

$$\sum_{i=1}^n \frac{x_i}{n} = \hat{M}$$

$$\int_{\alpha} g(x) dx = \alpha$$

$$H_1: f(x) = g(x)$$

$$H_2: f(x) = h(x)$$

Р. Л. МИЛЛЕР, ДЖ. С. КАН

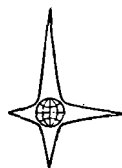
СТАТИСТИЧЕСКИЙ АНАЛИЗ В ГЕОЛОГИЧЕСКИХ НАУКАХ



Книга представляет собой первое пока единственное руководство по применению статистических методов для решения геологических задач.

Доступным для неподготовленного читателя языком в ней изложены основные понятия теории вероятностей и статистические методы, которыми пользуются для получения оценок неизвестных параметров. Наглядно, с привлечением большого количества рисунков и схем, показано применение способов статистической обработки при решении геологических задач.

Книга — учебное и методическое руководство, рассчитанное на геологов и студентов геологических вузов и техникумов, биологов, физико-географов, почвоведов.



ИЗДАТЕЛЬСТВО
« М И Р »

STATISTICAL ANALYSIS IN THE GEOLOGICAL SCIENCES

Robert L. Miller

Department of the Geophysical Sciences
University of Chicago

James Steven Kahn

Lawrence Radiation Laboratory
University of California

JOHN WILEY AND SONS, INC.

NEW YORK — LONDON

1962

Р. Миллер и Дж. Кан

СТАТИСТИЧЕСКИЙ АНАЛИЗ В ГЕОЛОГИЧЕСКИХ НАУКАХ

Перевод с английского
д. а. родионова

Под редакцией д-ра физ.-мат. наук
ю. в. прохорова

ИЗДАТЕЛЬСТВО «МИР»
Москва 1965

Книга представляет собой первое и пока единственное руководство по применению статистических методов для решения геологических задач.

Доступным для неподготовленного читателя языком в ней изложены основные понятия теории вероятностей и статистические методы, которыми пользуются для получения оценок неизвестных параметров. Наглядно, с привлечением большого количества рисунков и схем, показано применение способов статистической обработки при решении геологических задач.

Книга — учебное и методическое руководство, рассчитанное на геологов и студентов геологических вузов и техникумов, биологов, физико-географов, почвоведов.

ПРЕДИСЛОВИЕ

С каждым годом математические методы все интенсивнее проникают в различные области геологии. Если прежде математика применялась широко только в геофизике, в настоящее время она прочно вошла в геохимию, палеонтологию и другие разделы наук о Земле. Более широкое применение математических методов сказалось не только на увеличении объема вычислительных работ, но и на формулировке основных понятий, которые теперь уточняются в такой степени, что становятся возможными формальные определения, положенные в основу этих методов.

Многие геологические выводы основаны на использовании изменчивых эмпирических данных, например результатов химического анализа проб, число которых бывает весьма ограничено. Чтобы избежать ошибочных выводов из этих данных, необходимо применять методы математической статистики. Несомненно, конечно, бывают случаи, что можно прийти к правильному ответу и без привлечения тонких статистических методов, но теперь это стало скорее исключением, чем правилом (см. пример: табл. 6.4, где сравнивается радиоактивность свежих и выветрелых гранитов, когда невозможно решение «на глаз»).

Математическая статистика — наука о математических методах систематизации и обработки статистических данных, о способах извлечения из них наибольшей информации и принятия на этой основе наилучших в том или ином смысле решений. Применяя методы математической статистики в различных областях знания и техники, мы преследуем различные цели, но все они должны быть подчинены основным задачам соответствующих областей. Поэтому стало необходимым выде- **5**

лить такие дисциплины, как социально-экономическая статистика, биометрия и т. п.

Миллер и Кан, авторы предлагаемой вниманию читателей книги, видные американские ученые — геологи, которые заинтересовались возможностью решать геологические задачи с помощью математики. Предлагаемая книга — их первая попытка подвести итоги своей работы, поэтому допущенные ими некоторые ошибки вполне понятны и простительны и отнюдь не умаляют ценности полученных ими результатов. Авторы поставили перед собой цель изложить статистические методы в применении их к геологии; в этом отношении настоящая книга чуть ли не единственное руководство. Правда, имеются книги (такие, как книга Ф. Чейза¹), в которых освещается та или иная специальная проблема, но ни одна из них не может сравниться с книгой Миллера и Кана по степени охвата материала. Согласно замыслу авторов, для понимания книги не требуется предварительного знакомства с математической статистикой или теорией вероятностей. Все необходимые сведения авторы старались сообщить в книге. Однако им это удалось не в полной мере. Так, вопросы, связанные с понятием мощности статистических критериев, изложены подробно и понятно. Но в разделах, посвященных дисперсионному и мультивариантному анализам, материал изложен порой столь конспективно, что становится труден даже для подготовленного читателя. Поэтому в процессе чтения необходимо обращаться к специальной литературе по математической статистике. Следует также критически относиться к рекомендациям, даваемым авторами для малых выборок. Принятие или отклонение проверяемой статистической гипотезы может сильно зависеть от априорных допущений об исходных законах распределения (нормальность, логарифмическая нормальность, равномерность и т. п.), а обоснование этих последних в условиях малых выборок — задача подчас нелегкая.

Некоторые примеры подобраны авторами неудачно. В одних ответ ясен и без обращения к математической статистике (например, добавление о критерии Колмогорова — Смирнова); в других конечная цель применения статистических методов

¹ См. Ф. Чейз, Количественно-минералогический анализ шлифов под микроскопом, М., ИЛ, 1963.

останется не ясной, если мы не обратимся к источнику, из которого пример заимствован (пример в конце гл. 16).

Кроме того, к недостаткам книги следует отнести отсутствие необходимых статистических таблиц, что тоже затрудняет ее использование в полной мере. Как можно понять по ссылкам, авторам следовало бы поместить таблицы, приложенные к книге Диксона и Мессе (W. J. Dixon, F. J. Massey, *Introduction to Statistical Analysis*, New York, 1957). Из числа изданных у нас таблиц можно рекомендовать книгу Я. Янко, *Математико-статистические таблицы*, М., Госстатиздат, 1961.

Вероятно, поучительны были бы примеры неправильного применения статистических методов к геологическим задачам и их разбор.

Книга имеет и некоторые другие недостатки, на которых, однако, здесь нет необходимости останавливаться.

Переводчик и редактор позволили себе сделать небольшие сокращения некоторых несущественных мест и исправления замеченных неточностей и ошибок.

В редактировании книги принимал участие А. Г. Модель, который также внес исправления в ряд выкладок и формулировок.

Несмотря на сделанные выше критические замечания, выход в свет перевода книги Миллера и Кана, несомненно, принесет большую пользу широкому кругу советских геологов.

*Ю. Прохоров
Д. Родионов*

ПРЕДИСЛОВИЕ АВТОРОВ

Историю применения статистических методов в науке о Земле можно условно разделить на три этапа.

Первый этап охватывает период приблизительно с 1890 г. до начала тридцатых годов нашего века. В это время в Европе пользовались большим влиянием многочисленные работы Карла Пирсона. Сборник его статей, объединенных под общим названием «Вопросы математической теории эволюции», был опубликован в серии А «Философских трудов» [208]. Эта книга весьма интересна для палеонтологов, так как в ней содержится много задач, основанных на изучении ископаемых организмов. Довольно подробная библиография о европейских и американских работах этого раннего периода, посвященных вопросам палеобиометрии, приведена в статье Шмида [242]. Этот автор одним из первых пытался рассмотреть поверхность двумерной нормальной плотности, спроектированной на плоскость в виде эллипсов равных вероятностей. Из известных авторов русских работ наиболее ранние принадлежат Ф. Ю. Левинсону-Лессингу, который начиная с 1914 г. на протяжении почти 20 лет систематически применял статистические методы в петрологии и геохимии. Он изучал распределения «магматических коэффициентов», представляющие собой видоизмененные параметры магматических пород по Ниггли, Шенду, Вольфу и другим. При помощи статистического анализа этих коэффициентов Левинсон-Лессинг пытался установить границы между различными типами магматических пород, например между дацитами и андезитами, андезитами и базальтами и т. д. При анализе он использовал самые простые формы статистических приемов изучения больших выборок, включая вычисление оценок средних, дисперсий и показателей асимметрии распределений магматических коэффициентов¹. Отдельные работы, связанные с

¹ Ф. Ю. Левинсон-Лессинг, О пределах подразделения семейства андезитов, Изв. геол. ком., 43, № 6, 1923; О разделении базальтов и андезитов, Изв. геол. ком., 44, № 4, 1925; О разделении дацитов и липаритов, ДАН СССР, сер. А, 8, 1930; Статистические характеристики химизма трахитов, Изв. АН СССР, 7, № 1, 1930.

применением статистических методов, появились в двадцатых годах. Так, например, в работе Пирсона [207] применялись средние арифметические, коэффициент вариации и критерий Стьюдента. Заслуживает внимания также и многотомная монография Бринкмана [34]. Начало тридцатых годов ознаменовалось появлением работ по биометрии, число которых резко увеличилось после книги Симпсона и Роя «Количественная зоология» [255], вышедшей в свет первым изданием в 1939 г. Однако активное распространение палеобиометрических методов, разрабатываемых на основе этого учебника, началось лишь в послевоенный период, начиная с 1945 г., и продолжается до настоящего времени. В качестве примера работ первой половины тридцатых годов можно привести статью Гирша¹. До середины тридцатых годов другие отрасли геологии освещались в интересующем нас аспекте лишь в разрозненных работах. Некоторые из них были посвящены задачам, позднее формализованным специалистами по статистике и использованным геологами в послевоенный период. Примером может служить работа Юинга [90], в которой рассматривается вопрос о влиянии субъективного фактора на изменчивость.

Второй этап охватывает период от середины тридцатых годов до 1942 г. В середине тридцатых годов после опубликования нестатистической работы Уэнтурта и других, посвященной графическому анализу частотных диаграмм размеров осадочных частиц, значительно чаще стали применять статистические методы при изучении данных о размерах частиц осадка. В этот период стали известны имена Крамбейна, Отто и Драйдена. Из них сейчас только Крамбейн активно продолжает вести работы по применению статистического анализа при решении задач осадкообразования и стратиграфии. В 1935 г. появилась работа Эйзенхарта [84], посвященная вопросам изучения процесса осадкообразования и имеющая большое значение. Одно из первых приложений методов статистического контроля качества связано с работой Отто². В течение второго этапа работы, посвященные применению статистического анализа в геологии, помимо седиментологии и палеонтологии, появлялись относительно редко. Например, в работе Корна [145] рассмотрены вопросы применения сглаживания и анализа Фурье при изучении периодичности изменений мощности слоев. В 1940 г. в русских журналах появились две работы Разумовского³,

¹ A. H. Hersh, Evolutionary relative growth in the Titanotheres, Am. Nat., 68, 1934.

² G. Otto, The use of statistical methods in effecting improvements in a Jones sample splitter, Journ. Sed. Petrol., 7, 1937.

³ Н. К. Разумовский, Характер распределения содержаний металла в рудных месторождениях, ДАН СССР, 28, № 9, 1940; Логарифмически-нормальная функция распределения и ее свойства, ЛГУ, XX, 1940.

посвященные исследованию распределений содержаний элементов.

Третий этап можно считать начавшимся в 1945 г., т. е. после окончания второй мировой войны, когда статистические методы практически стали проникать во все области геологии. За эти годы появилось множество работ, посвященных статистическому анализу в свете применения его в различных отраслях геологии. К таким исследованиям относятся, например, работы Шоу и Бэнкьера [247], Крамбейна [159], Стралера [265] и других.

В 1953 и 1954 гг. журнал «Геология» (*Journ. Geol.*) посвятил два специальных выпуска вопросам применения статистики в геологии. В то же время под руководством Комитета по применению статистики в физических науках Американской статистической ассоциации было организовано Общество наук о Земле, представленное такими отраслями геологии, как метеорология, геофизика и геохимия. В 1945 г. в Лос-Анжелосе был проведен симпозиум по применению статистики в геологии, на котором были представлены результаты исследований практически во всех отраслях геологии. В 1956 г. вышла в свет книга Чейза [61], посвященная в основном количественно-минералогическому анализу. В этой книге для обоснования выводов широко используются статистические методы. В 1958 г. была издана вторая книга¹, посвященная вопросам эволюции и палеонтологии. Значительную часть этой книги занимает изложение методов статистического анализа.

В 1960 г. появились две работы Крамбейна² общетеоретического характера. В первой из них рассмотрены вопросы геологического опробования, а также трудности, связанные с одновременным изучением многих признаков, с чем, как правило, приходится сталкиваться при решении геологических задач; вторая посвящена общему обзору применения статистических методов в геологии, природы статистических совокупностей, рассматриваемых в геологических задачах, и методики опробования.

Настоящая книга написана с целью дать общий обзор методов статистического анализа, показать применение этих методов при решении самых разнообразных геологических задач и, наконец, познакомить читателя с литературой, в которой затронуты вопросы применения статистики в науке о Земле.

¹ E. C. Olsen, R. L. Miller, *Morphological Integration*, University of Chicago Press, Chicago, 1958.

² W. C. Krumbein, Some problems in applying statistics to geology, *App. Stat.*, IX, № 2, 1960; The geological population as a framework for analysing numerical data in geology, *Liverpool and Manchester, Geol. Journ.*, 2, part 3, 1960.

Авторы выражают надежду, что предлагаемая ими книга будет служить общим руководством по применению статистических методов в геологии для исследователей, работающих в различных областях науки о Земле, а также учебником для студентов.

В этой книге не рассматривается применение статистических методов в геофизике — эти вопросы требуют специального труда. В книгу также не включены статистические таблицы — их можно найти в любом руководстве по математической статистике.

При работе над этой книгой (с 1954 по 1962 г.) авторы пользовались советами и помощью очень многих людей, которым они выражают свою благодарность. Сведения о русской литературе, посвященной применению статистических методов, были любезно предоставлены А. Б. Вистелиусом, за что авторы считают долгом выразить ему свою признательность.

Роберт Л. Миллер,
Чикаго, Иллинойс.

Джемс С. Кан,
Ливермор, Калифорния.

НЕКОТОРЫЕ ЗАМЕЧАНИЯ

Как отмечалось в предисловии, в данной книге рассматриваются с геологической точки зрения методы исследования, называемые *статистическими*. Основу последних составляет теория вероятностей. Поэтому, чтобы лучше понять теорию статистических выводов, следует кратко разобрать основные положения теории вероятностей. Краткость разбора объясняется тем, что, во-первых, по этому вопросу имеется обширная литература, часть которой приведена в списке, помещенном в конце книги, с целью осветить те наиболее важные стороны проблемы, которые могут заинтересовать читателя; во-вторых, авторы стремились ограничиться лишь теми вопросами, которые непосредственно относятся к решению задач, поставленных в этой книге.

НЕКОТОРЫЕ СВЕДЕНИЯ О ТЕОРИИ ВЕРОЯТНОСТЕЙ

Представим, например, что на берегу озера Мичиган собрано 1000 галек. В табл. 1.1 приведены результаты их классификации по цвету.

Допустим, что цвет этих галек резко выражен. Геолог положил их в мешок и решил взять наудачу одну гальку из 1000. Какова возможность того, что выбранная галька окажется коричневой? Интуитивный ответ и его обоснование скорее всего будут следующими: имеется 1000 галек, из них 852 штуки коричневые; возможность, или вероятность, того, что одна извлеченная наудачу галька окажется коричневой, будет $852/1000$, или $0,852$. Теперь, вынув гальку, посмотрим, какого она цвета. Наблюдение, т. е. результат выборочного эксперимента, называется *исходом* или *событием*. Все 1000 галек образуют *генеральную совокупность*, или *выборочное пространство*. Одна вынутая галька представляет собой *выборку*, а возможный исход одной выборки называется *выборочной точкой*. В нашем при-

мере выборочных точек всего три: коричневая галька, черная галька и зеленая галька.

Заклучив, что вероятность вынуть коричневую гальку равна 0,852, мы сделали два предположения. Первое — события рассматриваются как исключающие друг друга, второе — они равновозможны. Другими словами, любая из галек имеет одинаковые шансы быть вынутой. Первое из этих предположений выполнено, так как мы заранее оговорились, что коричневые гальки четко отличаются от черных и зеленых. Второе предположение еще не рассматривалось. Допустим, что оно также выполнено. В этом случае мы приходим к определению, которое

Т а б л и ц а 1.1
Классификация 1000 галек
по цвету

Цвет	Частота
Коричневый . . .	852
Черный	93
Зеленый	55
Всего	1000

называется *классическим определением вероятности*: *вероятность события есть отношение числа возможных исходов, благоприятствующих событию, к общему числу возможных исходов, при условии что события взаимно исключают друг друга и что они являются равновозможными.*

Недостаток классического определения вероятности сразу же становится очевидным, как только геолог начинает рассматривать с геологических позиций предположение, согласно которому каждая галька имеет одинаковые шансы быть вынутой. Разница в удельном весе или размерах различных галек может, например, привести к тому, что черные гальки окажутся на дне мешка, а зеленые — сконцентрируются сверху. С позиций этой классификации вынимание одной гальки исключает вынимание другой, но не все гальки имеют одинаковые шансы быть вынутыми. В подобных случаях, когда необходимые предположения не выполняются, классическое определение оказывается несостоятельным. Бросание монеты служит хорошо известным тому примером. Исходы каждого единичного бросания взаимно исключаются: выпадет либо герб, либо решетка. Таким образом, если события равновозможны, вероятность того, что выпадет герб, будет равна $\frac{1}{2}$. Однако в действительности монета не бывает идеальной, в результате чего одной стороной она будет падать несколько чаще, чем другой. Следовательно, когда события не равновозможны (или не несовместимы), классическое определение для вычисления рассматриваемых вероятностей неприменимо. Его можно использовать лишь в идеальных случаях.

Как правило, решение задач, встающих перед геологом, не может быть основано на классическом определении вероятности. Конечно, встречаются случаи, когда классическая модель с успехом используется в качестве теоретической основы.

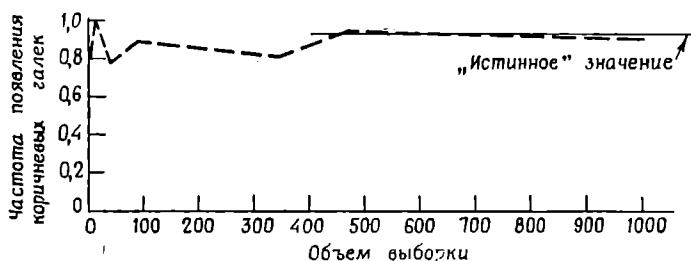
Этот вопрос более детально рассмотрен в других главах. При геологических исследованиях необходимо сделать те или иные статистические выводы по результатам наблюдений. Разрыв

Т а б л и ц а 1.2

**Распределение частот коричневых, черных и зеленых галек
в случайных выборках объемом 5, 10, 50, 100, 400, 500 и 600
из генеральной совокупности 1000 галек**

	Объем выборки	Частота события			
		коричневые	черные	зеленые	всего
	5	4	1	0	5
	10	10	0	0	10
	50	38	7	5	50
	100	86	9	5	100
	400	319	60	21	400
	500	430	43	27	500
	600	515	56	29	600
Генеральная совокуп- ность	1000	850	90	60	1000

между реальными задачами и идеальными моделями можно ликвидировать, если классическую теорию вероятностей заменить на так называемую *частотную теорию вероятностей* [17].



Ф и г. 1.1. График зависимости относительной частоты появления коричневых галек от объема выборки.

В свете этой теории рассмотрим 1000 галек. Возьмем из генеральной совокупности выборки объемом 5, 10, 50, 100, 500 и 600. Сначала выберем наудачу 5 галек, запишем результат и вернем их обратно в генеральную совокупность, затем также поступим с 10 гальками и наконец с 600. Результаты подобного эксперимента можно свести в таблицу, аналогичную табл. 1.2. В целом при увеличении объема выборки отношения числа ко-

ричных, черных и зеленых галек к числу наблюдений (так называемые *относительные частоты*, или *частости*) приближаются к определенным константам. При этом частости принимают значения, близкие к этим константам. Например, относительная частота исхода, заключающегося в появлении коричневой гальки, принимает значения в интервале от 0,76 до 1,00. Величина, принимающая при повторении эксперимента те или иные, заранее неизвестные значения, называется *случайной величиной*. Колебания относительной частоты появления коричневых галек приведены на фиг. 1.1.

ОПРЕДЕЛЕНИЕ ВЕРОЯТНОСТИ

Понятие относительной частоты события, стремящейся к некоторому фиксированному значению, ведет к определению вероятности, которое предлагают Крамер [16], Филлер [18] и Дуб¹. Это определение сводится к следующему: *вероятность появления события в серии экспериментов есть число P , поставленное в соответствие этому событию. При неограниченном повторении эксперимента относительная частота события будет сколь угодно близка к P* . В предыдущем примере значение P будет близким к 0,85. Это записывается как

$$P \text{ (появление коричневой гальки)} \approx 0,85. \quad (1.1)$$

В этой книге в связи с особенностями геологических задач будет принято частотное определение вероятности. В дальнейшем термин *частота* будет употребляться вместо понятия *относительная частота*, если заранее нет каких-либо оговорок. Под термином *относительная частота* следует понимать абсолютное значение частоты события (или общее число наблюдений частотного события), деленное на общее число наблюдений всех событий в выборке. Например, абсолютная частота появления черных галек в выборке объемом 400 равна 60 (табл. 1.2) при относительной частоте 0,15 (табл. 1.3).

Нетрудно видеть, что вероятность событий принимает значение в интервале от 0 до 1. Она не принимает значений больше 1, но может быть равной 1 или 0. Это записывается так:

$$0 \leq P \leq 1. \quad (1.2)$$

Если событие обязательно происходит при каждом эксперименте, то $P=1$, и такое событие называется *достоверным*. Примером может служить вероятность достоверного события, заключающегося в том, что одна галька, взятая наудачу из сово-

Таблица 1.3

Относительная частота появления коричневой, черной и зеленой галек
в случайных выборках объемом 5, 10, 50, 100, 400, 500 и 600
из генеральной совокупности 1000 галек

	Объем выборки	Частота события			
		коричневая	черная	зеленая	всего
	5	0,80	0,20	0,00	1,00
	10	1,00	0,00	0,00	1,00
	50	0,76	0,14	0,10	1,00
	100	0,86	0,09	0,05	1,00
	400	0,80	0,15	0,05	1,00
	500	0,86	0,09	0,05	1,00
	600	0,86	0,10	0,04	1,00
Генеральная совокуп- ность	1000	0,85	0,09	0,06	1,00

купности галек, окажется галькой. Ясно, что вероятность такого события равна 1. Однако, если $P=1$, событие не обязательно будет достоверным. Предположим, что вместо 852 коричневых галек у нас 996, а среди оставшихся — равное количество как черных, так и зеленых. При повторении эксперимента вероятность появления коричневой гальки будет почти равна 1. Однако в процессе отбора не обязательно должны быть отобраны только коричневые гальки, так как не исключена возможность, что будет взята черная или зеленая.

Вероятность невозможного события равна 0. Таким образом, вероятность вытянуть синюю гальку из рассматриваемой совокупности 1000 галек будет равна 0. Однако если вероятность события и равна 0, то это не значит, что событие невозможно. Из этого краткого обзора видно, насколько трудно предсказать единичное событие, тогда как группу событий можно предсказать с той или иной степенью надежности.

Следует особо отметить, что когда выбирается наудачу галька из группы 1000 галек, то речь идет о *вероятности a priori*. После того как галька выбрана, вероятность того, что взята галька конкретного типа, будет равна или 0, или 1. Наступившее событие заключается в том, что галька или коричневая (или черная, или зеленая), или не коричневая (или не черная, или не зеленая).

Условные и безусловные вероятности отдельных событий

Представим, что геолог вернулся с полевых работ и классифицирует образцы осадочных пород, располагая их по двум **17**

признакам: сначала по типам осадочных пород¹, а затем по характеру зернистости. Табл. 1.4 дает представление о подобном распределении образцов.

Таблица 1.4

Распределение 2000 образцов по типам осадочных пород и зернистости

Типы пород	Зернистость				
	грубая	средняя	тонкая	очень тонкая	всего
Аркозы	320	150	130	50	650
Граувакки	50	500	450	100	1100
Кварциты	80	100	20	50	250
Всего	450	750	600	200	2000

Таким образом, R_i может представлять любой из трех типов пород.

Такой тип представления данных можно обобщить путем записи в строку значений одного признака (тип породы), обозначенного через R , и записи в столбце значений другого признака (размер зерен), обозначенного через C . Вся совокупность объектов разделяется на части (подмножества) $R_1, R_2, \dots, R_i, \dots, R_r$ соответственно возможным значениям признака R (табл. 1.5). Величина R_r является последним подмножеством множества R и в нашем примере будет представлять кварциты. Так как R_i рассматривается как любое подмножество множества R , то можно говорить, что i изменяется в интервале от 1 до r ; в нашем примере, в частности, $r=3$. Здесь R_1 соответствует аркозам, R_2 — грауваккам, а $R_3=R_r$ и соответствует кварцитам.

Точно так же по признаку C выборка или совокупность разделяется на подмножества $C_1, C_2, \dots, C_j, \dots, C_s$. Так, если C — зернистость пород, то C_1 — грубозернистые, C_2 — среднезернистые, C_3 — тонкозернистые и $C_4=C_s$ — очень тонкозернистые. Эта классификация приведена в табл. 1.5, где n_{ij} — число наблюдений на пересечении строки i и столбца j .

Любая прямоугольная таблица, подобная табл. 1.4 и 1.5, называется *матрицей*. Порядок матрицы определяется числом образующих ее строк и столбцов. Матрица, представленная в табл. 1.5, имеет порядок $r \times s$. Таким образом, табл. 1.4 представляет собой матрицу порядка 3×4 , символическая запись

¹ P. D. Krynine, The megascopic study and field classification of sedimentary rocks, Journ. Geol., 56, 1948.

Распределение N событий при классификации по R и C

		C_1	C_2	C_3	C_4	C_j	C_s	Сумма
R	R_1	n_{11}	n_{12}	n_{13}	n_{14}	n_{1j}	n_{1s}	$n_{1.}$
	R_2	n_{21}	n_{22}	n_{23}	n_{24}	n_{2j}	n_{2s}	$n_{2.}$
	R_3	n_{31}	n_{32}	n_{33}	n_{34}	n_{3j}	n_{3s}	$n_{3.}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	R_i	n_{i1}	n_{i2}	n_{i3}	n_{i4}	n_{ij}	n_{is}	$n_{i.}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	R_r	n_{r1}	n_{r2}	n_{r3}	n_{r4}	n_{rj}	n_{rs}	$n_{r.}$
Сумма		$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	$n_{.j}$	$n_{.s}$	$N = n_{..}$

которой приведена в табл. 1.6. Первая цифра при указании порядка матрицы всегда означает число строк, а вторая — число столбцов.

Таблица 1.6

Обобщение распределения $N=2000$ образцов по типам пород R и зернистости C

		C_1	C_2	C_3	C_4	Сумма
R	R_1	n_{11}	n_{12}	n_{13}	n_{14}	$n_{1.}$
	R_2	n_{21}	n_{22}	n_{23}	n_{24}	$n_{2.}$
	R_3	n_{31}	n_{32}	n_{33}	n_{34}	$n_{3.}$
	Сумма	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	$N = n_{..}$

Таким образом, n_{12} следует понимать как число наблюдений на пересечении первой строки и второго столбца, или как число образцов, которые представляют собой среднезернистые аркозы. Здесь $n_{12}=150$, а n_{34} , представляющее число очень тонкозернистых кварцитов, равно 50.

Общее число аркозов получается посредством суммирования числа наблюдений для всех четырех разновидностей. Это общее число называется *безусловной (marginal) абсолютной частотой появления аркозов*, которая в данном случае равна 650.

Безусловная частота для аркозов может быть записана как

$$\sum_{j=1}^4 n_{1j}, \quad (1.3)$$

где Σ — знак суммы. Это выражение означает сумму всех тех значений n , для которых номер строки остается постоянным (в данном случае 1), и тех, для которых j принимает значения от 1 до 4 включительно. Известно, что j означает любое подмножество в столбцах и что в данном случае оно соответствует любой группе зернистости. Таким образом, у нас всего четыре группы аркозов, подразделенные по зернистости: грубозернистые, среднезернистые, тонкозернистые и очень тонкозернистые.

Следовательно, $\sum_{j=1}^4 n_{1j}$ представляет собой сумму по группе R_1 , соответствующей аркозам. Обозначим сумму (1.3) по элементам первой строки (табл. 1.6) через n_1 , а аналогичную сумму для второй строки — через n_2 . Эта сумма будет представлять собой абсолютную безусловную частоту граувакков. Точно так же будет выражена и абсолютная безусловная частота очень тонкозернистых пород, которую обозначим через n_4 ; в данном примере эта величина будет равна 200. Сумму всех наблюдений можно обозначить через N или $n_{..}$ (см. табл. 1.5 и 1.6).

Если все элементы матрицы порядка $r \times s$, представляющие абсолютные безусловные вероятности, разделить на N , получим результаты, приведенные в табл. 1.7, где

$$\frac{n_{ij}}{N} \approx p_{ij}. \quad (1.4)$$

В соответствии с нашим определением вероятности вероятность события, заключающегося в принадлежности образца как к группе R_1 , так и к C_3 , будет $P\{R_1C_3\} = p_{13}$. Вообще вероятность события, заключающегося в одновременном выполнении условий R_i и C_j , дается выражением

$$P\{R_iC_j\} = p_{ij} \approx \frac{n_{ij}}{N}. \quad (1.5)$$

Проведя все перечисленные преобразования с данными табл. 1.4, получим результаты, приведенные в табл. 1.8.

Следовательно, вероятность того, что взятый наудачу образец окажется очень тонкозернистым аркозом, примет выражение $P\{R_1C_4\} \approx 50 : 2000 = 0,025$.

Величина $P\{R_iC_j\}$ называется *совместной вероятностью* (или *вероятностью совмещения*) R_i и C_j . В случае когда рассматривается значение только признака R (или только признака C), например R_i , мы говорим о *безусловной вероятности* R_i , которая определяется по формуле

$$P\{R_i\} = \sum_{j=1}^s p_{ij} = p_{i.}. \quad (1.6)$$

Таблица 1.7

		C_1	C_2	C_3	...	C_j	...	C_s	Сумма
R	R_1	p_{11}	p_{12}	p_{13}	...	p_{1j}	...	p_{1s}	$p_{1.}$
	R_2	p_{21}	p_{22}	p_{23}	...	p_{2j}	...	p_{2s}	$p_{2.}$
	R_3	p_{31}	p_{32}	p_{33}	...	p_{3j}	...	p_{3s}	$p_{3.}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	R_l	p_{l1}	p_{l2}	p_{l3}	...	p_{lj}		p_{ls}	$p_{l.}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	R_r	p_{r1}	p_{r2}	p_{r3}		p_{rj}	...	p_{rs}	$p_{r.}$
Сумма		$p_{.1}$	$p_{.2}$	$p_{.3}$	...	$p_{.j}$	...	$p_{.s}$	$p_{..} = 1$

Таблица 1.8

Тип породы	Зернистость				
	C_1	C_2	C_3	C_4	Сумма
R_1	0,160	0,075	0,065	0,025	0,325
R_2	0,025	0,250	0,225	0,050	0,550
R_3	0,040	0,050	0,010	0,025	0,125
Сумма	0,225	0,375	0,300	0,100	1,000

Точно так же безусловная вероятность C_j примет вид

$$P\{C_j\} = \sum_{i=1}^r p_{ij} = p_{.j}. \quad (1.7)$$

Следовательно, безусловная вероятность события, заключающегося в том, что взятый наудачу образец окажется грубозернистым песчаником, выразится как

$$\begin{aligned}
 P\{C_1\} &= \sum_{i=1}^3 p_{i1} = p_{11} + p_{21} + p_{31} = P\{R_1 C_1\} + P\{R_2 C_1\} + \\
 &\quad + P\{R_3 C_1\} = \sum_{i=1}^3 P\{R_i C_1\} = \\
 &= 0,160 + 0,025 + 0,040 = 0,225 = p_{.1}. \quad (1.8)
 \end{aligned}$$

Аналогично безусловная вероятность того, что образец окажется кварцитом, равна 0,125, а среднезернистой поро-

дой — 0,375. Заметим, что

$$p_1 \approx \sum_{i=1}^3 \frac{n_{i1}}{N} = \frac{n_{.1}}{N} \quad (1.9)$$

и

$$P\{C_1\} \approx \frac{n_{.1}}{N}. \quad (1.10)$$

Вообще

$$P\{C_j\} \approx \frac{n_{.j}}{N} \quad (1.11)$$

и

$$P\{R_i\} \approx \frac{n_{i.}}{N}. \quad (1.12)$$

При рассмотрении безусловных вероятностей не принимаются во внимание какие-либо классификации внутри интересующего нас признака. Это равносильно переходу от системы классификации по двум признакам (таблицы с двумя входами)

Т а б л и ц а 1.9

R_1	0,325
R_2	0,550
R_3	0,125
Сумма	1,000

к классификации по одному признаку. Например, если в табл. 1.8 рассматривать только типы пород, получим табл. 1.9, в которой совершенно не учитывается размер зерен. Теперь ясно, почему $P\{R_3\} \approx 0,125$.

Предположим, что из группы 2000 образцов взят образец кварцита. Какова вероятность, что этот образец окажется грубозернистым? Иными словами, какова вероятность того, что взятый наудачу образец окажется грубозернистым, при условии, что будет взят именно кварцит? Это может быть записано в виде выражения

$$P\{C_1|R_3\}, \quad (1.13)$$

которое следует читать как вероятность C_1 при условии выполнения R_3 . Эта величина называется *условной вероятностью* C_1 при данном R_3 . Известно, что перед нами кварцит; кроме того, из табл. 1.4 известно также, что из 250 образцов кварцитов 80 грубозернистые. Следовательно, при этих условиях вероятность того, что образец окажется грубозернистым, будет 80/250. Это можно представить в виде формулы

$$P\{C_1|R_3\} = \frac{P\{C_1R_3\}}{P\{R_3\}}. \quad (1.14)$$

Условная вероятность события C_1 при условии выполнения R_3 равна вероятности совместного наступления событий C_1 и R_3 , деленной на безусловную вероятность события R_3 . Вообще

$$P\{C_j|R_i\} = \frac{P\{C_jR_i\}}{P\{R_i\}}. \quad (1.15)$$

Аналогично

$$P\{R_i|C_j\} = \frac{P\{C_j R_i\}}{P\{C_j\}}. \quad (1.16)$$

Так как

$$P\{R_i C_j\} \approx \frac{n_{ij}}{N}, \quad (1.17)$$

$$P\{C_j\} \approx \frac{n_{.j}}{N} \quad (1.18)$$

и

$$P\{R_i\} \approx \frac{n_{i.}}{N}. \quad (1.19)$$

то

$$P\{R_i|C_j\} \approx \frac{n_{ij}}{n_{.j}} \quad (1.20)$$

и

$$P\{C_j|R_i\} \approx \frac{n_{ij}}{n_{i.}}. \quad (1.21)$$

Из (1.6) и (1.7) получим

$$P\{R_i|C_j\} = \frac{P_{ij}}{P_{.j}} \quad (1.22)$$

и

$$P\{C_j|R_i\} = \frac{P_{ij}}{P_{i.}}. \quad (1.23)$$

ГРАФИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ МНОЖЕСТВ

Результаты предшествовавшего обзора можно представить другим, может быть, более наглядным способом. Представим события R_2 , R_3 , C_1 в виде подмножеств, как это показано на фиг. 1.2. Вероятность события, заключающегося в том, что наступит или R_3 , или C_1 , представляет собой

$$P\{R_3 \text{ или } C_1\} = P\{R_3\} + P\{C_1\} - P\{C_1 R_3\}. \quad (1.24)$$

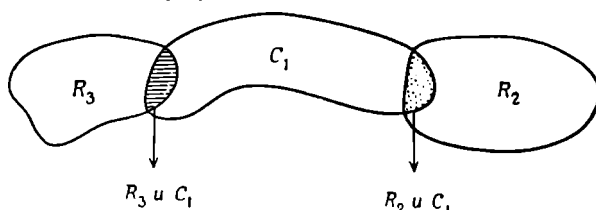
Следовательно, вероятность того, что образец окажется или грубозернистой породой, или кварцитом, равна сумме безусловной вероятности, что он окажется кварцитом, и безусловной вероятности, что он грубозернистый, минус вероятность одновременного появления этих событий. Это показано на фиг. 1.2. Все пространство в пределах контуров представляет безусловные вероятности, соответствующие R_2 , R_3 и C_1 . Они отдельно показаны на фиг. 1.3. Отсюда становится ясной справедливость равенства (1.24). При суммировании безусловных вероятностей величина $P\{R_3 C_1\}$ приплюсовывается дважды: один раз — в виде слагаемого в $P\{R_3\}$, а второй — в $P\{C_1\}$. Этого можно

избежать, вычитая величину $P\{R_3C_1\}$:

$$P\{R_i \text{ или } C_j\} = P\{R_i\} + P\{C_j\} - P\{R_iC_j\}. \quad (1.25)$$

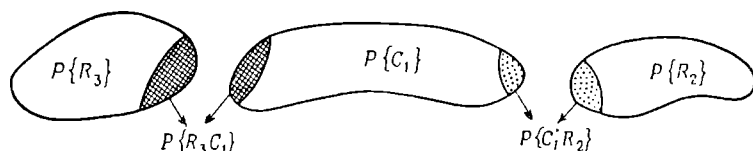
Это выражение называется *формулой сложения вероятностей*.

Какова вероятность события $\{R_2 \text{ или } R_3\}$? Из табл. 1.8 видно, что $P\{R_3\}=0,125$, а $P\{R_2\}=0,550$.



Фиг. 1.2. R_3 — кварциты, R_2 — граувакки (R_2 и R_3 — подмножества множества R); C_1 — грубозернистые песчаники (C_1 — подмножество множества C). Заштрихованная площадь является областью пересечения R_3 и C_1 (она общая для обоих множеств). Аналогично площадь, покрытая точками, является пересечением R_2 и C_1 . Следовательно, заштрихованная область соответствует грубозернистым кварцитам, а область, покрытая точками, — грубозернистым грауваккам.

Так как R_3 и R_2 — исключающие друг друга события, т. е. граувакки не кварциты, то $P\{R_3R_2\}$ — вероятность невозможного события и, следовательно, равна 0. Это показано на фиг. 1.2, из



Фиг. 1.3.

которой видно, что R_2 и R_3 не пересекаются и не имеют общей области. Таким образом,

$$\begin{aligned} P\{R_2 \text{ или } R_3\} &= P\{R_2\} + P\{R_3\} - P\{R_2R_3\} = \\ &= P\{R_2\} + P\{R_3\} - 0 = 0,675. \end{aligned}$$

Это ведет к следующему обобщению. Если два события несовместимы, то вероятность того, что наступит хотя бы одно из них (безразлично какое), равна просто сумме их безусловных вероятностей. Поэтому

$$P\{R_i \text{ или } R_j\} = P\{R_i\} + P\{R_j\}. \quad (1.26)$$

Следовательно,

$$P\{C_2 \text{ или } C_4\} = 0,475.$$

Чему равна вероятность $P\{R_1, \text{ или } R_2, \text{ или } R_3\}$? Так как

$$P\{R_i R_j\} = 0 \quad \text{и} \quad P\{R_i R_j R_k\} = 0,$$

то

$$P\{R_1, \text{ или } R_2, \text{ или } R_3\} = \sum_{i=1}^3 P\{R_i\} = 1.$$

К этому результату легко прийти геометрически.

Вероятностная независимость

Ларсен и Бермен¹ табулировали распределение анизотропных минералов в зависимости от их двусосности или одноосности и от оптического знака. Обозначим через U присутствие одноосного минерала, а через $\bar{U}$ его отсутствие. Тогда $\bar{U}$ представит двусосный минерал. Величины U и $\bar{U}$ называют *противоположными событиями*. Взятый наудачу анизотропный минерал обязательно окажется или одноосным, или двусосным (неодноосным). Следовательно,

$$P\{U\} + P\{\bar{U}\} = 1. \quad (1.27)$$

Это равенство выражает вероятность появления или изучаемого события, или противоположного ему. Аналогично обозначим наличие положительного анизотропного минерала через S , а через $\bar{S}$ — наличие отрицательного анизотропного минерала.

Используя (1.13) и (1.14), найдем условную вероятность события, заключающегося в том, что минерал окажется одноосным при условии, что он положительный, из равенства

$$P\{U|S\} = P\{US\} / P\{S\} = \frac{0,082}{0,448} = 0,183.$$

Заметим, что $P\{U|S\} \neq P\{U|\bar{S}\}$. Это значит вероятность того, что минерал окажется одноосным при условии, что он положителен, не совпадает с вероятностью события, заключающегося в появлении одноосного минерала, при условии, что он отрицательный. Кроме того, $P\{U|\bar{S}\} \neq P\{U\}$, т. е. вероятность появления одноосного минерала при условии, что он отрицательный, не равна безусловной вероятности этого события.

Если условная вероятность события R_i (при условии, что событие C_j осуществилось) равна условной вероятности события R_i при условии, что совершилось противоположное событие $\bar{C}_j$, события C_j и R_i называются *независимыми в вероятностном, или статистическом, смысле*. Это можно записать следующим образом:

$$P\{R_i|C_j\} = P\{R_i|\bar{C}_j\}. \quad (1.28)$$

¹ E. S. Larsen, H. Berman, The microscopic determinations of the monophase minerals, U. S. Dept. Inter. Bull. 848, 1934.

