классами, о способности переменных точно различать классы.

Наиболее распространенные методы многомерной статистики реализованы в различных статистических программах. Наиболее математически обоснованными являются методы факторного анализа. В принципе, методы факторного анализа можно применить и в кластерном анализе, только вместо группировки признаков использовать группировку объектов. Многие специализированные методы имеют свою аббревиатуру, часто пока еще не переведенную на русский язык, поскольку в русском языке еще нет устоявшихся терминов.

Глава 3
Практический анализ экспериментальных данных

3.1. Форматы представления данных. Визуализация

При обработке данных зачастую приходится проводить процедуру импорта и экспорта файлов различных форматов. Современные приборы практически всегда имеют выход на компьютер. Форматы выдачи данных могут быть самыми разными. С приборами приходят специализированные программы, предназначенные для работы со своими специфическими форматами. Достаточно универсальным является представление данных в текстовом формате. Например, выдача с четырехкружного дифрактометра STADI 4 фирмы STOE сразу идет в текстовом формате, файлы можно просмотреть в любом текстовом редакторе. Выдача информации с D8 Bruker идет в формате, который читается только специализированными программами, но есть отдельно программы преобразования в различные форматы. Электронные таблицы читают текстовые файлы и могут разбить текст по колонкам. Числа при этом имеют соответствующий числовой формат и доступны для дальнейшей обработки. Универсальным для современных приборов является то, что все «сырые» данные имеют вид числовых таблиц. Некоторые старые приборы или собранные в единичном экземпляре экспериментальные установки имеют аналоговую выдачу в виде непрерывного графика. Обработка таких данных включает дополнительный этап «оцифровки», т. е. перевод во все тот же текстовый табличный формат для дальнейшего анализа.

Важный этап анализа данных – *визуализация* или графическое представление. Лучше один раз увидеть. Таблицы чисел человеку трудно анализировать. В электронных таблицах и в других статистических программных пакетах всегда есть средства построения графиков различных видов и типов по табулированным данным.

Каждый объект можно представить в виде точки в многомерном пространстве признаков. Чтобы хотя бы качественно проанализировать структуру данных, т. е. посмотреть, как распределены объекты-точки в этом многомерном пространстве, насколько близки они друг к другу, образуют ли одну компактную группу или распадаются на несколько групп, можно построить различные проекции этих точек-объектов на отдельные признаки. Проекцию объектов на один признак можно построить в виде *гистограммы*. На рис. 17 приведен пример построения гистограммы распределения 200 объектов по 12 различным значениям какого-то абстрактного признака. Особенно полезен этот тип визуализации для большого количества наблюдений.

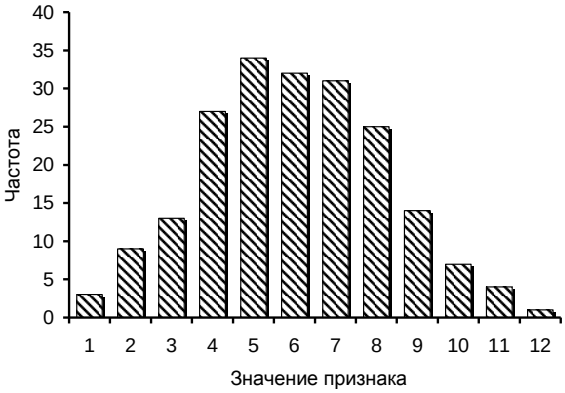

Рис. 17. Гистограмма, построенная для 200 объектов

Гисторамма, или частотное распределение объектов по оси значений признака, наглядно показывает, какие значения всречаются чаще всего, насколько сильно они различаются между собой, как сконцентрировано большинство наблюдений. Если признак классификационный, то достаточно для каждого класса или категории посчитать количество попавших в него объектов и по частотной таблице построить гистограмму. Если признак количественный и непрерывный, тогда ось признаков разбивают на интервалы. Обычно для 100 и более значений выбирается 10–15 интервалов или «карманов». Если выбрать слишком мало интервалов, то график получится непредставительным, трудно будет «увидеть», группируются ли значения вокруг одного или нескольких центров. Разбивку на интервалы нужно делать так, чтобы не было пустых или нулевых интервалов. Слишком мелкое или слишком крупное дробление на интервалы может привести к потере наглядности.

Построение гистограмм. Сначала оценивается шаг, с которым будут построены интервалы или диапазоны. Для этого определяется минимальное и максимальное значение признака, находится их разность и делится на выбранное количество диапазонов. Шаг обычно округляется до удобного значения. В случае нормального распределения шаг должен быть равен примерно 0,4*s*. Затем нужно построить колонку диапазонов («карманов»). В электронных таблицах принято ставить в ячейку только одну границу диапазона. Затем подсчитывается количество наблюдений, попавших в каждый «карман». И уже по этой новой частотной таблице строится гистограмма. Пример построения колонки диапазонов («карманов») и интерпретации частот приведен на рис. 18.

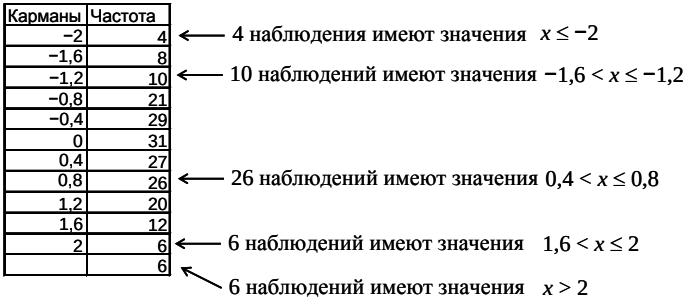

Рис. 18. Интерпретация частотной таблицы

В электронных таблицах существует и программная реализация алгоритма подсчета частот в виде отдельной процедуры или в виде функции. Как правило, в электронных таблицах необходимо вручную подготовить колонку «карманов» и затем использовать соответствующую команду или функцию. В статистических пакетах программ этот процесс автоматизирован, нужно только указать требуемое количество интервалов.

Огибающая, построенная по вершинам столбцов гистограммы называется *полигоном частот*. Если полигон частот имеет один максимум, то распределение называют унимодальным, два максимума – бимодальным и т. д. Если переменная дискретная или измерена в неколичественных шкалах (порядковой или классификационной), то по частотной таблице можно определить моду. Например, распределение, показанное на рис. 17, унимодальное, модой является значение 5.

Одно из наиболее часто встречаемых графических представлений – *диаграмма рассеяния* двух признаков. Это проекция всех объектов на плоскость. По координатным осям откладываются значения выбранных признаков. Если весь массив собранных данных представить как n-мерное пространство объектов-точек, где n – количество признаков, то диаграммы рассеяния дают двумерные срезы структуры этого n-мерного пространства. Построение диаграмм рассеяния часто позволяет определить дальнейший ход анализа данных, выбрать подходящий метод обработки. Иногда построение диаграмм рассеяния позволяет определить новые направления в исследованиях, предсказать те или иные свойства. Если один признак функционально зависит от второго, то по диаграмме рассеяния можно увидеть характер этой зависимости. Если она нелинейная, то можно попытаться подобрать нужную модель, используя функциональные преобразования признаков и метод наименьших квадратов.

Некоторые диаграммы рассеяния даже носят имена их создателей. Приведем два примера: диаграмма Музера – Пирсона и диаграмма Рамачадрана. Диаграмма Музера – Пирсона (рис. 19) позволяет соотнести структурный тип соединения и ионность связи. По осям откладываются две величины: 1) среднее значение главных квантовых чисел атомов, входящих в соединение; 2) разность их электроотрицательности.

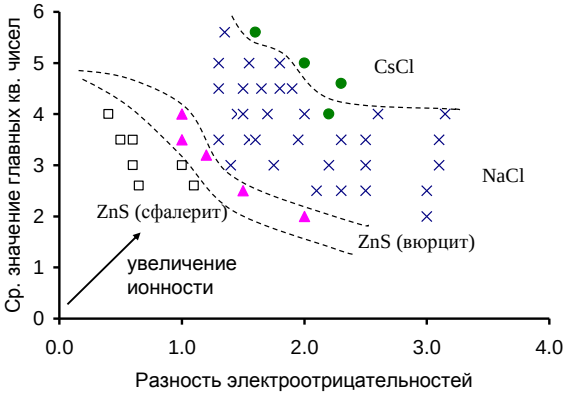

Рис. 19. Диаграмма Музера – Пирсона для соединений типа AB

Все соединения типа AB на диаграмме хорошо разделяются на четыре группы, соответствующие структурным типам сфалерита, вюрцита, NaCl, CsCl. Ионность связи на такой диаграмме возрастает от левого нижнего угла к правому верхнему. С помощью диаграммы Музера – Пирсона можно предсказать и объяснить структурный тип, к которому относится то или иное соединение. Подобные диаграммы строятся и для соединений AB_2, AB_3 и т. д.