Из (1.27) известно, что

$$P\{R_i\} + P\{\bar{R}_i\} = 1$$

и точно так же

$$P\{C_j\} + P\{\bar{C}_j\} = 1.$$

Следовательно, если верно (1.28), то

$$P\{R_i|C_j\} = P\{R_i\}. \quad (1.29)$$

Доказательство

Так как

$$P\{R_i\} = P\{R_i C_j\} + P\{R_i \bar{C}_j\},$$

то

$$P\{R_i\} = P\{R_i|C_j\}P\{C_j\} + P\{R_i|\bar{C}_j\}P\{\bar{C}_j\}.$$

Из (1.28) известно, что $P\{R_i|C_j\} = P\{R_i|\bar{C}_j\}$, поэтому

$$P\{R_i\} = P\{R_i|C_j\}[P\{C_j\} + P\{\bar{C}_j\}] \quad (1.30)$$

и

$$P\{R_i\} = P\{R_i|C_j\}. \quad (1.31)$$

Кроме того, так как

$$P\{R_i|C_j\} = \frac{P\{R_i C_j\}}{P\{C_j\}},$$

то, согласно (1.31),

$$P\{R_i C_j\} = P\{R_i\}P\{C_j\}. \quad (1.32)$$

Равенство (1.32) выражает статистическую независимость. События R_i и C_j независимы, если вероятность их совместного наступления равна произведению их безусловных вероятностей. Это определение называется *законом умножения для независимых событий*. Если два события независимы в вероятностном (статистическом) смысле, то вероятность их совместного появления равна произведению соответствующих безусловных вероятностей.

Пример

Игральная кость представляет собой куб, на каждой стороне которого показано определенное количество очков — от 1 до 6. Пусть в результате первого бросания кости выпали четыре очка. Естественно поставить вопрос: какова вероятность, что при втором бросании выпадут два очка, если известно, что при первом выпали четыре, т. е. $P(2|4)$ = ?

Нетрудно понять, что, если известен результат первого бросания, это не может оказать нам никакой помощи в определении вероятности результата второго бросания, что является следствием независимости. Вероятность того, что при втором

бросании выпадут два очка, равна $\frac{1}{6}$, что совпадает с условной вероятностью этого события. Следовательно,

$$P(2|4) = P(2) = \frac{1}{6}.$$

Вероятность того, что в результате двух бросаний выпадут два и четыре очка, следует выразить как произведение

$$P(2, 4) = P(2)P(4) = \frac{1}{36},$$

потому что если события независимы, то вероятность их совместного наступления равна произведению их безусловных вероятностей.

ДИСКРЕТНЫЕ РАСПРЕДЕЛЕНИЯ

В предыдущей главе в основном рассматривались вероятности отдельных событий. Теперь обратимся к сложным событиям. Примером может служить тяжелая фракция осадочной породы. Для петрографа, изучающего эту породу, особенно важным признаком является прозрачность или, наоборот, непрозрачность минералов. Предположим, что из анализа ранее рассмотренных образцов петрографу известно, что на каждые 20 рассмотренных под микроскопом зерен обычно приходится 2 непрозрачных. Следовательно, он может сказать, что вероятность появления непрозрачного зерна равна $1/10$. Обозначим через x число непрозрачных зерен минерала, попавших в поле зрения, содержащего n зерен. Так как в различных полях величина x будет меняться, ее можно рассматривать как случайную. Далее, если $p=1/10$ выражает вероятность появления непрозрачного минерала при одном наблюдении, то величина $1-p=9/10$ выразит вероятность противоположного события Q , т. е. появления прозрачного минерала. Допустим, что наши наблюдения независимы в статистическом смысле.

Если в поле зрения микроскопа будут видны три зерна ($n=3$), то вероятность того, что все три зерна непрозрачные, согласно теореме умножения (1.32) и в предположении статистической независимости, будет определена как

$$P(x=3) = p \cdot p \cdot p = p^3 = \left(\frac{1}{10}\right)^3.$$

Точно так же вероятность того, что все три зерна будут прозрачными, равна

$$P(x=0) = (1-p)^3 = \left(\frac{9}{10}\right)^3.$$

Конечно, возможные исходы выборки этим не ограничиваются. Может случиться, что одно зерно будет непрозрачным, а два — прозрачными или два — непрозрачными, а одно — про-

зрачным. Возможно, что первым из трех будет отмечено непрозрачное зерно, а прозрачные — последними; или первым — прозрачное, затем — непрозрачное и третьим — опять прозрачное; возможна и такая комбинация: первые два зерна прозрачны, а третье непрозрачно. Следовательно, вероятность того, что из трех зерен, попавших в поле зрения микроскопа, одно окажется непрозрачным, равна

$$\begin{aligned} P(x=1) &= P(Q\bar{Q}\bar{Q}, \text{ или } \bar{Q}Q\bar{Q}, \text{ или } \bar{Q}\bar{Q}Q) = \\ &= P(Q\bar{Q}\bar{Q}) + P(\bar{Q}Q\bar{Q}) + P(\bar{Q}\bar{Q}Q) = \\ &= p(1-p)(1-p) + (1-p)p(1-p) + (1-p)(1-p)p = \\ &= 3p(1-p)^2 = 3(1/10)(9/10)^2. \end{aligned}$$

Легко, кроме того, доказать, что

$$P(x=2) = 3(1/10)^2(9/10)$$

(это предоставляется читателю сделать самостоятельно).

Таким образом, если x — число непрозрачных минеральных зерен среди находящихся в поле зрения трех зерен ($n=3$), то вероятности событий, заключающихся в том, что $x=3$, $x=2$, $x=1$, $x=0$ при $p=1/10$, будут соответственно равны:

$$\begin{aligned} P(x=3) &= \left(\frac{1}{10}\right)^3 = p^3, \\ P(x=2) &= 3 \left(\frac{1}{10}\right)^2 \left(\frac{9}{10}\right) = 3p^2(1-p), \\ P(x=1) &= 3 \left(\frac{1}{10}\right) \left(\frac{9}{10}\right)^2 = 3p(1-p)^2, \\ P(x=0) &= \left(\frac{9}{10}\right)^3 = (1-p)^3. \end{aligned}$$

Так как одно из событий ($x=3$, 2, 1 или 0) должно обязательно произойти, то, следовательно, $\sum_{x=0}^3 P(x) = 1$. В этом легко убедиться:

$$\begin{aligned} \sum_{x=0}^3 P(x) &= p^3 + 3p^2(1-p) + 3p(1-p)^2 + (1-p)^3 = \\ &= p^3 + 3p^2 - 3p^3 + 3p - 6p^2 + 3p^3 + 1 - 3p + 3p^2 - p^3 = 1. \end{aligned}$$

Нетрудно показать, что вероятности, связанные с появлением зерен непрозрачного минерала, можно получить из разложения бинома $[p + (1-p)]^3$, что позволит найти вероятности любого события, связанного с x в выборках объемом $n=3$. Мы получим следующее выражение:

$$[p + (1-p)]^3 = P(x=3) + P(x=2) + P(x=1) + P(x=0) = 1. \quad \mathbf{29}$$

В общем случае при разложении бинома $[p + (1 - p)]^n$ становятся известными вероятности событий, заключающихся в том, что случайная величина x принимает значение $0, 1, 2, \dots, n$.

Пример

Вычислим вероятность появления непрозрачного минерала x раз, а прозрачного $n - x$ раз в выборке объемом n . Иными словами, нужно определить вероятность появления события Q x раз, а $\bar{Q}$ $n - x$ раз. При условии независимости вероятность того, что событие Q будет наблюдаться x раз, равна вероятности появления Q при одном испытании $P(Q) = p$ в степени x . Точно так же вероятность появления противоположного события $n - x$ раз равна $(1 - p)^{n-x}$. Однако не следует забывать, что данные события могут быть построены как различные сочетания из n по x . Число таких сочетаний равно

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \quad \text{при } x = 0, 1, 2, \dots, n,$$

где $n!$ (факториал) представляет произведение

$$n! = n(n-1)(n-2) \dots 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1,$$

например $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$, $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$, а $0!$ определяется равным 1. Вернувшись к поставленной задаче, найдем, что искомая вероятность равна

$$\binom{n}{x} p^x (1 - p)^{n-x} \quad \text{при } x = 0, 1, 2, \dots, n.$$

Вероятность того, что случайная величина x примет значение, равное x_0 , в n независимых испытаниях, дается выражением

$$P\{x = x_0\} = \binom{n}{x_0} p^{x_0} (1 - p)^{n-x_0} \quad \text{при } x_0 = 0, 1, 2, \dots, n,$$

где p — вероятность появления события при единичном испытании.

Случайная величина x принимает значения в интервале от 0 до n . В данном примере x принимает значения от 0 до 3.

Генеральная совокупность, в которой случайная величина может принимать только отдельные значения, называется *дискретной*. Дискретная совокупность может быть ограниченной, как, например, в случае тяжелых минералов (или неограниченной, например совокупность всех положительных целых чисел, или чисел больше 21,458).

Для дискретной случайной величины x определим функцию **30** $f(x)$, такую, что каждому значению $x = x_0$ должно соответство-

вать $f(x_0) = p(x_0)$. Тогда $f(x)$ будет называться *плотностью вероятностей случайной величины x* ¹.

Если генеральная совокупность дискретна, то $f(x)$ будет дискретной плотностью вероятностей.

Итак, если S — дискретное множество значений, а $f(x)$ — функция, заданная на S , такая, что $f(x) \geq 0$ и $\sum_{x \in S} f(x) = 1$ для всех значений x , содержащихся в S ($x \in S$), то $f(x)$ называют *дискретной плотностью вероятностей*. Теперь нетрудно видеть, что функция

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x},$$

где $x = 0, 1, 2, \dots, n$, действительно является дискретной плотностью вероятностей, так как $f(x)$ — функция на дискретном пространстве, большая или равная 0, и

$$\sum_{x=0}^n f(x) = [p + (1-p)]^n = 1.$$

Функция

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad (2.1)$$

называется *биномиальным распределением случайной величины x* . Это распределение может быть использовано в качестве теоретической модели в таких ситуациях или экспериментах, которые имеют два исключаяющих друг друга исхода, например герб или решетка, жизнь или смерть, гранат или не гранат, контакт между зернами или его отсутствие, успех или неудача.

Вероятность p успешного исхода единичного испытания и число испытаний n будут принимать различные значения в разных биномиальных распределениях. Например, вероятность, что выпадет герб при бросании монеты, равна $1/2$, тогда как вероятность, что будет наблюдаться контакт между зернами в аркозах, равна $6/10^2$. В первом примере n может быть равно 30, а во втором — порядка нескольких сотен. Таким образом, n и p являются фиксированными характеристиками частных биномиальных распределений. Эти характеристики называются *параметрами*, и их можно рассматривать как переменные величины в том смысле, что они меняют значения в различных биномиальных распределениях. Следовательно, чтобы установить

¹ Термин *плотность вероятностей* обычно применяется для непрерывных случайных величин, тогда как в случае дискретных случайных величин говорят о вероятности события, заключающегося в том, что случайная величина x приняла значение x_0 . — *Прим. перев.*

² J. S. Kahn, The analysis and distribution of the properties of packing in sand size sediment; The distribution of the packing measurements and an exemple of packing analysis, Journ. Geol., 64, 1956.

частное биномиальное распределение, необходимо знать точные значения n и p . Если же их точные значения неизвестны, нужно пользоваться по крайней мере их оценками. Если параметры и функция распределения случайной величины известны, можно определить вероятности любых событий для генеральной совокупности.

Мы уже упоминали, что вероятность наблюдения в шлифах контакта между зернами при структурном анализе аркозов была определена как 0,6. При последующем обсуждении структуры аркозов с группой геологов возник вопрос: что произойдет при изучении шлифа по линии пересечения 20 зерен? Можно ли увидеть точно три контакта между зернами? Ответы были таковы: «совершенно невозможно», «да», «может быть», «иногда», «почему бы и нет». Эти разнородные ответы свидетельствуют о необходимости общего подхода к постановке задачи. Ответ на этот вопрос точен. Вопрос, поставленный в вероятностной форме, должен ставиться относительно самого события, т. е. какова вероятность того, что мы увидим три контакта между зернами при пересечении 20 зерен аркозов, если вероятность события, заключающегося в том, что одно зерно соприкасается с другим, равна 0,6? Здесь уже нет неопределенности. Рассмотрим ход рассуждения, который поможет ответить на этот вопрос.

Число соприкосновений зерен в аркозе распределено как

$$f(x) = \begin{cases} \binom{n}{x} 0,6^x 0,4^{n-x} & \text{при } x = 0, 1, 2, \dots, n, \\ 0 & \text{вне этого интервала} \end{cases}$$

где x — случайная величина, представляющая число соприкосновений между зернами, а n — объем выборки или число просмотренных зерен. Зная закон распределения случайной величины x , а также параметры n и p , мы можем вычислить вероятность любого события, заключающегося в том, что x примет одно из значений от 0 до 20. В табл. 2.1 приведены вероятности событий $x = 0, 1, 2, \dots, n$ при $n=20$ и $p=0,6$.

Теперь вычислим вероятности некоторых сложных событий.

1) Вероятность того, что число зерен, имеющих контакты с другими зернами, находится в интервале от 7 до 19 включительно, равна

$$P\{7 \leq x \leq 19\} = \sum_{x=7}^{19} f(x) = f(7) + f(8) + \dots + f(19) = 0,994.$$

2) Вероятность того, что число контактов равно 12 при условии $x < 18$, равна

$$P\{x = 12 | x < 18\} = \frac{P\{x = 12\}}{P\{x < 18\}} = \frac{f(12)}{\sum_{x=0}^{17} f(x)} = \frac{0,180}{0,997} = 0,1805.$$

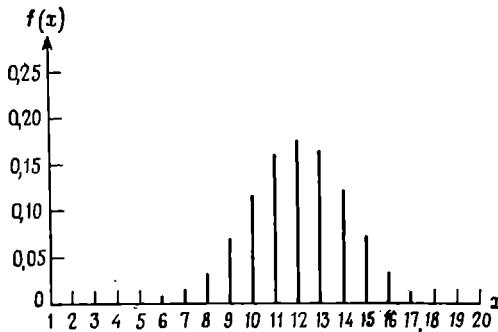
Вероятности биномиального распределения при $p = 0,6$ и $n = 20$

Событие x	Вероятность события $P\{x\} = f(x)$	Событие x	Вероятность события $P\{x\} = f(x)$
0	~ 0	11	0,160
1	~ 0	12	0,180
2	~ 0	13	0,166
3	~ 0	14	0,124
4	~ 0	15	0,075
5	0,001	16	0,035
6	0,005	17	0,012
7	0,015	18	0,003
8	0,036	19	~ 0
9	0,071	20	~ 0
10	0,117		

3) Вероятность того, что число контактов меньше 10 при условии $x < 14$, равна

$$P\{x < 10 | x < 14\} = \frac{P\{x < 10\}}{P\{x < 14\}} = \frac{0,128}{0,751} = 0,1705.$$

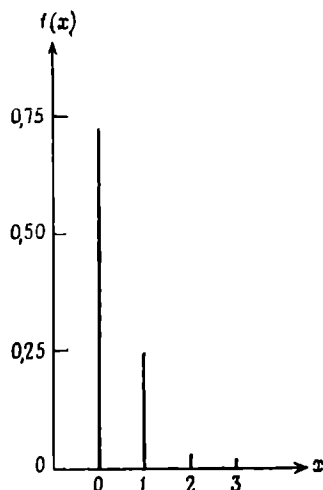
Теперь вернемся к задаче о вероятности того, что в пересечении шлифа 20 зерен мы увидим три контакта. На основании данных табл. 2.1 это событие очень редко. Вероятность появления трех контактов приблизительно равна 0. Событие, кажущееся вполне возможным, практически невероятно!



Фиг. 2.1. Биномиальное распределение при $n=20$ и $p=0,6$.

Рассмотренное биномиальное распределение (при $n=20$ и $p=0,6$) представлено на фиг. 2.1, а описанный ранее пример с тремя гальками ($n=3$, $p=0,1$) — на фиг. 2.2. На этих графиках

вероятность $f(x_0)$ любого события $x=x_0$ равна длине перпендикуляра, параллельного оси ординат и восстановленного из соответствующей точки $x=x_0$ на оси абсцисс.



Фиг. 2.2. Биномиальное распределение при $n=3$ и $p=0,1$.

В случае дискретных распределений следует иметь в виду, что выражения $P\{x \leq b\}$ и $P\{x < b\}$ не равны между собой. Например, если $n=20$, а $p=0,6$, то

$$P\{x \leq 9\} = 0,128,$$

тогда как

$$P\{x < 9\} = 0,057.$$

Первое выражение равно сумме

$$\sum_{x=0}^9 f(x),$$

а второе — сумме

$$\sum_{x=0}^8 f(x).$$

ФУНКЦИЯ РАСПРЕДЕЛЕНИЯ ДИСКРЕТНОЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

Выше отмечалось, что если $f(x)$ — дискретная плотность вероятностей случайной величины x , такой, что $a_1 \leq x \leq a_n$, то величина $P\{a_i \leq x \leq a_j\}$ при условии, что $a_1 < a_2 < \dots < a_n$, равна

$$\sum_{x=a_i}^{a_j} f(x) = f(a_i) + f(a_{i+1}) + \dots + f(a_{j-1}) + f(a_j). \quad (2.1)$$

Так как

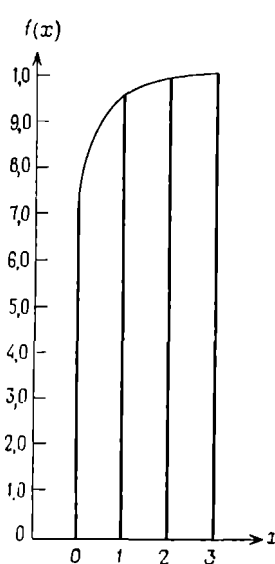
$$\sum_{x=a_1}^{a_{i-1}} f(x) = f(a_1) + f(a_2) + \dots + f(a_{i-2}) + f(a_{i-1})$$

и

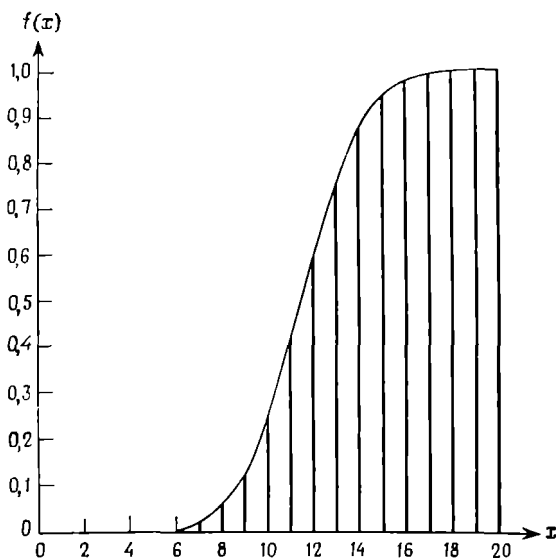
$$\sum_{x=a_1}^{a_j} f(x) = f(a_1) + f(a_2) + \dots + f(a_{i-2}) + f(a_{i-1}) + f(a_i) + \dots + f(a_{j-1}) + f(a_j),$$

то

$$\sum_{x=a_1}^{a_j} f(x) - \sum_{x=a_1}^{a_{i-1}} f(x) = f(a_i) + \dots + f(a_{j-1}) + f(a_j) = \sum_{x=a_i}^{a_j} f(x).$$



Фиг. 2.3. Функция биномиального распределения при $n=3$ и $p=0,1$.



Фиг. 2.4. Функция биномиального распределения при $n=20$ и $p=0,6$.

Ясно, что это выражение представляет собой вероятность

$$P\{a_i \leq x \leq a_j\}.$$

Следовательно,

$$P\{a_i \leq x \leq a_j\} = P\{x \leq a_j\} - P\{x < a_{i-1}\}.$$

Таким образом, сумму (2.1) можно подсчитать, если вычислить: 1) вероятность того, что x примет значение, меньшее или

равное a_j , 2) вероятность того, что x будет меньше чем a_i , и 3) разность между этими двумя вероятностями.

Функция, которая выражает вероятность того, что случайная величина x примет значение, меньшее или равное заданному $x=a_k$, называется *функцией распределения случайной величины x* и обозначается через $F(x)$.

В дальнейшем мы будем обозначать функцию распределения через $F(x)$, а плотность вероятностей через $f(x)$.

Итак,

$$F(x=a_k)=F(a_k)=P(x\leq a_k)=\sum_{x=a_1}^{a_k} f(x),$$

где

$$a_1\leq a_2\leq \dots \leq a_k\leq \dots \leq a_n$$

Т а б л и ц а 2.2

**Функция распределения
биномиального закона
с параметрами $n=3$ и $p=0,1$**

Событие x	$F(x)$
0	0,729
1	0,972
2	0,999
3	1,000

Т а б л и ц а 2.3

**Функция распределения биномиального закона
с параметрами $n=20$ и $p=0,6$**

Событие x	$F(x)$	Событие x	$F(x)$
0	~ 0	11	0,405
1	~ 0	12	0,585
2	~ 0	13	0,751
3	~ 0	14	0,875
4	~ 0	15	0,950
5	0,001	16	0,985
6	0,006	17	0,997
7	0,021	18	$\sim 1,000$
8	0,057	19	$\sim 1,000$
9	0,128	20	$\sim 1,000$
10	0,245		

и

$$x = a_1, a_2, \dots, a_k.$$

Таким образом,

$$0 = F(a_1) \leq F(a_2) \leq F(a_3) \leq \dots \leq F(a_k) \leq \dots \leq F(a_n) = 1.$$

Отсюда видно, что $F(x)$ — неубывающая положительная функция. Значения $F(x)$, соответствующие биномиальным распределениям, приведенным на фиг. 2.1 и 2.2, представлены в табл. 2.2 и 2.3, а их графическое выражение — на фиг. 2.3 и 2.4.

Следовательно,

$$P\{7 \leq x \leq 19\} = F(19) - F(6) = 0,994,$$

$$P\{10 \leq x \leq 12\} = F(12) - F(9) = 0,457,$$

$$P\{9 < x < 13\} = P\{9 < x \leq 12\} = F(12) - F(9) = 0,457.$$

ПОЛИНОМИАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

Если в результате эксперимента может произойти r возможных исходов, то каждый из них может произойти с вероятностями $p_1, p_2, \dots, p_r$. Например, эксперимент может состоять в определении присутствия некоторых редких элементов $l_1, l_2, \dots, l_t$. В данном случае исходами эксперимента могут быть $l_1, l_2, \dots, l_t$ с соответствующими вероятностями $p_1, p_2, \dots, p_t$.

Компания, изготавливающая керамические порошки, контролирует свою продукцию и при этом бракует ее, если обнаруживает 0,5% или более частиц с размерами, превышающими допустимый. Продукция бракуется и в том случае, если в ней окажется 3% или более частиц с размерами меньше допустимого. В данном примере возможны следующие исходы: 1) продукция бракуется в связи с крупной зернистостью с вероятностью p_1 ; 2) продукция бракуется в связи с недопустимой мелкозернистостью с вероятностью p_2 ; 3) продукция принимается с вероятностью p_3 . Предположим, что одна проба продукции не может быть одновременно мелкозернистой и крупнозернистой.

Если предположить, что события независимы в вероятностном смысле, то вероятность забраковать продукцию x_1 раз в связи с крупнозернистостью и x_2 раз в связи с мелкозернистостью в выборке объемом n будет представлять величину, равную

$$p_1^{x_1} p_2^{x_2} p_3^{n-x_1-x_2} = p_1^{x_1} p_2^{x_2} (1 - p_1 - p_2)^{n-x_1-x_2},$$

умноженную на общее число способов $(x_1, x_2, (n-x_1-x_2), \dots)$, которые могут появиться в выборке объемом n . Так, число способов возникновения первого исхода в выборке объемом n

равно $\binom{n}{x_1}$, x_2 в оставшихся $n-x_1$ членах выборки объемом n равно $\binom{n-x_1}{x_2}$, тогда как для x_3 остается только $n-x_1-x_2$. Следовательно, x_1 , x_2 и x_3 могут быть получены совместно из совокупности в n элементов различными путями, число которых $\binom{n}{x_1} \binom{n-x_1}{x_2}$. Однако

$$\binom{n}{x_1} \binom{n-x_1}{x_2} = \frac{n!}{x_1! (n-x_1)!} \frac{(n-x_1)!}{x_2! (n-x_1-x_2)!} = \frac{n!}{x_1! x_2! x_3!},$$

поэтому вероятность того, что в выборке объемом n будет забраковано x_1 крупнозернистых проб, x_2 мелкозернистых и x_3 будут признаны удовлетворительными, будет равна

$$P\{x_1 x_2 x_3\} = \frac{n!}{x_1! x_2! x_3!} p_1^{x_1} p_2^{x_2} p_3^{x_3}.$$

Если через Π обозначить операцию произведения, то

$$P\{x_1 x_2 x_3\} = n! \prod_{i=1}^3 \frac{p_i^{x_i}}{x_i!} = f(x_1 x_2 x_3).$$

Таким образом, вероятность того, что в выборке объемом $n=200$ будут забракованы пять проб из-за крупнозернистости, а три из-за тонкозернистости при $p_1=0,005$ и $p_2=0,03$, равна $\frac{200!}{13! 192!} (0,005)^5 (0,03)^3 (0,965)^{192}$.

Вообще если в результате эксперимента может произойти один из r возможных исходов, каждый с вероятностью $p_1, p_2, \dots, p_r$, то вероятность того, что при n экспериментах первый исход появится x_1 раз, второй — x_2 раз, ..., r -й — x_r раз, окажется равной

$$f(x_1 x_2 \dots x_r) = n! \prod_{i=1}^r \frac{p_i^{x_i}}{x_i!},$$

где

$$n = \sum_{i=1}^r x_i.$$

Это распределение называется *полиномиальным*.

Так как $\sum_{i=1}^r x_i = n$, то из всех переменных произвольно могут изменяться только $r-1$. Величина с номером r фиксируется автоматически. Следовательно, $f(x_1 x_2 \dots x_r)$ можно также записать как $g(x_1 x_2 \dots x_{r-1})$. Поэтому в случае $r=2$

$$\begin{aligned} f(x_1 x_2 \dots x_r) &= g(x_1) = \frac{n!}{x_1! (n-x_1)!} p_1^{x_1} (1-p_1)^{n-x_1} = \\ &= \binom{n}{x_1} p_1^{x_1} (1-p_1)^{n-x_1}. \end{aligned}$$

Этот частный случай полиномиального распределения называется *биномиальным распределением*.

Пример применения дискретного распределения в геологии

Крамбейн и Слак¹ показали, что x — число α -частиц, излучаемых поверхностью, представляющей собой круг диаметром 38 мм, — распределено как

$$f(x) = \begin{cases} \frac{\exp(-np)(np)^x}{x!} & \text{при } x = 0, 1, 2, \dots, \\ 0 & \text{для всех других } x. \end{cases}$$

В этом примере n — объем выборки или число наблюдений, тогда как p — вероятность исхода, заключающаяся в излучении α -частицы в единичном испытании. Функция $f(x)$ называется *распределением Пуассона*. Если принять, что $\lambda = np$, то

$$f(x) = \frac{\exp(-\lambda)\lambda^x}{x!}.$$

Таким образом, теоретическое дискретное распределение служит соответствующей моделью геологического эксперимента. Применение модели Пуассона рассматривается ниже, в гл. 16.

Чейз [61] получил выражение стандартного отклонения, исходя из биномиальной модели, в применении к петрографическому количественно-минералогическому анализу.

НЕПРЕРЫВНЫЕ РАСПРЕДЕЛЕНИЯ

Водородный показатель pH раствора меняется в пределах от 0 до 14 и любое значение из этого интервала будет теоретически возможным. Скорость течения реки может быть от 0 до 20 фут/сек. Величина скорости может принимать любое значение в этих пределах. Показатель отражения минералов изменяется в пределах от 1,40 до 3,22; при этом предполагается, что значения этого показателя для разных минералов находятся в указанных пределах².

Нельзя говорить, что $1\frac{1}{4}$ зерна гематита непрозрачны или что $3\frac{1}{2}$ α -частицы были получены за первую секунду наблюде-

¹ H. A. Slack, W. C. Krumbein, Measurement and statistical evaluation of low level radioactivity in rocks, Trans. A. G. U., 36, 1955; см. также W. C. Krumbein, H. A. Slack, Statistical analysis of low-level radioactivity in Pennsylvanian black fissile schale in Illinois, Bull. G. S. A., 67, 1956.

² A. F. Rogers, P. F. Kerr, Optical mineralogy, New York, 1959.

ний, но можно сказать, что вагон движется со скоростью 97,6 *миля/час*, что минерал имеет показатель отражения 1,546, что процентное содержание CaCO_3 в образце равно 33,25%, что модуль разрыва природы равен 25 020,9 *фунт/кв. дюйм*. Если переменная величины может принимать любое значение из данной области, то она называется *непрерывной переменной*. Так же как и дискретные случайные величины, непрерывные переменные характеризуются функциями распределения или плотностями вероятности.

Если $f(x)$ определена в некоторой области R , т. е. $f(x)$ определена в интервале от 0 до 123 990, или от $-\frac{\pi}{2}$ до -27π и т. п., и если $f(x) \geq 0$ для всех x , принадлежащих R , а также $\int_{-\infty}^{\infty} f(x) dx = 1$, то $f(x)$ называется *плотностью вероятностей непрерывного распределения* или *плотностью вероятностей непрерывной случайной величины* x .

Вероятность события, заключающегося в том, что случайная величина x примет значение, меньшее или равное заранее заданной величине t , может быть записана как

$$P\{x \leq t\} = \int_{-\infty}^t f(x) dx. \quad (2.1a)$$

Пусть $P\{x \leq t\} = F(t)$; $F(t)$ — функция распределения случайной величины x . Если $F(t)$ — непрерывная функция, имеющая непрерывную первую производную (за возможным исключением конечного числа точек), тогда x и ее плотность вероятностей непрерывны. Следует обратить внимание на некоторые свойства функции распределения.

1. $F(-\infty) = 0$, т. е. $\lim_{x \rightarrow -\infty} F(x) = 0$.
2. $F(\infty) = 1$, т. е. $\lim_{x \rightarrow \infty} F(x) = 1$.
3. $F(-\infty) \leq F(x_1) \leq F(x_2) \leq \dots \leq F(\infty)$,

где $-\infty \leq x_1 \leq \dots \leq \infty$.

В любом случае отношение величины изменения функции распределения к величине изменения x эквивалентно плотности вероятности $f(x)$ случайной величины. Это означает, что первая производная $F(x)$ по x есть $f(x)$. Таким образом, $dF(x)/dx = f(x)$ представляет плотность вероятностей, которой соответствует функция распределения $F(x)$. Так как

$$\int_a^a f(x) dx = 0,$$

то вероятность события $x=a$ равна 0. Следовательно, при определении вероятности в случае непрерывных случайных величин мы приписываем ее интервалу, а не точке, как это было при рассмотрении дискретных распределений. Практически не имеет значения, какая запись используется — строгое или нестрогое неравенство, потому что

$$P\{c < x \leq b\} = P\{c \leq x \leq b\} = P\{c < x < b\} = \int_c^b f(x) dx.$$

В случае непрерывных случайных величин вероятность неравенства $c \leq x \leq b$ можно рассматривать как площадь, ограниченную кривой $f(x)$, осью x и ординатами $x=c$, $x=b$. Поэтому многие функции могут быть превращены в плотность вероятности при соответствующем выборе констант. Выбор этих констант

осуществляется простым вычислением $\int_{-\infty}^{\infty} g(x) dx$. Если значение

этого выражения не 1 и не 0, то плотность вероятности случайной величины x будет

$$\frac{g(x)}{\int_{-\infty}^{\infty} g(x) dx}.$$

Например,

$$g(x) = \begin{cases} \frac{1}{2a} \left(\frac{x}{2a} + 1 \right) = \frac{x+2a}{4a^2} & \text{для } -2a \leq x \leq 2a, \\ 0 & \text{вне этого интервала} \end{cases}$$

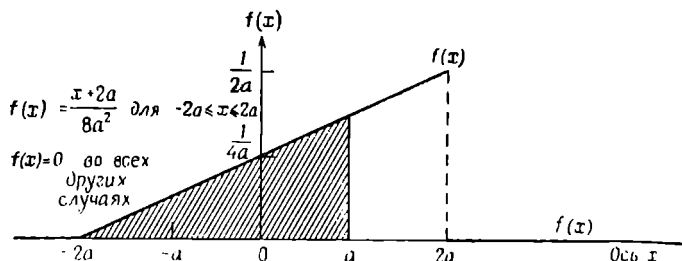
не является плотностью, так как

$$\begin{aligned} \int_{-\infty}^{\infty} g(x) dx &= \int_{-\infty}^{-2a} g(x) dx + \int_{-2a}^{2a} g(x) dx + \int_{2a}^{\infty} g(x) dx = \\ &= \int_{-\infty}^{-2a} 0 \cdot dx + \int_{-2a}^{2a} \frac{x+2a}{4a^2} dx + \int_{2a}^{\infty} 0 \cdot dx = \frac{1}{4a^2} \left(\frac{x^2}{2} + 2ax \right) \Big|_{-2a}^{2a} = \\ &= \frac{1}{4a^2} [(2a^2 + 4a^2) - (2a^2 - 4a^2)] = 2. \end{aligned}$$

Однако

$$f(x) = \begin{cases} \frac{1}{2} \cdot \frac{1}{2a} \left(\frac{x}{2a} + 1 \right) = \frac{x+2a}{8a^2} & \text{для } -2a \leq x \leq 2a, \\ 0 & \text{вне этого интервала} \end{cases}$$

является функцией плотности от x . Эта функция графически выражена на фиг. 2.5. Жирная прямая линия представляет функцию плотности.



Фиг. 2.5. График плотности вероятности $f(x) = (x+2a)/8a^2$. Можно удостовериться, что площадь, ограниченная $f(x)$, осью x и прямой $x=2a$, равна единице, т. е. $\frac{1}{2} \cdot 4a \cdot \frac{1}{2a} = 1$.

Следовательно, вероятность, например, $P(-2a \leq x \leq a)$ можно вычислить двумя способами. Первый способ:

$$\begin{aligned} \frac{1}{8a^2} \int_{-2a}^a (x+2a) dx &= \frac{1}{8a^2} \left(\frac{x^2}{2} + 2ax \right) \Big|_{-2a}^a = \\ &= \frac{1}{8a^2} \left[\left(\frac{a^2}{2} + 2a^2 \right) - (2a^2 - 4a^2) \right] = \frac{9}{16}, \end{aligned}$$

а второй заключается в измерении площади, которая заштрихована на фиг. 2.5.

Примеры непрерывных распределений

Равномерное распределение

Если случайная величина x распределена так, что

$$f(x) = \begin{cases} \frac{1}{\beta - \alpha} & \text{для } \alpha \leq x \leq \beta, \\ 0 & \text{вне этого интервала,} \end{cases}$$

то говорят, что x равномерно распределена, или имеет прямоугольное распределение:

$$P\{x \leq c\} = \frac{1}{\beta - \alpha} \int_{\alpha}^c dx = \frac{c - \alpha}{\beta - \alpha} \quad \text{при } \alpha \leq c \leq \beta,$$

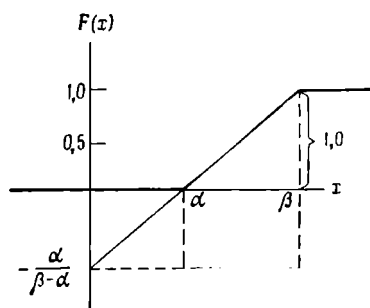
$$P\{c < x < d\} = \frac{d - c}{\beta - \alpha} \quad \text{при } \alpha \leq c \leq d \leq \beta;$$

из уравнения (2.1) (фиг. 2.6)

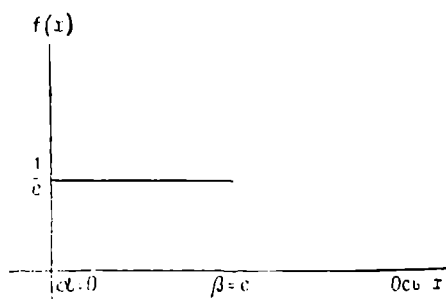
$$F(x) = \int_a^x f(x) dx,$$

$$F(x) = \frac{x-a}{\beta-a}.$$

Следовательно, если случайная величина x распределена равномерно, то вероятность того, что x примет значение, принадлежащее некоторому подмножеству области возможных значений этой величины, равна отношению абсолютной длины этого



Фиг. 2.6. График функции равномерного распределения $F(x) = \frac{x-a}{\beta-a}$, где область R есть $a \leq x \leq \beta$; $F(x) = 1$ при $x \geq \beta$, $F(x) = 0$ при $x \leq \alpha$.



Фиг. 2.7. График плотности вероятности равномерного распределения $f(x) = \frac{1}{\beta-a} = \frac{1}{e}$ при $0 \leq x \leq e$. $f(x) = 0$ во всех остальных случаях.

подмножества к абсолютной длине области возможных значений.

График равномерного распределения

$$f(x) = \begin{cases} \frac{1}{\beta-a} & \text{для } a \leq x \leq \beta, \\ 0 & \text{вне этого интервала} \end{cases}$$

при $\alpha=0$ и $\beta=e$ представлен на фиг. 2.7.

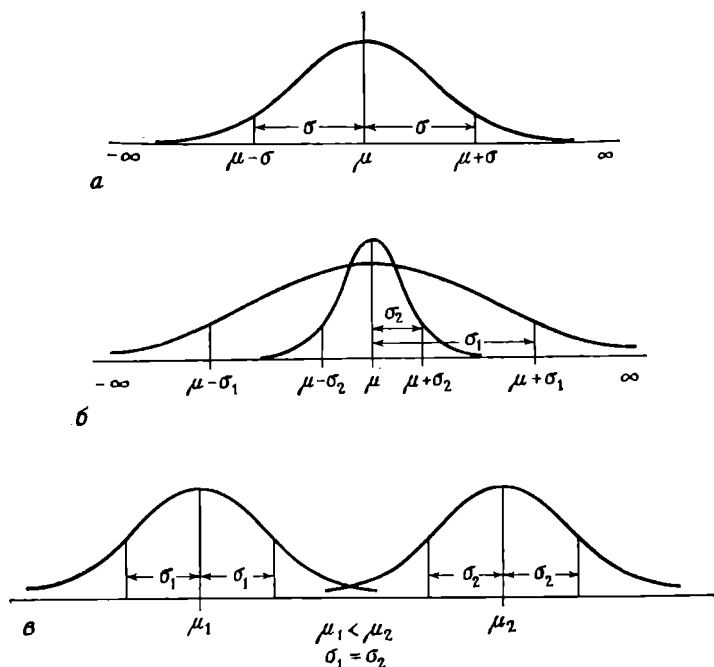
Нормальное распределение

Если случайная величина x распределена как

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2},$$

т. е. $N(\mu, \sigma^2)$, то говорят, что величина x распределена нормально. В данном случае μ и σ представляют параметры в этом двухпараметрическом семействе распределений.

На фиг. 2.8 приведены некоторые формы и положения кривых нормального распределения. Различия в форме и положении кривых распределений на оси x обусловлены изменениями значений параметров μ и σ . Из этих фигур ясно, что изменение μ при постоянном значении σ приводит к простому перемещению



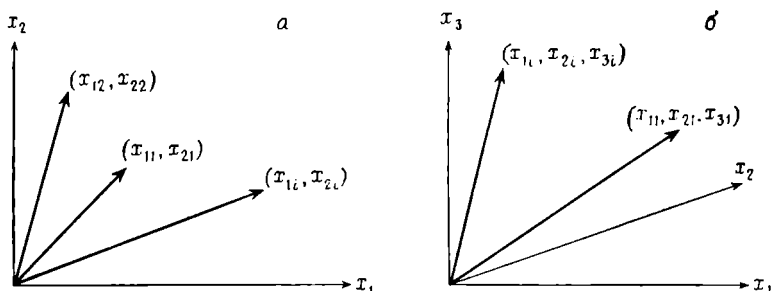
Фиг. 2.8. Некоторые формы и положения кривых нормального распределения. а — типичная кривая плотности вероятности нормального распределения с параметрами μ и σ ; б — зависимость формы кривой от значения величины σ при $\mu_1 = \mu_2$; в — зависимость положения кривой на оси абсцисс от значения μ при постоянной величине σ .

кривой по оси x , причем это перемещение осуществляется без изменения формы кривой. Изменение значений σ при постоянном μ приводит к изменению крутизны кривой распределения и к ее более заостренной форме. Таблицы значений нормального распределения легко найти, например, у Хальда¹.

¹ A. Hald, Statistical tables and formulas, New York, 1952. [Кроме того, см. Я. Янко, Математико-статистические таблицы, Госстатиздат, 1963. Краткие таблицы нормального распределения имеются в любом учебнике теории вероятностей и статистики. — Прим. ред.]

При решении задач современной физики часто бывает необходимо рассматривать несколько переменных. В примерах, которые мы приводили, была одна независимая переменная и одна зависимая переменная.

Если изучается только одно свойство объекта, т. е. производится эксперимент, и наблюдается одно событие, то каждый результат можно изобразить как точку на оси x (оси абсцисс). В случае повторных наблюдений некоторые точки в действительности будут включать несколько наблюдений. В этом случае каждому наблюдению можно приписать значение $1/n$, где n — общее число наблюдений. Следовательно, каждое наблюдение, или исход, можно однозначно описать одним значением, помещенным в некоторую точку оси x .