Диаграмма Рамачадрана показывает, как распределены полипептиды в координатах двух торсионных углов (рис. 20). Эти углы характеризуют вращение вокруг пептидной связи. На диаграмме четко выделяются две группы объектов (полипептидов). В одной группе цепи свернуты в α-спирали, в другой – соответствуют β-складкам.

Иногда весь анализ данных заканчивается на этапе построения графиков. От правильного и грамотного построения и оформления графиков зависти и качество выводов. Напомним, что при анализе многомерных данных также широко используются диаграммы рассеяния, только в качестве координатных осей используются либо факторные шкалы, либо дискриминантные функции (рис. 16).

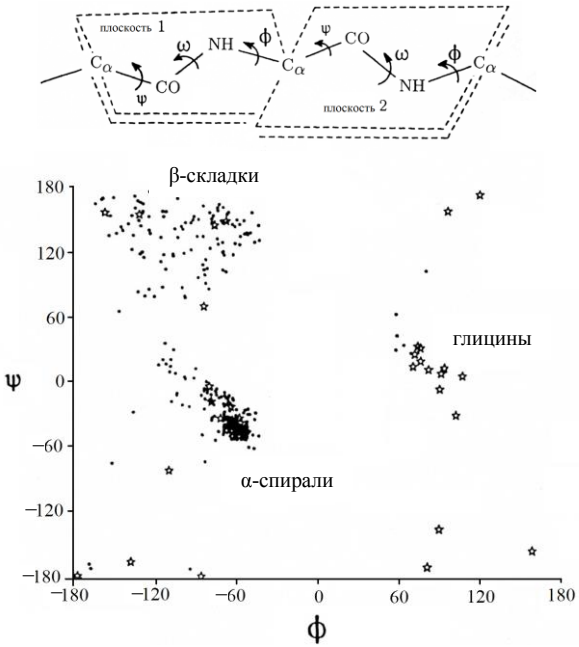

Рис. 20. Диаграмма Рамачадрана

Рассмотрим еще один пример построения диаграмм рассеяния, который в последнее время получил широкое распространение при анализе межмолекулярных контактов в кристаллических структурах. Этот метод реализован в бесплатно распространяемой программе CrystalExplorer. По известной структуре молекулярного кристалла для отдельной молекулы строится поверхность Хиршфельда. Она определяется по соотношению электронных плотностей (ρ) отдельной молекулы и молекулы в кристалле:

$$\omega(\vec{r}) = \sum_{a \in mol} \rho_a(\vec{r}) \Big/ \sum_{a \in cryst} \rho_a(\vec{r}) = \rho_{promol}(\vec{r}) / \rho_{procyst}(\vec{r}),$$

суммирование идет по всем атомам молекулы. Поверхность охватывает область вокруг молекулы, для которой выполняется неравенство $\omega(\vec{r}) \geq 0{,}5$. Далее выбираются два признака: d_i – расстояние от поверхности Хиршфельда до ближайшего ядра атома внутренней молекулы, d_e – расстояние от поверхности Хиршфельда до ближайшего ядра атома других соседних молекул. Все точки поверхности проецируются на плоскость

этих двух признаков. На рис. 21 приведен пример таких диаграмм для молекулы хлорпропамида в разных полиморфных модификациях.

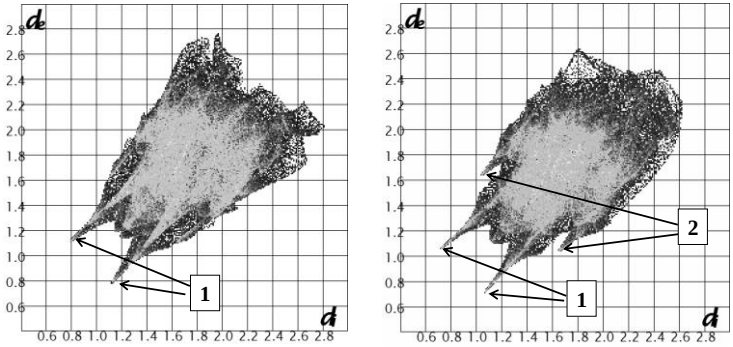

Рис. 21. Двумерная проекция поверхности Хиршфельда молекулы хлорпропамида в высокотемпературной ε-форме (*слева*) и в низкотемпературной ε'-форме (*справа*): *1* – водородные связи N–H...O; *2* – короткие контакты N–H...Cl

Из диаграмм рассеяния видно, как изменилось межмолекулярное взаимодействие после полиморфного перехода. В низкотемпературной форме появились новые короткие контакты (*2*). В программе есть возможность выделить контакты отдельных типов атомов, например, O...H, H...H и т. д. Форма и распределение плотности точек на диаграммах характеризуют каждую молекулу в кристаллической структуре так же, как отпечатки пальцев человека. В англоязычной литературе эти диаграммы называют «Hishfeld fingerprint plots». Сравнительный анализ таких диаграмм позволяет выделить характерные особенности межмолекулярных взаимодействий в твердом состоянии. Особенно полезна такая визуализация при анализе взаимодействия молекул одного сорта, но находящихся в разном кристаллическом окружении (полиморфы, сольваты, гидраты, сокристаллы).

Напомним некоторые *правила представления графиков*. Оси обязательно должны быть подписаны (название или обозначение признака и единиц измерения). При представлении нескольких зависимостей на одном графике необходима «легенда», т. е. указание того, каким образом они обозначены (цвет, тип линии, вид символа). Легенду можно помещать и в подрисуночную подпись. Масштаб по осям выбирают так, чтобы как можно меньше оставалось пустого пространства. Если на графике представлены экспериментальные данные в виде отдельных точек, то для каждой точки нужно указать погрешность измерения (нарисовать «усы»). Если «усы» ошибок не нарисованы, то по умолчанию считается, что

погрешность не превышает размера символа, которым отмечается экспериментальная точка.

3.2. Оценка точности измерений, распространение погрешностей

Цель проведения измерения – получить правильную оценку «истинного» значения. Выше в п. 1.3. мы уже обсуждали причины возникновения ошибок и пришли к выводу, что любое экспериментально измеренное значение является случайной величиной. Напомним, что есть две категории ошибок, из которых складывается общая погрешность измерения: случайная и систематическая.

Систематические ошибки могут быть аддитивными и мультипликативными. Основные источники аддитивных ошибок: влияние селективности метода, когда другие компоненты системы реагируют так, что дают ложно высокое значение измеряемого компонента; матричный эффект, когда источник ошибок в присутствии компонентов, которые сами не реагируют, но которые сдерживают или увеличивают измеряемое значение; неадекватная поправка на бланк или смещение нуля. Мультипликативные ошибки происходят из-за ошибок в калибровке, неправильного предположения о линейности в измеряемой области.

Существуют и другие источники ошибок, которые не так-то просто классифицировать. Например, загрязнение автоматической системы анализа предыдущим образцом. С одной стороны, это является ошибкой прибора, систематической. С другой стороны, она до некоторой степени случайная. Случайная и систематическая ошибки вместе дают ошибку индивидуального измерения, что является одним из важнейших критериев, по которому судят о результате. Должны быть учтены все источники вариации.

Исследователь всегда должен задаваться вопросом, нужно ли стремиться уменьшить случайную ошибку. Например, при определении содержания свинца в сельхоз. лаборатории было обнаружено, что 87,8 % ошибки идет за счет пробоотбора, 9,4 % – межлабораторная ошибка, 1,4 % – подготовка образца и только 1,4 % – случайная ошибка. Ясно, что уменьшение случайной ошибки вызывает слабый интерес. И наоборот, очевидно, не стоит пытаться получить ошибку метода 0,01 %, когда область разброса нормального распределения случайной ошибки порядка 20 %.

На практике разделить две составные части погрешности очень трудно. В соответствии с видами ошибок выделяют и два вида точности измерений, точность как меткость, воспроизводимость (*precision*) и точность как правильность, корректность (*accuracy*). В первом случае мы можем иметь малый разброс данных вокруг среднего значения, но не обязательно это

значение является истинным. Эта точность характеризует неопределенность измерений и связана со случайной ошибкой (чем меньше разброс, тем точнее измерение). Во втором случае точность определяется как получение правильного, истинного значения даже при большом случайном разбросе. Со вторым типом точности связана систематическая ошибка (корректность результата, отсутствие смещения от истинного значения). Наглядный пример разного типа точности приведен на рис. 22.