Фиг. 2.9. Графическое представление точек в двухмерном (а) и в трехмерном (б) пространствах.

Пусть излучаемый объект характеризуется еще одним свойством, за которым также ведется наблюдение. Тогда каждое наблюдение, или исход, будет характеризоваться двумя факторами и может быть представлено как точка на плоскости, образованной двумя взаимно перпендикулярными осями (x_1 и x_2). Таким образом, подобный исход можно представить как точку в двухмерном пространстве. Это рассуждение легко распространяется на 3, 4, ..., k факторов или свойств, которые наблюдаются в каждом из n изучаемых объектов. В данном случае каждое наблюдение будет представлено как точка в 3, 4, ..., k -мерном пространстве.

Пример двухмерного графика приведен на фиг. 2.9, а. Здесь каждая точка представлена двумя значениями. Первый индекс — рассматриваемое свойство, второй — номер наблюдения. Так, выражение (x_{17}, x_{27}) означает, что над первым и вторым свойствами проведены наблюдения в седьмом объекте.

Линия, проведенная из начала координат ($x_{1i} = x_{2i} = 0$) в любую точку двухмерного пространства, представляет вектор,

наклон и направление которого определяются положением точки в пространстве. Точно так же при помощи трехмерного графика можно представить исходы в трехмерном пространстве. В данном случае любая линия, соединяющая точку в пространстве с началом координат, будет вектором, полностью определяющим положение этой точки по отношению к осям в трехмерном пространстве, а результаты наблюдения — векторами в этом пространстве.

Это определение теперь можно обобщить на случай k -мерного пространства, где каждую точку в нем можно рассматривать и как k -мерный вектор.

В примере, приведенном на фиг. 2.9, каждому значению x_0 оси x соответствует значение $F(x_0)$, совокупность которых представляет одномерное многообразие в двухмерном пространстве. Если исходам (x_{1i}, x_{2i}) поставить в соответствие $f(x_1, x_2)$ в точках (x_{1i}, x_{2i}) , то получим поверхность в трехмерном пространстве.

Аналогично в случае k переменных соответствующая им плотность вероятностей представит k -мерную поверхность в $(k+1)$ -мерном пространстве, хотя сами исходы изображаются в виде точек в k -мерном пространстве. Обозначим случайные величины $(x_1, x_2, \dots, x_k)$ через X_k . Тогда $f(X_k)$ будет плотностью вероятностей, если

$$f(X_k) \geq 0 \quad \text{при} \quad X_k \in R,$$

$f(X_k)$ непрерывна и если

$$\int_R f(X_k) dX_k = 1,$$

где

$$dX_k = dx_1 dx_2 \dots dx_k.$$

Если для $f(X_k)$ выполнены все перечисленные условия, то она называется *совместной плотностью вероятностей случайных величин* $(x_1, x_2, \dots, x_k) = X_k$.

Понятия безусловных и условных вероятностей легко распространить и на случай k -мерных распределений.

Рассмотрим функцию

$$g(x_1, x_2) = 3 - x_1 + x_2 \quad \text{при} \quad R: \begin{cases} -3 \leq x_1 \leq 0, \\ 0 \leq x_2 \leq 3. \end{cases}$$

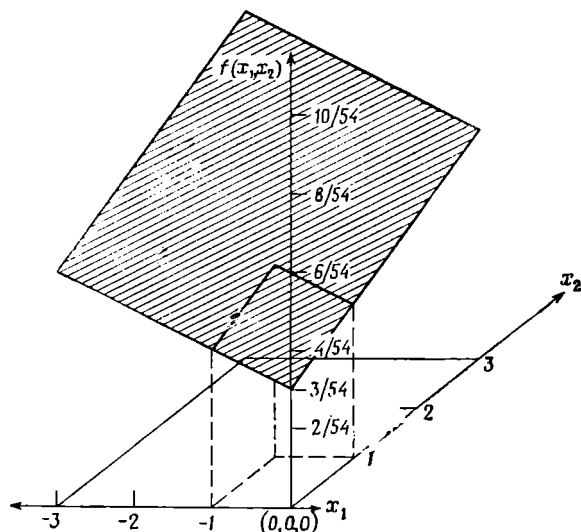
Получить из этой функции плотность вероятностей можно, проинтегрировав $g(x_1, x_2)$ по области R :

$$\int_0^3 \int_{-3}^0 (3 - x_1 + x_2) dx_1 dx_2 = \int_0^3 \left(\frac{27}{2} + 3x_2 \right) dx_2 = 54,$$

откуда

$$f(x_1, x_2) = \begin{cases} \frac{1}{54}(3 - x_1 + x_2) & \text{при } R: \begin{cases} -3 \leq x_1 \leq 0, \\ 0 \leq x_2 \leq 3, \end{cases} \\ 0 & \text{во всех других случаях.} \end{cases} \quad (2.2)$$

Из этого выражения следует, что $f(x_1, x_2)$ является плотностью вероятностей (фиг. 2.10).



Фиг. 2.10.

Если плотность известна, то, например,

вероятность $P\{-1 \leq x_1 \leq 0, 1 < x_2 < 2\}$

будет равна

$$\frac{1}{54} \int_{-1}^0 \int_1^2 (3 - x_1 + x_2) dx_1 dx_2 = \frac{1}{54} \int_1^2 \left(\frac{7}{2} + x_2\right) dx_2 = \frac{5}{54}.$$

Иногда из всех k переменных исследователю может потребоваться только одна. Иными словами, он должен исключить все случайные величины, кроме x_i , т. е. следует найти такую функцию от x_i , с помощью которой можно было бы определить $P\{b \leq x_i \leq a\}$. Эта функция определяется как $f_i(x_i)$ и называется *условной (частной) (marginal) плотностью вероятности случайной величины x_i* , которая может быть получена путем интегрирования совместной плотности $f(X_k)$ по всем пере-

менным, за исключением x_i :

$$f_i(x_i) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(X_k) dx_1 dx_2 \dots dx_{i-1} dx_{i+1} \dots dx_k.$$

Для случая

$$f(x_1, x_2) = \frac{1}{54} (3 - x_1 + x_2)$$

плотность $f_1(x_1)$ для x_1 равна

$$f_1(x_1) = \frac{1}{54} \int_0^3 (3 - x_1 + x_2) dx_2 = \frac{1}{54} \left(-3x_1 + \frac{27}{2} \right) = -\frac{x_1}{18} + \frac{1}{4}.$$

Следовательно,

$$P[a \leq x_1 \leq b] = \int_a^b \left(-\frac{x_1}{18} + \frac{1}{4} \right) dx_1 = \frac{(a-b)(a+b-9)}{36}.$$

Если $f(x_1, x_2, \dots, x_k)$ — k -мерная поверхность в $(k+1)$ -мерном пространстве, то совместная плотность вероятностей x_i, x_j, x_k будет равна

$$f_{ijk}(x_i, x_j, x_k) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(X_k) dx_1 \dots dx_{i-1} dx_{i+1} \dots \dots dx_{j-1} dx_{j+1} \dots dx_{k-1}.$$

Совместная функция распределения $(x_1, x_2, \dots, x_k)$ определится как

$$F(X_k) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \dots \int_{-\infty}^{x_k} f(X_k) dX_k,$$

а совместное распределение $(x_i, x_j, \dots, x_k)$ — как

$$F_{ijk}(x_i, x_j, x_k) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \int_{-\infty}^{x_i} \dots \int_{-\infty}^{x_j} \dots \int_{-\infty}^{\infty} \int_{-\infty}^{x_k} f(X_k) dX_k.$$

Из определения условной вероятности (1.14) следует, что

$$P[a \leq x_i \leq b | c \leq x_j \leq d] = \frac{\int_a^b \int_c^d f_{ij}(x_i x_j) dx_i dx_j}{\int_c^d f_j(x_j) dx_j};$$

поэтому условная плотность вероятностей примет вид

$$f(x_i | x_j) = \frac{f_{ij}(x_i x_j)}{f_j(x_j)}.$$

Таким образом, условная плотность вероятностей одного набора случайных величин при условии, что даны значения другого набора, представляет собой частное от деления совместной плотности обоих наборов на безусловную плотность первого набора.

Двухмерное нормальное распределение

1. Пусть x_1 и x_2 — независимые нормально распределенные случайные величины с параметрами μ_1 , σ_1^2 и μ_2 , σ_2^2 соответственно. Обозначим плотности вероятностей для x_1 через $g(x_1)$, а для x_2 через $g(x_2)$; тогда вследствие независимости совместная плотность x_1 и x_2 окажется равной

$$f(x_1, x_2) = g(x_1)g(x_2) = \frac{1}{2\pi\sigma_1\sigma_2} \exp \left\{ -\frac{1}{2} \left[\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 \right] \right\}.$$

2. Пусть x_1 и x_2 — зависимые нормально распределенные случайные величины с параметрами μ_1 , σ_1^2 и μ_2 , σ_2^2 соответственно и с ковариацией, равной σ_{12} (гл. 3).

В данных условиях совместная плотность вероятностей случайных величин x_1 и x_2 выразится в виде

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2 \left(\frac{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}{\sigma_1^2\sigma_2^2} \right)^{\frac{1}{2}}} \times \\ \times \exp \left\{ -\frac{1}{2} \left[\frac{\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 - 2 \frac{\sigma_{12}}{\sigma_1\sigma_2} \cdot \frac{(x_1 - \mu_1)}{\sigma_1} \cdot \frac{(x_2 - \mu_2)}{\sigma_2}}{\frac{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}{\sigma_1^2\sigma_2^2}} \right] \right\}.$$

Положим $\rho = \sigma_{12}/\sigma_1\sigma_2$; тогда

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2 \sqrt{1 - \rho^2}} \times \\ \times \exp \left\{ -\frac{1}{2} \left[\frac{\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 - 2\rho \left(\frac{x_1 - \mu_1}{\sigma_1} \right) \left(\frac{x_2 - \mu_2}{\sigma_2} \right)}{(1 - \rho^2)} \right] \right\}.$$

Безразмерный параметр ρ называется *коэффициентом корреляции случайных величин* x_1 и x_2 и служит мерой силы зависимости между этими величинами.

Упражнения

1. а) Найти функцию плотности из функции $6 - 2x_1 - x_2$, если

$$R: \begin{cases} 0 \leq x_1 \leq 2 \\ -1 \leq x_2 \leq 2 \end{cases}$$

$$\text{Ответ: } \frac{1}{26} (6 - 2x_1 - x_2).$$

48

Найти:

b) $f(x_1, x_2)$

c) $f_2(x_2)$

d) $f_1(x_1)$

Ответ: $\frac{1}{21}(33 - 12x_1)$;

e) $F_2(x_2)$

f) $F(x_1, x_2)$

g) $F_1(x_2)$

Ответ: $\frac{1}{21}(6x_1 - x_1^2 - 2x_1x_2)$.

Вычислить:

h) $P\left\{-\frac{1}{2} \leq x_1 \leq 1, 0 \leq x_2 \leq 2\right\}$

Ответ: $\frac{1}{6}$;

i) $P\{0 \leq x_2 \leq 2\}$

j) $P\left\{\frac{1}{2} \leq x_1 \leq 1\right\}$

k) $P\left\{\frac{1}{2} \leq x_1 \leq 1 \mid 0 \leq x_2 \leq 2\right\}$

Ответ: $\frac{7}{24}$.

2. а) Найти плотность вероятностей для $g(x_1, x_2, x_3) = 2x_1x_2x_3$ при

$$R: \begin{cases} 0 \leq x_1 \leq 2, \\ 0 \leq x_2 \leq 2, \\ 0 \leq x_3 \leq 2. \end{cases}$$

Найти:

b) $f_i(x_i)$ при $i = 1, 2, 3$

c) $f_{ij}(x_i, x_j)$

Ответ: $x_ix_j/4$;

d) $f(x_1 | x_2x_3)$

e) $f(x_1x_2 | x_3)$

Ответ: $x_1x_2/4$.

3. Являются ли x_1, x_2 и x_3 в задаче 2 независимыми случайными величинами?

Ответ: Да.

4.

$$f(x) = \begin{cases} \frac{3}{8}x_2, & \text{при } R: 0 \leq x \leq 2, \\ 0 & \text{во всех других случаях.} \end{cases}$$

а) Найти число a , при котором

$$P\{x \geq a\} = 0,9.$$

Напоминаем, что

$$P\{x \geq a\} = 1 - P\{x \leq a\}.$$

б) Найти число b , при котором

$$P\{x \leq b\} = 0,95.$$

с) Начертить графики плотности вероятностей и функции распределения и показать на них a и b .

д) Найти число c , при котором

$$P\{x \leq c\} = 0,5.$$

Величина c называется *медианой распределения* и является таким значением случайной величины, для которого

$$P\{x \leq c\} = P\{x \geq c\} = 0,5.$$

$$5) f(x) = \frac{3x^2 - 6x + 9}{20} \quad \text{при } R: 1 \leq x \leq 3.$$

Найти:

а) $P\{x \geq 1\}$;

б) $P\{x \geq 2\}$;

с) $P\{x \geq 3\}$.

Каковы вероятности, что каждое из событий, указанных в а, б и с настоящей задачи, появится три раза подряд?

$$\text{Ответ: } 1, \left(\frac{13}{20}\right)^3, 0.$$

е) Какова вероятность, что в результате четырех наблюдений будет получено хотя бы одно значение, меньшее 2?

Напоминаем, что $P(\text{хотя бы одно значение} > 2) = (1, \text{ или } 2, \text{ или } 3, \text{ или } 4 \text{ значения} > 2) = 1 - P(\text{ни одно из четырех значений не превышает } 2) = 1 - \left(\frac{13}{14}\right)^4$.

6. а) Показать, что если функция $\frac{1}{54}(3 - x_1 + x_2)$ определена в области $\begin{cases} -3 \leq x_1 \leq 0, \\ 0 \leq x_2 \leq 3, \end{cases}$ то

$$f(x_1 | x_2) = \frac{6 - 2x_1 + 2x_2}{27 + 6x_2^2}.$$

б) Найти $P\{-2 \leq x_1 \leq -1 | 0 \leq x_2 \leq 1\}$.

$$\text{Ответ: } 1/3.$$

7. Случайные величины x_1 и x_2 нормально распределены с математическими ожиданиями μ_1 и μ_2 , дисперсиями σ_1^2 и σ_2^2 и ковариацией σ_{12} . Определить коэффициент корреляции между случайными величинами $\frac{x_1 - \mu_1}{\sigma_1}$ и $\frac{x_2 - \mu_2}{\sigma_2}$.

Термин *момент* хорошо известен из физики (например, момент инерции). Центр тяжести также является моментом. Моменты определяются как интегралы или суммы в зависимости от того, какие переменные рассматриваются — непрерывные или дискретные. Если некоторая сила F приложена к стержню длиной r , закрепленному в какой-то точке или на оси, вокруг которой он может вращаться, и если общая масса тела состоит из очень малых частиц массы Δm_i на расстоянии r_i от оси вращения, то момент инерции стержня определяется как $\int r^2 dm$.

Математическим ожиданием функции от случайной величины (или самой случайной величины) называется среднее взвешенное значение этой функции (или величины), полученное в результате усреднения по всем их возможным значениям.

Моменты случайной величины представляют собой математические ожидания степеней этой величины.

Так, например, дискретная случайная величина принимает значения $x=2, 4, 6, 8, 10$ с вероятностями $1/20, 2/20, 4/20, 6/20, 7/20$ соответственно. Следовательно, математическое ожидание $E(x)$ случайной величины x есть

$$E(x) = 2/20 + 8/20 + 24/20 + 48/20 + 70/20 = 152/20 = 7,6.$$

Аналогично первым моментом случайной величины x является

$$E(x^1) = E(x) = 7,6.$$

Вообще если $f(x)$ — функция плотности вероятности дискретной случайной величины, то

$$E(x) = \sum_{x=a_1}^{a_n} x f(x), \quad (3.1)$$

52 где $x = a_1, a_2, \dots, a_n$.

Если $f(x)$ — функция плотности непрерывной случайной величины x , то математическое ожидание x определяют как

$$E(x) = \int_R x f(x) dx. \quad (3.2)$$

Моментами относительно произвольно заданной точки x_0 называют величину

$$E[(x - x_0)^r] = \int_R (x - x_0)^r f(x) dx. \quad (3.3)$$

Моменты относительно точки $x_0 = 0$ равны

$$\mu'_k = E[(x)^k] = \int_R x^k f(x) dx. \quad (3.4)$$

Штрих у величины μ_k означает моменты относительно нуля (начальные моменты). Отсюда $\mu'_1 = E(x) = \int_R x f(x) dx$. Первый начальный момент называется *средним значением распределения* $f(x)$. Моменты относительно точки $x_0 = \mu_1 = E(x)$, или среднего значения распределения, обозначаются через μ_k и равны

$$\mu_k = E[x - E(x)]^k = E[(x - \mu'_1)^k] = \int_R (x - \mu'_1)^k f(x) dx, \quad (3.5)$$

или

$$\sum_{x=a_1}^{a_n} (x - \mu'_1)^k f(x),$$

когда величина x дискретна.

Третий центральный момент $\mu_3 = E[x - E(x)]^3$ называется *асимметрией*, а четвертый $\mu_4 = E[x - E(x)]^4$ — *эксцессом*¹.

НЕКОТОРЫЕ ОБОБЩЕНИЯ И СВОЙСТВА МАТЕМАТИЧЕСКИХ ОЖИДАНИЙ

Если $f(x_1, x_2, \dots, x_k)$ — совместная плотность вероятности случайной величины $X_k = (x_1, x_2, \dots, x_k)$, а $g = g(X_k)$ — заданная функция, то

$$E(g) = E[g(X_k)] = \int_R \dots \int g(X_k) f(X_k) dX_k. \quad (3.6)$$

¹ Под терминами *асимметрия* и *эксцесс* в советской литературе имеются в виду величины $\mu_3: \mu_2^{3/2}$ и $\mu_4: \mu_2^2 - 3$. — Прим. перев.

Далее, если

$$E(g+a) = \int_R \dots \int (g+a) f(X_k) dX_k,$$

где a — некоторая константа, то

$$\begin{aligned} \int_R \dots \int (g+a) f(X_k) dX_k &= \int_R \dots \int g f(X_k) dX_k + \\ &+ a \int_R \dots \int f(X_k) dX_k = E(g) + a, \end{aligned} \quad (3.7)$$

так как

$$\int_R \dots \int f(X_k) dX_k = 1$$

при условии, что $f(X_k)$ — плотность вероятности,

$$Ega = a \int_R \dots \int g f(X_k) dX_k = aE(g), \quad (3.8)$$

$$E[g - E(g)] = \int_R [g - E(g)] f(X_k) dX_k = \int_R g f(X_k) dX_k - E(g).$$

Следовательно,

$$E[g - E(g)] = 0. \quad (3.9)$$

Если $h = h(X_k)$ — также случайная величина, то

$$\begin{aligned} E(h+g) &= \int_R \dots \int (h+g) f(X_k) dX_k = \\ &= \int_R \dots \int h f(X_k) dX_k + \int_R \dots \int g f(X_k) dX_k \end{aligned}$$

и, таким образом,

$$E(h+g) = E(h) + E(g). \quad (3.10)$$

Если x_i и x_j — две статистически независимые случайные величины и их совместная плотность вероятности есть $f_{ij}(x_i, x_j)$, то

$$f_{ij}(x_i, x_j) = f_i(x_i) f_j(x_j)$$

и

$$E(x_i x_j) = \int \int x_i f_i(x_i) x_j f_j(x_j) dx_i dx_j = E(x_i) E(x_j). \quad (3.11)$$

Дисперсия

Второй центральный момент случайной величины называется *дисперсией* этой величины; она равна

$$\begin{aligned} \text{var}(x) &= E[x - E(x)]^2 = E[x^2 - 2xE(x) + [E(x)]^2] = \\ &= E x^2 - 2E(x)E(x) + [E(x)]^2 = \\ &= E x^2 - [E(x)]^2 = \mu_2 = \mu_2' - \mu_1'^2. \end{aligned} \quad (3.12)$$

Дисперсию случайной величины x часто обозначают через σ^2 ; величина $\sqrt{\sigma^2} = \sqrt{\text{var}(x)} = \sigma$ называется *стандартным отклонением*.

Некоторые свойства дисперсии

Если a — константа, то

$$\text{var}(x + a) = E[(x + a) - E(x + a)]^2,$$

где

$$E(x + a) = E(x) + E(a) = \mu'_1 + a.$$

Следовательно,

$$\text{var}(x + a) = E(x + a - \mu'_1 - a)^2 = E(x - \mu'_1)^2 = \mu_2 = \text{var } x, \quad (3.13)$$

$$\begin{aligned} \text{var}(ax) &= E[ax - E(ax)]^2 = E[ax - a\mu'_1]^2 = \\ &= E(a^2x^2) - 2a^2\mu'_1 E(x) + a^2(\mu'_1)^2 = \\ &= a^2Ex^2 - a^2(\mu'_1)^2 = a^2E[x^2 - (\mu'_1)^2] = a^2\mu_2 = a^2 \text{var } x. \end{aligned} \quad (3.14)$$

КОВАРИАЦИЯ И КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ

Ковариацией двух случайных величин x_1 и x_2 называется величина

$$\text{cov}(x_1x_2) = E\{[x_1 - E(x_1)][x_2 - E(x_2)]\}. \quad (3.15)$$

Переписав (3.15) в виде

$$Ex_1x_2 - E(x_1)E(x_2) = E(x_1)(x_2) + E(x_1)E(x_2),$$

получим¹

$$\text{cov}(x_1x_2) = E(x_1x_2) - E(x_1)E(x_2).$$

Ковариацию обычно обозначают через σ_{12} .

Коэффициент корреляции ρ между двумя случайными величинами x_1 и x_2 определяется по формуле

$$\rho_{x_1x_2} = \rho_{12} = \frac{\text{cov}(x_1x_2)}{[\text{var}(x_1)\text{var}(x_2)]^{1/2}} = \frac{\sigma_{x_1x_2}}{\sigma_{x_1}\sigma_{x_2}}.$$

Соответственно

$$\rho_{12}^2 = \frac{\sigma_{x_1x_2}^2}{\sigma_{x_1}^2\sigma_{x_2}^2}. \quad (3.16)$$

Если x_1 и x_2 — независимые случайные величины, то, согласно (3.11),

$$E(x_1x_2) = E(x_1)E(x_2),$$

¹ $E[E(x_1)x_2] = E(x_1)E(x_2)$, так как $E(x_1)$ — константа, т. е. $E(ax) = aE(x)$. **55**

и тогда

$$\sigma_{x_1 x_2} = 0,$$

а в соответствии с (3.15)

$$\rho_{12} = 0.$$

Обратное утверждение необязательно справедливо, т. е. если $\rho_{12} = 0$, то из этого необязательно должна следовать статистическая независимость x_1 и x_2 . Теперь можно определить дисперсию суммы случайных величин x и y как

$$\text{var}(x + y) = \text{var}(x) + \text{var}(y) + 2 \text{cov}(xy).$$

НЕРАВЕНСТВО ШВАРЦА И ВЕРХНЕЕ ЗНАЧЕНИЕ ЛИНЕЙНОГО КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ

Можно показать¹, что

$$E(x_1^2)E(x_2^2) \geq [E(x_1 x_2)]^2. \quad (3.17)$$

Это выражение называется *неравенством Шварца*. Покажем, что значения линейного коэффициента корреляции ρ_{12} двух случайных величин x_1 и x_2 не превышают единицы¹.

Вместо пары случайных величин x_1 и x_2 введем новые случайные величины $x_1 - E(x_1)$ и $x_2 - E(x_2)$. После подстановки этих величин в выражение (3.17) получим

$$E[x_1 - E(x_1)]^2 E[x_2 - E(x_2)]^2 \geq E\{[x_1 - E(x_1)][x_2 - E(x_2)]\}^2.$$

Если разделить обе части этого неравенства на величину

$$E[x_1 - E(x_1)]^2 E[x_2 - E(x_2)]^2,$$

которая не равна нулю, то

$$\frac{\{E[x_1 - E(x_1)][x_2 - E(x_2)]\}^2}{E[x_1 - E(x_1)]^2 E[x_2 - E(x_2)]^2} \leq 1. \quad (3.18)$$

Однако левая часть выражения (3.18) идентична (3.16), поэтому

$$\rho_{12}^2 \leq 1.$$

Таким образом,

$$|\rho_{12}| \leq 1$$

и

$$-1 \leq \rho_{12} \leq 1.$$

¹ См., например, М. Е. Мунroe, Theory of probability, New York, 1951.

Равномерное, или прямоугольное, распределение¹

$$f(x) = \begin{cases} \frac{1}{\beta - \alpha} & \text{при } R: \alpha \leq x \leq \beta, \\ 0 & \text{во всех других случаях.} \end{cases}$$

$$\mu'_1 = E(x) = \frac{1}{\beta - \alpha} \int_{\alpha}^{\beta} x dx = \frac{\beta^2 - \alpha^2}{2(\beta - \alpha)} = \frac{\beta + \alpha}{2},$$

$$\mu'_2 = E(x^2) = \frac{1}{\beta - \alpha} \int_{\alpha}^{\beta} x^2 dx = \frac{\beta^3 - \alpha^3}{3(\beta - \alpha)} = \frac{\beta^2 + \alpha^2 + \alpha\beta}{3},$$

$$\begin{aligned} \text{var } x = \mu'_2 - (\mu'_1)^2 &= \frac{4(\alpha^2 + \beta^2 + \alpha\beta) - 3(\beta^2 + \alpha^2 + 2\alpha\beta)}{12} = \\ &= \frac{\alpha^2 + \beta^2 - 2\alpha\beta}{12} = \frac{(\beta - \alpha)^2}{12}. \end{aligned}$$

Биномиальное распределение

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{при } R: x = 0, 1, 2, \dots, n, \\ 0 & \text{во всех других случаях.} \end{cases}$$

$$\mu'_1 = E(x) = \sum_{x=0}^n x f(x) = \sum_{x=0}^n x \binom{n}{x} p^x (1-p)^{n-x}.$$

Теперь можно записать это выражение в виде

$$0 + \sum_{x=1}^n x \frac{n(n-1)!}{(n-x)! x (x-1)!} p p^{x-1} (1-p)^{n-x}.$$

Вынося np за знак суммы, получим

$$np \sum_{x=1}^n \frac{(n-1)!}{(n-x)!(x-1)!} p^{x-1} (1-p)^{n-x}.$$

Выражение

$$\sum_{x=1}^n \binom{n-1}{x-1} p^{x-1} (1-p)^{n-x}$$

равно сумме всех вероятностей в биномиальном распределении с параметрами p и $n-1$, т. е. единице. Поэтому

$$\mu'_1 = Ex = np.$$

¹ Griffiths, Rosenfeld, A further test of dimensional orientation of quartz grain in Bradford sand, Am. Journ. Sci., 251, 1953.

Теперь

$$\text{var } x = E(x)^2 - (\mu'_1)^2,$$

что можно выразить в виде

$$E[x(x-1) + x] - (\mu'_1)^2 = E[x(x-1)] + E(x) - (\mu'_1)^2,$$

где

$$E[x(x-1)] =$$

$$= 0 + 0 + \sum_{x=2}^n x(x-1) \frac{n(n-1)(n-2)!}{(n-x)! x(x-1)(x-2)!} p^2 p^{x-2} (1-p)^{n-x} =$$
$$= n(n-1)p^2.$$

Следовательно,

$$\mu'_2 = \text{var } x = n(n-1)p^2 + np - n^2p^2 = np(1-p).$$

Упражнения

1. Число контактирующих зерен в аркозах распределено биномиально с параметром $p=0,6$.

Найти: а) среднее число контактирующих зерен в шлифе с 2000 зернами; б) дисперсию числа контактирующих зерен и с) стандартное отклонение.

Ответ: а) 1200; б) 480; с) 21,8.

2. Если $h(x) = e^{xt}$ и величина x распределена как $f(x)$ при $x \in R$, найти $E(e^{xt})$.

$$\text{Ответ: } \int_R e^{xt} f(x) dx.$$

3. Пусть $g(t) = E(e^{xt})$. Найти:

а) $\frac{dg(t)}{dt}$.

$$\text{Ответ: } \int_R x e^{xt} f(x) dx.$$

б) $\left. \frac{dg(t)}{dt} \right|_{t=0}$ (т. е. первая производная от $g(t)$ по t в точке $t=0$).

$$\text{Ответ: } \int_R x f(x) dx.$$

с) $\left. \frac{d^2 g(t)}{dt^2} \right|_{t=0}$.

$$\text{Ответ: } \int_R x^2 f(x) dx.$$

д) $\left. \frac{d^r g(t)}{dt^r} \right|_{t=0}$.

$$\text{Ответ: } \int_R x^r f(x) dx.$$

е) Каково функциональное соотношение между $g(t)$ и начальными моментами $f(x)$?

Ответ: Производная порядка r от $g(t)$ по t при $t=0$ равна μ_r ; функция $g(t)$ называется *производящей функцией моментов распределения*.

4. Найти: а) производящую функцию моментов; б) среднее значение; в) дисперсию; г) стандартное отклонение следующих распределений:

$$4(1). f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}.$$

Ответ: а) $\exp\left(\frac{2t\mu + \sigma^2 t^2}{2}\right)$;

б) μ ;

в) σ^2 ;

г) σ .

$$4(2). f(x) = \begin{cases} \frac{x^\gamma e^{-x/\lambda}}{\gamma! \lambda^{\gamma+1}} & \text{при } x > 0, \\ 0 & \text{при } x \leq 0. \end{cases}$$

Ответ: а) $(1 - \lambda \varepsilon)^{-(\gamma+1)}$;

б) $\lambda(\gamma + 1)$;

в) $\lambda^2(\gamma + 1)$;

г) $\lambda(\gamma + 1)^{1/2}$.

Распределение 4(2) называется *гамма-распределением*.

5. Найти: а) среднее значение; б) дисперсию бета-распределения

$$f(x) = \begin{cases} \frac{\alpha + \beta + 1}{\alpha! \beta!} x^\alpha (1-x)^\beta & \text{при } R: 0 < x < 1, \\ 0 & \text{во всех других случаях.} \end{cases}$$

Ответ: а) $\frac{(\alpha + \beta + 1)! (\alpha + 1)!}{(\alpha + \beta + 2)! \alpha!} = \frac{\alpha + 1}{\alpha + \beta + 2}$;

б) $\frac{(\alpha + 1)(\beta - 1)}{(\alpha + \beta + 2)^2 (\alpha + \beta + 3)}$.

6. Найти μ'_2 бета-распределения.

Ответ: $\frac{(\alpha + \beta + 1)! (\alpha + r)!}{(\alpha + \beta + r + 1)! \alpha!}$.

7. Доказать, что среднее значение и дисперсия распределения Пуассона равны (гл. 2, стр. 39).

8. Написать дискретные аналоги к формулам (3.6) — (3.11). **59**

В задачах экспериментальных исследований редко бывают заранее точно известны функции распределения изучаемых случайных величин. Имеется много примеров, в которых теоретическая модель служит основой для плана выборки; эта модель дает представление о значениях, которые в ней могут принимать параметры. Иногда такие модели бывают удачными и соответствующими действительности¹, иногда же — неудачными [183]. Далее, в большинстве случаев эксперименты проводятся с целью сделать заключения не об изучаемой выборке, а о той совокупности, из которой она взята. Научный метод должен приводить к выводам о совокупности по взятой из нее выборке. Статистика служит средством получения объективного вывода о совокупностях по выборкам. Какого типа выводы могут быть сделаны? Экспериментатору может понадобиться оценить параметры распределения, которому подчинены исследуемые случайные величины, или проверить гипотезу по материалу, с которым он работает. Таким образом, статистические выводы можно разделить на следующие две не обязательно взаимно независимые группы: 1) статистические выводы типа оценки параметров и 2) статистические выводы типа проверки гипотез о значениях этих параметров.

МЕТОД МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ

Один из методов получения оценок параметров — метод максимального правдоподобия, предложенный и развитый Фишером. Мы рассмотрим теорию этого метода только на примере распределения с непрерывной плотностью вероятности, так как приложение теории максимального правдоподобия к дискретным случайным величинам очевидно.

Рассмотрим выборку объемом n , взятую из совокупности, подчиненной непрерывному закону распределения с плотностью вероятности $f(x_i, \theta)$, где $\theta = \theta_1, \theta_2, \dots, \theta_k$ — параметры распределения. Функция максимального правдоподобия представляет функцию от всех выборочных точек и определяется формулой

$$L = L(x_1, x_2, \dots, x_n, \theta) = f(x_1, \theta) f(x_2, \theta) \dots f(x_n, \theta) = \prod_{i=1}^n f(x_i, \theta). \quad (4.1)$$

Нередко вместо функции L используют ее натуральный логарифм. Допустим, что выборка произведена и значения $x_1, x_2, \dots, x_n$ стали известны, тогда L будет функцией только неизвестных параметров $\theta_1, \theta_2, \dots, \theta_k$. Основная идея метода максимального правдоподобия заключается в выборе тех значений $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$, при которых выражение (4.1) достигает максимума. Таким образом, будет справедливо неравенство

$$\prod_{i=1}^n f(x_i, \hat{\theta}) > \prod_{i=1}^n f(x_i, \hat{\theta}_i),$$

где $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$ — любая другая последовательность оценок.

Определение максимума функции правдоподобия

Теперь мы подготовлены к определению максимума функции правдоподобия. Это делается обычным способом, с помощью математического анализа. Сначала дифференцируем L по соответствующему параметру, а затем приравниваем производную или, когда параметров несколько, частные производные к нулю. Далее, решив полученное уравнение (или систему уравнений) относительно параметра (параметров), получим значение, соответствующее максимуму или минимуму функции L . Крайне важно уметь отличать значения, соответствующие максимуму, от минимального, так как нас интересует только первое из них. В итоге из уравнения

$$\frac{\partial L}{\partial \theta_i} = \frac{\partial \prod_{i=1}^n f(x_i, \theta)}{\partial \theta_i} = 0$$

находим

$$\hat{\theta}_i = g_i(x_1, x_2, \dots, x_n)$$

и проверяем, дает ли $\hat{\theta}_i$ максимум.

Среднее и дисперсия нормального распределения

В качестве примера рассмотрим задачу об оценке параметров μ и σ^2 нормального распределения. В данном случае все значения x_i имеют распределение

$$f(x_i, \theta = \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right].$$

Если из такой совокупности взята наудачу выборка объемом n , то вследствие независимости ее членов функция правдоподобия примет вид

$$L = \prod_{i=1}^n f(x_i, \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right].$$

Проследим теперь ход получения оценки параметра μ . Находим $\frac{\partial L}{\partial \mu}$ и приравниваем к нулю. Так как $\ln L$ и L достигают максимума при одном и том же значении μ , то можно как более простой вариант рассмотреть $\ln L = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$ вместо L . Получив

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0,$$

находим

$$\sum_{i=1}^n (x_i - \mu) = 0, \quad \sum_{i=1}^n x_i = n\mu, \quad \sum_{i=1}^n \frac{x_i}{n} = \hat{\mu}.$$

Следовательно, максимально правдоподобной оценкой неизвестного параметра μ является выборочная статистика $\hat{\mu}$, т. е. первый начальный момент, или выборочная средняя, которую чаще всего обозначают через $\bar{x}$.

Чтобы получить максимально правдоподобную оценку параметра σ^2 , вычисляем

$$\begin{aligned} \frac{\partial \ln L}{\partial \sigma^2} &= \frac{\partial}{\partial \sigma^2} \left[-\frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right] = \\ &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2. \end{aligned}$$

Заменив σ^2 и μ на $\hat{\sigma}^2$ и $\bar{x}$, получаем

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4} \sum_{i=1}^n (x_i - \bar{x})^2 = 0,$$

$$n\hat{\sigma}^2 = \sum_{i=1}^n (x_i - \bar{x})^2, \quad \hat{s}^2 = \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Таким образом, максимально правдоподобной оценкой σ^2 является второй выборочный центральный момент, или выборочная дисперсия $\hat{s}^2$ ¹.

МЕТОД МОМЕНТОВ

Для получения оценок параметров можно также использовать метод моментов, разработанный Пирсоном.

Этот метод заключается в следующем. Допустим, что из совокупности с распределением $f(x_i, \theta)$, где $\theta = \theta_1, \theta_2, \dots, \theta_k$, взята наудачу выборка объемом n . Допустим далее, что начальные моменты плотности вероятности случайной величины x существуют и являются функциями параметров $\theta = \theta_1, \theta_2, \dots, \theta_k$.

Обозначая $\mu'_i = E(x)^i$, получим

$$\begin{aligned}\mu'_1 &= g_1(\theta_1, \theta_2, \dots, \theta_k), \\ \mu'_2 &= g_2(\theta_1, \theta_2, \dots, \theta_k), \\ &\vdots \\ \mu'_k &= g_k(\theta_1, \theta_2, \dots, \theta_k).\end{aligned}\tag{4.2}$$

Метод моментов основан на том, что сначала подставляют оценки $\hat{\theta}_i$ для θ_i в (4.2), затем заменяют моменты совокупности μ'_i выборочными моментами m'_i . Таким образом получают ряд уравнений в терминах неизвестных оценок и известных выборочных моментов, причем число уравнений должно быть

¹ Выборочные моменты определяются аналогично определению моментов генеральной совокупности. Так, в выборке объемом n

$$m'_r = \sum_{l=1}^n \frac{x'_l}{n}$$

H

$$m_r = \sum_{i=1}^n \frac{(x_i - \bar{x})^r}{n}.$$

равным числу неизвестных параметров:

$$\begin{aligned} m'_1 &= g_1(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k) = g_1(x_1, x_2, \dots, x_n), \\ m'_2 &= g_2(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k) = g_2(x_1, x_2, \dots, x_n), \\ . &. \\ m'_k &= g_k(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k) = g_k(x_1, x_2, \dots, x_n). \end{aligned}$$

Среднее и дисперсия нормального распределения

Для иллюстрации этого метода рассмотрим задачу оценки параметров μ и σ^2 нормального распределения.

Пусть $\theta_1 = \mu$, а $\theta_2 = \sigma^2$. Известно, что

$$\begin{aligned}\mu'_1 &= \theta_1, \\ \mu'_2 &= \theta_1^2 + \theta_2.\end{aligned}$$

Следовательно,

$$\left. \begin{aligned} m'_1 &= \hat{\theta}_1, \\ m'_2 &= \hat{\theta}_1^2 + \hat{\theta}_2. \end{aligned} \right\} \quad (4.3)$$

Таким образом, $\hat{\theta}_1$ оказывается ничем иным, как

$$\bar{x} = m'_1 = \frac{1}{n} \sum_{i=1}^n x_i.$$

Заменяя в (4.3) $\hat{\theta}_1$ на m'_1 , получим

$$m'_2 - (m'_1)^2 = \hat{\theta}_2,$$

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

Итак, по крайней мере в этом случае, метод моментов дает оценки, совпадающие с максимально правдоподобными. В общем случае метод моментов приводит к оценкам, которые представляют собой функции выборочных моментов. Являются ли эти оценки «хорошими»? Этот вопрос требует дополнительного изучения в каждом отдельном случае.

КРИТЕРИИ КАЧЕСТВА ОЦЕНОК

Как уже отмечалось, может случиться, что существует одна, две или более оценок неизвестного параметра. Пусть θ — неизвестный параметр. Обозначим функцию от случайных ве-

личии (результатов наблюдений), которая будет служить оценкой для θ , через $\Phi(X_k = x_1, x_2, \dots, x_k)$. Теперь определим некоторые свойства оценок, которыми, как мы интуитивно угадываем, они должны обладать, и покажем, что оценки во многом сходны со случайными величинами, из которых они получены. Затем изучим распределение вероятностей оценок и некоторых функций от них.

Несмещенные оценки

Определение

$\Phi(X_k)$ является несмещенной оценкой для θ , если

$$E[\Phi(X_k)] = \theta,$$

где k — константа¹. Если $E[\Phi(X_k)] > \theta$, то $\Phi(X_k)$ называется *положительно смещенной оценкой*; если же $E[\Phi(X_k)] < \theta$, то $\Phi(X_k)$ называют *отрицательно смещенной оценкой*.

Основываясь на общем представлении о математическом ожидании, можно интерпретировать несмещенность следующим образом. Если $\Phi(X_k)$ очень много раз используется в качестве оценки θ , то среднее из значений $\Phi(X_k)$ будет равным θ . В любом единичном случае, т. е. в любой единичной выборке объема k , не обязательно $\Phi(X_k) = \theta$. Однако если оценка $\Phi(X_k)$ будет вычисляться во многих опытах, то среднее значение будет приближаться к θ . Таким образом, разумность предварительного требования несмещенности можно считать доказанной, так как в действительности было бы абсурдным пользоваться в качестве оценки для θ величиной $\Phi(X_k)$, для которой, например,

$$E[\Phi(X_k)] = \pi e^3,$$

тогда как $\theta = e/\pi$. Однако в подобном случае можно легко устранить это явное несоответствие, рассматривая новую оценку для θ , а именно $\xi(X_k) = \frac{1}{\pi^2 e^2} \Phi(X_k)$, математическое ожидание которой равно θ :

$$E[\xi(X_k)] = \frac{\pi e^3}{\pi^2 e^2} = \frac{e}{\pi}.$$

Таким образом, в примерах, столь же тривиальных, как приведенный, в смещенную оценку можно ввести поправку, если известна величина смещения. К сожалению, она не всегда известна.