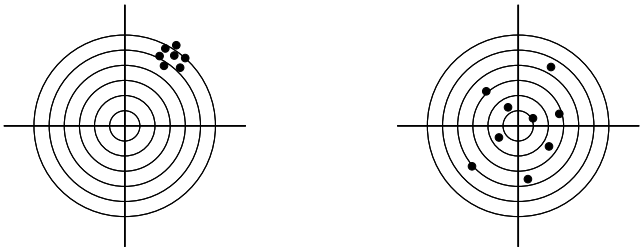

Рис. 22. Точность при стрельбе по мишени: *а* – большая систематическая ошибка и малая случайная; *б* – малая систематическая ошибка и большая случайная

Пусть x_i – индивидуальное измерение, μ – математическое ожидание оценки, μ_0 – истинное значение, тогда ошибка определяется как

$$e_i = x_i - \mu_o = x_i - \mu + \mu - \mu_0,$$

$(x_i - \mu)$ – случайная ошибка, $(\mu - \mu_0)$ – систематическая.

Случайная погрешность не может быть предсказана заранее. Однако можно высказать суждение о ее статистических свойствах. Для этого необходимо провести не одно измерение, а несколько, т. е. получить выборку и по ней оценить параметры (табл. 3). Построение доверительных интервалов позволяет сделать вероятностное суждение о том, насколько точно мы определили искомую величину. Точность можно повысить, увеличив объем выборки (см. п. 1.7). В каждом конкретном измерении знак случайной ошибки не определен.

Для каждого измерительного прибора указывается предел допустимой погрешности. Этот предел определяется по метрологическим измерениям, проводимым на стадии разработки прибора. При соблюдении всех правил и условий применения прибора погрешность измерения не должна превышать допустимую. Знак этой погрешности может быть неопределенным, так как включает действие множества неучтенных факторов при проведении измерения. В некоторых случаях определяют отдельно пределы погрешностей с положительным или отрицательным знаком. Ошибка кон-

кретного измерения обязательно содержит приборную погрешность, но не только. Например, ошибки, связанные с пробоотбором, в приборную погрешность не входят.

Систематическая ошибка, если она выявлена, в отличие от случайной имеет определенный знак. Если нам известны условия, которые привели к появлению систематической ошибки, то мы можем учесть это в итоговом результате. Более подробно о выявлении систематической ошибки мы будем говорить ниже. Приведем здесь только один пример. При измерении порошковой дифрактограммы в геометрии Брегга – Брентано важным фактором является расположение образца в прямом пучке. Не всегда удается точно соблюсти все необходимые условия при установке образца. Возникают ошибки, связанные со смещением образца или смещением нуля гониометра. Эти ошибки можно перевести в разряд систематических, используя внутренний стандарт. В измеряемый образец добавляют стандартное вещество с известной кристаллической структурой и по нему определяют систематическую ошибку каждого конкретного измерения.

Если нас интересует ошибка величины, вычисленной по нескольким экспериментально определенным величинам, то мы должны учесть погрешности всех измерений в итоговом результате. Пусть $y = f(x_1, x_2, \ldots, x_n)$ – функция, вычисленная по измеренным в эксперименте величинам x_i. Пусть нам известны все погрешности каждой отдельной величины x_i. Вычисление погрешности величины y называется *распространением погрешностей*. Рассмотрим распространение погрешностей в статическом случае.

При малых отклонениях отдельных измеренных значений результирующее отклонение можно рассчитать, используя первые члены ряда Тейлора:

$$\Delta y \cong \frac{\partial f}{\partial x_1}\Delta x_1 + \frac{\partial f}{\partial x_2}\Delta x_2 + \cdots + \frac{\partial f}{\partial x_n}\Delta x_n.$$

Определение систематической погрешности – это отдельная задача. Но если определена систематическая ошибка каждого измерения E_{sxi}, то при расчете суммарной систематической погрешности для y важно учесть знаки всех систематических ошибок. Тогда

$$E_{sy} = \frac{\partial f}{\partial x_1}E_{sx1} + \frac{\partial f}{\partial x_2}E_{sx2} + \ldots + \frac{\partial f}{\partial x_n}E_{sxn}.$$

Обозначим предел допустимой погрешности Δx_i. Превышать этот предел погрешность конкретного измерения не может, если соблюдены все описанные в инструкции правила и условия работы прибора. Пределы погрешностей могут иметь положительные, отрицательные или неопреде-

ленные знаки. При неопределенных знаках суммируют абсолютные значения пределов погрешностей отдельных измерений:

$$\Delta y = \left|\frac{\partial f}{\partial x_1}\Delta x_1\right| + \left|\frac{\partial f}{\partial x_2}\Delta x_2\right| + \ldots + \left|\frac{\partial f}{\partial x_n}\Delta x_n\right|.$$

Если знаки пределов погрешности известны, то положительные и отрицательные пределы погрешностей вычисляются отдельно.

Распространение случайной погрешности определяется дисперсией σ^2 или оценкой дисперсии s^2 каждой переменной. Если отдельные влияющие величины взаимно независимы, дисперсии известны и $\sigma_i << x_i$:

$$\sigma_y = \sqrt{\left(\frac{\partial f}{\partial x_1}\sigma_{x1}\right)^2 + \left(\frac{\partial f}{\partial x_2}\sigma_{x2}\right)^2 + \ldots + \left(\frac{\partial f}{\partial x_m}\sigma_{xm}\right)^2}.$$

Если вместо стандартных отклонений представлены их оценки, то

$$s_y = \sqrt{\left(\frac{\partial f}{\partial x_1}s_{x1}\right)^2 + \left(\frac{\partial f}{\partial x_2}s_{x2}\right)^2 + \ldots + \left(\frac{\partial f}{\partial x_m}s_{xm}\right)^2}.$$

3.3. Проверка распределений. Критерий согласия хи-квадрат

В теории статистического вывода широко используется предположение о том, что распределение случайной величины является нормальным. Подтвердить это в реальном эксперименте на все 100 % нельзя. Но можно проверить, насколько экспериментально полученное распределение близко к теоретическому. Существуют статистические критерии проверки. Одним из самых распространенных критериев является критерий согласия хи-квадрат. Этот критерий можно использовать для проверки близости любых двух распределений.

Общая идея критерия заключается в том, что в качестве меры расхождения наблюдаемой плотности вероятности и гипотетической используется некоторая статистика, приближенно подчиняющаяся хи-квадрат-распределению. Строится частотное распределение или гистограмма наблюдённых частот. Число наблюдений, попавших в i-й интервал, обозначим f_i. Число наблюдений, которые могли бы попасть в i-й интервал, если бы истинной плотностью была гипотетическая, обозначим F_i (ожидаемая частота i-го интервала). Строим выборочную статистику

$$X^2 = \sum_{i=1}^{K} \frac{(f_i - F_i)^2}{F_i},$$

где K – число интервалов.

Распределение величины X^2 приближенно совпадает с хи-квадрат распределением со степенью свободы n, равным числу интервалов минус число независимых линейных ограничений, наложенных на наблюдения. Одно такое ограничение связано с тем, что частота в последнем интервале связана с частотами во всех предыдущих интервалах. Если гипотетическая плотность – нормальная с неизвестным средним и с неизвестной дисперсией, то появляются еще два ограничения. Следовательно, в обычном случае проверки нормальности $n = K - 3$. Поскольку любое отклонение распределения от гипотетического вызовет увеличение критерия, то используем односторонний критерий (по верхней границе). Область принятия гипотезы имеет вид

$$X^2 \le \chi^2_{n,\alpha}.$$

Если значение попадает в область принятия гипотезы, то делается вывод о том, что гипотезу о совпадении распределений мы на заданном уровне значимости отвергнуть не можем. Следовательно, имеющиеся данные не противоречат гипотезе о принадлежности полученного эмпирического распределения к нормальному.

Аналогичным образом можно проверить соответствие эмпирического распределения любому другому распределению или совпадение двух разных эмпирических распределений. Только в каждом конкретном случае при выборе критической точки нужно корректировать число степеней свободы хи-квадрат-распределения. Так, при сравнении двух известных эмпирических распределений число степеней свободы $n = K - 1$.

В электронных таблицах существует функция хи-тест. Значение, выдаваемое этой фукцией, соответствует вероятности p хвоста хи-квадрат-распределения, отсекаемого вычисленным критерием X^2, причем число степеней свободы жестко задано $K - 1$. Интерпретация этого значения такова: если $p < \alpha$, то мы попадаем в область отвержения нулевой гипотезы на заданном уровне значимости α, т. е. два проверяемых распределения статистически значимо различаются. Эту функцию удобно использовать для сравнения двух эмпирических распределений. Для проверки распределения на нормальность ее использовать трудно, так как берется не то число степеней свободы ($K - 1$), тогда как нужно брать $n = K - 3$.