¹ Поскольку распределения X_k и $\Phi(X_k)$ зависят от неизвестного θ , то и соответствующие математические ожидания являются функциями θ . Поэтому правильнее было бы писать E_θ вместо E . Несмещенность означает, что $E_\theta \Phi(X) = \theta$ при всех θ . Аналогичное требование относится и к положительной (отрицательной) смещенности. Возможно, конечно, что оценка не входит ни в один из этих классов. — Прим. ред.

Пример

$E[\eta(X_k)] = \theta + \varepsilon$. Найти новую оценку, не имеющую смещения. *Ответ: $\eta(X_k) - \varepsilon$.*

Каково смещение оценки — положительное или отрицательное? *Ответ: положительное.*

Состоятельные оценки

Рассмотрим последовательность случайных выборок $X_1 = x_1$; $X_2 = x_1, x_2$; ...; $X_k = x_1, x_2, \dots, x_k$ объемом $j = 1, 2, \dots, k$. В каждой выборке оценивается неизвестный параметр θ при помощи оценки $\Phi(X_j)$.

Определение

Последовательность оценок

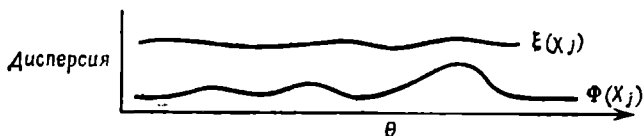
$$\Phi(X_1), \Phi(X_2), \dots, \Phi(X_k) = \{\Phi(X_j)\}$$

называют *состоятельной*, если вероятность того, что $\Phi(X_k)$, которое отличается от θ меньше, чем на сколь угодно малое число, стремится к единице при неограниченном возрастании j . Это можно записать в виде выражения $P\{\{\Phi(X_j)\}\}$, которое отличается от θ не более, чем на величину $\Delta\} \rightarrow 1$ при $j \rightarrow \infty$, где Δ — сколь угодно малое заданное число.

Эффективные оценки

Определение

Если оценка $\Phi(X_j)$ неизвестного параметра θ обладает минимальной дисперсией из всех возможных оценок, она называется *эффективной*. Такая оценка предпочтительнее любой

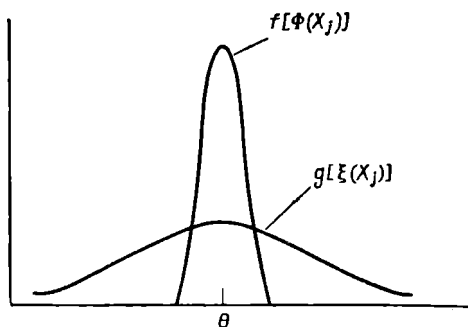


Фиг. 4.1.

другой несмещенной оценки $\xi(X_j)$, если $\Phi(X_j)$ не смещена и если дисперсия $\Phi(X_j)$ меньше или равна дисперсии $\xi(X_j)$.

Преимущества свойств эффективной оценки показаны на графике (фиг. 4.1). На этом графике изображены дисперсии оценок $\Phi(X_j)$ и $\xi(X_j)$ для всех значений θ . Нетрудно видеть, что $\Phi(X_j)$ предпочтительнее $\xi(X_j)$, так как при любом значении θ дисперсия $\Phi(X_j)$ меньше дисперсии $\xi(X_j)$. Это значит,

что в данном случае отклонения оценки $\Phi(X_j)$ от среднего значения в предположении несмещенности относительно θ меньше, чем у любой другой несмещенной оценки. Аналогичный пример приведен на фиг. 4.2. На этом графике сравниваются плотности вероятностей $f[\Phi(X_j)]$ и $g[\xi(X_j)]$ двух несмещенных оценок



Фиг. 4.2.

$\Phi(X_j)$ и $\xi(X_j)$ неизвестного параметра θ . Легко видеть, что вокруг θ значения оценки $\Phi(X_j)$ группируются теснее, чем значения $\xi(X_j)$.

Достаточные оценки

Определение

Оценка $\Phi(X_j)$ называется *достаточной оценкой неизвестного параметра* θ , если существует такая функция $h(X_j)$, не зависящая от θ , для которой выполняется равенство¹

$$\prod_{i=1}^n f(X_i, \theta) = L = g[\Phi(X_j)] h(X_j),$$

где L — функция правдоподобия для выборки объемом n , $g[\Phi(X_j)]$ — плотность распределения оценки $\Phi(X_j)$, $f(X_i, \theta)$ — плотность вероятностей случайной величины в точке x_i . Достаточная оценка включает всю информацию о неизвестном параметре, которую можно получить из выборки объемом n .

ВЫБОРОЧНЫЕ РАСПРЕДЕЛЕНИЯ; ИХ РОЛЬ И ПРИМЕНЕНИЕ

Оценка, или оценивающая функция, состоит из комбинации случайных величин. Следовательно, оценки являются функ-

¹ Здесь g может зависеть от θ , т. е. правильнее писать не $g[\Phi(X_j)]$, а $g[\Phi(X_j), \theta]$. — *Прим. ред.*

циями выборочных значений, и поэтому они изменяются при повторении выборок. Ясно, что это обстоятельство заставляет рассматривать оценку как случайную величину в том смысле, в котором она была определена в гл. 1. Так как нельзя заранее точно предсказать, какое значение примет случайная величина в результате единичного эксперимента, то в случае оценок также невозможно точно предсказать отдельный исход. Поэтому, чтобы говорить о надежности или точности оценки, необходимо изучить ее распределение как обычной случайной величины. Распределения оценок, как и любых других выборочных функций, называются *выборочными распределениями*. Мы рассмотрим распределения оценок некоторых параметров и распределения функций от них. В данном случае, как и при рассмотрении других плотностей вероятностей, будут изучаться такие характеристики, как средние значения, дисперсии и ковариации. Анализ выборочных распределений позволит сделать ряд вероятностных утверждений относительно наиболее интересных событий.

РАСПРЕДЕЛЕНИЕ χ^2

Пусть случайная величина x распределена нормально со средним $\mu=0$ и дисперсией $\sigma^2=1$, т. е. по закону $N(0, 1)$:

$$f(x) = \frac{1}{(2\pi)^{1/2}} e^{-x^2/2}, \quad -\infty < x < \infty.$$

Можно показать¹, что случайной величине $z = x^2$ соответствует плотность

$$\frac{1}{(2\pi)^{1/2}} \cdot \frac{e^{-z/2}}{\sqrt{z}}, \quad 0 \leq z < \infty,$$

а плотность вероятностей распределения случайной величины

$\chi_n^2 = \sum_{i=1}^n z$ определена как

$$\frac{1}{\left(\frac{n-2}{2}\right)!} \cdot \frac{1}{z^{n/2}} (\chi_n^2)^{\frac{n-2}{2}} e^{-\frac{\chi_n^2}{2}}, \quad \chi_n^2 > 0. \quad (4.4)$$

Это распределение называется *распределением χ^2 с n степенями свободы*.

Среднее и дисперсия χ_n^2

Среднее значение случайной величины χ_n^2 равно n , а дисперсия — $2n$. Это можно доказать с использованием производящей

функции моментов. Распределение χ_n^2 при различном числе степеней свободы можно найти в таблицах любого учебника по статистике.

ФУНКЦИЯ РАСПРЕДЕЛЕНИЯ

Функцию распределения χ_n^2 можно записать как

$$F(\chi_n^2) = \int_0^{\chi_n^2} \frac{1}{\left(\frac{n-2}{2}\right)! 2^{\frac{n}{2}}} (\chi_n^2)^{\frac{n-2}{2}} e^{-\frac{\chi_n^2}{2}} d\chi_n^2.$$

Функция $F(\chi_n^2)$ зависит от n — числа степеней свободы. Таким образом, при данном n

$$\begin{aligned} P\{b \leq \chi_n^2 \leq a\} &= a - F[b] = \\ &= P\{\chi_n^2 \leq a\} - P\{\chi_n^2 \leq b\}, \quad 0 < b \leq a \leq \infty. \end{aligned}$$

Следовательно, чтобы вычислить значение α для любой случайной величины, распределенной по закону χ^2 , необходимо знать a , b и n . Имеются и удобные для пользования таблицы распределения χ^2 двух типов. Одна из них предназначена для нахождения значения вероятности при заданном χ^2 и соответствующем числе степеней свободы, а другая служит для отыскания значения χ^2 при заданной вероятности и данном n . При пользовании таблицами следует обращать внимание на то, какая функция табулирована, т. е.

$$\int_0^{\chi_n^2} \frac{1}{\left(\frac{n-2}{2}\right)! 2^{\frac{n}{2}}} (\chi_n^2)^{\frac{n-2}{2}} e^{-\frac{\chi_n^2}{2}} d\chi_n^2 = F(\chi_n^2)$$

или

$$1 - \int_0^{\chi_n^2} f(\chi_n^2) d\chi_n^2 = \int_{\chi_n^2}^{\infty} f(\chi_n^2) d\chi_n^2.$$

Последнее выражение равно $1 - F(\chi_n^2)$.

Теорема сложения для распределений χ^2

Знать свойство сложения случайных величин, распределенных по закону χ^2 , особенно необходимо при изучении техники дисперсионного анализа (гл. 8). В связи с этим рассмотрим

плотность вероятностей случайной величины

$$v = \chi_{n_1}^2 + \chi_{n_2}^2,$$

где

$$\chi_{n_1}^2 = \sum_{i=1}^{n_1} z_i = \sum_{i=1}^{n_1} x_i^2,$$

$$\chi_{n_2}^2 = \sum_{i=n_1+1}^{n_1+n_2} z_i = \sum_{i=n_1+1}^{n_1+n_2} x_i^2.$$

Каждая из величин x_i распределена нормально с математическим ожиданием, равным нулю, и единичной дисперсией $N(0, 1)$, а случайные величины $\chi_{n_1}^2$ и $\chi_{n_2}^2$ статистически независи-

мы. Так как $\sum_{i=1}^{n_1} z_i + \sum_{i=n_1+1}^{n_1+n_2} z_i = \sum_{i=1}^{n_1+n_2} z_i$ — сумма $n_1 + n_2 = n$

случайных величин x_i^2 , то $\sum_{i=1}^{n_1+n_2} z_i$ представляет собой случайную величину, распределенную по закону χ^2 с $n_1 + n_2$ степенями свободы. Таким образом, сумма двух любых независимых χ^2 -распределенных случайных величин также распределена, как χ^2 с числом степеней свободы, равным сумме чисел степеней свободы слагаемых.

ТЕОРЕМА КОХРЕНА

Если сумма квадратов $x_1^2 + x_2^2 + \dots + x_n^2$ независимых одинаково распределенных по закону $N(0, 1)$ случайных величин разделена на m сумм квадратов $S_1^2, S_2^2, \dots, S_m^2$ с $n_1, n_2, \dots, n_m$ линейно независимыми слагаемыми каждая, т. е.

$$\chi^2 = \sum_{i=1}^n x_i^2 = S_1^2 + S_2^2 + \dots + S_m^2,$$

то необходимым и достаточным условием для того, чтобы $S_1^2, S_2^2, \dots, S_m^2$ были независимы в вероятностном смысле и распределены как χ^2 с $n_1, n_2, \dots, n_m$ степенями свободы, будет

$$\sum_{j=1}^m n_j = n.$$

F-РАСПРЕДЕЛЕНИЕ ФИШЕРА

Пусть x_1 и x_2 — независимые случайные величины, распределенные по закону χ^2 с n_1 и n_2 степенями свободы соответственно. Плотность вероятности отношения¹

$$z = \frac{x_1/n_1}{x_2/n_2}$$

равна

$$g(z) = \frac{n_1^{n_1/2} n_2^{n_2/2} z^{(n_1/2)-1} \left(\frac{n_1 + n_2 - 2}{2} \right)!}{\left(\frac{n_1 - 2}{2} \right)! \left(\frac{n_2 - 2}{2} \right)! (n_2 + n_1 z)^{(n_1 + n_2)/2}}$$

при $z > 0$.

Любая случайная величина с плотностью $g(z)$ называется *F-распределенной случайной величиной* с n_1 и n_2 степенями свободы, где n_1 — число степеней свободы числителя, а n_2 — знаменателя. Это записывается как $F \left[\begin{smallmatrix} n_1 \\ n_2 \end{smallmatrix} \right]$.

РАСПРЕДЕЛЕНИЕ ВЫБОРОЧНОЙ СРЕДНЕЙ

Выборочная средняя очень часто встречается как итог статистического изучения набора данных, и, конечно, она служит оценкой параметра μ нормального распределения¹. Ниже, изучая $\bar{x}$, мы рассмотрим два случая.

В первом случае $x_1, x_2, \dots, x_n$ — выборка объемом n из нормальной совокупности со средним μ и дисперсией σ^2 . Во втором случае будем предполагать только, что $\bar{x}$ является выборочной оценкой μ . Прежде чем перейти к дальнейшему рассмотрению, приведем без доказательства [19] следующую теорему, которая называется *теоремой сложения для нормально распределенных случайных величин*.

Теорема сложения для n независимых нормальных случайных величин

Если n независимых случайных величин распределены нормально с параметрами μ_i и σ_i^2 ($i=1, 2, \dots, n$), то любая линейная комбинация этих величин, например

$$l(x) = \tau_0 + \tau_1 x_1 + \dots + \tau_n x_n,$$

нормально распределена со средним

$$E[l(x)] = \tau_0 + \tau_1 \mu_1 + \dots + \tau_n \mu_n$$

и дисперсией

$$\text{var}[l(x)] = \tau_1^2 \sigma_1^2 + \dots + \tau_n^2 \sigma_n^2 = \sum_{i=1}^n \tau_i^2 \sigma_i^2.$$

¹ Авторы имеют в виду среднее арифметическое, являющееся хорошей оценкой математического ожидания случайной величины и в условиях распределений немного отличающееся от нормального распределения. — Прим. перев.

Возвращаясь к задаче о распределении выборочной средней из нормальной совокупности, положим

$$\tau_0 = 0, \quad \tau_1 = \tau_2 = \dots = \tau_n = \frac{1}{n}, \quad \mu_i = \mu \text{ и } \sigma_i^2 = \sigma^2;$$

тогда

$$l(x) = \bar{x} = \frac{x_1}{n} + \frac{x_2}{n} + \dots + \frac{x_n}{n}.$$

Согласно теореме сложения, средняя $\bar{x}$, будучи линейной комбинацией нормально распределенных случайных величин, также нормально распределена со средним

$$E[l(x)] = \sum_{i=1}^n \tau_i \mu_i = \sum_{i=1}^n \frac{1}{n} \mu = n \frac{1}{n} \mu = \mu$$

и дисперсией

$$\text{var}[l(x)] = \sum_{i=1}^n \tau_i^2 \sigma_i^2 = \sum_{i=1}^n \frac{1}{n^2} \sigma^2 = n \frac{1}{n^2} \sigma^2 = \frac{\sigma^2}{n}.$$

Таким образом, в вышеприведенных условиях случайная величина $\bar{x}$ распределена нормально со средним μ и дисперсией σ^2/n .

Упражнения

Используя теорему сложения для нормальных случайных величин, найти распределение случайной величины $(x - \mu)/\sigma$, если x — нормально распределенная величина с параметрами μ и σ^2 .

Указание: Пусть $\tau_0 = \frac{-\mu}{\sigma}$, $\tau_1 = \frac{1}{\sigma}$, а $\tau_2 = \dots = \tau_n = 0$. Тогда $l(x) = (x - \mu)/\sigma$, и, применяя упомянутую теорему, увидим, что величина $l(x)$ нормально распределена со средним

$$E[l(x)] = E\left(\frac{x - \mu}{\sigma}\right) = 0 \quad (4.5)$$

и дисперсией

$$\text{var}[l(x)] = \text{var}\left(\frac{x - \mu}{\sigma}\right) = \frac{1}{\sigma^2} \sigma^2 = 1. \quad (4.6)$$

Случайная величина $(x - \mu)/\sigma$, распределенная нормально с параметрами 0, 1, называется *центрированным и нормированным, или единичным, уклонением*. Любая нормальная величина x может быть преобразована в единичную:

$$\frac{x - \mu}{\sigma},$$

где μ и σ — среднее значение и стандартное отклонение величины x соответственно.

Значения единичной нормальной функции $\Phi[(x - \mu)/\sigma]$ для разных значений аргумента можно найти в таблицах любого учебника статистики.

1. Найти среднее и дисперсию случайной величины $\sum_{i=1}^m ix_i$, если x_i есть $N(i\mu_i, i^2\sigma^2)$.

$$\text{Ответ: } N \left[\sum_{i=1}^m i^2\mu_i, \sigma^2 \sum_{i=1}^m i^4 \right].$$

2. Найти среднее и дисперсию случайной величины $\frac{1}{m} \sum_{i=1}^m ix_i$, если распределение x_i есть $N(i\mu_i, i^2\sigma^2)$.

$$\text{Ответ: } N \left[\frac{1}{m} \sum_{i=1}^m i^2\mu_i, \frac{\sigma^2}{m} \sum_{i=1}^m i^4 \right].$$

3. Если величина x распределена как $N(0, 1)$, то вероятность события $x < 0$ можно определить следующим способом:

$$\begin{aligned} P\{x < 0\} &= P\left\{ \frac{x - \mu}{\sigma} < \frac{0 - \mu}{\sigma} \right\} = \\ &= P\left\{ \frac{x - \mu}{\sigma} < 0 \right\} = \Phi\left(\frac{x - \mu}{\sigma} = 0 \right); \end{aligned}$$

для

$$\begin{aligned} \frac{x - \mu}{\sigma} &= 0 \\ \Phi\left(\frac{x - \mu}{\sigma} = 0 \right) &= 0,5. \end{aligned}$$

Следовательно,

$$P\{x < 0\} = 0,5.$$

4. Найти $P\{x < 1\}$, если величина x распределена как $N(0, 1)$.

Ответ: 0,8413.

5. Найти $P\{x^2 < 1\}$, если величина x распределена как $N(0, 9)$.

Ответ: 0,1586.

6. Найти $P\{x < 2\sigma\}$, если величина x распределена как $N(16, 36)$.

Ответ: 0,2514.

Центральные предельные теоремы

Пусть $x_1, x_2, \dots, x_n$ — набор независимых случайных величин со средними μ_i , дисперсиями σ_i^2 и плотностями вероятностей g_i .

Случай 1

Все величины x_i имеют одну и ту же плотность g , среднее μ и дисперсию σ^2 . Если дисперсия существует и является конечной, то при возрастании n распределение суммы

$$x_1 + x_2 + \dots + x_n = \sum_{i=1}^n x_i$$

будет приближаться к нормальному со средним

$$E \left[\sum_{i=1}^n x_i \right] = n\mu$$

и дисперсией

$$\text{var} \left[\sum_{i=1}^n x_i \right] = n\sigma^2.$$

Поэтому выборочная средняя $\bar{x}$, полученная из n независимых одинаково (не обязательно нормально) распределенных случайных величин, со средним μ и дисперсией σ^2 будет распределена асимптотически нормально с параметрами

$$E \left[\frac{1}{n} \sum_{i=1}^n x_i \right] = \mu$$

и

$$\text{var} \left[\frac{1}{n} \sum_{i=1}^n x_i \right] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Случай 2

Все величины $x_1, \dots, x_n$ независимы, но распределены неодинаково.

При довольно общих условиях сумма $\sum_{i=1}^n x_i$ и среднее $\frac{1}{n} \sum_{i=1}^n x_i$ будет обладать распределениями, весьма близкими к нормальному, с параметрами $A_n = \sum_{i=1}^n \mu_i$, A_n/n , $B_n^2 = \sum_{i=1}^n \sigma_i^2$ и B_n^2/n^2 соответственно. Однако в этом случае, согласно Крамеру [16], для близости распределений к нормальному закону одного условия конечности дисперсии мало. В этой ситуации достаточно существование конечных третьих абсолютных моментов для всех случайных величин x_i , а также выполнение условия, заключающегося в том, что корень кубический из суммы третьих абсолютных центральных моментов, деленный на B , стремится к нулю

74 при $n \rightarrow \infty$.

Эти два сильных результата известны под названием *центральных предельных теорем для независимых случайных величин*. Более глубокие сведения по этому вопросу читатель найдет в гл. 17 («Нормальное распределение») указанной книги Крамера.

РАСПРЕДЕЛЕНИЕ ВЫБОРОЧНОЙ ДИСПЕРСИИ

Если $x_1, x_2, \dots, x_n$ — выборка объемом n из нормальной совокупности с известным средним значением μ и неизвестной дисперсией σ^2 , то удовлетворительной оценкой для σ^2 будет величина

$$\sum_{i=1}^n \frac{(x_i - \mu)^2}{n} = \hat{s}^2. \quad (4.7)$$

Распределение этой случайной величины может быть найдено следующим образом. В (4.5) и (4.6) было показано, что случайная величина $(x_i - \mu)/\sigma$ распределена как $N(0, 1)$, а величина

$$\chi_n^2 = \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma} \right)^2$$

распределена по закону χ^2 с n степенями свободы, так как на величины $(x_i - \mu)/\sigma$ не наложено никаких линейных связей.

Заменяя $\sum_{i=1}^n (x_i - \mu)^2$ на $n\hat{s}^2$, легко увидеть зависимость между $\hat{s}^2$ и χ_n^2 , а именно

$$\hat{s}^2 = \frac{\sigma^2}{n} \chi_n^2.$$

Точное распределение $\hat{s}^2$ можно получить, если в (4.3) заменить χ_n^2 на $n\hat{s}^2/\sigma^2$ и использовать равенство

$$d\hat{s}^2 = \frac{\sigma^2}{n} d\chi_n^2.$$

Плотность вероятностей распределения $\hat{s}^2$ равна

$$\frac{1}{\left(\frac{n-2}{2}\right)!} \frac{1}{2^{n/2}} \left(\frac{n \cdot \hat{s}^2}{\sigma^2}\right)^{\frac{n-2}{2}} \frac{n}{\sigma^2} \exp\left[-\frac{n}{2} \left(\frac{\hat{s}^2}{\sigma^2}\right)\right] \quad \text{при } \hat{s}^2 > 0.$$

Так как среднее и дисперсия χ^2 -распределенной случайной величины равны соответственно n и $2n$ (при n степенях

свободы), то среднее и дисперсия случайной величины $\hat{s}^2$ будут равны

$$E\hat{s}^2 = E\left[\frac{\sigma^2}{n} \chi_n^2\right] = \sigma^2,$$

$$\text{var } \hat{s}^2 = \text{var } \frac{\sigma^2}{n} \chi_n^2 = \frac{\sigma^4}{n^2} \cdot 2n = \frac{2\sigma^4}{n}.$$

Значение μ в действительных ситуациях обычно бывает неизвестным, и оценивается оно при помощи величины $\bar{x}$. В этом случае по аналогии функция от случайной величины $\sum_{i=1}^n (x_i - \bar{x})^2$ могла бы быть желательной оценкой для σ^2 . В основном используют как оценку величину

$$\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = s^2.$$

В отличие от $\hat{s}^2$ распределение этой величины нельзя найти столь же просто; в этом случае следует использовать теорему Кохрена.

Случайную величину $\sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma}\right)^2$, распределенную как χ^2 , можно представить в следующем виде:

$$\begin{aligned} \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma}\right)^2 &= \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sigma} + \frac{\bar{x} - \mu}{\sigma}\right)^2 = \\ &= \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sigma}\right)^2 + n \left(\frac{\bar{x} - \mu}{\sigma}\right)^2. \end{aligned}$$

Первая сумма квадратов имеет одно ограничение, возникающее как следствие линейной зависимости между x_i и $\bar{x}$. Поэтому данная сумма квадратов имеет $n-1$ степеней свободы, т.е. число переменных минус одно наложенное условие (линейная зависимость между x_i и $\bar{x}$).

Вторая сумма квадратов выражена через один квадрат, в котором нет линейной зависимости между $\bar{x}$ и константами μ , σ и n . Число степеней свободы для $n \left(\frac{\bar{x} - \mu}{\sigma}\right)^2$ равно единице.

Таким образом, величина $\sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma}\right)^2$, имеющая n степеней свободы, разделена на две суммы квадратов, обладающих $n-1$ и 1 степенями свободы соответственно. Согласно теореме Кох-

рена, получим, во-первых, что

$$\sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sigma} \right)^2 \text{ не зависит от } n \left(\frac{\bar{x} - \mu}{\sigma} \right)^2,$$

т. е. выборочная дисперсия нормальной совокупности не зависит от выборочного среднего, и, во-вторых, что $\sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sigma} \right)^2$ распределена как χ^2 с $n - 1$ степенями свободы.

Теперь можно случайную величину $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ записать как

$$s^2 = \frac{\sigma^2 \chi^2}{n-1}$$

и определить плотность вероятностей $g(s^2)$ ее распределения:

$$g(s^2) = \begin{cases} \frac{1}{\left(\frac{n-3}{2}\right)!} \frac{1}{2^{\frac{n-1}{2}}} \left(\frac{n-1}{\sigma^2} s^2\right)^{\frac{n-3}{2}} \frac{n-1}{\sigma^2} \exp\left[-\frac{(n-1)s^2}{2\sigma^2}\right] & \text{при } s^2 > 0, \\ 0 & \text{при } s^2 \leq 0; \end{cases}$$

не трудно видеть, что

$$Es^2 = \sigma^2$$

и

$$\text{var } s^2 = \frac{2\sigma^4}{n-1},$$

Заметим, что $Es \neq \sigma$.

ДРУГИЕ ОЦЕНКИ ДЛЯ μ И σ

1. Кроме среднего арифметического, в качестве оценок центра распределения иногда используют выборочную медиану и полусумму крайних членов вариационного ряда. Эти оценки бывают полезны в условиях выборок объемом 10 (или менее) наблюдений.

а) *Выборочная медиана (M)*. Определяется как срединное значение вариационного ряда. Если число наблюдений нечетно, в качестве медианы принимается срединное значение; если же число наблюдений четное, за медиану принимается середина отрезка между двумя срединными значениями.

б) *Полусумма крайних* (т. е. наибольшего и наименьшего) членов вариационного ряда (M_r).

Например, в выборке объемом 6

$$M_r = \frac{x_1 + x_6}{2},$$

где x_1 — наименьшее в вариационном ряду значение, x_6 — наибольшее.

2. Кроме выборочной статистики s , в качестве оценок для σ (стандартного отклонения) используются:

а) *Размах выборки w (для выборок объемом 10 и менее).*

Выборочный размах является не только неэффективной оценкой для σ , но и смещенной. Это смещение может быть устранено путем введения поправочного коэффициента. Для выборок, объем которых не превышает 10, эффективность размаха оказывается весьма высокой. Таблица для выборочного размаха как оценки параметра σ включает поправочный коэффициент, а также эффективность для объемов выборок от 2 до 10; эти таблицы приведены в книге Диксона и Мессе [79].

б) Приведенные в книге Диксона и Мессе другие оценки для σ , пригодные для выборок, объем которых не превышает 10.

ПРОВЕРКА ГИПОТЕЗ

Статистические выводы являются одним из наиболее важных вопросов, рассматриваемых в этой книге. Вычисление средних и дисперсий вообще бесполезно, если оно делается только с целью сокращения объема записи при недостатке места. Разве лишь в этом заключается все то, что можно сказать о выборке? Что можно сказать о генеральной совокупности по взятым из нее выборкам? Какие выводы могут быть сделаны по выборке о генеральной совокупности? Наши выводы также должны быть сделаны с определенной степенью уверенности или надежности.

Практика обычных наблюдений над серией экспериментальных результатов и заключения, основанные исключительно на интуитивных предположениях, показали, что, если мы хотим получить выводы из эмпирических данных, необходимо применять объективные критерии.

Таким образом, перед нами задача — получить способ действия, управляемый набором правил, применение которых всеми заинтересованными исследователями сведет к минимуму ошибки интерпретации, что в свою очередь окажет положительное влияние на науку и технику. Ниже мы рассмотрим способы проверки гипотез, которые ни в коей мере не являются универсальным средством на все случаи жизни: они позволят свести к минимуму неясность и неопределенность в выводах при обработке данных.

Неформализованная проверка гипотез; эвристический подход к задаче проверки гипотез

Допустим, что арестован человек, подозреваемый в преступлении. Суд на основе представленных улик должен решить, действительно ли обвиняемый совершил преступление, а затем ответить на вопрос, виновен ли обвиняемый или невиновен. Процесс установления невиновности или виновности обвиняемого нами рассматриваться не будет. Мы знаем, что суд оглашает

свое решение лишь после того, как им рассмотрены и взвешены все доводы «за» и «против». Формулировка вопроса, на который отвечают присяжные, должна быть более точной, чем «виновен ли этот человек (или не виновен)», «совершил ли этот человек то преступление, в котором его обвиняют». Гипотеза может быть сформулирована, как утверждение: «м-р X не совершал преступления, в котором его обвиняют».

Если имеющиеся данные противоречивы и свидетельствуют как за обвинение, так и против него, суд вынесет решение, что этот человек виновен или невиновен, в соответствии с которым обвиняемый будет осужден или оправдан.

Каково бы ни было решение суда, оно может быть ошибочным. Возможные ошибки будут заключаться в том, что суд признает обвиняемого виновным, тогда как на самом деле он невиновен, или, наоборот, оправдает его, тогда как в действительности он виновен. Сформулируем нулевую гипотезу, обозначив ее через H_1 , — «обвиняемый невиновен». В связи с этой гипотезой рассмотрим возможные решения суда в зависимости от того, является ли обвиняемый действительно виновным или нет (табл. 5.1).

Таблица 5.1

Возможные решения суда и действительные положения подсудимого; величины α и β — вероятности ошибок первого и второго рода

Решение суда	Действительное положение подсудимого	
	невиновен	виновен
Виновен	α	$1 - \beta$
Невиновен	$1 - \alpha$	β

«Обвиняемый или виновен, или невиновен». Если решение суда «невиновен», то принимается H_1 , если «виновен», H_1 отвергается. Это положение сведено в табл. 5.1, где по горизонтали даны возможные решения суда, а по вертикали — действительное положение обвиняемого. Допустим, что обвиняемый на самом деле невиновен (первый столбец). Суд, конечно, этого не знает. Нам как объективным наблюдателям хотелось бы, чтобы было принято решение «невиновен». Однако суд может ошибиться и решить «виновен», т. е. отвергнуть проверяемую нулевую гипотезу, тогда как на самом деле она верна. Эта ошибка называется *ошибкой первого рода*. Естественно желание сделать так, чтобы эта ошибка появлялась как можно реже, иными словами, свести к минимуму вероятность ее появления (α). С другой стороны, желательно, чтобы вероятность исхода, заключаю-

щегося в оправдании действительно невиновного человека, была большой. Эта вероятность равна $1 - \alpha$.

Теперь допустим, что м-р X виновен. Суд, как и в предыдущем случае, может принять одно из двух возможных решений — «невиновен» или «виновен». Но теперь нам как объективным наблюдателям хотелось бы, чтобы суд решил «виновен». В данной ситуации суд также может ошибиться. Он может решить «невиновен». Это значит, что принята гипотеза H_1 , которая на самом деле неверна. Такая ошибка называется *ошибкой второго рода*. Желательно также свести к минимуму вероятность ее появления (β). С другой стороны, если H_1 — ложная гипотеза, было бы желательно вынести решение «виновен», т. е. по возможности увеличить вероятность $1 - \beta$, называемую *мощностью критерия*.

БАЛАНС ОШИБОК; ПОДХОД К ПРАВИЛУ ПРИНЯТИЯ РЕШЕНИЯ¹

Обе приведенные выше ошибки влекут за собой серьезные последствия как для общества, так и для обвиняемого. Не касаясь здесь вопроса, какая из ошибок принесет больший вред, приходим к выводу, что следует уметь контролировать эти ошибки, т. е. свести к минимуму вероятности их появления. В некоторых случаях одна из ошибок влечет более серьезные последствия. Ниже мы вернемся к вопросу о проверке гипотез и о связанных с этой проверкой ошибках и рассмотрим его в более общем виде. Приведенный простой пример дает всего лишь наглядное представление о способе подхода к решению этого вопроса. Поставленная задача сформулирована в виде некоторой гипотезы, которая затем должна быть проверена. Когда проводится эксперимент или приводятся какие-либо другие доводы, они помогают выработать отношение к выдвинутой гипотезе, т. е. она принимается или отвергается. Ниже мы рассмотрим с вероятностных позиций типы ошибок и способы, которые помогают сводить их по возможности к минимуму.

Проверка гипотез; простая альтернатива — общий подход

Рассмотрим совокупность объектов, каждый из которых характеризуется k признаками. Пусть

$$X = X_i = x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{ik},$$

где

$$1 \leq i \leq \infty,$$

$$k \geq 1,$$

$$-\infty < x_{ij} < \infty.$$

¹ Мы будем говорить короче — «решающее правило». — Прим. ред.

Это значит, что любому объекту, например объему с номером i , соответствует k -мерная величина X_i , образованная значениями k признаков. Первое значение x_{i1} , второе x_{i2} и т. д. Примером подобной совокупности могут служить все образцы пород, взятые в Оклахоме за 1953 г. из буровых скважин. В данном случае x_{i1} может означать цвет образца с номером i , x_{i2} — содержание в нем CaCO_3 и т. д.

Если $x_{ij}=0$ при всех $j>1$, то совокупность характеризуется только одним признаком, т. е. является одномерной:

$$X_i = x_{i1}.$$

Обозначим плотность вероятностей случайной величины X через $f(X)$. Предположим, *a priori* известно, что может быть одно из двух равенств: $f(X)=g(X)$ или $f(X)=h(X)$. Выберем в качестве нулевой гипотезы равенство

$$H_1: f(X)=g(X)$$

при альтернативе

$$H_2: f(X)=h(X).$$

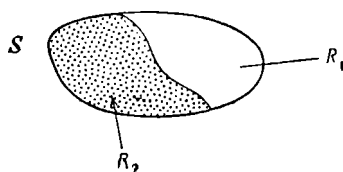
Критерием, с помощью которого принимается одна из гипотез (H_1 или H_2), служит единственная выборочная точка

$$X_i = x_{i1}, x_{i2}, \dots, x_{ik},$$

являющаяся результатом наблюдения; на ее основе и будет принята или отвергнута гипотеза H_1 . Достоинства и недостатки выборки такого объема здесь не рассматриваются. Таким образом, процедура состоит из формулировки гипотезы о данной совокупности и принятия решения, основанного на единичном наблюдении. Решение заключается в том, что нулевая гипотеза или принимается, или отвергается. При такой выборочной процедуре существует множество возможных исходов, и поэтому возникает вопрос: как принимать решение? Хорошей процедурой могла бы быть следующая: мы рассматриваем все возможные результаты выборки в предположении, что нулевая гипотеза H_1 справедлива, и отбираем те из них, которые делают H_1 правдоподобной. Действительно, если случайная величина X распределена как $g(X)$, то существуют такие значения X , которые ближе подходят к $g(X)$ чем к $h(X)$. Рассмотрев все возможные исходы, можно условиться, что гипотеза H_1 будет принята при условии, если X_i попадет в определенную область значений, и отвергнута, если X_i окажется в другой области.

Пусть R_1 — область, содержащая те значения X , которые обязывают принять гипотезу H_1 , а R_2 — область тех значений X , наблюдение которых требует ее отвергнуть; R_1 называется *областью принятия нулевой гипотезы*, а R_2 — *критической областью*. Эти области показаны на фиг. 5.1. Теперь было бы желательно принять гипотезу H_1 , если она верна (1), и отвергнуть

ее в случае, если она ошибочна 2). Требованию (1) соответствует попадание выборочной точки X_i в R_1 при условии, что верна гипотеза H_1 , а требованию (2) — попадание X_i в R_2 при условии, что она неверна. Наоборот, было бы нежелательно отвергнуть гипотезу H_1 , тогда как на самом деле она верна (3), или принять ее в случае, если она ложна (4).



Фиг. 5.1. Критическая область и область принятия нулевой гипотезы. S — выборочное пространство всех возможных исходов выбора из совокупности индивидуумов; R_1 — принимается гипотеза H_1 ; R_2 — гипотеза H_1 отвергается и принимается гипотеза H_2 .

Перечисленные четыре события с соответствующими им вероятностями приведены в табл. 5.2, в которой знак $\in$ означает «содержится в ...» или «принадлежит».

Таблица 5.2

Событие	Условие	Вероятность
(1) $X \in R_1$	Верна H_1	$P_{H_1}(X \in R_1) = \int_{R_1} g(X) dX = 1 - \alpha$
(2) $X \in R_2$	» H_2	$P_{H_2}(X \in R_2) = \int_{R_2} h(X) dX = 1 - \beta$
(3) $X \in R_2$	» H_1	$P_{H_1}(X \in R_2) = \int_{R_2} g(X) dX = \alpha$
(4) $X \in R_1$	» H_2	$P_{H_2}(X \in R_1) = \int_{R_1} h(X) dX = \beta$

Из предыдущего текста нетрудно понять, почему вероятности $1 - \alpha$ и $1 - \beta$ должны быть возможно большими, тогда как α и β , — наоборот, возможно малыми. Рассмотрим событие (1).

При условии, что верна гипотеза H_1 , случайная величина X будет распределена как $g(X)$, и тогда

$$\int_{-\infty}^{\infty} g(X) dX = 1.$$

Чтобы найти вероятность события

$$a < X \leq b,$$

следует просто вычислить

$$\int_a^b g(X) dX.$$

Таким образом, для *любой* области можно найти вероятность того, что наблюдаемое значение X будет ей принадлежать.

Например, вероятность попадания X в область R_1 равна

$$\int_{R_1} g(X) dX.$$

Если R_1 — область принятия гипотезы H_1 , то вероятность ее принятия при условии, что она верна, будет определена из равенства

$$P\{X \in R_1\} = \int_{R_1} g(X) dX = 1 - \alpha.$$

Аналогично если случайная величина X в действительности распределена как $h(X)$, то вероятность того, что значение X будет принадлежать R_1 , дается выражением

$$\int_{R_1} h(X) dX = \beta.$$

Так составлялась табл. 5.2.

РИСК, СВЯЗАННЫЙ С ОШИБКАМИ ПЕРВОГО И ВТОРОГО РОДА

Ошибка первого рода

При условии что верна гипотеза $H_1: f(X) = g(X)$, будут существовать такие значения X , которые появляются с малой вероятностью. Тем не менее этим же маловероятным значениям при правильной гипотезе H_1 могут соответствовать большие вероятности, если верна альтернатива $H_2: f(X) = h(X)$. Например, этими значениями могут быть $|X| > k$, $X > k$ или $X < k$. Таким образом, рассматриваются те значения X , которые встречаются часто, т. е. с частотой $1 - \alpha$, где α — малое значение, а область R_1 такова, что

$$\int_{R_1} g(X) dX = 1 - \alpha.$$

Эта величина представляет вероятность появления события **84** $X \in R_1$ при условии, что гипотеза H_1 верна. Фактически эта

модель приводит к тому, что значения X , имеющие малую вероятность (меньшую или равную α), в условиях распределения $g(X)$ из рассмотрения исключаются. Вероятность такого события будет равна

$$P_{H_1}\{X \in R_2\} = \int_{R_2} g(X) dX = \alpha. \quad (5.1)$$

Если гипотеза H_1 верна, а наблюдается событие $X \in R_2$, то, согласно выбранному решающему правилу, гипотеза H_1 должна быть отвергнута. В данном случае будет ошибочно отвергнута *правильная нулевая гипотеза*. Это и есть ошибка первого рода. Частое появление такой ошибки нежелательно, и, следовательно, нужно заранее определить, сколь малым должно быть значение величины α . Если α задана заранее, то R_1 будет функцией от α . Если же заранее задана область R_1 , то величине α нельзя приписать заданное значение, так как она будет функцией от R_1 .

Ошибка второго рода и мощность критерия

Выше было рассмотрено положение выборочной точки X при условии, что верна гипотеза H_1 : $f(X) = g(X)$. Рассмотрим противоположный случай: в действительности случайная величина X распределена как $h(X)$, т. е. проверяемая гипотеза H_1 неверна. Событие $X \in R_1$ имеет в этих условиях вероятность, которая выражается через

$$\int_{R_1} h(X) dX. \quad (5.2)$$

Согласно решающему правилу, если наблюдается событие $X \in R_1$, должна быть принята гипотеза H_1 несмотря на то, что она неверна. Таким образом, *будет ошибочно принята проверяемая нулевая гипотеза, которая на самом деле неверна*. Такая ошибка, безусловно, нежелательна, и последствия ее зависят от конкретных обстоятельств. Поэтому, когда гипотеза H_1 неверна, необходимо сводить к минимуму вероятность появления такого события. Обозначим величину (5.2) через β . Естественно, что нужно сделать β , называемую *вероятностью появления ошибки второго рода*, по возможности малой.

Если значение β мало, то величина $1 - \beta$, представляющая вероятность появления события $X \in R_2$ при условии, что гипотеза H_1 неверна, становится большой, что желательно:

$$1 - \beta = P_{H_2}\{X \in R_2\} = \int_{R_2} h(X) dX.$$

Эта вероятность называется *мощностью критерия относительно конкурирующей гипотезы* H_2 . При любой процедуре **85**

проверки гипотезы естественно делать величину α по возможности малой, тогда как мощность должна быть максимальной.

УРОВЕНЬ ЗНАЧИМОСТИ, МОЩНОСТЬ КРИТЕРИЯ И КРИТИЧЕСКИЕ ОБЛАСТИ

Выше случайную величину X предполагали k -мерной, т. е. образованной k характеристиками. В связи с тем что графическое представление многомерного распределения осуществить трудно, упростим задачу, сведя ее к одномерному случаю. Положим

$$X_i = x_{i1} = x_i,$$

а

$$f(X) = f(x).$$

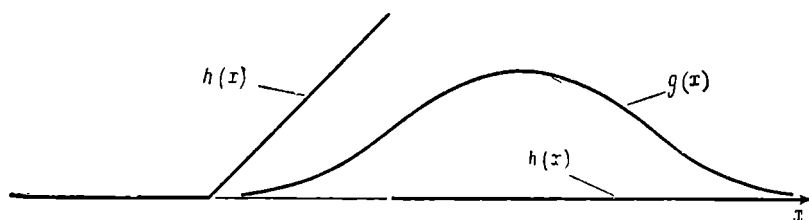
Сформулируем нулевую гипотезу в виде

$$H_1: f(X) = g(x)$$

при альтернативе

$$H_2: f(X) = h(x).$$

Если верна гипотеза H_1 , то случайная величина x будет распределена как $g(x)$, а если справедлива H_2 — как $h(x)$. Эти распределения приведены на фиг. 5.2.



Фиг. 5.2. Распределения случайной величины x при условиях: $g(x)$ — верна гипотеза H_1 ; $h(x)$ — верна гипотеза H_2 .

Сформулировав нулевую гипотезу и соответствующую ей альтернативу, необходимо определить область принятия гипотезы H_1 и критические области, где она отвергается, а также уровень значимости и мощность критерия. Процедура заключается в установлении или уровня значимости α , или мощности критерия $1 - \beta$. Если одна из этих величин задана, то возможные значения другой будут ограничены, так как для любой заданной пары R_1 и R_2 определена другая ошибка.

Как и раньше, над случайной величиной производится одно наблюдение, на основе которого гипотеза H_1 отвергается или

принимается. Прежде всего необходимо выбрать критическую область, которая является функцией заданного уровня значимости. Так как

$$\int_{R_2} g(x) dx = \alpha,$$

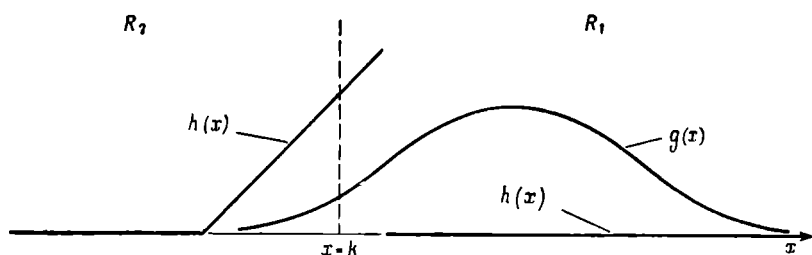
то резонно выбрать α малым. Будет вполне естественным определить положение области R_2 так, чтобы она находилась слева от $g(x)$, вернее, левее некоторой точки $x=k$, такой, что

$$\int_{R_2} g(x) dx = \alpha = \int_{-\infty}^k g(x) dx$$

и

$$G(k) = \alpha$$

(фиг. 5.3). Таким образом может быть найдено значение k как функции от α , т. е. $k = k(\alpha)$.



Фиг. 5.3. Критическая область R_2 и область R_1 , в которой гипотеза H_1 принимается.

Определить β можно следующим образом:

$$\int h(x) dx = \beta;$$

интегрируя, получим

$$\beta = 1 - H[k(a)].$$

Области, ограниченные кривыми $h(x)$ и $g(x)$, представляют собой α , β , $1 - \alpha$, $1 - \beta$ (фиг. 5.4).

Установление границы между R_1 и R_2

Критическое значение $x=k$ является произвольным по отношению к множеству значений x , которые могут быть использованы как границы R_2 при условии, что

$$\int_{R_1} g(x) dx = \alpha.$$

Например, рассмотрим некоторые другие значения x : k' , k'' , k''' и k^{IV} . Условимся принимать гипотезу H_1 , если

$$x < k',$$

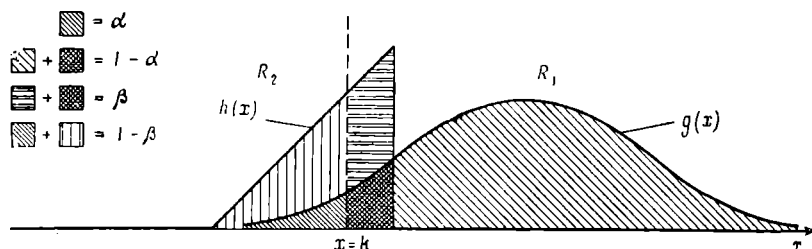
или

$$x > k^{IV},$$

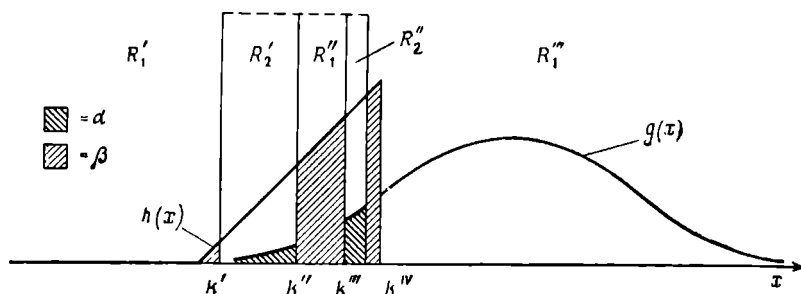
или

$$k'' < x < k'''.$$

Эти точки определяют другую критическую область. Пусть R'_2 — область, содержащая все значения x , удовлетворяющие



Фиг. 5.4. Площади, представляющие вероятности α и β ошибок первого и второго рода и их дополнения до единицы.



Фиг. 5.5. Вероятности α и β ошибок первого и второго рода как функции критических областей. В данном случае R'_1 , R''_1 и R'''_1 — области, в которых нулевая гипотеза принимается, а R'_2 и R''_2 — критические области.

неравенствам $k' < x < k''$. Пусть также R''_2 — область, содержащая все значения x , удовлетворяющие неравенствам $k''' < x < k^{IV}$. Тогда

$$\left[\int_{R'_2} g(x) dx = \int_{k'}^{k''} g(x) dx \right] + \left[\int_{R''_2} g(x) dx = \int_{k'''}^{k^{IV}} g(x) dx \right] = \alpha,$$

а если

$$R_2' + R_2'' = R_2,$$

то

$$\int_{R_2} g(x) dx = \alpha.$$

Следовательно, новые критические области сохраняют удовлетворительные свойства критерия по отношению к ошибке первого рода, так как вероятность ее появления будет равна α . Какая же критическая область должна быть выбрана? Ответить на этот вопрос можно только после рассмотрения вероятности (β) появления ошибки второго рода.

На фиг. 5.5 показаны площади, соответствующие значениям α и β . При сравнении фиг. 5.4 и 5.5 видно, что критическая область на фиг. 5.4 выбрана лучше, так как в этом случае вероятность появления ошибки второго рода окажется меньшей, чем на фиг. 5.5. Таким образом, наилучшей критической областью будет такая, которая удовлетворит условию сохранения α при максимальной мощности.

ЛЕММА НЕЙМАНА — ПИРСОНА И ВЫДЕЛЕНИЕ КРИТИЧЕСКИХ ОБЛАСТЕЙ

Выборка объемом $n = 1$

До сих пор мы оставались в рамках эвристического подхода к задаче проверки гипотез. Но существует общая процедура получения критериев и критических областей. Эту процедуру, показанную на приведенных выше примерах одного наблюдения, легко распространить на n наблюдений, которые можно представить в виде одного многомерного результата. Вернемся к задаче, поставленной выше, а именно

$$H_1: f(X) = g(x),$$

$$H_2: f(X) = h(x).$$

На фиг. 5.6 показаны ветви функций: левая $g(x)$ и правая $h(x)$. Выберем критическую область, включив в нее все точки, расположенные левее значения k . Тогда

$$\int_{R_1} g(x) dx = \int_{-\infty}^k g(x) dx. \quad (5.3)$$

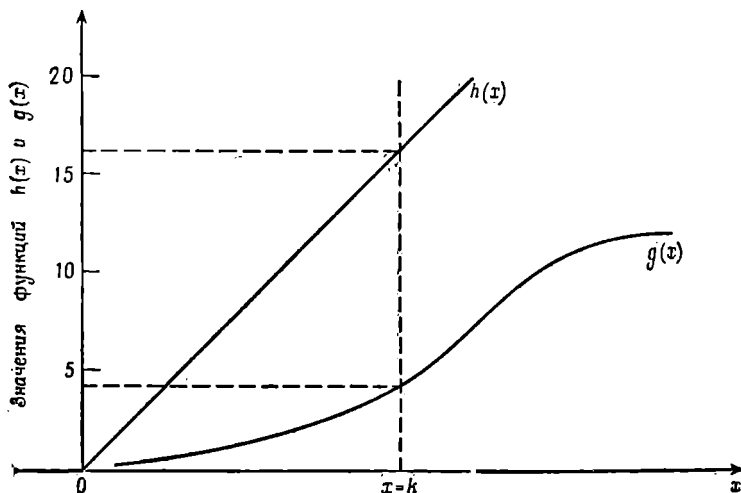
Отношение

$$\frac{h(k)}{g(k)}$$

есть некоторое число, в данном случае большее единицы:

$$\frac{h(k)}{g(k)} \approx \frac{16}{4,5} \approx 3,6.$$

Ясно, что для всех $x < k$ это отношение больше 3,6. Именно это нам и нужно, так как если указанное отношение превышает 3,6, то большая часть точек x , соответствующих $h(x)$,



Ф и г. 5.6. Функции $g(x)$ и $h(x)$ в области критического значения $x = k$.

будет находиться в R_2 . Следовательно, гипотеза H_1 будет отвергаться с большей вероятностью в том случае, когда она неверна. Вообще, чтобы построить критическую область при заданном уровне значимости α , следует выбирать те значения x , для которых

$$\frac{h(x)}{g(x)} > \text{const} = c.$$

Выборки объемом n

Полученный результат легко распространить на случай n выборочных значений одномерной случайной величины

$$X = x_i.$$

Так как результаты n наблюдений можно рассматривать как значения n одинаково распределенных случайных величин, то нулевую гипотезу H_1 и альтернативу H_2 можно записать в сле-

дующем виде:

$$H_1: f(X) = g(x_1)g(x_2) \dots g(x_n) = g(x), \quad (5.4)$$

$$H_2: f(X) = h(x_1)h(x_2) \dots h(x_n) = h(x). \quad (5.5)$$

Если выражение (5.4) записать как $\prod_{i=1}^n g(x_i)$, а (5.5) — как $\prod_{i=1}^n h(x_i)$, то неравенство, определяющее критическую область, примет вид

$$\frac{\prod_{i=1}^n h(x_i)}{\prod_{i=1}^n g(x_i)} \geq c.$$

Лемма Неймана — Пирсона

Ниже приведена без доказательства лемма Неймана — Пирсона: если критическая область R_2 такова, что

$$1) \int_{R_2} \prod_{i=1}^n g(x_i) dx_i = \alpha = \int_{R_2} g(x) dx,$$

2) а c — такая константа, что $h(x) \geq cg(x)$ для всех $x \in R_2$ и $h(x) \leq cg(x)$ для всех $x \in R_1$, то не существует более мощной критической области уровня α , чем R_2 . Эту лемму легко понять с помощью фиг. 5.6, согласно которой

$$g(X) = g(x),$$

$$h(X) = h(x).$$

Мы предполагали, что

$$\int_{R_2} f(x) dx = \alpha$$

и что существует число c , такое, что для всех значений x , принадлежащих критической области,

$$h(x) \geq cg(x),$$

а для всех x , содержащихся в R_1 ,

$$h(x) \leq cg(x).$$

Путем подстановки в $g(x)$ и $h(x)$ различных значений x можно легко проверить, что это требование выполняется.

**ПРОВЕРКА ГИПОТЕЗ О СРЕДНЕМ ЗНАЧЕНИИ НОРМАЛЬНОГО
РАСПРЕДЕЛЕНИЯ**

Очень важную группу задач представляет собой проверка различных гипотез по выборкам из нормальных совокупностей. Из них чаще всего приходится сталкиваться со следующими гипотезами.

Пусть выборочная точка X — результаты n -кратного выбора из нормальной совокупности с неизвестным средним μ и известной дисперсией σ^2 . Сформулируем нулевую гипотезу в виде

$$H_1: \mu = \mu_1$$

при альтернативе

$$H_2: \mu = \mu_2.$$

Таким образом, проверяемая нулевая гипотеза заключается в том, что неизвестное среднее μ нормальной совокупности имеет некоторое частное значение μ_1 . Если гипотеза H_1 отвергается, то принимается конкурирующая гипотеза H_2 , согласно которой μ равно другому значению μ_2 .

Положим вероятность ошибки первого рода равной α , а вероятность ошибки второго рода — β . Соответственно мощность критерия будет равна $1 - \beta$. Пусть также R_1 — область принятия нулевой гипотезы, а R_2 — критическая область.

Используя лемму Неймана — Пирсона, найдем такую константу c , что

$$\prod_{i=1}^n h(x) \geq c \prod_{i=1}^n g(x) \quad \text{для всех } x \in R_2$$

и

$$\prod_{i=1}^n h(x) \leq c \prod_{i=1}^n g(x) \quad \text{для всех } x \in R_1.$$

В поставленной частной задаче

$$h(x) = N[\mu_2, \sigma^2],$$

$$g(x) = N[\mu_1, \sigma^2],$$

и, следовательно,

$$\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2}\sum_{i=1}^n\left(\frac{x_i-\mu_2}{\sigma}\right)^2\right] \geqslant c \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2}\sum_{i=1}^n\left(\frac{x_i-\mu_1}{\sigma}\right)^2\right].$$

После сокращения получим

$$\exp\left[-\frac{1}{2}\sum_{i=1}^n\left(\frac{x_i-\mu_2}{\sigma}\right)^2\right] \geqslant c \exp\left[-\frac{1}{2}\sum_{i=1}^n\left(\frac{x_i-\mu_1}{\sigma}\right)^2\right],$$

откуда

$$\frac{\exp\left[-\frac{1}{2}\sum_{i=1}^n\left(\frac{x_i-\mu_2}{\sigma}\right)^2\right]}{\exp\left[-\frac{1}{2}\sum_{i=1}^n\left(\frac{x_i-\mu_1}{\sigma}\right)^2\right]} \geqslant c,$$

$$\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^n(x_i-\mu_2)^2-\sum_{i=1}^n(x_i-\mu_1)^2\right]\right\} \geqslant c,$$

$$-\frac{1}{2\sigma^2}\left[\sum_{i=1}^n(x_i-\mu_2)^2-\sum_{i=1}^n(x_i-\mu_1)^2\right] \geqslant \ln c.$$

Если принять

$$-2\sigma^2 \ln c = c',$$

то

$$\sum_{i=1}^n(x_i-\mu_2)^2-\sum_{i=1}^n(x_i-\mu_1)^2 \geqslant c',$$

$$\sum_{i=1}^n x_i^2 - 2\mu_2 \sum_{i=1}^n x_i + \mu_2^2 - \sum_{i=1}^n x_i^2 + 2\mu_1 \sum_{i=1}^n x_i - \mu_1^2 > c',$$

$$\sum_{i=1}^n x_i(-2\mu_2 + 2\mu_1) > c' + \mu_1^2 - \mu_2^2.$$

Положив

$$c'' = c' + \mu_1^2 - \mu_2^2,$$

получим

$$\sum_{i=1}^n x_i > \frac{c''}{2\mu_1 - 2\mu_2}.$$

Если обозначить

$$nc''' = \frac{c''}{2\mu_1 - 2\mu_2},$$

получим

$$\sum_{i=1}^n x_i > nc'''$$

и

$$\sum_{i=1}^n \frac{x_i}{n} = \bar{x} > c'''.$$

Таким образом, согласно решающему правилу, гипотеза H_1 будет отвергнута, если $\bar{x}$ превысит некоторую константу c''' , и принята, если $\bar{x} < c'''$.

Теперь нужно найти такую константу c''' , чтобы

$$P\{\bar{x} \geq c'''\} = \alpha.$$

Выше было показано, что среднее арифметическое n одинаково распределенных нормальных случайных величин со средним μ и дисперсией σ^2 распределено также нормально с параметрами μ и σ^2/n . Число n представляет собой объем выборки.

В результате если нулевая гипотеза верна, то $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ распределено как $N(\mu_1, \sigma^2/n)$, а случайная величина $(\bar{x} - \mu_1)/\sqrt{\sigma^2/n}$ также нормально распределена со средним 0 и дисперсией, равной 1. Поэтому

$$P\{\bar{x} \geq c'''\} = \alpha = P_{H_1}\left\{\frac{\bar{x} - \mu_1}{\sqrt{\sigma^2/n}} \geq \frac{c''' - \mu_1}{\sqrt{\sigma^2/n}}\right\}.$$

Из таблицы стандартной нормальной функции легко найти значение c''' , соответствующее α (см. упражнения „с“ и „f“ гл. 4).

Мощность критерия может быть определена как

$$1 - \beta = P_{H_2}\{\bar{x} \geq c'''\} = P_{H_2}\left\{\frac{\bar{x} - \mu_2}{\sqrt{\sigma^2/n}} \geq \frac{c''' - \mu_2}{\sqrt{\sigma^2/n}}\right\}.$$

Пример 1

Общее положение. Пусть x_i — выборочные значения случайной величины, распределенной нормально с неизвестным средним и известной дисперсией:

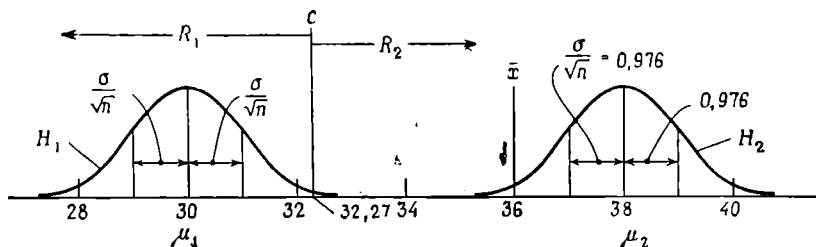
$$H_1: \mu = \mu_1,$$

$$H_2: \mu = \mu_2.$$

Частный случай. Предположим, что среднее значение абсолютной величины минимального различия между отсчетами азимута системы замеров распределено нормально со средним μ и дисперсией σ^2 . Пусть $\sigma^2 = 100$, объем выборки равен 105 и требуется проверить гипотезу $H_1: \mu = \mu_1 = 30$ при $\bar{x} = 36$ и конку-

рирующей гипотезе $H_2: \mu = \mu_2 = 38$. Величину α положим равной 0,01.

Теперь нужно выбрать решающее правило, т. е. установить значение, которое будет определять, отвергается или принимается гипотеза H_1 . На фиг. 6.1 иллюстрируется зависимость между нулевой и конкурирующей гипотезами и выборочным средним $\bar{x}$. На этой фигуре показаны распределения $\bar{x}$ при условии, что верна гипотеза H_1 , и при условии, что верна гипотеза H_2 .



Фиг. 6.1. Распределения случайной величины $\bar{x}$ при условии, что верна гипотеза H_1 , и при условии, что верна гипотеза H_2 .

Выборочное среднее $\bar{x}$ также отмечено на графике.

На этой диаграмме ясно видно, почему критическая область расположена справа от μ_1 . Найдём эту область.

Известно, что

$$P_{H_1} \left\{ \frac{\bar{x} - 30}{\sqrt{100/105}} \geq \frac{c - 30}{\sqrt{100/105}} \right\} = \alpha = 0,01.$$

Если Φ — нормальная стандартная функция распределения (гл. 4), то

$$1 - \Phi \left[\frac{c - 30}{\sqrt{100/105}} \right] = 0,01$$

и

$$\Phi \left[\frac{c - 30}{0,976} \right] = 0,99.$$

По таблице нормальной функции распределения находим, что значение $(c - 30)/0,976$, которому соответствует вероятность 0,99, равно 2,33, т. е.

$$\frac{c - 30}{0,976} = 2,33.$$

Тогда

$$c = 2,33 \cdot 0,976 + 30 = 32,27.$$

Судя по положению c на фиг. 6.1, выборочная средняя $\bar{x}$ оказывается в критической области, что обязывает отвергнуть нулевую гипотезу H_1 . Мощность этого критерия при данной

альтернативе будет

$$\begin{aligned} 1 - \beta &= P_{H_2} \{ \bar{x} \in R_2 \} = P_{H_2} \left\{ \frac{\bar{x} - 38}{0,976} > \frac{32,37 - 38}{0,976} \right\} = \\ &= P_{H_2} \left\{ \frac{\bar{x} - 38}{0,976} > -5,87 \right\} = 1 - \Phi[-5,87] = \\ &= \Phi[5,87] \approx 0,99999998, \end{aligned}$$

а это значит, что вероятность отвергнуть нулевую гипотезу, когда она неверна, очень высокая. В связи с тем что $1 - \beta$ также имеет большое значение, разумнее выбрать α еще меньше, чем 0,01.

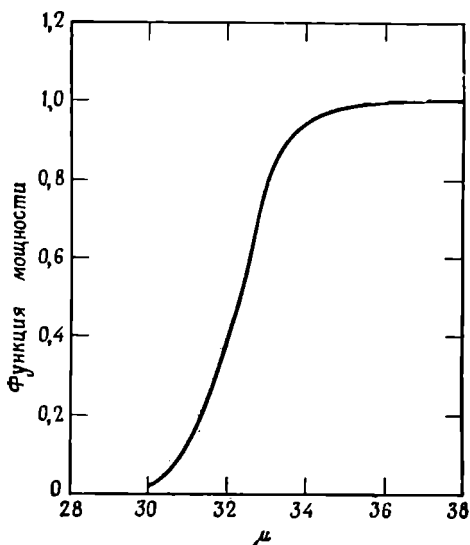
Пример 2

Общее положение. Пусть x_i — выборочное значение случайной величины, распределенной как $M[\mu, \sigma^2]$, где значение σ^2 известно.

$$H_1: \mu = \mu_1,$$

$$H_2: \mu > \mu_1.$$

Частный случай. Используя данные примера 1, найдем, как и в предыдущем случае, что гипотеза H_1 отвергается и что сле-



Фиг. 6.2. Функция мощности одностороннего нормально распределенного критерия с параметрами 0, 1.

дует принять гипотезу H_2 . Обратим внимание на мощность критерия при альтернативе $\mu > \mu_1$.

Функция мощности критерия была определена как $1 - \beta = P_{H_2} \{ \bar{x} \in R_2 \}$. Без сомнения, она зависит от альтернативных значений параметра. Так как $H_2: \mu > \mu_1$, функция мощности критерия для этого примера будет равна

$$P_{H_2} \{ \bar{x} \in R_2 \} = P_{H_2} \{ \bar{x} > c \} = P_{H_2} \left\{ \frac{\bar{x} - \mu}{0,976} > \frac{32,27 - \mu}{0,976} \right\} = \\ = 1 - \Phi \left[\frac{32,27 - \mu}{0,976} \right] = 1 - \beta.$$

Из этого равенства видно, что функция мощности полностью определяется параметром μ . Чтобы получить графическое представление зависимости между функцией мощности и альтернативными значениями μ при условии, что верна гипотеза H_2 , следует вычислить мощность для различных значений $\mu > \mu_1$, т. е. вычислить

$$1 - \Phi \frac{32,27 - \mu}{0,976} \text{ для всех } \mu > \mu_1.$$

В табл. 6.1 приведены значения мощности критерия, соответствующие различным значениям μ , а на фиг. 6.2 эти данные выражены графически. Так как рассматриваются только значения $\mu > \mu_1$, то задача называется односторонней. Соответственно и критерий также оказывается односторонним.

Таблица 6.1

Значения функции мощности

Конкурирующие гипотезы	Мощность	Конкурирующие гипотезы	Мощность
30	$1 - \Phi(2,32) = 0,01000$	34	$1 - \Phi(-1,77) = 0,96160$
31	$1 - \Phi(1,30) = 0,09850$	35	$1 - \Phi(-2,79) = 0,99740$
32	$1 - \Phi(0,28) = 0,39300$	36	$1 - \Phi(-3,81) = 0,99990$
33	$1 - \Phi(-0,75) = 0,77340$	37	$1 - \Phi(-4,84) = 0,99999$

Пример 3

Общее положение. Пусть x_i — выборочные значения случайной величины, распределенной как $N(\mu_1, \sigma^2)$, где дисперсия σ^2 известна. Требуется проверить нулевую гипотезу

$$H_1: \mu = \mu_1$$

при альтернативе

$$H_2: \mu \neq \mu_1$$

и заданном уровне значимости α .

Конкурирующая гипотеза H_2 в данном примере симметрична относительно μ_1 . Рассмотрим область, в которой гипотеза H_1

отвергается. Следует так выбрать R_2 , чтобы

$$\int_{R_2} f(x) dx = \alpha,$$

где $f(x)$ — плотность распределения статистики (в данном случае $\bar{x}$) при условии, что гипотеза H_1 верна. Область R_2 может быть удовлетворительно определена как состоящая из двух частей, одна из которых расположена слева от μ_1 , а другая — справа. Гипотеза H_1 будет отвергнута, если $\bar{x} > c_2$ или $\bar{x} < c_1$, где $c_1 < c_2$. Эти константы можно определить из следующего соотношения:

$$P_{H_1} \{ \bar{x} \in R_2 \} = P_{H_1} \{ \bar{x} > c_2 \} + P_{H_1} \{ \bar{x} < c_1 \} = \alpha.$$

Следовательно,

$$P_{H_1} \left\{ \frac{\bar{x} - \mu_1}{\sqrt{\sigma^2/n}} > \frac{c_2 - \mu_1}{\sqrt{\sigma^2/n}} \right\} + P_{H_1} \left\{ \frac{\bar{x} - \mu_1}{\sqrt{\sigma^2/n}} < \frac{c_1 - \mu_1}{\sqrt{\sigma^2/n}} \right\} = \alpha,$$

или, что равносильно,

$$1 = \Phi \left[\frac{c_2 - \mu_1}{\sqrt{\sigma^2/n}} \right] + \Phi \left[\frac{c_1 - \mu_1}{\sqrt{\sigma^2/n}} \right] = \alpha.$$

Критическая область представит собой две области, для которых

$$\int_{-\infty}^{c_1} f(x) dx + \int_{c_2}^{\infty} f(x) dx = \alpha, \quad (6.1)$$

где $f(x)$ — плотность распределения $\bar{x}$ при условии, что гипотеза H_1 верна.

Нетрудно видеть, что требованию (6.1) удовлетворяет множество комбинаций c_1 и c_2 . В их число входит и область, определенная как $\bar{x} > c_1$ и $\bar{x} < -c_1$. Чтобы решить, какая критическая область «лучше», полезно рассмотреть функции мощности в условиях альтернативной гипотезы при различных значениях c_1 и c_2 . В качестве примера положим, что

$$\frac{\sigma^2}{n} = \frac{1}{25}, \quad \mu_1 = 0, \quad \alpha = 0,05. \quad (6.2)$$

Область, в которой гипотеза H_1 принимается, выбрана как

$$\bar{x} > c_1, \quad \bar{x} < c_2.$$

Рассмотрим три частных случая, все удовлетворяющие требованию (6.1), но характеризующиеся различными функциями мощности. Во всех трех случаях мощность дается выражением

$$P(\mu) = 1 - \Phi \left[\frac{c_2 - \mu}{\sigma/\sqrt{n}} \right] + \Phi \left[\frac{c_1 - \mu}{\sigma/\sqrt{n}} \right].$$

В частности, при $\mu = \mu_1$

$$1 - \Phi \left[\frac{c_2 - \mu_1}{\sigma/\sqrt{n}} \right] + \Phi \left[\frac{c_1 - \mu_1}{\sigma/\sqrt{n}} \right] = 0,05 = \alpha.$$

Случай 1

$$1 - \Phi \left[\frac{c_2 - \mu_1}{\sigma/\sqrt{n}} \right] + \Phi \left[\frac{c_1 - \mu_1}{\sigma/\sqrt{n}} \right] = 0,05.$$

В условиях (6.2) это выражение можно записать в виде

$$1 - \Phi [5c_2] + \Phi [5c_1] = 0,05,$$

$$1 - 0,96 + 0,01 = 0,05;$$

следовательно,

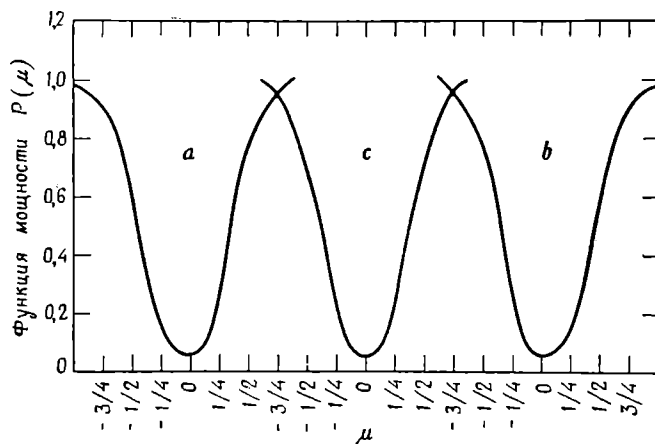
$$5c_2 = 1,75,$$

$$c_2 = 0,35$$

и

$$5c_1 = -2,326,$$

$$c_1 = -0,465.$$



Фиг. 6.3.

Функция мощности при этих значениях c_1 и c_2 (фиг. 6.3,а) будет

$$P(\mu) = 1 - \Phi [1,75 - 5\mu] + \Phi [-2,326 - 5\mu].$$

Случай 2

$$1 - \Phi [5c_2] + \Phi [5c_1] = 0,05,$$

$$1 - 0,98 + 0,03 = 0,05;$$

следовательно,

$$5c_2 = 2,06,$$

$$c_2 = 0,412$$

и

$$5c_1 = -1,88,$$

$$c_1 = -0,376.$$

Функция мощности этой критической области (фиг. 6.3, б) имеет вид

$$P(\mu) = 1 - \Phi[2,06 - 5\mu] + \Phi[-1,88 - 5\mu].$$

Случай 3

$$1 - \Phi[5c_2] + \Phi[5c_1] = 0,05,$$

$$1 - 0,975 + 0,025 = 0,05.$$

В этом случае

$$5c_2 = 1,96,$$

$$c_2 = 0,392$$

и

$$5c_1 = -1,96,$$

$$c_1 = -0,392.$$

Так как $c_1 = -c_2$, то критическая область симметрична и функция мощности определится как

$$P(\mu) = 1 - \Phi[1,96 - 5\mu] + \Phi[-1,96 - 5\mu],$$

где

$$P(\mu) = P(-\mu).$$

В табл. 6.2 приведены значения функций мощности для всех трех рассмотренных случаев при одном и том же значении μ . Из таблицы нетрудно видеть, что различным альтернативам соответствуют различные наилучшие критические области. Например, в случае альтернативы $\mu > \mu_1$ критическая область I лучше областей II и III. Таким образом, выбор критической области определяется условиями конкретной задачи, и в действительности не существует «лучшей» области. Экспериментатор сам или с помощью статистика правильнее других может судить о критической области. Если не оговорено противное, при альтернативе $\mu \neq \mu_1$ мы используем симметричную критическую область, применяя *двусторонний критерий*.

Частный случай. Пусть x_i — выборочные значения случайной величины, распределенной как $N(\mu, 100)$, и

$$H_1: \mu = 30,$$

$$H_2: \mu \neq 30,$$

$$\bar{x} = 31, \quad n = 10, \quad \alpha = 0,05.$$

**Значения функции мощности¹ для различных значений μ
при трех разных критических областях**

Нулевая гипотеза $H_1: \mu = 0$ при альтернативе $H_2: \mu \neq 0$; $\sigma^2/n = \frac{1}{25}$.

$\alpha = 0,05$.

Критические области: I. $c_1 = -0,465$, $c_2 = 0,35$; II. $c_1 = -0,376$, $c_2 = 0,412$;
III. $c_1 = -0,392$, $c_2 = 0,392$.

μ	Мощность		
	I	II	III
-1,00	0,9962	0,9991	0,9988
-0,75	0,9222	0,9693	0,9633
-0,50	0,5675	0,7324	0,7054
-0,25	0,1414	0,2648	0,2396
0,00	0,0500	0,0500	0,0500
0,25	0,3085	0,2099	0,2396
0,50	0,7774	0,6772	0,7054
0,75	0,9772	0,9545	0,9633
1,00	0,9994	0,9984	0,9988

¹ I. $P(\mu) = 1 - \Phi(1,75 - 5,0\mu) + \Phi(-2,33 - 5,0\mu)$;

II. $P(\mu) = 1 - \Phi(2,06 - 5,0\mu) + \Phi(-1,88 - 5,0\mu)$;

III. $P(\mu) = 1 - \Phi(1,96 - 5,0\mu) + \Phi(-1,96 - 5,0\mu)$.

В данном случае

$$\begin{aligned} \int_{-\infty}^{c_1} f(x) dx &= P_{H_1} \{ \bar{x} < c_1 \} = P_{H_1} \{ \bar{x} - 30 < c_1 - 30 \} = \\ &= \Phi [c_1 - 30] = \frac{\alpha}{2} = 0,025. \end{aligned}$$

Таким образом,

$$c_1 - 30 = -1,96,$$

$$c_1 = 28,04.$$

Аналогично

$$\int_{c_1}^{\infty} f(x) dx = P_{H_1} \{ \bar{x} > c_2 \} = \frac{\alpha}{2} = 0,025;$$

тогда

$$\Phi [c_2 - 30] = 0,975,$$

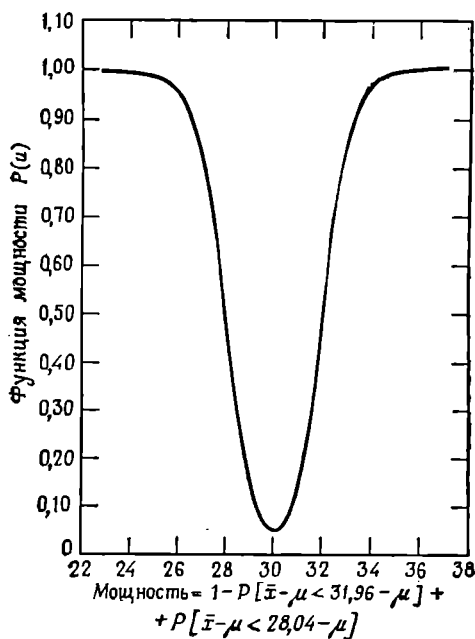
$$c_2 - 30 = 1,96,$$

$$c_2 = 31,96.$$

Следовательно, критическая область состоит из двух областей:

$$\bar{x} > 31,96 \text{ и } \bar{x} < 28,04.$$

Так как $\bar{x} = 31$, следует принять гипотезу H_1 . Мы рассматривали области, лежащие справа и слева от μ , т. е. имели дело не с простой, а со сложной альтернативой. Далее был применен



Ф и г. 6.4.

двусторонний критерий, функцию мощности которого можно записать как

$$1 - \beta = P_{H_2} \{ \bar{x} \in R_2 \} = P_{H_2} \{ \bar{x} - \mu > 31,96 - \mu \} + \\ + P_{H_2} \{ \bar{x} - \mu < 28,04 - \mu \} = 1 - P_{H_2} \{ \bar{x} - \mu < 31,96 - \mu \} + \\ + P_{H_2} \{ \bar{x} - \mu < 28,04 - \mu \} \text{ при } \mu = 30,$$

$$1 - \beta = 1 - P_{H_2} \{ \bar{x} - 30 < 1,96 \} + P_{H_2} \{ \bar{x} - 30 < -1,96 \} = 0,05 = \alpha.$$

Некоторые значения этой функции для отдельных альтернативных значений μ приведены в табл. 6.3. Графически ее изображение дано на фиг. 6.4. Читателю, возможно, интересно знать, что при надлежащем выборе масштаба по оси абсцисс (фиг. 6.4) будет отложена функция мощности любого двустороннего симметричного критерия с $\alpha = 0,05$. Это объясняется тем, что, ка-

ково бы ни было μ_1 , стандартизированная величина $(c_1 - \mu)/\sqrt{\sigma^2/n}$ будет принимать одни и те же значения.

Рассмотренные задачи были посвящены проверке гипотез о средних значениях нормально распределенных случайных величин при известной дисперсии σ^2 . Последнее предположение недостаточно реально, так как точное значение дисперсии σ^2 очень часто остается неизвестным. В подобной ситуации можно воспользоваться следующей оценкой для σ^2 :

$$\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = s^2,$$

где s^2 в силу нормальности распределения не зависит от $\bar{x}$.

Таблица 6.3

Значения функции мощности
 $1 - P\{\bar{x} - \mu < 31,96 - \mu\} + P\{\bar{x} - \mu < 28,04 - \mu\}$

μ	Мощность	μ	Мощность
22	0,99999	31	0,17010
24	0,99998	32	0,51600
26	0,97930	33	0,85080
27	0,85080	34	0,97930
28	0,51600	36	0,99998
29	0,17010	38	0,99999
30	0,05000		

Естественно, что в случае проверки гипотезы о средних условиях нормального распределения с применением одностороннего критерия ($H_1: \mu = \mu_1$ при альтернативе $\mu > \mu_1$) используется статистика

$$\frac{\bar{x} - \mu}{s/\sqrt{n}}, \quad (6.3)$$

а граница критической области определяется из уравнения

$$P_{H_2} \left\{ \frac{\bar{x} - \mu_1}{s/\sqrt{n}} > \frac{c_1 - \mu_1}{s/\sqrt{n}} \right\} = \alpha.$$

Если разделить числитель и знаменатель выражения (6.3) на $\sigma \neq 0$, получим

$$\frac{(\bar{x} - \mu)/\sigma}{\frac{1}{\sigma} \sqrt{\frac{1}{n(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2}} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}. \quad (6.3a)$$

Если гипотеза H_1 верна, числитель правой части равенства (6.3а) будет распределен нормально с параметрами $(0, 1)$, а знаменатель — эквивалентен выражению

$$\sqrt{\frac{\chi_{(n-1)}^2}{n-1}},$$

где $n-1$ — число степеней свободы, причем числитель и знаменатель независимы. Для величины, числитель которой нормально распределен с параметрами $(0, 1)$, а знаменатель есть корень квадратный из величины χ^2 с $n-1$ степенями свободы, деленной на $n-1$, известен закон распределения, называемый *законом Стьюдента*. Это распределение табулировано. Таким образом, величина

$$t = \frac{\bar{x} - \mu_1}{s/\sqrt{n}}$$

распределена по закону Стьюдента с $n-1$ степенями свободы (t_{n-1}) и

$$P_{H_1} \left\{ \frac{\bar{x} - \mu_1}{s/\sqrt{n}} > \frac{c_1 - \mu_1}{s/\sqrt{n}} \right\} = \alpha = P_{H_1} \left\{ t > \frac{c_1 - \mu_1}{s/\sqrt{n}} \right\} = \int_{t_{n-1}}^{\infty} f(t) dt.$$

Если H_2 : $\mu < \mu_1$, критическая область при заданном уровне значимости α такова, что

$$(\bar{x} - \mu_1)/\sqrt{s^2/n} \leq t_{n-1},$$

где

$$\int_{-\infty}^{t_{n-1}} f(t) dt = \alpha.$$

Если H_2 : $\mu \neq \mu_1$, для применяемого двустороннего критерия критическая область определится значениями t'_{n-1} и t''_{n-1} , для которых

$$\int_{-\infty}^{t'_{n-1}} f(t) dt + \int_{t''_{n-1}}^{\infty} f(t) dt = \alpha.$$

Если критическая область выбрана симметричной, гипотеза H_1 будет отвергнута и будет принята гипотеза H_2 : $\mu \neq \mu_1$ при

$$\frac{\bar{x} - \mu_1}{s/\sqrt{n}} > t_{n-1},$$

или

$$\frac{\bar{x} - \mu_1}{s/\sqrt{n}} < -t_{n-1},$$

т. е. гипотеза H_1 отвергается, если

$$\left| \frac{\bar{x} - \mu_1}{s/\sqrt{n}} \right| > t_{n-1}.$$

В этом случае функция мощности будет

$$P(\mu) = P_{H_2} \left\{ \frac{\bar{x} - \mu}{s/\sqrt{n}} > t_{n-1} \right\} + P_{H_2} \left\{ \frac{\bar{x} - \mu}{s/\sqrt{n}} < -t_{n-1} \right\}$$

при $\mu = \mu_1$, $P(\mu) = \alpha$.

Пример 4

Общее положение. Пусть x_i — выборочные значения случайной величины, распределенной как $N(\mu, \sigma^2)$, и значение σ^2 неизвестно:

$$H_1: \mu = \mu_1 = 30,$$

$$H_2: \mu < \mu_1 = 30.$$

Частный случай. Пусть

$$s^2 = 100,$$

$$\bar{x} = 26,$$

$$n = 25.$$

Положим $\alpha = 0,05$.

Критическое значение $t_{(n-1)=24}$ при $\alpha = 0,05$ таково, что

$$P_{H_1} \{t \in R_2\} = \alpha = 0,05.$$

В данном случае критическая область расположена слева от μ и, следовательно,

$$P_{H_1} \{t < -t_{24, 0,05}\} = \alpha,$$

где $t_{24, 0,05}$ — значение случайной величины, распределенной по закону Стьюдента с $n - 1$ степенями свободы при $\alpha = 0,05$.

В таблице значений t находим, что $-t_{24, 0,05} = -1,711$. Вычисленная нами величина $t = (\bar{x} - \mu_1) / \sqrt{s^2/n} = -\frac{6}{2} = -3 < -1,711$, из чего следует, что гипотеза H_1 должна быть отвергнута.

Пример 5

Общее положение. Пусть x_i — выборочные значения случайной величины, распределенной нормально с параметрами μ и σ^2 , причем значение σ^2 неизвестно:

$$H_1: \mu = \mu_1,$$

$$H_2: \mu \neq \mu_2.$$

Оценки s^2 и $\bar{x}$ вычислены по данным выборки объемом n .