3.4. Методы сравнения экспериментальных данных

Когда исследователь проводит несколько повторяющихся измерений одного и того же образца с использованием одной и той же процедуры, аппаратуры, реагентов и т. д., как правило, получается нормальное распределение ошибок. Стандартное отклонение этого распределения является случайной ошибкой процедуры.

Если процедура применяется как рутинная, могут играть роль другие источники ошибок и точность уменьшается. Например, часто наблюдается, что точность определения образцов, взятых из разных куч (пробоотбор) или измеренных в разные дни, хуже, чем образцов, взятых из одной кучи или измеренных в один день. В англоязычной литературе это явление называют «точность *day-to-day*». Дополнительные источники вариаций не всегда случайны. Источником ошибки может быть нестабильность реагентов или старение части оборудования, это уже относится к систематическим ошибкам. Такая зависимость ошибки от времени известна как *дрейф*.

Когда измерения проводятся с одним и тем же образцом в разных лабораториях, каждая со своим оборудованием, персоналом и т. д., то часто получают случайный разброс с нормальным распределением, но с большей дисперсией, чем при измерениях в одной лаборатории.

Очень часто вновь разработанный метод, который начинает использоваться в других лабораториях, дает плохой результат. Общая точность метода состоит из двух частей: внутрилабораторная ошибка и межлабораторная ошибка. Известно, что межлабораторная ошибка обычно больше, чем внутрилабораторная. Исследования показали, что никакое достаточно детальное описание метода не изменит этого соотношения ошибок.

Существуют различные варианты статистических методов выявления систематических ошибок. Например, можно сделать некоторые выводы при сравнении двух и более процедур измерения.

Допустим, разработана методика, позволяющая измерять некоторую величину с лучшей точностью. Внутри метода проверить наличие систематической ошибки мы не можем. Необходимо сравнение с ранее применявшимися методами. Подобная задача может возникнуть и при сравнении со стандартами. Сравнение двух методик измерения полезно не только для выявления систематической ошибки. Можно также выяснить, какой из методов является более точным.

Простейший способ выявления ошибки – проанализировать стандарт, для которого измеряемая величина известна с хорошей точностью. Если стандарта нет, то проводят сравнение с «тестовым методом» для одного и того же образца или с образцом сравнения.

На примере сравнения двух процедур можно продемонстрировать применение некоторых широко распространенных статистических методов.

Существуют разные подходы к сравнению.

- Нахождение корреляции. В идеале результаты, полученные обоими методами, должны сильно коррелировать, т. е. коэффициент корреляции должен быть близок к 1.
- Проверка статистической гипотезы относительно результатов измерений: а) *t-критерии* (предполагается, что ошибки распределены нормально); б) *непараметрические критерии* (когда нет уверенности в нормальном распределении).
- *Регрессионный анализ* используется для выявления взаимозависимости двух переменных. Позволяет определить вид систематической ошибки, является ли она аддитивной или мультипликативной.
- *Дисперсионный анализ.* Кроме сравнения двух процедур может возникнуть задача проверки адекватности множества процедур.

3.4.1. Использование *t*-теста

При проверке статистической гипотезы с помощью *t*-теста во всех нижеперечисленных вариантах предполагается, что выборочный *t*-критерий подчиняется распределению Стьюдента.

Сравнение со стандартом. Допустим, нам известно с высокой точностью значение интересующей нас величины для стандарта. На языке статистики дано μ_0, $\sigma \to 0$. Статистически нужно установить, можно ли выборку объемом N и со средним x и дисперсией s^2 рассматривать как взятую из генеральной совокупности с математическим ожиданием μ_0.

Выдвигается нуль-гипотеза H_0 и альтернативная ей H_1:

$$H_0: \; x = \mu_0;$$
$$H_1: \; x \neq \mu_0.$$

Так как нам не важно, в какую сторону результат отклоняется от истинного, используется двусторонняя альтернативная гипотеза. Если объем выборки меньше 30, то для проверки этой гипотезы необходимо брать распределение Стьюдента. Критерий находится по формуле

$$t = \frac{x - \mu_0}{s/\sqrt{N}}.$$

Критическое значение $t_{N\text{-}1,\alpha/2}$ находим по распределению Стьюдента с (N–1) степенями свободы и уровнем риска α. Так как гипотеза двусторонняя, берем двухвостовое распределение, т. е. вероятность распределена на два хвоста, площадь каждого $\alpha/2$.

Область принятия гипотезы:

$$-t_{N-1,\alpha/2} \le t < t_{N-1,\alpha/2}\,.$$

Если вычисленный критерий t попадает в область принятия гипотезы, то отвегнуть нуль-гипотезу мы не можем и можно считать, что на данном уровне значимости измеренное этим методом значение соответствует значению стандарта.

Если объем $N > 30$, можно вместо t-теста использовать z-тест. Критерий вычисляется по той же фомуле, только процентные точки для построения критической области находятся по нормальному стандартизованному распределению.

Измерение двух независимых образцов. Вопрос ставится так: взяты ли два образца из одной и той же генеральной совокупности с одинаковым средним? В этом случае стандартное отклонение образцов нельзя считать пренебрежимо маленьким. Здесь возможны два варианта: а) дисперсии одинаковые, б) дисперсии разные. Чаще всего при проверке новой методики выбирается случай с разными дисперсиями. Предполагается, что выборки независимые. Гипотеза выдвигается в следующем виде:

$$\mathrm{H}_0\colon\ \mu_1 = \mu_2\,;$$
$$\mathrm{H}_1\colon\ \mu_1 \ne \mu_2\,.$$

Критерий находится по приближенной формуле

$$t = \frac{x_1 - x_2}{\sqrt{\dfrac{s_1^2}{N_1} + \dfrac{s_2^2}{N_2}}}\,.$$

Количество степеней свободы для определения процентной точки вычисляется по формуле

$$n \cong \frac{\left(s_1^2/N_1 + s_2^2/N_2\right)^2}{\dfrac{\left(s_1^2/N_1\right)^2}{N_1 - 1} + \dfrac{\left(s_2^2/N_2\right)^2}{N_2 - 1}}\,,$$

затем округляется до целого.

Дальше все так же, как и в предыдущем пункте: находится критическое значение, строятся области принятия и отвержения гипотезы, сравнивается с вычисленным критерием, делаются статистические (можно или нет отвергнуть нулевую гипотезу) и практические выводы (дает ли новый метод правильный результат, есть или нет систематическая ошибка, можно ли новый метод использовать в эксперименте).

Сравнение двух парных выборок. Допустим, используются процедуры, не разрушающие образец, тогда один и тот же образец можно измерить двумя разными методами. В этом случае используется парный *t*-тест для средних, если объем выборки меньше 30. При большом объеме выборки можно использовать *z*-тест.

Вычисляется разность двух парных значений $d_i = x_{1i} - x_{2i}$ и по этой новой выборке находятся среднее $\bar{d}$ и стандартное отклонение s_d. Выдвигаются гипотезы:

$$H_0: \mu_d = 0;$$
$$H_1: \mu_d \neq 0.$$

Критерий находится по формуле

$$t = \frac{\bar{d}}{s_d / \sqrt{N}}.$$

По распределению Стьюдента с $(N - 1)$ степенями свободы находятся процентные точки для выбранного уровня значимости и строится критическая область. Принимается решение.

Сравнение дисперсий. Сопутствующей проверкой может быть проверка равенства дисперсий. Для этого используется *F*-тест. Выдвигается гипотеза:

$$H_0: \sigma_1 = \sigma_2;$$
$$H_1: \sigma_1 > \sigma_2.$$

Предполагается, что выборки независимые. Выдвигается односторонняя гипотеза, так как мы должны сделать вывод, достоверно ли одна дисперсия больше другой. Вычисляется критерий

$$F = \frac{s_1^2}{s_2^2}.$$

В числитель ставится большая дисперсия, в знаменатель – меньшая. Процентная точка находится по *F*-распределению со степенями свободы $N_1 - 1$ и $N_2 - 1$. Если вычисленный критерий лежит правее процентной точки, то нуль-гипотеза отвергается на уровне значимости α, если левее, то нуль-гипотезу отвергнуть нельзя.

Если выборки зависимы (коэффициент корреляции значимо отличается от нуля), то для сравнения дисперсий лучше использовать *t*-критерий, который проверяется по *t*-распределению:

$$t = \frac{s_1^2 - s_2^2}{\sqrt{\frac{4s_1^2 s_2^2}{N-2}(1-r_{12}^2)}},$$

r_{12} – коэффициент корреляции.

Можно проверять и двустороннюю гипотезу относительно равенства дисперсий. Тем не менее односторонняя гипотеза позволяет доказать, что одна из дисперсий больше другой, что существенно, если мы хотим сделать вывод о том, какой метод из двух сравниваемых методов точнее и дает лучший результат.

3.4.2. Непараметрические тесты

Все предыдущие тесты основаны на предположении о нормальном распределении величин. Но не во всех случаях можно утверждать, что измеряемая величина распределена нормально. Одной из причин отклонения от гауссианы может быть гетерогенность образцов, произведенных в виде гранул (ненепрерывное распределение по размерам) и измеренных вблизи предела чувствительности детектора. В таких случаях можно использовать методы, свободные от распределений. В этих методах не требуется вычислять обычные оценки, такие как среднее и дисперсия. Однако такие тесты менее эффективны и при прочих равных условиях требуется больший объем выборки для обеспечения определенного уровня доверия, чем при тестах, основанных на нормальном распределении. С другой стороны, непараметрические методы более общие и могут использоваться и для данных, измеренных не в количественных шкалах (порядковые, номинальные). А параметрические тесты в обязательном порядке требуют измерений в количественных шкалах (интервалов, отношений).

В основе непараметрических методов, как правило, лежит процедура ранжирования выборки. Выборка сортируется, и каждому значению приписывается ранг, одинаковым значениям присваиваются обязательно одинаковые ранги. Последний ранг должен совпадать с количеством объектов или измерений. В электронных таблицах существует функция ранжирования, которая особенно удобна для больших выборок.

В качестве примеров непараметрических тестов выбраны *U*-тест для независимых измерений и тест Вилкоксона для парных измерений.

U-тест. Две группы измерений (они могут быть разного объема) ранжируются как одна группа. Находится сумма рангов по первой (R_1) и по второй (R_2) группе. Затем вычисляются величины:

$$U_1 = N_1 N_2 + \frac{N_1(N_1+1)}{2} - R_1;$$

$$U_1 = N_1 N_2 + \frac{N_2(N_2+1)}{2} - R_2.$$

Выдвигается нулевая гипотеза о равенстве этих значений и альтернативная двусторонняя:

$$H_0: U_1 = U_2;$$
$$H_1: U_1 \neq U_2.$$

В качестве критерия берется наименьшее из двух значений:

$$U = \min(U_1, U_2).$$

Отметим, что $U_1 + U_2 = N_1 N_2$. Дальше находим критическое значение по таблице для U-теста (см. приложение 1) на заданном уровне значимости с соответствующими степенями свободы. Затем идет обычная процедура проверки гипотезы. Если вычисленный критерий U меньше критического значения, то гипотеза отвергается.

Тест Вилкоксона. Находится разность парных значений d_i.(i = 1, …, N) Новая выборка ранжируется по абсолютному значению разности (знак разности при этом не учитывается). Далее, каждому рангу приписывается знак разности d_i, находится сумма положительных рангов (T^+) и сумма отрицательных рангов (T^-). Если проверяемые методы одинаковы, то следует ожидать близких значений T^+ и T^-. Критическое значение находится по таблице Вилкоксона (см. приложение 2) и сравнивается с T:

$$T = \min(T^+, T^-).$$

Гипотеза отвергается, если вычисленный критерий меньше критического значения. Существуют справочники, в которых таблицы для теста Вилкоксона основаны на выборе максимального значения, в этом случае гипотеза отвергается, если критерий (максимальное значение из двух) больше критического значения, выбранного на заданном уровне значимости.

3.4.3. Сравнение двух процедур методами регрессионного анализа

Одно из возможных применений регрессионного анализа – сравнение двух методов. В идеале при отсутствии ошибок результаты измерения нескольких образцов тестовой процедурой y и проверяемой (новой) процедурой x должны ложиться на прямую $y = x$, наклон прямой b равен 1, а константа a равна 0. Рассмотрим действие различного рода ошибок. Наличие случайных ошибок ведет к рассеянию точек вокруг линии регрессии, коэффициенты регрессии при этом остаются истинными. Случайную

ошибку при этом можно оценить, вычислив стандартное отклонение y по x, $s_{y/x}$, т. е. отклонение y от линии регрессии.

Относительная или мультипликативная систематическая ошибка ведет к изменению наклона прямой. Отклонение b от единицы дает оценку относительной ошибки. Абсолютная или аддитивная систематическая ошибка приводит к отклонению коэффициента a от нуля. Таким образом, регрессионный анализ способен выявить различные виды систематических ошибок.

Построим линию регрессии по выборке с помощью метода наименьших квадратов. Если экспериментальная оценка α близка к нулю, а оценка β – к 1, можно сделать вывод об отсутствии систематических ошибок. Для этого требуется проверить две статистические гипотезы. Если объем выборки мал, можно использовать t-критерий. В данном конкретном случае для проверки гипотезы о том, что $\beta = 1$, вычисляем

$$t' = \frac{b - \beta}{\sqrt{1 - r^2}} \sqrt{n - 2},$$

где $n - 2$ – число степеней свободы распределения Стьюдента для данного критерия (фиксируется два параметра a и b), r – коэффициент корреляции.

Чтобы проверить значимость отклонения a от нуля, проверяем гипотезу $\alpha = 0$, вычисляем критерий

$$t'' = \frac{a - \alpha}{s_a}.$$

Для этого критерия число степеней свободы также $n - 2$. Дальше в соответствии с выбранным уровнем значимости строим критические области по распределению Стьюдента и делаем выводы.

Проверка гипотезы равенства констант линейного уравнения определенному значению может быть использована, например, при определении размеров частиц и микронапряжений по данным порошковой дифрактометрии.

3.4.4. Сравнение нескольких процедур методами дисперсионного анализа

Кроме сравнения двух процедур измерений может возникнуть задача проверки адекватности множества процедур. Простой пример – сравнение измерений, проведенных по одной и той же методике в разных лабораториях. В этом случае часто используют дисперсионный анализ ANOVA.

Основная проблема при применении ANOVA – определить, какая часть дисперсии в генеральной совокупности происходит вследствие системати-

ческих причин (так называемых факторов), а какая из-за случайного разброса. При сравнении нескольких процедур каждая из них может иметь систематическую ошибку. Кроме сравнения измерительных процедур дисперсионный анализ широко используется для определения источника вариации данных.

Допустим, есть предположение о том, какой именно фактор влияет на проведение измерений (влажность, освещенность, взаимное расположение измерительной аппаратуры, среднее внешнее давление в районе расположения лаборатории и т. д.). В этом случае мы фиксируем фактор и проводим измерения при различных уровнях проявления этого фактора (контролируемый фактор). Проверяем, влияет ли этот фактор на разброс данных. В этом случае используется однофакторный дисперсионный анализ.

Другой пример: проводятся измерения в разных лабораториях с различным образом приготовленными образцами. Проверяется одновременно и влияние лаборатории, и влияние подготовки образца. В этом случае мы имеем дело с двухфакторным дисперсионным анализом. Два варианта двухфакторного дисперсионного анализа (без повторений и с повторениями) различаются тем, как были проведены измерения на каждом уровне двух факторов. В приведенном примере может возникнуть вопрос перекрестного влияния лаборатории и подготовки образца. Допустим, в одной лаборатории лучше анализируют мелкодисперсный порошок, в другой – гранулированный образец. Дисперсионный анализ позволяет определить это влияние, если измерения были проведены с повторениями (см. табл. 9).

В специализированных программах различные варианты дисперсионного анализа реализованы в виде отдельных процедур. В электронных таблицах можно найти отдельные функции, которые позволяют провести дисперсионный анализ. В таблицах Exel есть возможность подключить пакет анализа данных, в котором есть процедуры однофакторного и двух вариантов двухфакторного дисперсионного анализа. В электронных таблицах OpenOffice такого пакета анализа нет и приходится делать расчеты по формулам, приведенным в табл. 7, 8, 10.

Дисперсионный анализ часто используется как сопутствующий метод в некоторых других методах многомерного анализа данных.

3.5. Калибровочные процедуры

Важным этапом практически любого эксперимента в химии является *калибровка* (иногда в русскоязычной литературе используют термин «градуировка»). Хорошая точность может быть получена только тогда, когда хорошо откалиброваны приборы и измерительные процедуры.

Как уже отмечалось, химия обычно имеет дело с косвенными измерениями. В редких случаях величина измеряется напрямую. Чаще всего

необходимо дополнительное преобразование сигнала в химическую информацию (например, измерение концентрации компонента в образце). Естественно, физический сигнал y, с помощью которого измеряется величина x, должен быть функционально связан с x, $y = g(x)$. Тогда можно построить обратную калибровочную функцию. Калибровочную функцию можно получить из экспериментальных данных, аппроксимируя их адекватной математической моделью.

Например, чтобы откалибровать термометр, соотносят 0 и 100 °С к высотам столбика термометра (ртутного или спиртового) в точках замерзания и кипения воды при нормальных условиях.