Частный случай. Сумма пяти показаний измерения давления, выраженного в динах на квадратный сантиметр, для звуковой волны с частотой 40 гц была определена в Нарраганзетской морской лаборатории Университета штата Род-Айленд числом

$3200 \text{ дин/см}^2 \cdot \sum_{i=1}^5 (x_i - \bar{x})^2 = 1200 \text{ (дин/см}^2)^2$. По данным Дитца

и Кана¹, полученным в 1962 г., отсчеты давления распределены как $N(485, \sigma^2)$. Спрашивается, противоречат ли эти новые данные общепринятым? Пусть $\alpha = 0,05$. В нашем случае

$$H_1: \mu = 485,$$

$$H_2: \mu \neq 485.$$

Для проверки нулевой гипотезы нужно выбрать такое значение t , для которого

$$\int_{t_{n-1}}^{\infty} f(t) dt = \frac{\alpha}{2} = \int_{-\infty}^{-t_{n-1}} f(t) dt.$$

Таким образом, при выбранных нами критических областях гипотеза H_1 будет отвергнута, если

$$t = \left| \frac{\bar{x} - \mu}{s/\sqrt{n}} \right| > 2,776.$$

В данной задаче

$$t = \left| \frac{640 - 485}{\sqrt{\frac{3100}{5}}} \right| = 6,22 > 2,776,$$

из чего следует, что нулевая гипотеза должна быть забракована как ложная.

Пример 6

Общее положение. Пусть x_i и y_i — выборочные значения двух независимых нормально распределенных случайных величин с параметрами μ_1, σ_1^2 и μ_2, σ_2^2 соответственно. Кроме того, обе дисперсии известны и $\sigma_1^2 \neq \sigma_2^2$. Требуется проверить нулевую гипотезу

$$H_1: \mu_1 = \mu_2$$

¹ F. T. Dietz, J. S. Kahn, Application of the theoretical model for the seasonal analysis of shallow water ambient sound, Submitted to Journ. Am., Acoustical Soc., 1962.

при следующих трех альтернативах:

$$1) H_2: \mu_1 < \mu_2,$$

$$2) H_3: \mu_1 > \mu_2,$$

$$3) H_4: \mu_1 \neq \mu_2.$$

Положим уровень значимости равным α и допустим, что $\bar{x}$ получено по данным выборки объемом n_1 , а $\bar{y}$ — объемом n_2 . Так как x_i и y_i — значения независимых нормально распределенных случайных величин, то математическое ожидание величины $\bar{x} - \bar{y}$ будет равно $\mu_1 - \mu_2$, а ее дисперсия — $\sigma_1^2/n_1 + \sigma_2^2/n_2$. Следовательно, случайная величина

$$\frac{(\bar{x} - \bar{y}) - E(\bar{x} - \bar{y})}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (6.4)$$

распределена нормально с параметрами (0, 1). Если нулевая гипотеза верна, то $\mu_1 - \mu_2 = 0$ и (6.4) будет иметь вид

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}};$$

так как распределение этой величины известно, она может быть эффективно использована для построения статистического критерия.

1) Критическая область при альтернативе $\mu_1 < \mu_2$ определяется константой c , для которой $\bar{x} - \bar{y} < c$ и

$$P_{H_1} \left\{ \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < \frac{c}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right\} = \alpha,$$

и, следовательно,

$$\Phi \left[\frac{c}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right] = \alpha.$$

2) Критическая область при альтернативе $\mu_1 > \mu_2$ может быть найдена аналогично, так как

$$P_{H_1} \left\{ \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > \frac{c}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right\} = \alpha,$$

$$\Phi \left[\frac{c}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right] = 1 - \alpha.$$

3) В случае третьей альтернативы ($\mu_1 \neq \mu_2$) критерий становится двусторонним, и критическая область в этом случае определяется константой c , для которой

$$P_{H_1} \left\{ \left| \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right| > \frac{c}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right\} = \alpha.$$

Упражнения

1. Построить статистический критерий для вышеприведенной задачи, если а) $\sigma_1^2 = \sigma_2^2$, $n_1 \neq n_2$; б) $\sigma_1^2 = \sigma_2^2$, $n_1 = n_2$.

2. Привести примеры проверки гипотезы $H_1: \mu_1 = \mu_2$, используя данные ваших собственных исследований при альтернативах а) $H_2: \mu_1 > \mu_2$; б) $H_2: \mu_1 \neq \mu_2$.

Пример 7

Общее положение. Пусть x_i и y_i — выборочные значения независимых нормально распределенных случайных величин с неизвестными параметрами μ_1 , σ_1^2 и μ_2 , σ_2^2 соответственно. Оценки для σ_1^2 и σ_2^2 обозначим через s_1^2 и s_2^2 . Нужно проверить нулевую гипотезу

$$H_1: \mu_1 = \mu_2$$

при трех следующих альтернативах:

$$H_2: \mu_1 < \mu_2,$$

$$H_3: \mu_1 > \mu_2,$$

$$H_4: \mu_1 \neq \mu_2.$$

Если верна гипотеза H_1 , то

$$E(\bar{x} - \bar{y}) = \mu_1 - \mu_2 = 0$$

и

$$\text{var}(\bar{x} - \bar{y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Тогда величина

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

будет распределена как $N(0, 1)$. Конечно, хорошо было бы знать значения σ_1^2 и σ_2^2 , но они, к сожалению, неизвестны, и вместо них придется воспользоваться оценками s_1^2 и s_2^2 . Подставляя значения оценок в предыдущее выражение вместо σ_1^2 и σ_2^2 , получим

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (6.5)$$

Чтобы использовать полученную статистику при нахождении критической области и выборе решающего правила в случаях трех вышеприведенных альтернатив, необходимо знать распределения этого критерия при условии, что гипотеза H_1 верна. Можно показать, что эта статистика не будет распределена по закону Стьюдента, так как подкоренное выражение $\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)$ не будет распределено как величина

$$\frac{\sigma^2 \chi^2}{(\text{число степеней свободы})}.$$

Однако если n_1 и n_2 — большие числа, то случайная величина (6.5) будет распределена приблизительно нормально. В подобной ситуации критические области определяются, как показано в примере 6. Теперь изменим условия нашей задачи, записав их в следующем виде:

$$H_1: \mu_1 = \mu_2 \text{ при } \sigma_1^2 = \sigma_2^2 = \sigma^2,$$

а конкурирующие гипотезы оставим прежними. Тогда выражение (6.5) запишется проще:

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\bar{x} - \bar{y}}{\sigma \sqrt{\frac{n_1 + n_2}{n_1 n_2}}}.$$

Затем нужно получить оценку для σ^2 . Так как $\sigma_1^2 = \sigma_2^2 = \sigma^2$, можно получить «объединенную» оценку для σ^2 , используя значения s_1^2 и s_2^2 .

Выше было показано, что если x_i — значения нормально распределенной величины с параметрами μ и σ^2 , вычисленная по этим данным оценка s^2 может быть представлена как функция от χ^2 , т. е.

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}.$$

Если из m нормальных совокупностей взяты m выборок различного объема n_i , для каждой оценки s_i^2 будет

существовать аналогичная связь с χ^2 . Таким образом,

$$\begin{aligned}\chi_1^2 &= \frac{(n_1 - 1) s_1^2}{\sigma_1^2}, \\ \chi_2^2 &= \frac{(n_2 - 1) s_2^2}{\sigma_2^2}, \\ &\dots \dots \dots \\ \chi_m^2 &= \frac{(n_m - 1) s_m^2}{\sigma_m^2}.\end{aligned}$$

Далее, величина

$$\sum_{i=1}^m \chi_i^2 = \sum_{i=1}^m \frac{(n_i - 1) s_i^2}{\sigma_i^2}$$

будет распределена как χ^2 с $\sum_{i=1}^m (n_i - 1)$ степенями свободы. Следовательно, если

$$\sigma_1^2 = \sigma_2^2 = \dots = \sigma_m^2 = \sigma^2,$$

то

$$\sum_{i=1}^m \chi_i^2 = \frac{1}{\sigma^2} \sum_{i=1}^m (n_i - 1) s_i^2 = \chi^2 \quad (6.6)$$

и

$$\frac{\sigma^2 \chi^2}{\sum_{i=1}^m (n_i - 1)} = \frac{\sum_{i=1}^m (n_i - 1) s_i^2}{\sum_{i=1}^m (n_i - 1)}.$$

В результате выражение

$$\frac{\sigma^2 \chi^2}{\sum_{i=1}^m (n_i - 1)}$$

будет равным некоторой взвешенной оценке для σ^2 . Если положить

$$s_p^2 = \frac{\sum_{i=1}^m (n_i - 1) s_i^2}{\sum_{i=1}^m (n_i - 1)},$$

то объединенная оценка для $\sigma_1^2 = \sigma_2^2 = \sigma^2$ будет иметь вид

$$s_p^2 = \frac{\sum_{i=1}^2 (n_i - 1) s_i^2}{\sum_{i=1}^2 (n_i - 1)} = \frac{(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2}{n_1 + n_2 - 2}$$

и статистика, по которой строится критерий, примет следующее выражение:

$$\frac{\bar{x} - \bar{y}}{s_p \sqrt{\frac{n_1 + n_2}{n_1 n_2}}}.$$

Полученная величина распределена (при гипотезе H_1) по закону Стьюдента с $n_1 + n_2 - 2$ степенями свободы. Исходя из этого, можно указать критические значения для трех выше сформулированных альтернатив:

1) $H_2: \mu_1 < \mu_2$. Критическая область определяется величиной $t_{n_1+n_2-2, \alpha}$, для которой

$$t = \frac{\bar{x} - \bar{y}}{s_p \sqrt{\frac{n_1 + n_2}{n_1 n_2}}} < t_{n_1+n_2-2, \alpha},$$

где $t_{n_1+n_2-2, \alpha}$ — отрицательное число.

2) $H_3: \mu_1 > \mu_2$. Гипотеза H_1 отвергается, если

$$t = \frac{\bar{x} - \bar{y}}{s_p \sqrt{\frac{n_1 + n_2}{n_1 n_2}}} > t_{n_1+n_2-2, 1-\alpha}.$$

3) $H_4: \mu_1 \neq \mu_2$. Гипотеза H_1 отвергается, если

$$|t| > t_{n_1+n_2-2, 1-\frac{\alpha}{2}}.$$

В качестве примера можно привести данные Смита и Фланегана¹, сведенные в табл. 6.4. Эти авторы изучали изменения радиоактивности в зависимости от степени выветривания в конвейских гранитах Нью-Гэмпшира. Радиоактивность измерялась при помощи β — γ -счетчиков.

Одна из интересных задач состоит в решении вопроса о наличии каких-либо существенных расхождений между средними значениями радиоактивности окрашенных выветрелых и свежих гранитов, в данном случае красных и зеленых. Таким образом, в качестве нулевой гипотезы относительно красных гранитов утверждается, что между средними значениями радиоактивности

¹ W. L. Smith, F. J. Flanagan, Use of statistical methods to detect radioactivity change due to weathering of a granite, Am. Journ. Sci., 254, 1956.

Значения радиоактивности в образцах из Красного карьера

Зеленые граниты				Красные граниты			
выветрелые		свежие		выветрелые		свежие	
№ образца	отсчет ¹	№ образца	отсчет ¹	№ образца	отсчет ¹	№ образца	отсчет ¹
135	909	113	710	123	836	101	942
136	(1418) ²	114	754	124	768	102	887
137	799	115	712	125	904	103	1041
138	831	116	750	126	1044	104	933
139	762	117	713	127	1143	105	969
140	716	118	767	128	853	106	1023
141	801	119	715	129	1030	107	1126
142	760	120	700	130	924	108	776
143	806	121	724	131	996	109	980
144	821	122	716	132	893	110	1051
				133	1100	111	1081
				134	873	112	1060
Количество образцов . .	9	10		12		12	
Среднее . . .	798,1	726,1		955,3		989,1	
Дисперсия . .	3109,0	507,0		12 260,0		9 234,0	
Стандартные отклонения	55,8	22,5		110,7		96,1	

¹ Продолжительность В — γ-отсчетов 5 мин.² Это значение из вычислений исключено.

выветрелых и свежих пород нет существенных различий и что их можно рассматривать как равные. Согласно конкурирующей гипотезе, представляющей больший интерес, среднее значение радиоактивности свежих гранитов больше, чем выветрелых. Формальная запись этих гипотез следующая:

$$H_1: \mu_f = \mu_w$$

при альтернативе

$$H_2: \mu_f > \mu_w,$$

где μ_f — неизвестное истинное среднее значение радиоактивности свежих красных пород, а μ_w — выветрелых. Вычисленные оценки μ_f и μ_w равны 989,1 и 955,3 соответственно, $n_f = n_w = 12$.

112 Применяя вышеописанную процедуру проверки нулевой гипо-

тезы и допуская, что необходимые для ее успешного применения условия выполнены, найдем, что

$$s_p = \left[\frac{11 \cdot 12260 + 11 \cdot 9234}{22} \right]^{\frac{1}{2}} = (10747)^{\frac{1}{2}} = 103,7$$

и что $t = -33,8/17,3 = -1,96$. Критическая область представляет собой множество точек, значения которых меньше $-2,06 = -t_{22, 0,05}$. Так как вычисленное значение t больше $-2,06$, нулевую гипотезу следует принять как согласующуюся с эмпирическими данными. Вывод, полученный в результате такой проверки, заключается в том, что радиоактивность свежих и выветрелых красных гранитов одна и та же.

ЗАМЕНА КРИТЕРИЯ СТЬЮДЕНТА ДРУГИМИ

Критерий τ_d для выборок малого объема

При сравнении выборочных средних отношение Стьюдента можно заменить другими статистиками, в частности величиной, содержащей $\bar{x}$ и выборочный размах ω . Этот критерий не рекомендуется для выборок, объем которых значительно превышает 10, но во всех остальных случаях он наилучший, так как обеспечивает быстроту работы, если имеются таблицы. Для выборок равного объема статистика τ_d определяется формулой

$$\tau_d = \frac{\bar{x}_1 - \bar{x}_2}{\omega_1 + \omega_2};$$

ее распределение табулировано в книге Диксона и Мэсси [79]. Например, в поле с двух разных участков собраны две коллекции образцов пород. В каждом образце на глаз определили количество кварца в единицах площади образца. Нулевую гипотезу, согласно которой средние количества кварца в пределах этих участков отличаются несущественно, можно проверить по имеющимся данным следующим способом:

Коллекция А	Коллекция В
3	6
5	8
6	11
7	12
9	13
$\omega_A = 9 - 3 = 6$	$\omega_B = 13 - 6 = 7$
$\bar{x}_A = 6$	$\bar{x}_B = 10$
$n_A = 5$	$n_B = 5$
$\tau_d = \frac{\bar{x}_A - \bar{x}_B}{\omega_A + \omega_B} = \frac{6 - 10}{6 + 7} = -0,308.$	

Так как для выборки объемом $n_1 = n_2 = 5$ критическое значение, соответствующее 5%-ному уровню значимости, равно 0,246, то нулевую гипотезу следует отвергнуть.

Критерий U Манна — Уитни; непараметрическая замена критерия Стьюдента

Этот критерий непараметрический и предназначен для сравнения двух независимых выборок. Его можно рассматривать как заменитель критерия Стьюдента, когда условия, необходимые для успешного применения последнего, не выполняются. Критерий U может быть применен во всех случаях, когда выборочные данные можно упорядочить. Статистика U строится следующим образом:

1. Ряд данных располагается по величине в единой упорядоченной последовательности без изменения первоначального написания отдельных значений.

2. Меньшая по объему выборка обозначается через N_1 , а большая — N_2 .

3. Одна из выборок, допустим N_1 , используется для описанной ниже процедуры подсчета.

4. Все члены упорядоченного ряда последовательно рассматриваются от наименьшего к наибольшему. После того как установлен первый член выборки N_1 , подсчитывается число всех членов выборки N_2 , которые ему предшествуют. Затем эта же операция повторяется для второго члена выборки N_1 начиная с начала. Процедура продолжается до тех пор, пока такой счет можно производить. Величина U представляет собой сумму всех результатов подсчета.

Этот метод можно проиллюстрировать на следующем примере. Исследователя интересуют плеохроичные дворики как показатели возрастного сходства пород. Диаметры плеохроичных двориков в биотите, окружающем включения, измерены в двух шлифах из образцов, отобранных из различных пород.

Предположим, что порода A стратиграфически старше породы B . В табл. 6.5 приведены измеренные диаметры двориков в единицах микрометра окуляра.

Исследователь хочет предварительно проверить справедливость утверждения, что значения, полученные из совокупности образцов породы A , в целом превышают значения из B . Рассмотрим применение критерия.

Шаг 1. Все данные располагаются в единой возрастающей последовательности, и каждое значение диаметра плеохроичного дворика ставится в соответствие с A или B в зависимости от того, какой выборке это значение принадлежит. Наименьшая выборка $N_1 = 3$, а $N_2 = 4$ (табл. 6.6).

Таблица 6.5

А	В
8	7
14	5
11	10
9	

Таблица 6.6

5	7	8	9	10	11	14
В	В	А	А	В	А	А

Шаг 2. Вычисляется величина U как общая сумма всех B , предшествующих каждому последовательно рассматриваемому A , т. е.

$$U = 2 + 2 + 3 + 3 = 10.$$

Шаг 3. Смотрим таблицу J^1 для одностороннего критерия нулевой гипотезы H_0 , заключающейся в том, что выборки A и B взяты из совокупностей с одним и тем же распределением; при альтернативе H_1 выборка A взята из совокупности, статистически большей, чем B . Для $N_2=4$ и $N_1=3$ находим, что вычисленное значение U больше наибольшего табличного значения U , равного 8. Пересчитываем U следующим образом:

$U_{\text{пересчитанное}} = N_1 N_2 - U = 3 \cdot 4 - U = 2$. Этот же результат можно получить другим путем, подсчитывая в табл. 6.6 общее число членов A , предшествующих каждому выбираемому последовательно члену B .

При $N_2=4$, $N_1=3$ и $U_{\text{пересчитанное}}=2$ находим, что вероятность, соответствующая такому малому значению U , как 2, есть 0,114. Если принять $\alpha=0,05$, то критическая область для гипотезы H_0 будет включать все значения U , которым соответствуют вероятности, меньшие или равные 0,05. Так как в данном случае вероятность, соответствующая $U=2$, превышает 0,05, то нулевая гипотеза должна быть принята. Эти данные не позволяют исследователю делать вывод о том, что в более древних образцах диаметры плеохроичных двори́ков в биотите, окружающем включения, больше.

Если бы альтернативная гипотеза H_2 состояла в том, что две выборки взяты из различных совокупностей (а не в том, что совокупность A статистически больше B), то следовало бы пользоваться двусторонним критерием и таблицами. Вопрос о проверке нулевой гипотезы (включающей случай больших N_1 и N_2) с общих позиций рассматривает Сигал². Вывод критерия и соответствующие таблицы приведены в работе Манна и Уитни³.

¹ S. Siegal, Nonparametric statistics for the behavioral sciences, New York, 1956, p. 271—273.

² См. сноску на стр. 111.

³ H. B. Mann, D. R. Whitney, On a test of whether one of two random variables in stochastically larger than the other, Ann. Math. Stats., 18, 1947,

Критерии для проверки гипотез о дисперсии нормального распределения

Рассмотренные выше критерии, принадлежащие к наиболее часто употребляемым, относятся к проверке гипотез только об одном из параметров нормального распределения — о среднем значении μ . Ниже будут рассмотрены критерии, применяемые для проверки гипотез о втором параметре нормального распределения — о дисперсии σ^2 .

Пример 8

Общее положение. Пусть x_i — выборочные значения случайной величины, распределенной как $N(\mu, \sigma^2)$, а $\bar{x}$ и s^2 определены по данным выборки объемом n . Выберем в качестве нулевой гипотезы $H_1: \sigma_1^2 = \sigma^2$, которую нужно проверить в условиях трех различных альтернатив:

$$H_2: \sigma^2 < \sigma_1^2,$$

$$H_3: \sigma^2 > \sigma_1^2,$$

$$H_4: \sigma^2 \neq \sigma_1^2.$$

Выше было показано, что выражение

$$\chi_{(n-1)}^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \bar{x})^2,$$

где x_i — значения нормально распределенной случайной величины с параметрами μ, σ^2 , можно записать в виде

$$\chi_{(n-1)}^2 = \frac{s^2 (n-1)}{\sigma^2}.$$

Если гипотеза H_1 , сформулированная выше, верна, то

$$\chi_{(n-1)}^2 = \frac{s^2 (n-1)}{\sigma_1^2}.$$

Таким образом, критические области можно определить значениями $\chi_{(n-1)}^2$, при которых H_1 будет отвергнута, а одна из альтернатив (H_2, H_3, H_4) — принята, если

$$\frac{s^2 (n-1)}{\sigma_1^2} < \chi_{(n-1), \alpha}^2,$$

$$\frac{s^2 (n-1)}{\sigma_1^2} > \chi_{(n-1), 1-\alpha}^2,$$

$$\frac{s^2 (n-1)}{\sigma_1^2} > \chi_{(n-1), 1-\frac{\alpha}{2}}^2$$

или

$$\frac{s^2(n-1)}{\sigma_1^2} < \chi_{(n-1), \frac{\alpha}{2}}^2$$

ФУНКЦИИ МОЩНОСТИ КРИТЕРИЕВ ДЛЯ ПРОВЕРКИ ГИПОТЕЗ О ДИСПЕРСИИ НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ

Критическое значение критерия при альтернативе (H_2) есть $\chi_{(n-1), \alpha}^2$; гипотеза H_1 должна быть отвергнута, если

$$\frac{s^2(n-1)}{\sigma_1^2} < \chi_{(n-1), \alpha}^2. \quad (6.7)$$

В условиях рассматриваемой альтернативы H_2 : $\sigma^2 < \sigma_1^2$ при $\sigma^2 \neq 0$ выражение (6.7) можно записать как

$$\frac{s^2(n-1)}{\sigma^2} < \frac{\sigma_1^2}{\sigma^2} \chi_{(n-1), \alpha}^2;$$

тогда функцию мощности $P(\sigma^2)$ можно представить в виде

$$P(\sigma^2) = 1 - \beta = P_{H_2} \left\{ \frac{s^2(n-1)}{\sigma^2} < \frac{\sigma_1^2}{\sigma^2} \chi_{(n-1), \alpha}^2 \right\}.$$

Если положить $\lambda = \frac{\sigma_1^2}{\sigma^2}$, предыдущее выражение можно записать короче:

$$P \left\{ \frac{s^2(n-1)}{\sigma^2} < \lambda \chi_{(n-1), \alpha}^2 \right\}.$$

Для других конкурирующих гипотез

$$P(\sigma^2) = P_{H_3} \left\{ \frac{s^2(n-1)}{\sigma^2} > \lambda \chi_{(n-1), 1-\alpha}^2 \right\},$$

$$P(\sigma^2) = P_{H_4} \left\{ \frac{s^2(n-1)}{\sigma^2} > \lambda \chi_{(n-1), 1-\frac{\alpha}{2}}^2 \right\} + \\ + P_{H_4} \left\{ \frac{s^2(n-1)}{\sigma^2} < \lambda \chi_{(n-1), \frac{\alpha}{2}}^2 \right\}.$$

Пример 9

Общее положение. Пусть x_i и y_i — выборочные значения независимых нормально распределенных случайных величин с параметрами μ_1 , σ_1^2 и μ_2 , σ_2^2 соответственно. Оценки s_1^2 и s_2^2 для неизвестных дисперсий σ_1^2 и σ_2^2 получены по выборкам объемами n_1 и n_2 .

Требуется проверить нулевую гипотезу

$$H_1: \sigma_1^2 = \sigma_2^2$$

при альтернативах

$$H_2: \sigma_1^2 < \sigma_2^2,$$

$$H_3: \sigma_1^2 > \sigma_2^2,$$

$$H_4: \sigma_1^2 \neq \sigma_2^2.$$

С этой задачей очень часто приходится сталкиваться в исследовательской работе. В данном случае нужно проверить справедливость утверждения, согласно которому две выборки, взятые из нормальных совокупностей, безразлично, с одинаковыми или равными средними, характеризуются равными дисперсиями, т. е. эти выборки однородны в смысле изменчивости. Напомним пример 7: при проверке гипотезы $\mu_1 = \mu_2$ в условиях неизвестных дисперсий σ_1^2 и σ_2^2 мы пришли к выводу, что для построения критической области t -статистики нужно предположить $\sigma_1^2 = \sigma_2^2$. Справедливость этого предположения и рассматривается в данном примере. При проверке гипотезы о равенстве средних при помощи t -критерия предполагалось, что гипотеза $\sigma_1^2 = \sigma_2^2$ уже проверена.

Как величина

$$\frac{1}{\sigma_1^2} \sum_{i=1}^{n_1} (x_i - \bar{x})^2 = \frac{s_1^2 (n_1 - 1)}{\sigma_1^2},$$

так и величина

$$\frac{1}{\sigma_2^2} \sum_{i=1}^{n_2} (y_i - \bar{y})^2 = \frac{s_2^2 (n_2 - 1)}{\sigma_2^2}$$

независимы и распределены по закону χ^2 с $n_1 - 1$ и $n_2 - 1$ степенями свободы соответственно. Отношение двух любых независимых случайных величин, распределенных как χ^2 , зависит только от n_1 и n_2 . Распределение этого отношения известно под названием *F-распределения*¹:

$$F_{(n_1-1, n_2-1)} = \frac{\frac{s_1^2 (n_1 - 1)}{\sigma_1^2 (n_1 - 1)}}{\frac{s_2^2 (n_2 - 1)}{\sigma_2^2 (n_2 - 1)}} = \frac{s_1^2 \sigma_2^2}{s_2^2 \sigma_1^2}.$$

Если гипотеза $H_1: \sigma_1^2 = \sigma_2^2$ верна, то

$$F_{(n_1-1, n_2-1)} = \frac{s_1^2}{s_2^2};$$

это отношение можно использовать в качестве критерия для проверки нулевой гипотезы. Критические области уровня значимости α для вышеперечисленных альтернатив определяются

как

$$F = \frac{s_1^2}{s_2^2} < F_{(n_1-1, n_2-1), \alpha},$$

$$F = \frac{s_1^2}{s_2^2} > F_{(n_1-1, n_2-1), 1-\alpha},$$

$$F = \frac{s_1^2}{s_2^2} > F_{(n_1-1, n_2-1), 1-\frac{\alpha}{2}}$$

или

$$F = \frac{s_1^2}{s_2^2} < F_{(n_1-1, n_2-1), \frac{\alpha}{2}}.$$

По данным Смита и Фланегена (см. сноску на стр. 111), проверим гипотезу об одинаковой изменчивости показаний радиоактивности свежих и выветрелых красных гранитов, т. е. проверим справедливость сделанного ранее допущения. Таким образом, нужно проверить нулевую гипотезу

$$H_1: \sigma_w^2 = \sigma_f^2$$

при альтернативе

$$H_2: \sigma_w^2 > \sigma_f^2.$$

Выборочные оценки для σ_w^2 и σ_f^2 приведены в табл. 6.4; они равны соответственно $s_w^2 = 12\,260$ и $s_f^2 = 9234$. Если гипотеза H_1 верна, то

$$F \binom{11}{11} = \frac{s_w^2}{s_f^2}.$$

Критическая область при уровне значимости $\alpha = 0,05$ представляет собой множество значений $s_w^2/s_f^2 > 2,82$. В данном случае $s_w^2/s_f^2 = 1,33$, из чего следует, что нулевую гипотезу можно принять и что сделанное предположение при проверке гипотезы $\mu_w = \mu_f$ было справедливым.

Если потребуется проверить гипотезу о равенстве средних значений радиоактивности в выветрелых и свежих зеленых гранитах, то, прежде чем перейти к ее проверке, следует прежде всего проверить гипотезу о равенстве дисперсий. Обозначим через ξ_w^2 неизвестную дисперсию радиоактивности выветрелых зеленых гранитов, а в свежих — ξ_f^2 . Их оценки будут соответственно $s_w^2 = 3109$ и $s_f^2 = 507$. Сформулируем нулевую гипотезу как

$$H_1: \xi_w^2 = \xi_f^2$$

при альтернативе

$$H_2: \xi_w^2 > \xi_f^2,$$

которая заключается в предположении, что изменчивость радиоактивности выветрелых зеленых гранитов больше, чем свежих. Критическая область будет представлять множество значений, больших 3,23. Так как

$$F = \frac{3109}{507} = 6,13 > 3,23,$$

то нулевая гипотеза отвергается и принимается конкурирующая H_2 . Это свидетельствует о том, что совокупности различны в смысле изменчивости, которая больше в выветрелых гранитах, чем в свежих. Проверять эту гипотезу не следует, если обе выборки взяты одним и тем же методом из заведомо одной совокупности.

Критерий для проверки гипотезы о равенстве более двух дисперсий¹

Критерий для проверки гипотезы о равенстве более двух дисперсий был предложен Бартлетом (1937 г.). Не останавливаясь на теоретическом выводе критерия, рассмотрим пример.

Пример 10

Общее положение. Пусть имеется m выборок объемом $n_1, n_2, \dots, n_m$ из нормальных совокупностей с неизвестными дисперсиями $\sigma_1^2, \sigma_2^2, \dots, \sigma_i^2, \dots, \sigma_m^2$. Обозначим через $s_1^2, s_2^2, \dots, s_i^2, \dots, s_m^2$ их статистические оценки, а через $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_i, \dots, \bar{x}_m$ оценки средних:

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, \quad s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2.$$

По имеющимся данным требуется проверить нулевую гипотезу $H_1: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_m^2 = \sigma^2$ при альтернативе H_2 , согласно которой хотя бы одно значение $\sigma_i^2 \neq \sigma^2$. Бартлет показал, что если гипотеза H_1 верна, то величина

$$\begin{aligned} \chi^2 &= \frac{1}{C} \left[\sum_{i=1}^m (n_i - 1) \ln s_i^2 - \sum_{i=1}^m (n_i - 1) \ln s_i^2 \right] = \\ &= \frac{2,3056}{C} \left[\sum_{i=1}^m (n_i - 1) \lg s_i^2 - \sum_{i=1}^m (n_i - 1) \lg s_i^2 \right] \quad (6.8) \end{aligned}$$

¹ Вывод критерия, приведенный авторами, опущен, так как в нем имеются ошибки. — Прим. ред.

распределена как χ^2 с $m-1$ степенями свободы и может быть использована в качестве критерия для проверки гипотезы H_1 . В выражении (6.8)

$$s_p^2 = \frac{1}{\left(\sum_{i=1}^m n_i\right) - 1} \sum_{i=1}^m (n_i - 1) s_i^2$$

представляет собой обобщенную оценку генеральной дисперсии σ^2 , а величина C равна

$$C = 1 + \frac{1}{3(m-1)} \left[\sum_{i=1}^m \frac{1}{(n_i - 1)} - \frac{1}{\sum_{i=1}^m (n_i - 1)} \right].$$

Таким образом, гипотеза H_1 должна быть отвергнута, если вычисленное значение χ^2 окажется больше допустимого $\chi_{(m-1), \alpha}^2$ при $m-1$ степенях свободы и уровне значимости α . Наоборот, гипотезу H_1 следует принять, если $\chi^2 < \chi_{(m-1), \alpha}^2$.

Пример 11

Общее положение. Пусть (x_1, x_2) — выборочное значение двумерной нормально распределенной случайной величины с параметрами $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ и σ_{12} . Неизвестный коэффициент корреляции обозначим через

$$\rho = \frac{\sigma_{12}}{\sigma_1 \sigma_2},$$

оценки для $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ — через $\bar{x}_1, \bar{x}_2, s_1^2, s_2^2$, а объем выборки — через n . По имеющимся данным требуется проверить нулевую гипотезу

$$H_1: \rho = 0$$

при альтернативе

$$H_2: \rho > 0.$$

Начнем с того, что оценкой для ρ служит величина

$$r = \frac{\sum_{i=1}^n (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\left[\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2 \sum_{i=1}^n (x_{i2} - \bar{x}_2)^2 \right]^{1/2}}. \quad (6.9)$$

Это выражение можно записать в следующем, хотя и довольно громоздком, но более удобном при работе на **121**

клавишных вычислительных машинах виде:

$$r = \frac{\sum_{i=1}^n x_{i1}x_{i2} - \frac{1}{n} \sum_{i=1}^n x_{i1} \sum_{i=1}^n x_{i2}}{\left\{ \left[\sum_{i=1}^n x_{i1}^2 - \frac{1}{n} \left(\sum_{i=1}^n x_{i1} \right)^2 \right] \left[\sum_{i=1}^n x_{i2}^2 - \frac{1}{n} \left(\sum_{i=1}^n x_{i2} \right)^2 \right] \right\}^{1/2}},$$

или в виде

$$r = \frac{n \sum_{i=1}^n x_{i1}x_{i2} - \sum_{i=1}^n x_{i1} \sum_{i=1}^n x_{i2}}{\left\{ \left[n \sum_{i=1}^n x_{i1}^2 - \left(\sum_{i=1}^n x_{i1} \right)^2 \right] \left[n \sum_{i=1}^n x_{i2}^2 - \left(\sum_{i=1}^n x_{i2} \right)^2 \right] \right\}^{1/2}}.$$

Распределение случайной величины r при условии, что гипотеза H_1 верна, довольно сложно, и ее трудно использовать для проверки гипотезы $\rho=0$. Однако, как показал Фишер¹, при $\rho=0$ случайная величина

$$\frac{r}{(1-r^2)^{1/2}} (n-2)^{1/2}$$

распределена по закону Стьюдента с $n-2$ степенями свободы. Таким образом, гипотеза $\rho=0$ должна быть отвергнута, если

$$t = \frac{r}{(1-r^2)^{1/2}} (n-2)^{1/2} > t_{(n-2), \alpha},$$

где $t_{(n-2), \alpha}$ — допустимое значение t при уровне значимости α и $n-2$ степенях свободы. В приведенном неравенстве для различных значений $t_{(n-2), \alpha}$ можно подсчитать соответствующие допустимые значения $r_{(n-2), \alpha}$, которые приведены в виде таблиц в некоторых учебниках². Наличие таких таблиц позволяет избежать вычисления величины t , и гипотеза α отвергается, если вычисленное значение r превышает $r_{(n-2), \alpha}$. В качестве примера рассмотрим результаты Силвермана и Бейтса³, которые пытались провести структурный контроль содержаний урана в глинистых сланцах Чаттануга при помощи проверки гипотезы равенства нулю коэффициента корреляции между степенью ориентированности слюд и иллитов, выраженной с помощью показателя ориентировки, и содержанием урана в процентах. С этой целью указанные авторы использовали 58 образцов глинистых сланцев, отобранных из керна одной скважины. Эти данные приведены в табл. 6.7.

¹ R. A. Fisher, Statistical methods for research workers, New York, 1954.

² Б. Л. Ван дер Варден, Математическая статистика, М., ИЛ., 1960.

³ E. N. Silverman, T. E. Bates, X-ray diffraction study of orientation in the Chattanooga shale, Am. Min., 45, 1960.

Показатели ориентировки и содержания урана в 58 образцах
керна скважины УВ-19

№ образца	Показатель ориентировки	Содержание урана, 10^{-4} %	№ образца	Показатель ориентировки	Содержание урана, 10^{-4} %
152,1	1,645	91	167,2	1,747	5
152,2	2,317	76	167,8	2,769	21
152,7	2,003	115	168,0	3,392	16
153,1	2,295	108	168,5	1,667	10
153,8	3,421	86	169,0	2,495	9
154,0	1,615	73	169,5	2,342	10
154,5	2,071	87	170,1	2,336	0
155,1	1,686	89	170,5	1,445	2
155,7	2,872	97	171,1	2,101	7
156,4	2,182	44	171,3	2,729	1
157,3	1,549	36	171,9	2,296	6
158,1	2,011	22	172,1	2,307	10
158,5	0,600	23	172,8	1,878	7
159,0	2,466	40	173,4	2,533	8
159,2	2,688	56	173,9	1,717	3
159,5	1,283	56	174,2	1,651	7
159,9	1,413	57	174,5	1,507	9
160,2	7,714	37	174,8	1,395	2
161,5	2,075	51	175,2	1,833	14
162,1	1,399	71	175,8	1,921	12
162,6	1,550	68	176,1	2,408	20
163,2	2,882	47	176,7	3,188	19
164,1	2,006	71	177,0	2,572	38
164,4	2,104	63	177,3	2,084	32
165,0	2,468	70	177,7	2,699	55
165,3	3,224	71	179,0	2,687	46
166,0	1,993	57	179,9	2,089	27
166,2	3,603	39	180,7	2,137	17
166,8	1,755	14	181,5	0,015	5

Вычисленная по этим данным оценка коэффициента корреляции r равна 0,143 при $n - 2 = 56$ степенях свободы. Из таблиц допустимых значений r для 56 степеней свободы и уровня значимости $\alpha = 0,05$ следует: для того чтобы гипотеза $H_1: \rho = 0$ была отвергнута, нужно чтобы оценка r превысила 0,26. Но, так как $r = 0,143 < 0,26$, гипотезу принимаем. Отсюда делаем вывод о независимости содержаний урана от ориентировки слюды и иллитов.

ПРОВЕРКА ГИПОТЕЗЫ О СОГЛАСИИ ВЫБОРОЧНОГО РАСПРЕДЕЛЕНИЯ И НЕКОТОРОГО ИЗВЕСТНОГО ТЕОРЕТИЧЕСКОГО РАСПРЕДЕЛЕНИЯ

Случай, когда параметры известны

Допустим, что исследователю нужно провести эксперимент и на основании полученных результатов проверить гипотезу при помощи критерия Стьюдента. В данном случае одним из предположений, которое он должен или безоговорочно принять, или проверить, является предположение нормальности распределения. Как бы поступил исследователь, используя полученные им данные? При решении этой задачи следует рассмотреть такие величины, которые служили бы мерой отклонения выборочных распределений от теоретических.

Приоритет введения величины χ^2 как объективной меры отклонения принадлежит К. Пирсону. Вообще, если дана выборка объемом $n = \sum_{i=1}^m f_i$, где f_i — число появлений события с номером i , теоретическая вероятность которого p_i , а $i = 1, 2, \dots, m$, то в качестве подходящей меры отклонения выборочного распределения от теоретического разумно было бы рассматривать некоторую функцию от $(f_i - np_i)$. В частности, было показано, что при $n \rightarrow \infty$ и при условии, что нулевая гипотеза H_1 верна, случайная величина

$$\sum_{i=1}^m \frac{(f_i - np_i)^2}{np_i}$$

распределена как χ^2 с $m - 1$ степенями свободы. Заметим, что в этом случае величина p_i известна *a priori* и что она не является оценкой, полученной по данным выборки. Таким образом, чтобы получить критическую область, нужно из таблиц χ^2 выбрать значение, соответствующее заданному уровню значимости и числу степеней свободы. Это значение используется в дальнейшем как критическое, т. е. если

$$\chi^2 = \sum_{i=1}^m \frac{(f_i - np_i)^2}{np_i} \geq \chi_{(m-1), 1-\alpha}^2$$

то нулевая гипотеза отвергается.

Случай, когда параметры теоретических распределений оцениваются по выборке

В большинстве случаев точные значения параметров распределения остаются неизвестными, и, следовательно, они должны оцениваться по выборке. Как показал Крамер [16], свойства вы-

борочного распределения χ^2 зависят от используемых оценок параметров. Так как для каждого параметра существует множество оценок, то возникает задача о распределении статистического критерия.

Допустим, как это делалось выше, существует m непересекающихся событий и каждое при эксперименте появилось f_i раз, где $i=1, 2, \dots, m$. Введем k неизвестных параметров $(\theta_1, \theta_2, \dots, \theta_k)$, которые оцениваются по выборке. Оказывается, что естественный критерий, аналогичный приведенному выше, будет распределен как χ^2 , только на этот раз с $m - k - 1$ степенями свободы. Каждому оцениваемому параметру соответствует потеря одной степени свободы, т. е. при k параметрах остается $(m - 1) - k$ степеней свободы. Таким образом, статистику, на которой основан критерий, можно записать как

$$\sum_{i=1}^m \frac{(f_i - e_i)^2}{e_i}, \quad (6.10)$$

где $e_i = p_i(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k)$, а $\hat{\theta}_i$ — оценка для θ_i , полученная методом «максимума правдоподобия». В качестве примера применения этого критерия рассмотрим обоснование предположения о нормальности по данным Кана (1956), который сообщил о следующих результатах изучения распределения содержаний терригенного материала по шлифам трех различных осадочных пород. Каждая порода была представлена шестью шлифами, в которых производились подсчеты в пяти пересечениях, приходящихся на один шлиф.

Процедура проверки пригодности нормального распределения как модели этих эмпирических данных заключается в следующем.

Так как значения параметров μ и σ неизвестны, следует найти их оценки $\bar{x}$ и s . В данном случае, согласно нулевой

Т а б л и ц а 6.8

Класс, %	f_i
90—100	11
80— 90	23
70— 80	23
60— 70	21
50— 60	9
40— 50	3
	90
	$m = 6$
	$\bar{x} = 74,67$
	$s = 12,95$

гипотезе, выборка взята из нормальной совокупности со средним $\mu=74,67$ и дисперсией $\sigma^2=(12,95)^2$. Если нулевая гипотеза верна, то случайная величина

$$\frac{x - 74,67}{12,95}$$

также распределена нормально с параметрами $(0, 1)$. Здесь x следует рассматривать в качестве любого наблюдения из имеющихся 90. Так как данные сгруппированы, можно подсчитать значения теоретической функции распределения (при данных значениях оценок параметров), соответствующие пределам классов. Так, $\Phi[(x - 74,67)/12,95]$ — теоретическое значение доли индивидуумов в исходной совокупности с величиной признака, меньшей или равной x .

Если эту процедуру провести для всех нижних пределов сгруппированных данных и из каждого последующего значения полученной функции распределения вычесть предыдущее, то полученные разности будут представлять собой теоретические значения вероятности того, что случайная величина примет значение, принадлежащее данному классу. Затем, если эти вероятности умножить на объем выборки $n \sum_{i=1}^m f_i$, получим ожидаемую частоту e_i для каждого класса с номером i .

Наконец, если подсчитать

$$\chi^2_{m-k-1} = \sum_{i=1}^m \frac{(f_i - e_i)^2}{e_i}, \quad (6.11)$$

то получим значение нашей статистики.

Предложенный пример детально рассмотрен в табл. 6.9. В данном случае $k=2$, так как вместо неизвестных параметров распределения μ и σ^2 были использованы их статистические оценки $\bar{x}$ и s^2 , т. е. наложены два линейных ограничения.

Читатель, очевидно, обратил внимание, что числа наблюдений в классах от 40 до 50 и от 50 до 60 объединены в одну величину — 12. Это было сделано в соответствии с общепринятыми правилами [16]¹. Вообще, если ожидаемая частота окажется меньше 5, то группа объединяется с соседней так, чтобы вновь полученное число было больше 5.

Так как $\chi^2 = \sum_{i=1}^5 \frac{(f_i - e_i)^2}{e_i} = 1,093$ и не превышает критического значения $\chi^2_{2, 0,05} = 5,99$, то нулевую гипотезу следует принять как согласующуюся с результатами наблюдений.

¹ См. гл. 16, где этот вопрос рассматривается в связи с распределением Пуассона.

Т а б л и ц а 6.9

Вычисление критерия согласия с нормальным распределением

i	Класс, %	$\frac{x_i - \bar{x}}{s}$	$\Phi\left(\frac{x_i - \bar{x}}{s}\right) = \Phi_i$	$\Phi_{i-1} - \Phi_i$ (при $i > 0$)	$n(\Phi_{i-1} - \Phi_i) = e_i$	f_i	$f_i - e_i$	$\frac{(f_i - e_i)^2}{e_i}$
0	> 100	—	1,0000	—	10,71	11	0,29	0,0079
1	90—100	1,184	0,8810	0,1190	19,97	23	3,03	0,4597
2	80— 90	0,412	0,6591	0,2219	26,97	23	—3,97	0,5844
3	70— 80	—0,361	0,3594	0,2997	20,72	21	0,28	0,0038
4	60— 70	—0,133	0,1292	0,2302	9,097	12	0,65	0,0372
5	50— 60	—1,905	0,0281	0,1011	2,25			
6	< 50	—2,736	0,0031	0,0250				
								Вычисленное $\chi^2 = 1,093$
								Табличное $\chi^2_{2, 0,05} = 5,991$

¹ Объединенная частота.

Часто при планировании экспериментов желательно заранее фиксировать размер риска, на который исследователь готов пойти, так как в очень многих случаях об этом забывают, и оказывается, что напрасно затрачены время, энергия и деньги. Известно, что вероятность появления той или иной ошибки есть функция эксперимента. Напомним, что уровень значимости α , т. е. вероятность появления ошибки первого рода, есть вероятность отвергнуть проверяемую нулевую гипотезу, тогда как на самом деле она верна. Вероятность β ошибки второго рода есть вероятность ошибочного принятия нулевой гипотезы, в то время как в действительности она ложная. Следовательно, мощность $1 - \beta$ представляет собой вероятность забраковать ошибочную нулевую гипотезу. Многие исследователи рассматривают только ошибку первого рода и мало уделяют внимания ошибке второго рода или даже вообще не обращают на нее внимания. Такое упущение может оказаться опасным, особенно в тех задачах, где ошибка второго рода играет важную роль. Рассмотрим следующие примеры.

Недавно происходила полемика о связи между курением табака и заболеванием раком. Пусть, согласно нулевой гипотезе, доля курящих, умерших от рака легких за определенный отрезок времени, равна доле некурящих, умерших от рака легких за тот же отрезок времени, т. е.

$$H_1: \Delta_S = \Delta_N$$

при альтернативе

$$H_2: \Delta_S > \Delta_N.$$

Этим вопросом могут заниматься два типа исследователей: одни — по поручению табачной промышленности, а другие — независимо от нее. Допустим, что исследователи каждой группы пытаются проверить нулевую гипотезу. Постановка задачи в такой форме должна бы приводить к отчетливому ответу на вопрос. Однако это не всегда так. Может случиться, что группа исследователей, не зависящих от табачной промышленности, отвергнет нулевую гипотезу, а зависящая группа примет ее как подтвердившуюся. Если данные были одни и те же для обеих групп исследователей, почему же могло возникнуть такое противоречие? В связи с этим рассмотрим ошибки первого и второго рода.

Понятно, что группа, связанная с табачной промышленностью, хочет сделать уровень значимости α по возможности малой величиной, т. е. ошибка первого рода для них очень нежелательна. С другой стороны, группа независимых исследователей может уделить меньше внимания ошибке первого рода и сосредоточить его на мощности критерия, сделав последнюю

по возможности большей, а это значит, что они сильно зашита-ресованы в принятии альтернативной гипотезы, если она верна. Следовательно, эта группа исследователей хочет, чтобы вероятность отвергнуть нулевую гипотезу, если она ложна, была бы очень большой. Таким образом, две группы исследователей, работающих над решением одной и той же задачи, могут получить различные критические области вследствие разного понимания важности того или иного вида ошибки.

В качестве другого примера влияния ошибок первого и второго рода рассмотрим следующий возможный, хотя и гипотетический, случай. Допустим, что при разведочных работах на нефть изучен потенциально перспективный для бурения район. После многократных обсуждений решили не бурить. Но к этому времени уже были произведены затраты, которые выразились в сумме $x'000$ долларов. Поэтому возможно, что, несмотря на такое решение, правильнее было бы принять противоположное. Формально нулевой гипотезе H_1 можно придать приблизительно следующую формулировку: характеристики потенциальной нефтеносности данного района не отличаются значимо от характеристик известных нефтеносных площадей при конкурирующей гипотезе, согласно которой они различны. Теперь, если бурить не рекомендовано, в то время как в данном районе есть нефть, будет допущена ошибка первого рода, т. е. нулевая гипотеза отвергается, будучи правильной. С другой стороны, допустим, что принято решение бурить, а нефти на самом деле нет. Значит, принимается ложная нулевая гипотеза, т. е. допущается ошибка второго рода.

Какая из этих ошибок обойдется дороже? Это, без сомнения, экономическая задача, так как должны быть уравновешены уже не просто «академические» ошибки первого и второго рода, а ошибки, связанные со стоимостью, выраженной в долларах и центах.

$\alpha, \beta, 1 - \beta$ и n

Предположим, что производится серия определений модуля разрыва нового твердого топлива с целью выяснить вопрос, существует ли разница между усилиями, требуемыми для его разрушения и разрушения старого испытанного топлива, для которого среднее значение усилия составляет $\mu = 0,18 \cdot 10^4 \text{ кг/см}^2$. Определим объем выборки, необходимый для обеспечения $\alpha = 0,05$ и $1 - \beta = 0,99$ при альтернативе $\mu = 0,25 \cdot 10^4 \text{ кг/см}^2$. Допустим, что стандартное отклонение σ приблизительно равно $0,14 \cdot 10^4 \text{ кг/см}^2$ и что результаты измерения распределены нормально, и определим функцию мощности:

$$P(\mu) = 1 - \Phi \left[\frac{c - \mu}{\sigma/\sqrt{n}} \right].$$

При $\mu = 3,5$

$$P(\mu) = 1 - \Phi \left[\frac{\sqrt{n}(c - 3,5)}{2} \right] = 0,99.$$

Так как

$$\Phi \left[\frac{\sqrt{n}(c - 3,5)}{2} \right] = 0,01,$$

то

$$\frac{\sqrt{n}(c - 3,5)}{2} = -2,33,$$

$$c\sqrt{n} - 3,5\sqrt{n} = -4,66.$$

Далее, при $\alpha = 0,05$

$$1 - \alpha = \Phi \left[\frac{\sqrt{n}(c - 2,5)}{2} \right] = 0,95$$

и, следовательно,

$$\frac{\sqrt{n}(c - 2,5)}{2} = 1,645.$$

Таким образом, решив систему уравнений

$$c\sqrt{n} - 3,5\sqrt{n} = -4,66,$$

$$c\sqrt{n} - 2,5\sqrt{n} = 3,29,$$

найдем, что n равно приблизительно 63, т. е. нужно исследовать 63 пробы нового топлива.

Теперь определим объем выборки, необходимый для того, чтобы обеспечить мощность критерия 85%, при проверке нулевой гипотезы о том, что пористость солоноватоводных отложений формации Моррисон равна 25% при альтернативе 22% и при уровне значимости 5%. Используем симметричный двусторонний критерий и примем, что стандартное отклонение приблизительно равно 10,0%. Тогда

$$1 - \Phi \left[\frac{c_2\sqrt{n} - 25\sqrt{n}}{10} \right] + \Phi \left[\frac{c_1\sqrt{n} - 25\sqrt{n}}{10} \right] = 0,05,$$

откуда

$$\Phi \left[\frac{c_1\sqrt{n} - 25\sqrt{n}}{10} \right] = 0,025,$$

$$c_1\sqrt{n} - 25\sqrt{n} = -1,96 \cdot 10,$$

$$c_1\sqrt{n} = 25\sqrt{n} - 1,96 \cdot 10, \quad (6.12)$$

$$c_2\sqrt{n} = 25\sqrt{n} + 1,96 \cdot 10.$$

Так как мощность $P(\mu)$ при $\mu = 22$ равна 0,85, то, следовательно,

$$1 - \Phi \left[\frac{c_2\sqrt{n} - 22\sqrt{n}}{10} \right] + \Phi \left[\frac{c_1\sqrt{n} - 22\sqrt{n}}{10} \right] = 0,85. \quad (6.13)$$

Подставив результаты (6.12) в (6.13), получим

$$1 - \Phi \left[\frac{19,6 + 3\sqrt{n}}{10} \right] + \Phi \left[\frac{-1,96 + 3\sqrt{n}}{10} \right] = 0,85.$$

Подставляя в (6.13) различные значения n , найдем, что приблизительно при $n=100$ наши требования будут удовлетворены.

КОЭФФИЦИЕНТ ВАРИАЦИИ

В тех случаях, когда сравнивается изменчивость признака, измеренного в выборке из одной совокупности, с изменчивостью этого же признака в другой совокупности, результаты измерения необходимо выражать в сопоставимых единицах. Так, например, при сравнении разброса веса мышей и разброса веса слонов было бы бессмысленным использовать F -отношение, выражая результаты измерения в граммах, фунтах и т. п. В геологической практике часто сталкиваются с тем, что стандартное отклонение и среднее находятся в прямой пропорциональной зависимости. Естественной шкалой для сравнения двух групп подобных результатов бывает логарифмическая шкала. Также и в других случаях мы можем поставить задачу отыскания преобразования первоначальных данных, обладающего теми или иными стабилизирующими свойствами¹. После преобразования мы придем к задаче сравнения групповых средних. Так, логарифмическое преобразование часто приводит к приближенно нормальным распределениям², для которых задача сравнения средних рассматривалась выше. Высказанное замечание относится к случаям, не столь крайним, как упомянутый выше (с мышами и слонами); например, его следует иметь в виду при сравнении степени изменчивости в нескольких группах наблюдений, когда переход к новой шкале дает определенный выигрыш (такие ситуации возникают в дисперсионном анализе).

Безразмерная величина $s/\bar{x}=v$, называемая *коэффициентом вариации*, является адекватной мерой изменчивости, если выборка взята из приблизительно нормальной совокупности. Эта мера рассмотрена в работе Симпсона и Роя³. Обозначив коэффициент вариации совокупности через V , можно различными методами проверить гипотезу $V=V_0$. Известно (см., например, книгу Хальда [19] и Шеффе⁴, что $\mu(v)=\sigma/\mu=V$ и дисперсия

¹ Как можно понять, здесь имеется в виду такое преобразование двух групп данных, после которого они будут иметь одинаковые дисперсии и, возможно, отличаться средними значениями. — *Прим. ред.*

² Таково положение в размерном анализе осадочных образований.

³ G. G. Simpson, A. Roe, *Quantitative zoology*, New York, 1939.

⁴ H. Scheffé, *A method for judging all contrasts in the analysis of variance*, *Biometrika*, 40, 1953.

$\text{var}(v)$ для оценки v коэффициента вариации дается выражением

$$\text{var}(v) = \frac{V^2(1+2V^2)}{2(n-1)}, \quad (6.14)$$

где n — объем выборки, который полагается большим, а значение V мало. Таким образом, критерий для проверки гипотезы $V=V_0$ можно строить по статистике

$$Z = \frac{v - V_0}{\sqrt{V_0(1+2V_0^2)/2(n-1)}}, \quad (6.15)$$

которая, если проверяемая гипотеза верна, распределена приблизительно нормально с параметрами $(0, 1)$. Устойчивость этого критерия, насколько известно, изучена мало.

Другой метод, предложенный Хальдом [19], основан на рассмотрении статистики $u = s - \bar{x}C_p$, где $P\{\frac{s}{\bar{x}} < C_p\} = p$. Можно показать, что случайная величина u распределена приблизительно нормально с параметрами

$$\begin{aligned} E(u) &= \sigma - \mu C_p, \\ \text{var}(u) &= \sigma^2 \left[\frac{1}{2(n-1)} + \frac{C_p}{n} \right]. \end{aligned} \quad (6.16)$$

Распределение этой случайной величины ближе к нормальному, чем распределение Z . При помощи статистики u можно построить критерий для проверки гипотезы $H_1: V=V_0$ следующим образом. Допустим, что нулевую гипотезу $H_1: V=V_0$ нужно проверить при альтернативе $V>V_0$ и уровне значимости $\alpha=0,05$. Величина

$$C_{0,95} = \frac{1 + Z_{0,95} \sqrt{\frac{1}{2(n-1)} \left[1 - \frac{V_0^2 Z_{0,95}^2}{n} + \frac{V_0^2}{n} \right]}}{1 - \frac{V_0^2 Z_{0,95}^2}{n}}, \quad (6.17)$$

где $Z_{0,95}$ — 95%-ная квантиль нормального распределения. Если $s/\bar{x} < C_{0,95}$, то гипотеза H_1 принимается, а при $s/\bar{x} > C_{0,95}$ она отвергается.

Существует еще один метод проверки гипотезы $H_1: V=V_0$, который, однако, требует таблиц нецентрального распределения Стьюдента. Правда, эти таблицы менее доступны, чем, скажем, таблицы χ^2 . Метод подробно рассмотрен в работе Джонсона и Уилча [136]. Обычно при проверке этой гипотезы достаточно воспользоваться приближением Мак-Кэя [177], который показал, что величина $\frac{nv^2(1/V^2+1)}{1+v^2}$ распределена прибли-

132 тельно как χ^2 с $n-1$ степенями свободы. Это приближение

Мак-Кэя может быть использовано в тех случаях, когда изучаемое распределение не очень асимметрично, и при условии, что $V \leq 1/3$. Гипотеза $H_1: V = V_0$ при альтернативе $H_2: V > V_0$ проверяется следующим образом. Находим, что $\chi_0^2 = \frac{nv(1/V_0^2 + 1)}{1 + v^2}$.

Если $\chi_0^2 > \chi_{(n-1), 0,05}^2$, нулевая гипотеза отвергается. Этот критерий может быть использован и в случае выборок очень малого объема (около 10), если обеспечены условия, сформулированные выше. Оказывается, что применение статистик v или χ_0^2 обеспечивает наименьшие погрешности в задаваемом уровне значимости.

Если задача состоит в оценке неизвестного значения V в совокупности, то наиболее точной оценкой будет $v' = (1 + 1/4n)v$, так как в нее введена поправка на систематическую ошибку [116]. Однако, если требуется охарактеризовать только выборку, оценка v вполне пригодна. Иногда может потребоваться оценить V или V^2 при помощи интервальной оценки. Такую оценку можно получить различными способами. Если применяется вышеописанный метод χ^2 , то 95%-ный доверительный интервал будет

$$\frac{nv^2}{\chi_{(n-1), 0,025}^2 [1 + v^2] - nv^2} < V^2 < \frac{nv^2}{\chi_{(n-1), 0,975}^2 [1 + v^2] - nv^2}. \quad (6.18)$$

Если же используется способ Джонсона и Уилча, то доверительный интервал для V можно построить, применив таблицы нецентрального распределения Стьюдента. В том случае, когда применяется u -критерий Хальда, можно найти величины $C_{0,025}$ и $C_{0,975}$, такие, что неравенство $C_{0,025} < \frac{s}{x} < C_{0,975}$ будет выполняться с надежностью 0,95.

Если применяется критерий Z , доверительный интервал строится аналогично.

В ряде случаев требуется сравнить неизвестные коэффициенты вариации по данным двух независимых выборок. Если объем выборок большой, для этой цели можно воспользоваться Z -критерием:

$$Z = \frac{v_1 - v_2}{\sqrt{\text{var}(v_1) + \text{var}(v_2)}}.$$

Этот критерий можно применять только при очень больших выборках, так как даже в условиях одной выборки он требует многочисленных наблюдений. Поэтому для сравнения коэффициентов вариации по двум выборкам следует настоятельно рекомендовать применение распределения Мак-Кэя. Напомним, что даже в условиях очень малых выборок (от 6 до 10) и при $V < 1/3$ случайная величина $\frac{nv^2}{1 + v^2} \left(\frac{1}{V^2} + 1 \right)$ распределена как

χ^2 с $n - 1$ степенями свободы. Таким образом, целесообразно для проверки гипотезы о равенстве коэффициентов вариации по двум выборкам, объем которых не сильно превышает 10, применять способ, изложенный ниже (правда, этот способ следовало бы подвергнуть экспериментальной проверке). Так как случайные величины $\frac{n_1 v_1^2}{1 + v_1^2} \left(1 + \frac{1}{V_1^2}\right)$ и $\frac{n_2 v_2^2}{1 + v_2^2} \left(1 + \frac{1}{V_2^2}\right)$ независимы и распределены как χ^2 с $n_1 - 1$ и $n_2 - 1$ степенями свободы соответственно, то при условии, что гипотеза $H_0: V_1^2 = V_2^2$ верна, получим величину

$$F = \left[\frac{n_1 - 1}{n_2 - 1} \right] = \frac{\frac{v_1^2}{1 + v_1^2} \left(\frac{n_1}{n_1 + 1} \right)}{\frac{v_2^2}{1 + v_2^2} \left(\frac{n_2}{n_2 + 1} \right)}, \quad (6.19)$$

которая распределена как F -отношение Фишера с $n_1 - 1$ и $n_2 - 1$ степенями свободы. Если в знаменателе, как обычно, будет стоять меньшее из выражений $\frac{v^2}{1 + v^2} \frac{n}{n + 1}$, то гипотеза H_0 будет браковаться при условии

$$F = \left[\frac{n_1 - 1}{n_2 - 1} \right] > F \left[\frac{n_1 - 1, 0,95}{n_2 - 1, 0,95} \right]$$

при уровне значимости $\alpha = 0,10$. Если $\alpha = 0,05$, критическое значение будет $F \left[\frac{n_1 - 1, 0,975}{n_2 - 1, 0,975} \right]$. Доверительный интервал для V_1^2/V_2^2 не так просто построить.

Другой возможный способ сравнения двух квадратов коэффициентов вариации связан с областью перекрытия их доверительных интервалов. Об этом способе в литературе почти нет данных.

Если поставить задачу о проверке гипотезы о равенстве K -коэффициентов вариации, то придется столкнуться с трудностями, во всяком случае не меньшими, чем те, которые возникли у Бартлета при построении им критерия для проверки гипотезы о равенстве нескольких дисперсий. Применение его критерия в нашей задаче оказывается невозможным вследствие высокой чувствительности этого критерия к виду изучаемого распределения. Таким образом, сравнение K -коэффициентов вариации является очень трудной, еще нерешенной задачей. Один из возможных способов ее решения состоит в построении набора F -критериев по методу, рассмотренному выше, для всех возможных пар коэффициентов вариации при уровне значимости

134 $\frac{2\alpha}{K(K-1)}$. Этот способ требует большой осторожности.

Средние нормального распределения

Обзор оценок в гл. 4 был посвящен так называемым точечным оценкам, которые характеризуют неизвестный параметр одним конкретным значением. Однако во многих случаях интервальные оценки оказываются предпочтительнее. Интервальная оценка представляет собой интервал, которому поставлено в соответствие число, выражающее вероятность того, что истинное значение параметра накрывается этим интервалом. Вышеизложенный метод точечных оценок может быть естественно развит в концепцию интервальных оценок.

Доверительные интервалы для среднего μ нормального распределения*Дисперсия σ^2 известна*

Так как среднее арифметическое $\bar{x}$ распределено нормально с параметрами μ , σ^2/n , то

$$P\left\{-c \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq c\right\} = 1 - \alpha,$$

где

$$\Phi(c) = 1 - \frac{\alpha}{2}.$$

Таким образом, в случае $100(1 - \alpha)\%$ *доверительного интервала для μ* получим

$$P\left\{\bar{x} + \frac{\sigma}{\sqrt{n}}c \geq \mu \geq \bar{x} - \frac{\sigma}{\sqrt{n}}c\right\} = 1 - \alpha.$$

Симметричный 95%-ный доверительный интервал определяется из выражения

$$P\left\{-1,96 \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq 1,96\right\} = 0,95,$$

где

$$P\left\{\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq 1,96 = c\right\} = 1 - \frac{\alpha}{2}$$

и

$$\Phi[c = 1,96] = 1 - \frac{\alpha}{2} = 0,975.$$

Вообще пределы этого интервала для μ равны

$$\bar{x} - \frac{\sigma}{\sqrt{n}} \cdot 1,96, \quad \bar{x} + \frac{\sigma}{\sqrt{n}} \cdot 1,96.$$

Дисперсия σ^2 неизвестна

Пусть s^2 — оценка неизвестной дисперсии σ^2 . Согласно (6.3а),

$$P \left\{ -t_{n-1, 1-\frac{\alpha}{2}} \leq \frac{\bar{x} - \mu}{s/\sqrt{n}} \leq t_{n-1, 1-\frac{\alpha}{2}} \right\} = 1 - \alpha;$$

поэтому $100(1 - \alpha)\%$ -ный интервал имеет вид

$$\bar{x} - t_{n-1, 1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1, 1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}.$$

Пример

Мелтон¹ проверил и принял как подтвердившуюся гипотезу о несущественном различии углов наклона северных и южных склонов долин Южной Аризоны. Измерив углы крутизны склонов, расположенных один против другого, по их разности, он нашел среднее арифметическое разностей $\bar{\theta} = -0,0255^\circ$, оценив стандартное отклонение ($1,97^\circ$) по выборке из 47 разностей. Каков же интервал, в котором с вероятностью 0,95 находится истинное значение среднего? В данном случае $\bar{\theta} = \bar{x} = -0,0255^\circ$, $s = 1,97$, $\sqrt{n} = \sqrt{47} = 6,85$, $t = 2,01$.

Таким образом, $-2,01 \leq \frac{-0,025 - \mu}{1,97/6,85} \leq 2,01$ и, следовательно, $-0,60^\circ \leq \mu \leq 0,55^\circ$.

Итак, при многократном повторении случайных выборок объемом $n=47$ нужно ожидать, что определенные таким способом интервалы будут покрывать истинное значение среднего приблизительно в 95 случаях из 100.

Дисперсия нормального распределения

Выше было показано, что

$$P \left\{ \chi^2_{n-1, 1-\frac{\alpha}{2}} \geq \frac{s^2(n-1)}{\sigma^2} \geq \chi^2_{n-1, \frac{\alpha}{2}} \right\} = 1 - \alpha.$$

Таким образом,

$$P \left\{ \frac{\chi^2_{n-1, 1-\frac{\alpha}{2}}}{s^2(n-1)} \geq \frac{1}{\sigma^2} \geq \frac{\chi^2_{n-1, \frac{\alpha}{2}}}{s^2(n-1)} \right\} = 1 - \alpha$$

и $100(1 - \alpha)\%$ -ный доверительный интервал для σ^2 имеет вид

$$\frac{s^2(n-1)}{\chi^2_{n-1, 1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{s^2(n-1)}{\chi^2_{n-1, \frac{\alpha}{2}}}.$$

¹ M. A. Melton, Intravalley variation in slope angles related to microclimate and erosional environment, Bull. Geol. Soc. Am., 71, 1960.

Доверительные пределы для параметров распределения это пределы области, накрывающей неизвестные значения параметров с заданной надежностью, т. е. в заданном проценте всех случаев.

С другой стороны, толерантные пределы не имеют отношения к параметрам. Те толерантные пределы, которые мы рассмотрим здесь, суть концы интервалов случайной длины, центрированных выборочным средним $\bar{X}$. Эти интервалы будут с заданной надежностью содержать по крайней мере долю P всех индивидуумов совокупности при данном виде распределения. Соответственно заданный процент будущих наблюдений с выбранной степенью надежности попадет в построенные таким способом интервалы.

Толерантные пределы записываются в следующей форме:

$$\bar{X} \pm Ks,$$

где $\bar{X}$ — выборочная средняя, s — выборочное стандартное отклонение, K — коэффициент. Толерантные пределы определяются по соответствующим таблицам, содержащим входные параметры: долю P , уровень надежности γ и объем выборки n . Такие таблицы с соответствующими пояснениями приведены в работе Эйзенхарта и др. «Техника статистического анализа» (1947 г.), а также в книге Диксона и Мэсси [79]. Для этих таблиц предполагается, что распределение в совокупности нормальное. Толерантные пределы для совокупностей, распределения которых отличаются от нормального, а также непараметрические толерантные пределы рассмотрены Вальдом (1946 г.) и Уилксом (1942 г.). Однако до сих пор нет ответа общей формулировки на вопрос, какого типа толерантные пределы следует использовать в прикладных задачах. Если нельзя сделать никаких предположений о виде распределения в изучаемой совокупности, то можно воспользоваться таблицами непараметрических толерантных пределов Мэрфи¹. В этом случае сколько-нибудь полезные пределы получаются при очень больших объемах выборки. Например, если нужно определить пределы, содержащие 95% всей совокупности, с допустимой ошибкой 0,10, то минимальный требуемый объем выборок составит 79 [79]. К сожалению, коллекции данных в различных разделах геологии, например в геохимии или палеонтологии, редко достигают такого объема.

В качестве примера из палеонтологии рассмотрим случай, когда найдена ископаемая форма несколько больших размеров по сравнению с теми же особями в обычной коллекции. В данном

¹ R. B. Murphy, Nonparametric tolerance limits, Ann. of Math. Stat., 19, 1948.

случае представляется возможным, если мы используем пределы, полученные в предположении нормального распределения, решить не только вопрос о принадлежности найденной формы к данному виду, но и охарактеризовать «таксономический размах» вокруг среднего значения исходной коллекции, в соответствии с которым все найденные в дальнейшем формы будут или причислены к данному виду, или же отнесены к другому. Допустим, что среднее арифметическое, полученное по результатам измерения пяти особей из основной коллекции, равно 10,00, а оценка стандартного отклонения $s=2,916$. Из таблиц находим, что 99% всей совокупности при $n=5$ находятся в области $\bar{X} \pm 6,6345s$, чему соответствует интервал $10 \pm 19,345$ с риском 0,01; 90% совокупности находятся в области $\bar{X} \pm 3,494s$ с риском 0,10, т. е. $10,00 \pm 189$. Следовательно, нет оснований рассматривать особь, укладывающуюся по размерам в эти пределы, как принадлежащую другому виду.

Однако такой размах оказывается столь большим, что сравнение близких видов практически невозможно, хотя во многих случаях такой подход целесообразен. Правда, для выборок объемом 50 значение коэффициента при s снижается до 1,916 ($\bar{X} \pm 1,916s$ при $P=90\%$, $\gamma=0,90$).

Допустим, что при описании «статистического экотипа» (биотипа) станет необходимым (наряду с оценками параметров распределения) определить верхний или нижний предел (или же тот и другой одновременно) так, чтобы 99% всех будущих индивидуумов находились в этих пределах. Например, если измерены зубы взрослых млекопитающих, то в таксономическом отношении будут весьма важны как верхний, так и нижний толерантные пределы зубных пропорций.

В приведенном примере можно с успехом применить критерий Манна — Уитни, хотя возможно, что при этом встанет вопрос о независимости выборок. Этот критерий предназначен для сравнения двух независимых выборок. В данном случае одна из них была бы выборка объемом $n_1=1$, а другая — основная коллекция. В следующем разделе этой главы рассматривается критерий Манна — Уитни в связи с применениями его к выборкам, одна из которых имеет объем $n=1$.

КРАЙНИЕ ЗНАЧЕНИЯ

Непараметрический критерий для отбрасывания крайних значений (или «больших ошибок») в условиях малых выборок

Предположим, что в результате эксперимента получена серия наблюдений со следующими значениями, которые расположены в порядке возрастания: 2, 7, 8, 10, 12, 13. Первое значе-

ние значительно дальше отстоит от следующего, т. е. дальше чем все смежные пары отстоят друг от друга. Спрашивается, можно ли рассматривать значение 2 как «большую ошибку» или же его следует считать принадлежащим к той же совокупности, к которой относятся остальные пять значений? Диксон и Мэсси рассмотрели возможный критерий для решения этого вопроса и составили таблицу соответствующих критических значений.

Обозначим наименьшее значение упорядоченной последовательности K различных величин через X_1 , а наибольшее через X_K . Если же вызывает сомнение наибольшее значение, обозначим его через X_1 , а наименьшее через X_K .

В нашем примере получим:

$$\begin{array}{cccccc} 2 & 7 & 8 & 10 & 12 & 13 \\ X_1 & X_2 & X_3 & X_4 & X_5 & X_K \end{array}$$

Нулевая гипотеза заключается в том, что значение 2 принадлежит той же совокупности, к которой относятся остальные пять значений. Для проверки этой гипотезы нужно вычислить отношение, которое зависит от объема выборки. Если вычисленное значение отношения превысит критическое, соответствующее уровню значимости $\alpha=0,05$ или $\alpha=0,01$, нулевая гипотеза отвергается в зависимости от того, какое значение α будет выбрано. В данном случае $K=6$ и отношение

$$\frac{X_2 - X_1}{X_K - X_1} = \frac{7 - 2}{13 - 2} = \frac{5}{11} = 0,455.$$

Критическое значение при $\alpha=0,05$ и $K=6$ равно 0,560, из чего следует, что нулевая гипотеза должна быть принята.

В качестве примера применения этого критерия можно привести сравнение содержания обломочного материала в одном образце песчаников, залегающих в Пасифике (Миссури), и в серии из 22 образцов тех же песчаников, но развитых в Ведроне (Иллинойс). Наибольшее значение этой величины в образцах песчаников Ведрона равно 85,1%, а в образце песчаников Пасифика — 89,6%. Все 23 значения, включая и образец из Пасифика, расположены в убывающем порядке с *наибольшим* X_1 и *наименьшим* X_K . Для случая $K=23$ отношение вычисляется по следующей формуле:

$$\frac{|X_3 - X_1|}{X_{K-2} - X_1} = \frac{|82,7 - 89,6|}{60,1 - 89,6} = -0,234.$$

Так как вычисленное значение не превышает допустимого, которое для 95%-ной квантили равно 0,421, то нулевую гипотезу следует принять как подтвердившуюся.

Однако мы не уверены в законности такого применения этого критерия, так как экстремальное значение было отмечено в образце, отобранном не в том же районе, где все остальные; следовательно, по сути дела этот образец представляет собой независимую выборку. В подобной ситуации, но в случае двух независимых выборок более подходящим критерием может служить критерий U Манна — Уитни, для которого $N_1 = 1$ и $N_2 = 22$. В данном случае величину U можно вычислить одним из двух методов, описанных в работе Сегала¹, так как значение N_2 велико. Мы применим один из этих методов, что не исключает возможности применения другого. Пусть

$$U = N_1 N_2 + \frac{N_1 (N_1 + 1)}{2} - R_1, \quad (6.20)$$

где R_1 — сумма рангов, приписанных членам выборки объемом N_1 . При этом ранг определяется как номер, приписанный всем членам обеих выборок в общей упорядоченной возрастающей последовательности.

Вернувшись к нашему примеру и расположив в возрастающем порядке выборочные данные образцов песчаников Ведрона и единичное значение образца песчаников Пасифика, которое равно 89,6, получим следующую последовательность: 58,0; 59,2; 60,1; 63,4; 64,7; 65,1; 65,6; 66,3; 66,3; 66,5; 66,6; 67,2; 69,6; 70,6; 72,7; 73,2; 75,4; 77,8; 77,9; 81,5; 82,7; 85,1; 89,6. В данном случае $N_1 = 1$ и, следовательно, соответствующее значение $R_1 = 23$. Тогда, согласно выражению (6.20), $U = 1 \cdot 22 + \frac{1 \cdot (1 + 1)}{2} - 23 = 0$.

По мере возрастания N_1 и N_2 выборочное распределение случайной величины U приближается к нормальному; это приближение становится хорошим уже при $N_2 > 20$. В данном случае среднее значение U будет близко к $a = N_1 N_2 / 2 = 22 / 2 = 11$, а стандартное отклонение приблизительно можно выразить как

$$s = \sqrt{\frac{N_1 N_2 (N_1 + N_2 + 1)}{12}} = \sqrt{\frac{1 \cdot 22 \cdot (1 + 22 + 1)}{12}} = 6,63.$$

Вычислив $Z = \frac{U - a}{s} = \frac{0 - 11}{6,63} = -1,66$, получим значение

случайной величины ξ , которая при условии правильности нулевой гипотезы должна быть распределена нормально с параметрами (0, 1). Согласно нулевой гипотезе, обе выборки взяты из совокупностей с одним и тем же средним при конкурирующей гипотезе H_2 , состоящей в том, что выборка объемом N_1 взята из совокупности с большим средним значением, чем выборка объемом N_2 . В условиях гипотезы H_2 вероятность собы-

тия $\xi \leq -1,66$ равна 0,0485. Для конкурирующей гипотезы H_2 требуется применение одностороннего критерия, а при уровне значимости $\alpha=0,05$ нулевая гипотеза должна быть отвергнута и следует принять гипотезу H_2 , так как вычисленное значение вероятности (0,0485) меньше 0,05. Таким образом, приняв альтернативную гипотезу, мы должны сделать следующий вывод: образец песчаников Пасифика взят из совокупности с существенно большим средним содержанием обломочного материала по сравнению с совокупностью, из которой взята группа образцов песчаников Ведрона. Вследствие сомнительной применимости критерия в предыдущем случае вывод был обратным.

ОДНОФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ**Модели I и II; постановка задачи**

Для введения в технику дисперсионного анализа, идущего рука об руку с задачей планирования экспериментов, рассмотрим две проблемы, которые представлены двумя коренным образом отличающимися в эмпирическом отношении моделями, но весьма сходными по характеру математических действий¹.

Модель I

Фабриканту, изготавливающему пружины на 14 машинах, поступили жалобы от потребителей на случаи снижения силы пружин. После изучения этого вопроса фабрикант приходит к выводу, что брак является следствием разладки некоторых машин. Возникает вопрос: каких машин из имеющихся 14? Для получения ответа на этот вопрос фабрикант отбирает из продукции каждой машины одинаковое число образцов и проверяет их годность, применяя соответствующие критерии.

Модель II

С другой стороны, рассмотрим коллекцию, состоящую из 14 образцов, отобранных с некоторого гранитного массива штата Нью-Йорк. Каждый образец представлен одним шлифом. Предполагая, что 14 образцов достаточно для изучения данной площади и что образцы являются представительными, мы можем поставить вопрос: всюду ли гранит однороден по своему составу? Условимся, что возможными ответами на этот вопрос могут быть только «да» или «нет». Для этого нужно каждый шлиф изучить по многочисленным линиям пересечений и затем полученные оценки минерального состава сравнить между со-

¹ Применяемый в этой главе подход к задаче похож на схему, используемую Эйзенхартом. (C. Eisenhart, The assumptions underlying the analysis of variance, *Biometric*, 3, 1947.

бой. Условимся также, что мы учитываем лишь следующие компоненты: кварц, несдвойникованный и сдвойникованный полевой шпат и биотит. Итак, мы хотим выяснить, имеются ли различия в минералогическом составе 14 изучаемых шлифов.

Различия между моделями

По характеру постановки вопроса обе приведенные задачи аналогичны. Однако между ними существует коренное отличие, с помощью которого в дальнейшем будут выделены два основных типа дисперсионного анализа. Рассмотрим сначала пример, связанный с изучением гранитного массива. Четырнадцать образцов, по которым будут получены оценки минерального состава, следует рассматривать как один вариант, выбранный из бесчисленного множества возможных групп того же объема; они могли быть отобраны из гранитного массива. Шлиф, изготовленный из образца, является одним из бесчисленного множества вариантов возможных его шлифов. Таким образом, эти шлифы можно рассматривать как случайную выборку из множества всех возможных шлифов. Пересечения, по которым будет подсчитываться минеральный состав, также следует рассматривать как наудачу выбранные из бесчисленного множества возможных пересечений.

В примере с пружинами машины, их производящие, конечно, не случайны. Их всего 14, и каждая производит равное число пружин. При данной выборочной схеме из продукции каждой из 14 машин отбирается одно и то же число образцов. Машины, таким образом, фиксированы и неслучайны, так как все элементы этой конечной совокупности машин использованы. Пружины, которые будут подвергнуты испытаниям, можно рассматривать как случайную выборку из очень большого числа пружин, произведенных машинами за определенный период времени.

ДИСПЕРСИОННЫЙ АНАЛИЗ ОДНОФАКТОРНЫЙ

Модель I

Результаты контрольного измерения силы пружин могут быть представлены так, как это сделано в табл. 7.1. В данном случае x_{ij} представляет собой результат испытания j -й пружины, произведенной i -й машиной. Таким образом, величину x_{ij} следует рассматривать как значения случайных величин, обладающих фиксированным средним, или математическим ожиданием, являющимся постоянным для каждой машины, т. е.

$$E(x_{ij}) = \mu. \quad (7.1) \quad 143$$

Запись результатов

Машины	Измерения						Сумма	Среднее
1	x_{11}	x_{12}	. . .	x_{1j}	. . .	x_{1c}	$x_{1.}$	$\bar{x}_{1.}$
2	x_{21}	x_{22}	. . .	x_{2j}	. . .	x_{2c}	$x_{2.}$	$\bar{x}_{2.}$
.	.	.		.		.	.	.
.	.	.		.		.	.	.
.	.	.		.		.	.	.
i	x_{i1}	x_{i2}	. . .	x_{ij}	. . .	x_{ic}	$x_{i.}$	$\bar{x}_{i.}$
.	.	.		.		.	.	.
.	.	.		.		.	.	.
.	.	.		.		.	.	.
r	x_{r1}	x_{r2}	. . .	x_{rj}	. . .	x_{rc}	$x_{r.}$	$\bar{x}_{r.}$

При фиксированном i

$$E(x_{ij}) = \mu_i,$$

и при всех i и j

$$\text{var}(x_{ij}) = \sigma^2.$$

Ниже будут использованы следующие обозначения:

$$\sum_{j=1}^c x_{ij} = x_{i.}, \quad (7.1a)$$

$$\frac{1}{c} \sum_{j=1}^c x_{ij} = \bar{x}_{i.}, \quad (7.1b)$$

$$\sum_{i=1}^r \sum_{j=1}^c x_{ij} = x_{..}, \quad (7.1c)$$

$$\frac{1}{rc} \sum_{i=1}^r \sum_{j=1}^c x_{ij} = \bar{x}_{..} = \bar{x}, \quad (7.1d)$$

$$\sum_{i=1}^r \mu_i = \mu_{.}, \quad (7.1e)$$

$$\frac{1}{r} \sum_{i=1}^r \mu_i = \bar{\mu}_{.} = \mu. \quad (7.1f)$$

Любая величина μ_i может быть представлена в виде следующего выражения:

$$\mu_i = \mu + (\mu_i - \mu). \quad (7.2)$$

Это значит, что истинное среднее значение силы растяжения пружин, произведенных i -й машиной, равно истинному значению генеральной средней μ плюс отклонение μ_i от μ . Уравнение (7.2) выражает условие аддитивности. В результате измерений никогда не будут получены точные значения μ_i , а будут лишь значения x_{ij} случайной величины, похожей на μ_i . Значения x_{ij} можно представить как

$$x_{ij} = \mu + (\mu_i - \mu) + \varepsilon_{ij}, \quad (7.3)$$

где ε_{ij} — значения случайной величины, математическое ожидание которой равно нулю. Эта величина представляет собой сумму погрешностей измерения и колебаний в качестве пружин, произведенных одной и той же машиной. Дисперсия этой величины σ^2 является мерой рассеяния значений силы растяжения пружин, произведенных одной машиной.

До сих пор мы еще не получили ответа на поставленный вопрос: какие машины производят брак? Но мы и не могли его получить. На основании всего вышесказанного можно суммировать результаты эксперимента так, как это сделано в табл. 7.1.

Теперь мы произведем некоторые алгебраические преобразования имеющихся данных. Это в первую очередь разложение суммы квадратов эмпирических данных.