Как уже говорилось, анализ остатков позволяет сделать некоторые выводы о пригодности модели. В частности, можно утверждать, что линейная модель адекватна, если дисперсии остатков не зависят от величины x (гомоскедактичность). Если мы имеем дело с гетероскедактичностью, то возможны варианты преобразования первичных данных и перехода к гомоскедактичному случаю. Один из вариантов преобразования – это функциональное преобразование, например, $y = a + b\exp(x)$ или $\ln(y) = a + b\ln(x)$ и т. д.

Другой вариант поправки на гетероскедактичность – приписывание каждому значению x_i определенного веса. В качестве весов, например, можно взять значения, обратно пропорциональные дисперсиям ($w_i = 1/s_{y_i}^2$). В этом случае

$$Q = \sum_i w_i (y_i - a - bx_i)^2 ,$$

$$x = \frac{\sum_i w_i x_i}{\sum_i w_i}, \quad y = \frac{\sum_i w_i y_i}{\sum_i w_i}.$$

При построении градуировочной прямой строятся доверительные интервалы для линии регрессии и для индивидуальных значений.

Этапы процедуры градуировки

1. Выбор модели.
2. Нахождение параметров модели с помощью МНК.
3. Проверка достоверности модели.
4. Вычисление доверительных интервалов параметров.

Отклонения от линейности не всегда очевидны. Один из способов проверки линейности регрессионной модели – дисперсионный анализ. В этом случае общая сумма квадратов отклонений (SS_t) разделяется на три слага-

емых, первое соответствует отклонению индивидуального значения от среднего значения в группе (SS_w), второе – отклонению среднего по группе от оценки по модели (SS_{lof}) и третье – отклонение оценки по модели от общего среднего (SS_{reg}):

$$\underset{SS_t}{\sum_i \sum_j (y_{ij} - y)^2} = \underset{SS_w}{\sum_i \sum_j (y_{ij} - y_i)^2} + \underset{SS_{lof}}{\sum_i n_i (y_i - \hat{y}_i)^2} + \underset{SS_{reg}}{\sum_i n_i (\hat{y}_i - y)^2},$$

$i = 1, \ldots, k$; $j = 1, \ldots, n_i$.

Тест на линейность соответствует отношению дисперсии средних значений по группе относительно линии регрессии к дисперсии в группе (см. табл. 11). Для проведения такого теста необходимо, чтобы в каждой точке по x было проведено несколько измерений y.

Таблица 11

Дисперсионный анализ в случае линейной регрессии

Источник вариации	*SS*	*df*	*MS*	*F*	*F критическое*
Из-за регрессии	SS_{reg}	1	MS_{reg}	$\frac{MS_{reg}}{MS_r}$	$F_{1-\alpha,dfr,dfw}$
Отклонение от линии регресии в группе	SS_{lof}	$k-2$	MS_{lof}	$\frac{MS_{lof}}{MS_w}$	$F_{1-\alpha,k-2,dfw}$
Внутри групп	SS_w	$\sum_i n_i - k$	MS_w		
Итого (все группы вместе)	SS_t	$\sum_i n_i - 1$			

MS_r соответствует $SS_r = SS_t - SS_{reg} = SS_{lof} + SS_w$.

Значение критерия F во второй строке таблицы позволяет проверить гипотезу о линейности модели. Если значение попадает в область отвержения нулевой гипотезы, то калибровочная зависимость не является линейной и необходимо провести дополнительные преобразования перехода к гомоскедактичности.

Глава 4

Примеры заданий

Задание 1. Построение гистограммы. Дано 200 измерений некоторой величины (например, длины однотипных межатомных связей в разных кристаллических структурах):

1,160	1,175	1,181	1,185	1,189	1,193	1,197	1,200	1,205	1,211
1,163	1,176	1,181	1,185	1,189	1,193	1,197	1,201	1,206	1,211
1,164	1,176	1,181	1,185	1,189	1,193	1,197	1,202	1,206	1,212
1,165	1,177	1,182	1,186	1,189	1,193	1,197	1,202	1,207	1,213
1,166	1,177	1,182	1,186	1,189	1,193	1,197	1,202	1,207	1,213
1,166	1,178	1,182	1,186	1,190	1,193	1,197	1,202	1,207	1,214
1,167	1,178	1,183	1,186	1,190	1,193	1,197	1,202	1,208	1,215
1,168	1,178	1,183	1,187	1,190	1,194	1,198	1,202	1,208	1,215
1,168	1,178	1,183	1,187	1,190	1,194	1,198	1,202	1,208	1,215
1,169	1,178	1,183	1,187	1,191	1,194	1,198	1,203	1,208	1,216
1,169	1,178	1,183	1,187	1,191	1,194	1,198	1,203	1,208	1,217
1,170	1,179	1,183	1,187	1,191	1,194	1,199	1,203	1,208	1,218
1,171	1,179	1,184	1,188	1,191	1,195	1,199	1,203	1,208	1,219
1,172	1,179	1,184	1,188	1,191	1,195	1,199	1,204	1,209	1,219
1,173	1,180	1,184	1,188	1,191	1,196	1,199	1,204	1,209	1,220
1,173	1,180	1,184	1,188	1,192	1,196	1,199	1,204	1,210	1,221
1,174	1,180	1,184	1,188	1,192	1,196	1,199	1,205	1,210	1,222
1,175	1,180	1,185	1,189	1,192	1,196	1,200	1,205	1,210	1,223
1,175	1,181	1,185	1,189	1,192	1,197	1,200	1,205	1,210	1,225
1,175	1,181	1,185	1,189	1,193	1,197	1,200	1,205	1,210	1,230

Определить среднее значение, дисперсию и стандартное отклонение величины (использовать соответствующие функции в электронных таблицах).

Построить частотную таблицу. Для построения интервалов (карманов) нужно определить шаг. Это можно сделать двумя способами: либо размах данных (макс. значение минус мин. значение) разделить на количество интервалов (~12), либо оценить как 0,4 от стандартного отклонения. Создается новая колонка, в которую записываются верхние границы интервалов. Граница первого кармана выбирается так, чтобы в него попало несколько значений (больше одного). В электронных таблицах принято считать, что в данный интервал (карман) попадают все измерения, значения которых меньше или равны значению в соседней ячейке, но больше значения в предыдущей ячейке. Первый интервал имеет нижнюю границу

–∞. Последний карман имеет верхнюю границу +∞ и также должен содержать несколько значений.

По частотной таблице построить гистограмму. Является ли распределение одномодальным?

Задание 2. Элементарные вычисления и построение диаграмм рассеяния. Дана таблица изменения параметров элементарной ячейки ромбической модификации парацетамола в зависимости от температуры. Провести анализ этих данных:

- внести данные в электронную таблицу (стандартные отклонения, которые стоят в скобках, внести в отдельную таблицу);
- построить график изменения объема в зависимости от температуры, правильно оформить;
- вычислить относительные изменения параметров ячейки (за исходное значение взять значение при T = 300 К); вычислить ошибки относительных изменений с учетом законов распространения погрешностей (взять формулу для предела погрешности);
- построить на одном графике относительные изменения всех параметров ячейки, правильно оформить.

Параметры элементарной ячейки ромбического парацетамола при разных температурах

T, К	a, Å	b, Å	c, Å	V, Å^3
100	7,1986(17)	11,782(10)	17,183(4)	1457,4(10)
150	7,2433(18)	11,793(11)	17,175(5)	1467,1(11)
200	7,2927(14)	11,806(8)	17,169(3)	1478,3(8)
250	7,3467(14)	11,818(8)	17,165(3)	1490,2(8)
260	7,3583(12)	11,826(7)	17,163(3)	1493,5(7)
270	7,3697(15)	11,823(9)	17,165(4)	1495,6(9)
280	7,3812(16)	11,830(9)	17,164(4)	1498,8(9)
300	7,4049(13)	11,835(8)	17,162(3)	1504,3(8)
320	7,4308(13)	11,839(7)	17,163(3)	1509,9(7)
330	7,4430(16)	11,847(9)	17,160(4)	1513,2(9)
340	7,4556(16)	11,853(9)	17,159(4)	1516,3(9)
350	7,4683(15)	11,855(9)	17,158(4)	1519,1(9)
360	7,4807(15)	11,853(9)	17,160(4)	1521,6(9)

Формула расчета относительных изменений параметра в процентах:

$$\Delta l / l(\%) = \frac{l - l_0}{l_0} * 100.$$

Сделать выводы: носит ли изменение объема линейный характер в исследуемом диапазоне температур, положительные или отрицательные коэффициенты термического расширения (объемный и линейные), в каком направлении структура деформируется сильнее всего, существует ли направление, в котором деформация отсутствует?

Задание 3. Работа с текстовой информацией, сортировка и фильтрация. Импортировать текстовый файл с исходными данными с монокристального четырехкружного дифрактометра STADI4 в таблицу. Обратить внимание на условия импорта. Чтобы данные распределились по колонкам, нужно поставить разделителем пробел.

Ниже приведено значение каждой колонки данных в файле дифрактометра и расшифровка статуса рефлекса.

Формат сырых данных:

H	K	L	Stat	Int	Sigma	Psi	W	T	и т. д.
4	0	0	40	4800.6	20.7	0	128	16	и т. д.
0	4	0	40	4482.6	19.9	0	128	16	и т. д.
0	0	2	40	3747.7	17.9	0	128	16	и т. д.

где H, K, L – индексы Миллера, Stat – статус рефлекса (см. ниже), Int – интегральная интенсивность, Sigma – стандартное отклонение интенсивности, Psi – угол, W – ширина сканирования (× 100 °), T – общее время съемки рефлекса в секундах.

Статус рефлекса в файле данных – это число, состоящее из двух слагаемых (Тип + Код), где Код может быть комбинацией (суммой) более чем одного условия:

Тип =	Стандартный рефлекс	0 × 40
	Psi-сканирование	0 × 20
	Фриделевская пара при –2θ	0 × 10
Код =	плохая апроксимация профиля	0 × 01
	неодинаковый фон	0 × 02
	слишком широкий рефлекс	0 × 04
	рефлекс не в центре	0 × 08

Дальнейшая информация в строке несущественна в плане выполнения задания. Отметим только, что после буквы P в строке идет профиль рефлекса (интенсивность в зависимости от угла сканирования). По импортированным данным выполнить следующие задания.

- Скопировать рефлексы без профилей и углов (до колонки угла пси) на отдельный лист, поставить автофильтр. Пользуясь автофильтром, посчитать, сколько рефлексов имеют неодинаковый фон, сколько слишком широких рефлексов, сколько находилось не в центре при сканировании.
- Посчитать общее количество измеренных рефлексов без учета стандартов (отсортировав по условию, скопировать этот массив на новый лист и воспользоваться функцией «счет»), найти отношение средней интенсивности к среднему значению sigma. Найти количество сильных рефлексов (I/sigma > 4). Для этого в колонке рядом вычислить отношение I/sigma и поставить фильтр. Отсортировав по условию, скопировать этот массив на новый лист и воспользоваться функцией «счет».
- Навести статистику по сильным рефлексам (по каким группам рефлексов есть погасания). По погасаниям определить возможные ПГС кристаллической структуры.

Задание 4. Графическое представление нормального распределения. Построить график стандартного нормального распределения. Задать значения по x от –3 до 3 с шагом 0,1. Построить интегральное распределение, воспользовавшись функцией нормального распределения в электронных таблицах. Построить дифференциальное распределение (плотность вероятности), поставив соответствующий статус в функции нормального распределения.

Задание 5. Определение численных значений нормального распределения по таблицам и их интерпретация. Используя функции электронных таблиц, ответить на следующие вопросы.

- Какова вероятность того, что случайное измерение (нормальное распределение) попадет в интервал $\pm1\sigma$, $\pm2\sigma$, $\pm3\sigma$?
- Пусть среднее значение равно 100, стандартное отклонение равно 15 (нормальное распределение). Какова вероятность того, что мы получили значение 120 и выше случайно? А 140 и выше?
- Для стандартного нормального распределения $N(0, 1)$ найти значение процентной точки z_α, если известна вероятность α случайно получить значение $x > z_\alpha$: $\alpha = 0{,}1$; $\alpha = 0{,}05$; $\alpha = 0{,}01$?
- Найти значение процентной точки $z_{\alpha/2}$, если известна вероятность случайно получить значение x в интервалах либо больше $z_{\alpha/2}$, либо меньше $-z_{\alpha/2}$ (двусторонний случай) для тех же вероятностей α.

Задание 6. Различные виды распределений случайных величин, основанных на нормальном, и их представление в электронных таблицах.

Построить график хи-квадрат-распределения (степень свободы задать в отдельной ячейке и использовать ее как аргумент в формуле). Задать значения по *x* от 0 до 10 с шагом 0,1. Меняя степень свободы, проследить, как меняется положение максимума.

Какова вероятность того, что случайное измерение при объеме выборки 8 попадет в интервал $\pm 1\sigma$, $\pm 2\sigma$, $\pm 3\sigma$ (использовать распределение Стьюдента)? Как будет меняться эта вероятность при увеличении объема выборки? При каком объеме выборки различие между нормальным распределением и распределением Стьюдента станет меньше 1 % для интервала $\pm 2\sigma$?

Допустим, что дисперсии двух выборок одинакового объема одинаковы. Какова вероятность случайно получить оценку отношения дисперсий 1, 3, 12 (взять объем выборки 10)? Для ответа воспользоваться табличными значениями *F*-распределения.

Задание 7. Интервальное оценивание (построение доверительных интервалов). Анализируется международный стандарт вещества новым методом. Известна концентрация $\mu_0 = 0{,}500$ мг/г. Получили следующие результаты:

0,559; 0,594; 0,531; 0,465; 0,509; 0,487; 0,421; 0,328; 0,428; 0,601; 0,450; 0,440; 0,409; 0,419; 0,357.

Построить доверительные интервалы для среднего и для дисперсии. Попадает ли истинное значение в этот доверительный интервал? Можно ли считать, что новый метод дает удовлетворительный результат на уровне значимости 5 %, 1 % (проверка статистической гипотезы, сравнение со стандартом)?

Задание 8. Проверка соответствия распределений. Проверить по критерию согласия хи-квадрат соответствие экспериментального и нормального распределений (по таблице из задания 1). Сделать это двумя способами: а) вычислить по формулам, которые даны в теоретической части и проверить соответствующую статистическую гипотезу по хи-квадрат-распределению; б) воспользоваться статистической функцией хи-тест. Сравнить результаты. Пояснение: хи-тест выдает значение вероятности, соответствующее критерию (площади «хвоста» функции плотности вероятности распределения).

Задание 9. Ошибка второго рода при проверке статистической гипотезы. Предположим, есть основания считать, что среднее значение случайной величины x известно (μ_x = 8). Предположим далее, что дисперсия x также известна (σ_x^2 = 2). Найти размер выборки, позволяющей построить критерий проверки гипотезы с 5 %-м уровнем значимости и 5 %-й ошибкой второго рода для выявления 10 %-х отклонений от гипотетического значения. Построить область принятия гипотезы для данного критерия.

Задание 10. Проверка статистических гипотез, сравнение двух серий измерений. Результаты анализа двухфазного образца с известным содержанием фаз 1 : 1 двумя разными методами (одна из фаз в %):

Проба	Метод А	Метод В
1	49,9	50,1
2	49,9	49,5
3	50,0	50,3
4	50,1	49,6
5	50,0	49,9
6	50,0	50,2

Дают ли методы одинаковый результат? Есть ли систематические ошибки методов? Какой метод точнее? (Проверить, одинаковы ли дисперсии этих измерений по *F*-тесту).

Задание 11. Проверка статистических гипотез, сравнение двух серий измерений. Содержание полиморфной модификации А в 10 анализируемых образцах, определенное двумя методами (*R* и *T*), дано в таблице:

Образец	X_R	X_T
1	57	58
2	25	21
3	50	48
4	10	5
5	45	47
6	53	50
7	50	48
8	48	51
9	80	75
10	55	52

Дает ли проверяемый метод *R* такой же результат, как и тестовый (*t*-тест)? Точнее ли он?

Задание 12. Проверка статистических гипотез, сравнение двух серий измерений. Концентрация активного вещества в 10 пробах до и после обработки дана в таблице:

Проба	До	После
1	20,5	21,3
2	19,9	20,4
3	21,1	20,8
4	21,5	22,1
5	20,0	20,7
6	19,1	18,9
7	20,1	20,5
8	20,5	20,4
9	19,4	19,7
10	19,1	19,4

Эффективна ли обработка, если цель обработки увеличить содержание активного вещества (сравнение двух парных измерений)? Проверить гипотезу на разных уровнях значимости.

Задание 13. Непараметрические методы сравнения, тест Вилкоксона. Содержание полиморфной модификации А в 10 анализируемых образцах, определенное двумя методами (*R* и *T*), дано в таблице задания 11. Проверить по тесту Вилкоксона, дает ли проверяемый метод *R* такой же результат, как и тестовый *T*?

Задание 14. Непараметрические методы сравнения, *U*-тест. Сравнить две группы измерений, дают ли они одинаковый результат?

A: 1,34; 1,64; 1,78; 1,33; 1,80; 1,93; 2,08; 1,31; 1,30; 1,40.
B: 1,31; 1,34; 1,45; 1,49; 1,86; 1,75; 1,62; 1,30.

Для проверки использовать *U*-тест.

Задание 15. Сравнение двух процедур, регрессионный анализ. По таблице из задания 11 проверить с помощью регрессионного анализа, дает ли проверяемый метод *R* такой же результат, как и тестовый? Есть ли систематическая ошибка? Проанализировать возможный тип систематической

ошибки (проверить соответствующую статистическую гипотезу на каждый тип).

Задание 16. Регрессионный анализ. Сделать выборку измеренных стандартов из задания 3. Вычислить относительные изменения интенсивностей стандартов (отсчитывать от первого стандарта из группы). Есть или нет тренд (тенденция падения или возрастания) стандартов? Сравнить с усредненной по всем стандартам зависимостью (вычислить наклоны линий регрессии и ошибки, сделать выводы).

Задание 17. Однофакторный дисперсионный анализ. Определено содержание действующего компонента М в семи лабораториях:

а	б	в	г	д	е	ж
16	46	12	15	60	62	33
29	28	19	27	39	38	38
35	30	29	34	43	55	55
18	45	11	20	58	42	49
22	31	29	34	40	53	45

Дают ли все лаборатории одинаковый результат (есть или нет систематическая лабораторная ошибка)?

Дополнительно. Провести ANOVA в программе Statistica. В этой программе данные должны быть организованы по-иному, в две колонки: колонка измерений и колонка факторов (это переменная, по которой данные группируются).

Задание 18. Двухфакторный дисперсионный анализ без повторений. Концентрация загрязняющего вещества вблизи участка размещения промышленных отходов в зависимости от расстояния от города и от глубины взятия пробы дана в таблице:

Глубина, м	Расстояние от города, км			
	1	2	3	4
0,0	25,0	15,3	10,1	5,2
0,5	23,0	15,2	9,0	4,0
1,0	22,5	13,8	12,5	3,0

Влияет ли местоположение взятия пробы, глубина взятия пробы? Можно ли определить перекрестное влияние?

Задание 19. Двухфакторный дисперсионный анализ с повторениями. Содержание активного вещества в пробах разное. Пробы обрабатываются двумя разными методами. Проверить, одинаковые ли результаты дают разные методы, влияет ли содержание активного вещества в пробах на результат и существует ли перекрестное влияние «метод и концентрация активного вещества»?

	Проба 1	Проба 2	Проба 3
Метод 1	92	101	89
	114	118	120
	107	98	110
	99	105	115
Метод 2	91	101	79
	74	68	88
	65	59	55
	90	70	93

Задание 20. Градуировочные процедуры. Построить градуировочную прямую по данным таблицы, построить доверительные интервалы для линии регрессии.

Данные калибровки:

Конц., x	Пик, y
0,144	21,333
0,280	40,521
0,044	7,702
0,018	3,707
0,232	32,165
0,120	17,094
0,035	4,758
0,022	3,396
0,035	3,829
0,058	9,057
0,104	14,765
0,070	10,285
0,208	28,149
0,128	19,396
0,080	11,340
0,144	21,333

Задание 21. Градуировочные процедуры, анализ причин вариации данных. Провести дисперсионный анализ по таблице для проверки гипотезы о линейности модели. Сделать выводы.

Результаты градуировочного эксперимента:

X	2	4	6	10	20
Y	13,28	17,44	24,80	39,44	65,92
	10,96	17,76	24,24	41,12	68,72
	11,84	17,44	25,84	41,28	68,48
	11,44	17,44	24,96	41,28	67,20

Рекомендуемая литература

Brereton R. G. Chemometrics. Data analysis for the laboratory and chemical plant. Wiley, Chichester, UK, 2003. 489 p.

Massart D. L., Vandeginste B. G. M., Deming S. N., Michotte Y., Kaufman L. Chemometrics: a textbook. Elsevier, 1988. 488 p.

Александров В. В., Алексеев А. И., Горский Н. Д. Анализ данных на ЭВМ (на примере системы СИТО). М.: Финансы и статистика, 1990. 192 с.

Бендат Дж., Пирсол А. Прикладной анализ случайных данных: Пер. с англ. М.: Мир, 1989. 540 с.

Брандт З. Анализ данных. Статистические и вычислительные методы для научных работников и инженеров: Пер. с англ. М.: Мир; АСТ, 2003. 686 с.

Боровков В. STATISTICA: искусство анализа данных на компьютере. Для профессионалов. СПб.: Питер, 2001. 656 с.

Загоруйко Н. Г. Прикладные методы анализа данных и знаний. Новосибирск: Изд-во Ин-та математики, 1999. 270 с.

Измерения в промышленности: Справ. изд: В 3 кн.: Пер. с нем. / Под ред. П. Профоса. 2-е изд., перераб. и доп. М.: Металлургия, 1990. Кн. 1: Теоретические основы. 492 с.

Ким Дж.-О., Мьюллер Ч. У., Клекка У. Р. и др. Факторный, дискриминантный и кластерный анализ: Пер. с англ. М.: Финансы и статистика, 1989. 215 с.

Родионова О. Е., Померанцев А. Л. Хемометрика в аналитической химии. 2006. URL: http://www. chemometrics.ru/materials/articles/ chemometrics_review.pdf.

Родионова О. Е. Хемометрический подход к исследованию больших массивов химических данных // Российский химический журнал (Ж. Рос. хим. об-ва им. Д. И. Менделеева). 2006. Т. 50, № 2. С. 128–144.

Шараф М. А., Иллмэн Д. Л., Ковальски Б. Р. Хемометрика: Пер. с англ. Л.: Химия, 1989. 272 с.

Hughes I. G., Hase T. P. A. Measurements and their uncertainties. Oxford Univ. Press Inc., 2010. 136 p.

Приложение 1

Статистические таблицы для *U*-теста

Если наблюдаемое значение U меньше или равно соответствующему табличному значению, то нуль-гипотеза может быть отклонена на соответствующем уровне значимости.

Критические значения U на уровне риска 5 %.

N_1 \ N_2	3	4	5	6	7	8	9	10	11
3			0	1	1	2	2	3	3
4		0	1	2	3	4	4	5	6
5	0	1	2	3	5	6	7	8	9
6	1	2	3	5	6	8	10	11	13
7	1	3	5	6	8	10	12	14	16
8	2	4	6	8	10	13	15	17	19
9	2	4	7	10	12	15	17	20	23
10	3	5	8	11	14	17	20	23	26
11	3	6	9	13	16	19	23	26	30

Критические значения U на уровне риска 10 %.

N_1 \ N_2	3	4	5	6	7	8	9	10	11
3	0	0	1	2	2	3	3	4	5
4	0	1	2	3	4	5	6	7	8
5	1	2	4	5	6	8	9	11	12
6	2	3	5	7	8	10	12	14	16
7	2	4	6	8	11	13	15	17	19
8	3	5	8	10	13	15	18	20	23
9	3	6	9	12	15	18	21	24	27
10	4	7	11	14	17	20	24	27	31
11	5	8	12	16	19	23	27	31	34

Приложение 2

Статистические таблицы для теста Вилкоксона

Если наблюдаемое значение *T* меньше или равно соответствующему табличному значению, то нуль-гипотеза может быть отклонена на соответствующем уровне значимости (α).

	N					
α	5	6	7	8	9	10
0,10	1	2	4	6	8	11
0,05		1	2	4	6	8
0,02			0	2	3	5
0,01				0	2	3

	N					
α	11	12	13	14	15	16
0,10	14	17	21	26	30	36
0,05	11	14	17	21	25	30
0,02	7	10	13	16	20	24
0,01	5	7	10	13	16	19

	N					
α	17	18	19	20	21	22
0,10	41	47	54	60	68	75
0,05	35	40	46	52	59	66
0,02	28	33	38	43	49	56
0,01	23	28	32	37	43	49

	N					
α	23	24	25	26	27	27
0,10	83	92	101	110	120	130
0,05	73	81	90	98	107	117
0,02	62	69	77	85	93	102
0,01	55	61	68	76	84	92

	N					
α	29	30	31	32	33	34
0,10	141	152	163	175	188	201
0,05	127	137	148	159	171	183
0,02	111	120	130	141	151	162
0,01	100	109	118	128	138	149

Учебное издание

Дребущак Татьяна Николаевна

ВВЕДЕНИЕ В ХЕМОМЕТРИКУ

Учебное пособие

Редактор *К. В. Шмугурова*

Подписано в печать 19.08.2013
Формат 60 × 84 1/16. Офсетная печать.
Уч.-изд. л. 5,6. Усл.-печ. л. 5,2. Тираж 50 экз.

Заказ № 211
Редакционно-издательский центр НГУ.
ул. Пирогова, 2, 630090, Новосибирск.

Т. Н. Дребущак

ВВЕДЕНИЕ В ХЕМОМЕТРИКУ