Разложение сумм квадратов

Полная сумма квадратов в данном эксперименте определяется следующим образом:

$$\sum_i \sum_j (x_{ij} - \bar{x})^2;$$

это выражение может быть записано как

$$\sum_i \sum_j [(x_{ij} - \bar{x}_{i.}) + (\bar{x}_{i.} - \bar{x})]^2. \quad (7.4)$$

Разложив квадрат суммы в скобках, получим

$$\sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2 + \sum_i \sum_j (\bar{x}_{i.} - \bar{x})^2 + 2 \sum_i \sum_j (x_{ij} - \bar{x}_{i.})(\bar{x}_{i.} - \bar{x}).$$

Выражение

$$2 \sum_i \sum_j (x_{ij} - \bar{x}_{i.})(\bar{x}_{i.} - \bar{x})$$

при суммировании по j превращается в уравнение

$$2c \sum_i (\bar{x}_{i.} - \bar{x}_{i.})(\bar{x}_{i.} - \bar{x}) = 0.$$

Таким образом, формула (7.4) примет более краткий вид:

$$\sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2 + \sum_i \sum_j (\bar{x}_{i.} - \bar{x})^2.$$

В этом выражении первая сумма квадратов является мерой уклонения результатов наблюдения над продукцией от соответствующих каждой машине оценок средних значений $\bar{x}_i$, а вторая сумма¹ характеризует уклонения оценок средних $\bar{x}_i$ от оценки генеральной средней $\bar{x}$.

При работе удобно располагать материал в виде табл. 7.2, в которой перечисляются источники изменчивости результатов эксперимента и указываются соответствующие им суммы квадратов и выборочные дисперсии.

Т а б л и ц а 7.2

Разложение сумм квадратов; модель I

Тип дисперсии ¹	Сумма квадратов	Оценка дисперсии
Между машинами . . .	$\sum_i \sum_j (\bar{x}_i - \bar{x})^2 = S_m^2$	$S_m^2 / (r - 1)$
По отдельным машинам	$\sum_i \sum_j (x_{ij} - \bar{x}_i)^2 = S_r^2$	$S_r^2 / r (c - 1)$
Общая дисперсия . . .	$\sum_i \sum_j (x_{ij} - \bar{x})^2 = S_T^2$	$S_T^2 / (rc - 1)$

¹ Здесь и в табл. 7.3 термин *дисперсия* соответствует термину *изменчивость*. — Прим. ред.

Выше мы допустили, что

$$x_{ij} = \mu + (\mu_i - \mu) + \varepsilon_{ij},$$

где ε_{ij} — значение случайной величины, математическое ожидание которой равно нулю, а дисперсия — σ^2 , из чего следует, что x_{ij} — значение случайной величины, математическое ожидание которой равно μ_i . Если предположить, что каждой паре значений i и j соответствует одна и та же дисперсия и что все значения x_{ij} независимы, то получим, что

$$\text{var}(x_{ij}) = E(x_{ij} - \mu_i)^2 = \sigma^2. \quad (7.5)$$

Таким образом,

$$\bar{x}_i = \mu + (\mu_i - \mu) + \frac{1}{c} \sum_{j=1}^c \varepsilon_{ij},$$

где

$$E(\bar{x}_i) = \mu_i$$

¹ Вторая сумма равна просто $c \sum_i (\bar{x}_i - \bar{x})^2$. — Прим. ред.

и

$$\text{var}(\bar{x}_{i.}) = E(\bar{x}_{i.} - \mu_i)^2 = \frac{\sigma^2}{c}. \quad (7.6)$$

Наконец,

$$\bar{\bar{x}}_{..} = \bar{x} = \mu + \frac{1}{rc} \sum_i \sum_j \varepsilon_{ij},$$

где

$$E(\bar{x}) = \mu$$

и

$$\text{var} \bar{x} = E(\bar{x} - \mu)^2 = \frac{\sigma^2}{rc}. \quad (7.7)$$

Итак, несмещенными оценками неизвестных истинных средних μ_i (которые в нашем примере представляют собой неизвестные средние значения силы пружин, произведенных i -й машиной) и генеральной средней μ являются выборочные средние $\bar{x}_{i.}$ и $\bar{x}$ соответственно. Все вышесказанное позволяет суммировать наши данные в форме таблиц дисперсионного анализа (анализа изменчивости), а также получить несмещенные оценки неизвестных средних μ_i и, более того, их оценки любой линейной комбинации. Так, например, оценкой для $\mu_i - \mu_{i+1}$ будет $\bar{x}_i - \bar{x}_{i+1}$. Получив оценки для всех μ_i , мы пока еще не можем решить, есть ли между неизвестными нам значениями μ_i существенные различия. Этот вопрос будет освещен ниже, а сейчас сосредоточим внимание на получении оценки дисперсии σ^2 , характеризующей погрешности эксперимента. Было бы вполне естественным искать такую оценку, используя рассмотренные выше суммы квадратов.

Математические ожидания сумм квадратов

Рассмотрим

$$E \sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2. \quad (7.8)$$

Запишем это выражение в следующем виде:

$$E \sum_i \sum_j (x_{ij} - \mu_i - \bar{x}_{i.} + \mu_i)^2 = E \sum_i \sum_j [(x_{ij} - \mu_i) - (\bar{x}_{i.} - \mu_i)]^2.$$

Возведя в квадрат выражение, стоящее под знаком суммы в правой части, получим

$$\begin{aligned} E \sum_i \sum_j [(x_{ij} - \mu_i)^2 + (\bar{x}_{i.} - \mu_i)^2 - 2(x_{ij} - \mu_i)(\bar{x}_{i.} - \mu_i)] = \\ = \sum_i \sum_j E(x_{ij} - \mu_i)^2 + \sum_i \sum_j E(\bar{x}_{i.} - \mu_i)^2 - \\ - 2E \sum_i \sum_j (x_{ij} \bar{x}_{i.} - x_{ij} \mu_i - \bar{x}_{i.} \mu_i + \mu_i^2). \end{aligned}$$

Вычисляя математические ожидания, суммируя и используя результаты (7.5) и (7.6), получим

$$\begin{aligned}rc \operatorname{var} x_{ij} + rc \operatorname{var} \bar{x}_i - 2 \sum_i E(c\bar{x}_i^2 - c\bar{x}_i\mu_i - c\bar{x}_i\mu_i + c\mu_i^2) = \\= rc\sigma^2 + rc \frac{\sigma^2}{c} - 2c \sum_i E(\bar{x}_i - \mu_i)^2 = \\= rc\sigma^2 + r\sigma^2 - 2r\sigma^2 = r(c-1)\sigma^2. \quad (7.9)\end{aligned}$$

Следовательно,

$$E \sum_i \sum_j (x_{ij} - \bar{x}_i)^2 = r(c-1)\sigma^2,$$

откуда

$$\frac{1}{r(c-1)} E \sum_i \sum_j (x_{ij} - \bar{x}_i)^2 = \sigma^2.$$

Таким образом, сумма квадратов уклонений от оценок средних для продукции каждой отдельной машины, деленная на соответствующее *число степеней свободы*, является несмещенной оценкой дисперсии σ^2 . Величина σ^2 (и ее оценка) отражает степень неоднородности продукции, выпускаемой одной и той же машиной.

Теперь рассмотрим выражение

$$E \sum_i \sum_j (\bar{x}_i - \bar{x})^2,$$

которое запишем следующим образом:

$$\begin{aligned}E \sum_i \sum_j [(\bar{x}_i - \mu_i) - (\bar{x} - \mu) + (\mu_i - \mu)]^2 = \\= E \sum_i \sum_j [(\bar{x}_i - \mu_i)^2 + (\bar{x} - \mu)^2 + (\mu_i - \mu)^2 - \\- \overbrace{2(\bar{x}_i - \mu_i)(\bar{x} - \mu)}^A + \overbrace{2(\bar{x}_i - \mu_i)(\mu_i - \mu)}^B - \overbrace{2(\bar{x} - \mu)(\mu_i - \mu)}^C].\end{aligned} \quad (7.10)$$

Рассмотрим математические ожидания сумм смешанных произведений A , B и C , не принимая во внимание знаки и коэффициент 2.

$$A) E \sum_i \sum_j (\bar{x}_i - \mu_i)(\bar{x} - \mu).$$

Суммируя по i , получим

$$rE \sum_j (\bar{x} - \mu)^2.$$

Найдя математическое ожидание суммы, получим, что это выражение равно $r \operatorname{cvar} \bar{x}$, из чего следует, что

$$E \sum_i \sum_j (\bar{x}_{i.} - \mu_i)(\bar{x} - \mu) = \sigma^2.$$

$$\begin{aligned} B) \quad E \sum_i \sum_j (\bar{x}_{i.} - \mu_i)(\mu_i - \mu) &= \sum_i \sum_j E(\bar{x}_{i.} - \mu_i)(\mu_i - \mu) = \\ &= \sum_i \sum_j (\mu_i - \mu) E(\bar{x}_{i.} - \mu_i). \end{aligned}$$

Так как

$$E\bar{x}_{i.} = \mu_i,$$

то

$$E(\bar{x}_{i.} - \mu_i) = 0$$

и, следовательно,

$$E \sum_i \sum_j (\bar{x}_{i.} - \mu_i)(\mu_i - \mu) = 0.$$

$$C) \quad E \sum_i \sum_j (\bar{x} - \mu)(\mu_i - \mu).$$

Переписав это выражение в виде

$$\sum_i \sum_j (\mu_i - \mu) E(\bar{x} - \mu),$$

получим

$$\sum_i \sum_j (\mu_i - \mu) \cdot 0 = 0.$$

Первые три квадрата в выражении (7.10) имеют следующие математические ожидания:

$$E(\bar{x}_{i.} - \mu_i)^2 = \operatorname{var} \bar{x}_{i.} = \frac{\sigma^2}{c},$$

$$E(\bar{x} - \mu)^2 = \operatorname{var} \bar{x} = \frac{\sigma^2}{rc},$$

$$E(\mu_i - \mu)^2 = (\mu_i - \mu)^2.$$

Следовательно,

$$E \sum_{i=1}^r \sum_{j=1}^c (\bar{x}_{i.} - \mu_i)^2 = r\sigma^2,$$

$$E \sum_{i=1}^r \sum_{j=1}^c (\bar{x} - \mu)^2 = \sigma^2$$

и

$$E \sum_{i=1}^r \sum_{j=1}^c (\mu_i - \mu)^2 = c \sum_{i=1}^r (\mu_i - \mu)^2.$$

Используя полученные результаты, выражение (7.10) можно записать в виде

$$r\sigma^2 + \sigma^2 + c \sum_{i=1}^r (\mu_i - \mu)^2 - 2\sigma^2 = (r-1)\sigma^2 + c \sum_{i=1}^r (\mu_i - \mu)^2.$$

Следовательно,

$$E \left[\frac{\sum_{i=1}^r \sum_{j=1}^c (\bar{x}_{ij} - \bar{x})^2}{r-1} \right] = \sigma^2 + \frac{c}{r-1} \sum_{i=1}^r (\mu_i - \mu)^2 = \sigma^2 + \frac{c}{r-1} \sum_{i=1}^r \xi_i^2,$$

где

$$\xi_i = \mu_i - \mu.$$

Таким образом, табл. 7.2 может быть переписана в форме табл. 7.3, в которую включены числа степеней свободы оценок дисперсий наряду с их математическими ожиданиями.

Т а б л и ц а 7.3

Дисперсионный анализ; модель I

Тип дисперсии	Число степеней свободы	Сумма квадратов	Оценка дисперсий	Математическое ожидание ошибок
Между машинами	$r-1$	$\sum_i \sum_j (\bar{x}_{ij} - \bar{x})^2 = S_m^2$	$S_m^2/(r-1)$	$\sigma^2 + \frac{c}{r-1} \sum_{i=1}^r \xi_i^2$
По отдельным машинам	$r(c-1)$	$\sum_i \sum_j (x_{ij} - \bar{x}_{ij})^2 = S_r^2$	$S_r^2/r(c-1)$	σ^2
Общая дисперсия	$rc-1$	$\sum_i \sum_j (x_{ij} - \bar{x})^2 = S_T^2$	$S_T^2/(rc-1)$	—

УПРАЖНЕНИЕ. Показать, что $S_T^2/(rc-1)$ не является несмещенной оценкой для σ^2 .

Модель II

Вернувшись к примеру, в котором требуется решить вопрос о гомогенности гранита, мы можем свести эмпирические данные в таблицу, аналогичную табл. 7.1.

В данном случае имеющиеся r шлифов рассматриваются как случайная выборка объемом r из бесконечной совокупности возможных вариантов шлифов. Любой результат наблюдения x_{ij} рассматривается как

$$x_{ij} = \mu + (y_i - \mu) + \varepsilon_{ij},$$

где $(y_{i.} - \mu)$ — значения случайной величины, представляющей отклонения среднего состава i -го шлифа от генерального среднего μ состава данной породы, полученного по всем возможным вариантам шлифов.

Т а б л и ц а 7.4

Разложение сумм квадратов; модель II

Тип дисперсии	Сумма квадратов	Оценка дисперсий
Между шлифами	$\sum_i \sum_j (\bar{x}_{i.} - \bar{x})^2 = S_s^2$	$S_s^2/(r-1)$
По пересечениям в шлифах	$\sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2 = S_t^2$	$S_t^2/r(c-1)$
Общая дисперсия	$\sum_i \sum_j (x_{ij} - \bar{x})^2 = S_T^2$	$S_T^2/(rc-1)$

Если $E(y_{i.} - \mu) = 0$, то величина $\text{var}(y_{i.} - \mu) = \sigma_s^2$ является мерой неоднородности состава в различных шлифах. Величина ε_{ij} — значение ошибки эксперимента — представляет собой сумму ошибки отсчета и разницы в значениях подсчета в отдельных пересечениях в одном и том же шлифе. Если $E(\varepsilon_{ij}) = 0$, величина $\text{var} \varepsilon_{ij} = \sigma^2$ выражает неоднородность состава по данным различных пересечений в одном и том же шлифе.

Таким образом, для случайных величин x_{ij}

$$Ex_{ij} = \mu + 0 + 0 = \mu$$

и

$$\text{var } x_{ij} = \text{var } \mu + \text{var}(y_{i.} - \mu) + \text{var } \varepsilon_{ij} = \sigma_s^2 + \sigma^2,$$

$$\bar{x}_{i.} = \mu + (y_{i.} - \mu) + \bar{\varepsilon}_{i.},$$

$$\text{var } \bar{x}_{i.} = \sigma_s^2 + \frac{\sigma^2}{c},$$

$$\bar{\bar{x}}_{..} = \bar{x} = \mu + \frac{1}{r} \sum_i (y_{i.} - \mu) + \bar{\bar{\varepsilon}}_{..},$$

$$\text{var } \bar{x} = \frac{1}{r} \sigma_s^2 + \frac{1}{rc} \sigma^2.$$

Разложение общей суммы квадратов

Как и в случае модели I, общая сумма квадратов результатов эксперимента может быть определена как

$$\sum_{i=1}^r \sum_{j=1}^c (x_{ij} - \bar{x})^2 = \sum_{i=1}^r \sum_{j=1}^c (x_{ij} - \bar{x}_{i.})^2 + \sum_{i=1}^r \sum_{j=1}^c (\bar{x}_{i.} - \bar{x})^2.$$

Для удобства можно снова составить таблицу типов дисперсии и соответствующих им оценок по данным выборки (табл. 7.4).

Рассмотрим математические ожидания приведенных выше сумм квадратов и оценок дисперсий.

Как и в (7.9),

$$E \sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2 = r(c-1)\sigma^2,$$

так как

$$\begin{aligned} E \sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2 &= E \sum_i \sum_j [(x_{ij} - \mu) - (\bar{x}_{i.} - \mu)]^2 = \\ &= rc \operatorname{var} x_{ij} + rc \operatorname{var} \bar{x}_{i.} - 2E \sum_i \sum_j (x_{ij} \bar{x}_{i.} - \mu \bar{x}_{i.} + \mu^2 - x_{ij} \mu) = \\ &= rc \operatorname{var} x_{ij} + rc \operatorname{var} \bar{x}_{i.} - 2rc \operatorname{var} \bar{x}_{i.} = \\ &= rc \operatorname{var} x_{ij} - rc \operatorname{var} \bar{x}_{i.} = rc \left(\sigma_s^2 + \sigma^2 - \sigma_s^2 - \frac{\sigma^2}{c} \right). \end{aligned}$$

Следовательно,

$$\frac{1}{r(c-1)} E \sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2 = \sigma^2.$$

Далее,

$$\begin{aligned} E \sum_i \sum_j (\bar{x}_{i.} - \bar{x})^2 &= E \sum_i \sum_j [(\bar{x}_{i.} - \mu) - (\bar{x} - \mu)]^2 = \\ &= rc \operatorname{var} \bar{x}_{i.} + rc \operatorname{var} \bar{x} - 2rc \operatorname{var} \bar{x} = \\ &= rc \operatorname{var} \bar{x}_{i.} - rc \operatorname{var} \bar{x} = rc [\operatorname{var} \bar{x}_{i.} - \operatorname{var} \bar{x}] = \\ &= rc \left(\sigma_s^2 + \frac{\sigma^2}{c} - \frac{\sigma_s^2}{r} - \frac{\sigma^2}{rc} \right) = \\ &= rc \sigma_s^2 \left(\frac{r-1}{r} \right) + rc \frac{\sigma^2}{c} \left(\frac{r-1}{r} \right) = \\ &= c \sigma_s^2 (r-1) + (r-1) \sigma^2 = \sigma^2 (r-1) + c(r-1) \sigma_s^2. \end{aligned}$$

Таким образом,

$$\frac{1}{r-1} E \sum_i \sum_j (\bar{x}_{i.} - \bar{x})^2 = \sigma^2 + c \sigma_s^2.$$

Итак, таблица дисперсионного анализа может быть составлена и для случая модели II (табл. 7.5).

Т а б л и ц а 7.5

Дисперсионный анализ; модель II

Тип дисперсии	Число степеней свободы	Сумма квадратов	Оценка дисперсий	Математическое ожидание оценок
Между шлифами	$r - 1$	$\sum_i \sum_j (\bar{x}_{i.} - \bar{x})^2 = S_s^2$	$\frac{S_s^2}{r - 1}$	$\sigma^2 + c\sigma_s^2$
Между пересечениями в шлифах	$r(c - 1)$	$\sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2 = S_t^2$	$\frac{S_t^2}{r(c - 1)}$	σ^2
Общая дисперсия	$rc - 1$	$\sum_i \sum_j (x_{ij} - \bar{x})^2 = S_T^2$	$\frac{S_T^2}{rc - 1}$	—

ДВУХФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

К двухфакторному дисперсионному анализу обычно приводит такой эксперимент, результаты которого можно классифицировать по двум признакам. Этот вид дисперсионного анализа при некоторых дополнительных условиях может быть назван *дисперсионным анализом со вполне случайными блоками*. В качестве примера может служить приведенная в гл. I классификация осадочных пород по двум признакам — составу и зернистости. Результаты наблюдения над любым третьим признаком изучаемых осадочных пород (например, цветом образца в изломе, удельным весом, коэффициентом фильтрации и т. д.) могут быть выражены в таблице, клетки которой соответствуют значениям состава и зернистости. В качестве другого примера двухфакторной схемы можно привести случай, когда из различных рудных жил отобраны серии образцов, которые отданы в разные лаборатории на химический анализ. Полученные результаты могут быть классифицированы в соответствии с двумя признаками — лабораторией, выполнявшей анализ, и жилой, из которой взят образец.

Для обработки подобных данных требуется составить таблицу, аналогичную табл. 7.6. Пусть эта таблица содержит r строк и c столбцов, причем на пересечении одной строки и одного столбца расположен результат только одного наблюдения.

Введем следующие обозначения:

$$\sum_{i=1}^r x_{ij} = x_{.j}, \quad (7.11a)$$

$$\frac{1}{r} \sum_{i=1}^r x_{ij} = \frac{x_{.j}}{r} = \bar{x}_{.j}, \quad (7.11b)$$

$$\sum_{j=1}^c \bar{x}_{.j} = c\bar{x}, \quad (7.11c)$$

$$\sum_{i=1}^r \bar{x}_{i.} = r\bar{x}, \quad (7.11d)$$

$$\frac{1}{rc} \sum_{i=1}^r \sum_{j=1}^c x_{ij} = \bar{\bar{x}} = \bar{x}, \quad (7.11e)$$

$$\sum_{i=1}^r \mu_{ij} = \mu_{.j}, \quad (7.11f)$$

$$\frac{1}{r} \sum_{i=1}^r \mu_{ij} = \bar{\mu}_{.j}, \quad (7.11g)$$

$$\frac{1}{rc} \sum_{i=1}^r \sum_{j=1}^c \mu_{ij} = \mu, \quad (7.11h)$$

$$\frac{1}{c} \sum_{i=1}^r \sum_{j=1}^c \mu_{ij} = r\mu, \quad (7.11i)$$

$$\frac{1}{r} \sum_{i=1}^r \sum_{j=1}^c \mu_{ij} = c\mu. \quad (7.11j)$$

Модель I. Статистические выводы о систематических компонентах

Для подобного эксперимента возможно (так же, как и ранее) предложить одну из трех моделей: модель I, содержащую только фиксированные объекты; модель II, в которой исследуемые объекты выбираются наудачу из некоторого множества, и модель III, которая представляет собой смешанный вариант моделей I и II. Техника составления таблиц дисперсионного анализа одинакова для всех трех моделей. Каждая модель допускает выполнение следующих операций: сведение эмпирических данных в таблицу дисперсионного анализа, получение

применение методов проверки гипотез с выводами об изменчивости (табл. 7.6). Одни задачи остаются идентичными для всех моделей, тогда как другие могут решаться различными способами. Пример двухфакторной классификации будет способствовать выявлению упомянутых выше сходства и различия, а также ясному пониманию основных предположений, которые будут сделаны в последующем изложении дисперсионного анализа.

Таблица 7.6

Результаты наблюдений

№ строк	№ столбцов					Сумма	Среднее
	1	2	...	j	...	c	
1	x_{11}	x_{12}	...	x_{1j}	...	x_{1c}	$\bar{x}_{1.}$
2	x_{21}	x_{22}	...	x_{2j}	...	x_{2c}	$\bar{x}_{2.}$
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{ic}	$\bar{x}_{i.}$
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
r	x_{r1}	x_{r2}	...	x_{rj}	...	x_{rc}	$\bar{x}_{r.}$
Сумма	$x_{.1}$	$x_{.2}$	...	$x_{.j}$	...	$x_{.c}$	$x_{..}$
Среднее	$\bar{x}_{.1}$	$\bar{x}_{.2}$	...	$\bar{x}_{.j}$	...	$\bar{x}_{.c}$	$\bar{x}_{..}$

Рассмотрим одну из задач, упомянутых во введении к этому разделу, а именно измерение коэффициента фильтрации в сериях образцов из керна. Предположим, что каждый образец керна может быть классифицирован в соответствии с двумя признаками — с зернистостью и типом породы (по составу). Допустим также, что каждому элементу табл. 7.6 соответствует результат измерения коэффициента фильтрации. Так как число классов зернистости предполагается конечным и полученные наблюдения представляют все рассматриваемые типы пород и классов зернистости, будет справедливым считать имеющуюся классификацию результатов измерения коэффициента фильтрации фиксированной, где одному признаку соответствуют столбцы, а другому — строки. Таким образом, числа x_{ij} рассматриваются как значения случайных величин, математические ожидания которых зависят от положения в таблице. Если очень много раз повторять измерения x_{ij} , то полученные в результате

эмпирические распределения будут обладать средним, сколь угодно близким к μ_{ij} .

Кроме введенных предположений, мы допустим, что любое среднее значение находится в аддитивной зависимости от генеральной средней μ , от отклонения средних по строкам от μ (эффект строк) и от отклонения средних по столбцам от μ (эффект столбцов). Иначе говоря,

$$\mu_{ij} = \mu + \xi_i + \eta_j, \quad (7.12)$$

где

$$\mu = \frac{1}{rc} \sum_i \sum_j \mu_{ij},$$

$$\xi_i = \bar{\mu}_{.i} - \mu,$$

$$\eta_j = \bar{\mu}_{.j} - \mu.$$

Уравнение (7.12) называется *условием аддитивности*, без выполнения которого нельзя получить наиболее важные линейные комбинации несмещенных оценок, а также нельзя получить оценки дисперсии для x_{ij} .

Таким образом, мы можем составить табл. 7.7, похожую на табл. 7.6, но содержащую математические ожидания изучаемых случайных величин. Эмпирические значения x_{ij} следует рассматривать как суммарный результат действия на генеральную среднюю μ двух факторов — типа породы и зернистости. Кроме того, необходимо учесть еще и действие случайного фактора, которое выражается в добавлении к сумме значения ϵ_{ij} , для которого

$$E(\epsilon_{ij}) = 0,$$

$$\text{var } \epsilon_{ij} = \sigma^2.$$

В результате

$$x_{ij} = \mu + \xi_i + \eta_j + \epsilon_{ij},$$

и значение x_{ij} , состоящее из фиксированных компонент и случайной компоненты, следует рассматривать как значение случайной величины, без чего невозможно применение статистики для обоснования выводов.

Главные параметры распределения этой величины определяются как

$$E(x_{ij}) = \mu + \xi_i + \eta_j = \mu_{ij}, \quad (7.13)$$

$$\text{var } x_{ij} = E(x_{ij} - \mu_{ij})^2 = \sigma^2. \quad (7.14)$$

Таким образом, уравнение (7.14) показывает, что дисперсия всех величин x_{ij} предполагается одной и той же, т. е. совершенно не зависит ни от i , ни от j . Это условие обычно называют *предположением о равенстве дисперсий* или *условием гомоскедастичности*. В дальнейшем для получения несмещенных

Математические ожидания

	1 2 ... j ... c	Сумма	Среднее
1	$\mu_{11} \ \mu_{12} \ \dots \ \mu_{1j} \ \dots \ \mu_{1c}$	$\mu_{1.}$	$\bar{\mu}_{1.}$
2	$\mu_{21} \ \mu_{22} \ \dots \ \mu_{2j} \ \dots \ \mu_{2c}$	$\mu_{2.}$	$\bar{\mu}_{2.}$
.	.	.	.
.	.	.	.
.	.	.	.
i	$\mu_{i1} \ \mu_{i2} \ \dots \ \mu_{ij} \ \dots \ \mu_{ic}$	$\mu_{i.}$	$\bar{\mu}_{i.}$
.	.	.	.
.	.	.	.
.	.	.	.
r	$\mu_{r1} \ \mu_{r2} \ \dots \ \mu_{rj} \ \dots \ \mu_{rc}$	$\mu_{r.}$	$\bar{\mu}_{r.}$
Сумма	$\mu_{.1} \ \mu_{.2} \ \dots \ \mu_{.j} \ \dots \ \mu_{.c}$	$\mu_{..}$	
Среднее	$\bar{\mu}_{.1} \ \bar{\mu}_{.2} \ \dots \ \bar{\mu}_{.j} \ \dots \ \bar{\mu}_{.c}$		$\bar{\mu}_{..} = \mu$

оценок σ^2 нужно ввести предположение об отсутствии корреляции между рассматриваемыми случайными величинами, т. е.

$$E[(x_{ij} - \mu_{ij})(x_{kl} - \mu_{kl})] = 0$$

для всех $i \neq k$ и $j \neq l$.

Так как

$$x_{ij} = \mu + \xi_i + \eta_j + \varepsilon_{ij},$$

то

$$\bar{x}_{i.} = \mu + \xi_i + 0 + \bar{\varepsilon}_{i.},$$

где

$$\frac{1}{c} \sum_{j=1}^c (\bar{\mu}_{.j} - \mu) = 0 = \frac{1}{c} \sum_{j=1}^c \eta_j$$

и

$$\bar{\varepsilon}_{i.} = \frac{1}{c} \sum_{j=1}^c \varepsilon_{ij}.$$

Тогда

$$E \bar{x}_{i.} = \mu + \xi_i = \bar{\mu}_{i.}, \quad (7.15)$$

$$\text{var } \bar{x}_{i.} = E(\bar{x}_{i.} - \bar{\mu}_{i.})^2 = \text{var } \frac{1}{c} \sum_{j=1}^c x_{ij} = \frac{\sigma^2}{c}. \quad (7.16)$$

Точно так же

$$E \bar{x}_{.j} = \mu + \eta_j = \bar{\mu}_{.j}, \quad (7.17)$$

$$\text{var } \bar{x}_{.j} = E (\bar{x}_{.j} - \bar{\mu}_{.j})^2 = \frac{\sigma^2}{r}. \quad (7.18)$$

Наконец,

$$E \bar{x} = \mu \quad (7.19)$$

и

$$\text{var } \bar{x} = E (\bar{x} - \mu)^2 = \frac{1}{rc} \sigma^2. \quad (7.20)$$

Используя выражения (7.1a), (7.11b) и (7.11e), получим

$$\sum_{j=1}^c (x_{ij} - \bar{x}_{i.}) = 0, \quad (7.21)$$

$$\sum_{i=1}^r (x_{ij} - \bar{x}_{.j}) = 0, \quad (7.22)$$

$$\sum_i \sum_j (x_{ij} - \bar{x}) = 0. \quad (7.23)$$

Разложение сумм квадратов

Для данных, приведенных в табл. 7.6, имеется следующее алгебраическое соотношение:

$$x_{ij} - \bar{x} = (\bar{x}_{i.} - \bar{x}) + (\bar{x}_{.j} - \bar{x}) + (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}).$$

Возведем обе части равенства в квадрат:

$$\begin{aligned} (x_{ij} - \bar{x})^2 = & (\bar{x}_{i.} - \bar{x})^2 + (\bar{x}_{.j} - \bar{x})^2 + (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2 + \\ & + 2(\bar{x}_{i.} - \bar{x})(\bar{x}_{.j} - \bar{x}) + 2(\bar{x}_{i.} - \bar{x})(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}) + \\ & + 2(\bar{x}_{.j} - \bar{x})(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}). \end{aligned}$$

Производя сложение по i и j и используя формулы (7.21) и (7.22), получим следующие суммы квадратов:

$$\begin{aligned} \sum_{i=1}^r \sum_{j=1}^c (x_{ij} - \bar{x})^2 = & \sum_{i=1}^r (\bar{x}_{i.} - \bar{x})^2 + \sum_{j=1}^c (\bar{x}_{.j} - \bar{x})^2 + \\ & + \sum_{i=1}^r \sum_{j=1}^c (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2. \end{aligned}$$

Выражение

$$\sum_{i=1}^r \sum_{j=1}^c (x_{ij} - \bar{x})^2$$

является, как и раньше, общей суммой квадратов отклонений каждого из значений x_{ij} от выборочной генеральной средней $\bar{x}$.

тов в виде суммы трех слагаемых:

$$\sum_{i=1}^r (\bar{x}_{i.} - \bar{x})^2, \quad (7.24)$$

$$\sum_{j=1}^c (\bar{x}_{.j} - \bar{x})^2, \quad (7.25)$$

$$\sum_{i=1}^r \sum_{j=1}^c (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2. \quad (7.26)$$

Первое из них [уравнение (7.24)] представляет собой сумму квадратов отклонений выборочных средних в строках от выборочной генеральной средней. В рассматриваемом примере эта величина служит характеристикой отклонений средних значений коэффициента фильтрации для различных типов пород от среднего для всех образцов керна, вычисленного вне зависимости как от типов пород, так и от зернистости. Второе [уравнение (7.25)] отвечает аналогичной сумме квадратов отклонений выборочных средних в столбцах от оценки генеральной средней. Эта сумма в данном случае является мерой отклонения выборочных характеристик средних значений коэффициента фильтрации для различных групп зернистости от генеральной средней. И, наконец, третье [уравнение (7.26)] плохо поддается физической интерпретации. Часто его называют смешанной суммой квадратов. Возможно, что эту сумму квадратов удобнее рассматривать не в виде (7.26), а в терминах ее математического ожидания с учетом того, что она является оценкой.

Мы можем расположить полученные суммы квадратов, как это сделано в табл. 7.8, в которой каждому типу дисперсии соответствует сумма квадратов, число степеней свободы и оценка дисперсии.

Т а б л и ц а 7.8

Разложение сумм квадратов

Тип дисперсии	Сумма квадратов	Число степеней свободы	Оценка дисперсии
Между средними строк	$\sum_{i,j} (\bar{x}_{i.} - \bar{x})^2 = S_r^2$	$r - 1$	$S_r^2 / (r - 1)$
Между средними столбцов	$\sum_{i,j} (\bar{x}_{.j} - \bar{x})^2 = S_c^2$	$c - 1$	$S_c^2 / (c - 1)$
Остаточная дисперсия	$\sum_{i,j} (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2 = S_{rc}$	$(r - 1)(c - 1)$	$\frac{S_{rc}^2}{(r - 1)(c - 1)}$
Общая дисперсия	$\sum_{i,j} (x_{ij} - \bar{x})^2 = S_T^2$	$rc - 1$	$S_T^2 / (rc - 1)$

На основании развитых выше положений можно разработать способы систематизации эмпирических данных (табл. 7.6) и разложения общей суммы квадратов на составляющие (табл. 7.8). Естественно, что статистики x_{ij} , $\bar{x}_i$, $\bar{x}_j$ и $\bar{x}$ могут быть использованы для получения оценок параметров совокупности. Эти величины, согласно (7.13), (7.15), (7.17) и (7.19), можно рассматривать как оценки для μ_{ij} , μ_i , μ_j и μ . Мы можем оценить и линейные комбинации этих параметров. Так, например, оценкой для $\mu_j - \mu_{j+1}$ будет величина $\bar{x}_j - \bar{x}_{j+1}$.

Теперь мы снова приходим к задаче получения оценки для ошибки эксперимента σ^2 . Чтобы найти эту оценку, нужно рассмотреть данные табл. 7.8, в частности математические ожидания сумм квадратов.

Математические ожидания сумм квадратов и средних квадратов при двухфакторном анализе

Рассмотрим сначала математическое ожидание суммы квадратов по строкам. В данном случае

$$E \sum_{i,j} (\bar{x}_i - \bar{x})^2 = E \sum_{i,j} [(\bar{x}_i - \bar{\mu}_i) - (\bar{x} - \mu) + (\bar{\mu}_i - \mu)]^2. \quad (7.27)$$

Возведя в квадрат выражение в правой части равенства, получим

$$\begin{aligned} \sum_{i,j} E (\bar{x}_i - \bar{\mu}_i)^2 + \sum_{i,j} E (\bar{x} - \mu)^2 + \sum_{i,j} E (\bar{\mu}_i - \mu)^2 - \\ - 2 \sum_{i,j} E (\bar{x}_i - \bar{\mu}_i)(\bar{x} - \mu) + 2 \sum_{i,j} E (\bar{x}_i - \bar{\mu}_i)(\bar{\mu}_i - \mu) - \\ - 2 \sum_{i,j} E (\bar{x} - \mu)(\bar{\mu}_i - \mu). \end{aligned}$$

Рассмотрим в этом выражении три первых слагаемых. Ясно, что первое из них, согласно (7.16), есть дисперсия $\bar{x}_i$, а именно $\text{var } \bar{x}_i = \sigma^2/c$. Второе, согласно (7.20), есть $\text{var } \bar{x} = \sigma^2/rc$, тогда как третье, образованное константами, само является константой, т. е. $\sum_{i,j} \xi_i^2$. Таким образом, в результате суммирования по i и по j сумма трех первых слагаемых будет

$$rc \frac{\sigma^2}{c} + rc \frac{\sigma^2}{rc} + \sum_i \sum_j \xi_i^2.$$

Теперь рассмотрим два последних смешанных произведения, которые, как мы покажем, равны нулю. Действительно,

$$E (\bar{x}_i - \bar{\mu}_i)(\bar{\mu}_i - \mu) = (\bar{\mu}_i - \bar{\mu}_i)(\bar{\mu}_i - \mu),$$

но

$$\sum_{i,j} (\bar{\mu}_{i.} - \bar{\mu}_{i.})(\bar{\mu}_{i.} - \mu) = 0.$$

Далее,

$$\sum_{i,j} E(\bar{x} - \mu)(\bar{\mu}_{i.} - \mu) = \sum_{i,j} (\mu - \mu)(\bar{\mu}_{i.} - \mu) = 0.$$

Теперь рассмотрим $E \sum_{i,j} (\bar{x}_{i.} - \bar{\mu}_{i.})(\bar{x} - \mu)$. Перемножив скобки, запишем сумму в следующем виде:

$$\begin{aligned} E \left[\sum_{i,j} (\bar{x}_{i.}\bar{x} - \bar{x}_{i.}\mu - \bar{\mu}_{i.}\bar{x} + \bar{\mu}_{i.}\mu) \right] &= \\ &= E \left(\bar{x} \sum_{i,j} \bar{x}_{i.} - \mu \sum_{i,j} \bar{x}_{i.} - \bar{x} \sum_{i,j} \bar{\mu}_{i.} + \mu \sum_{i,j} \bar{\mu}_{i.} \right) = \\ &= rcE(\bar{x}^2 - 2\mu\bar{x} + \mu^2) = rcE(\bar{x} - \mu)^2 = rc \text{ var } \bar{x} = \sigma^2. \end{aligned}$$

Следовательно, формулу (7.27) можно записать как

$$r\sigma^2 + \sigma^2 - 2\sigma^2 + \sum_{i,j} \xi_i^2 = (r-1)\sigma^2 + \sum_{i,j} \xi_i^2$$

и

$$E \left[\sum_{i,j} (\bar{x}_{i.} - \bar{x})^2 \right] = (r-1)\sigma^2 + \sum_{i,j} \xi_i^2.$$

Точно таким же способом получим

$$E \left[\sum_{i,j} (\bar{x}_{.j} - \bar{x})^2 \right] = (c-1)\sigma^2 + \sum_{i,j} \eta_j^2. \quad (7.28)$$

Рассмотрим математическое ожидание остаточной суммы квадратов

$$\begin{aligned} E \sum_{i,j} (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2 &= \\ &= E \sum_{i,j} [(x_{ij} - \mu_{ij}) - (\bar{x}_{i.} - \bar{\mu}_{i.}) - (\bar{x}_{.j} - \bar{\mu}_{.j}) + \\ &\quad + (\bar{x} - \mu) + (\mu_{ij} - \bar{\mu}_{i.} - \bar{\mu}_{.j} + \mu)]^2. \quad (7.29) \end{aligned}$$

Это выражение получено путем одновременного добавления и вычитания величин μ_{ij} , $\bar{\mu}_{i.}$, $\bar{\mu}_{.j}$ и μ . Применяя условие аддитивности, т. е.

$$\mu_{ij} = \mu + (\bar{\mu}_{i.} - \mu) + (\bar{\mu}_{.j} - \mu),$$

и раскрыв скобки, получим для (7.29) выражение

$$\sum_{i,j} E(x_{ij} - \mu_{ij}) + \sum_{i,j} E(\bar{x}_{i.} - \bar{\mu}_{i.}) + \sum_{i,j} E(\bar{x}_{.j} - \bar{\mu}_{.j})^2 + \sum E(\bar{x} - \mu)^2$$

плюс ряд смешанных произведений, которые будут рассмотрены ниже. Используя формулы (7.14), (7.16), (7.18) и (7.20), полу-

чим математические ожидания сумм первых четырех квадратов, которые в сумме составят

$$rc\sigma^2 + rc\frac{\sigma^2}{c} + rc\frac{\sigma^2}{r} + rc\frac{\sigma^2}{rc} = \\ = (rc + r + c + 1)\sigma^2 = (r+1)(c+1)\sigma^2.$$

Рассмотрим теперь математические ожидания сумм шести смешанных произведений.

$$(1) \quad E \sum_{i,j} (x_{ij} - \mu_{ij})(\bar{x}_{i.} - \bar{\mu}_{i.}).$$

Суммируя по j , получим

$$c \sum_i E(\bar{x}_{i.} - \bar{\mu}_{i.})(\bar{x}_{i.} - \bar{\mu}_{i.}) = c \sum_i E(\bar{x}_{i.} - \bar{\mu}_{i.})^2 = \\ = cr \text{ var } \bar{x}_{i.} = cr \frac{\sigma^2}{c} = r\sigma^2.$$

$$(2) \quad E \sum_{i,j} (x_{ij} - \mu_{ij})(\bar{x}_{.j} - \bar{\mu}_{.j}) = r \sum_j E(\bar{x}_{.j} - \bar{\mu}_{.j})^2 = cr \text{ var } \bar{x}_{.j} = c\sigma^2.$$

$$(3) \quad E \sum_{i,j} (x_{ij} - \mu_{ij})(\bar{x} - \mu).$$

Суммируя по i , а затем по j , получим

$$rcE(\bar{x} - \mu)^2 = rc \text{ var } \bar{x} = \sigma^2. \\ (4) \quad E \sum_{i,j} (\bar{x}_{i.} - \bar{\mu}_{i.})(\bar{x}_{.j} - \bar{\mu}_{.j}).$$

После суммирования по i и j

$$Erc(\bar{x} - \mu)^2 = \sigma^2; \\ (5) \quad E \sum_{i,j} (\bar{x}_{i.} - \bar{\mu}_{i.})(\bar{x} - \mu) = rc \text{ var } \bar{x} = \sigma^2. \\ (6) \quad E \sum_{i,j} (\bar{x}_{.j} - \bar{\mu}_{.j})(\bar{x} - \mu) = rc \text{ var } \bar{x} = \sigma^2.$$

Следовательно,

$$E \sum_{i,j} (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2 = \\ = \sigma^2(rc + r + c + 1 - 2r - 2c + 2 + 2 - 2 - 2) = \sigma^2(r-1)(c-1).$$

Таким образом, если выполнено условие аддитивности¹, математические ожидания средних квадратов из табл. 7.8

¹ В случае невыполнения условия аддитивности

$$E \sum_{i,j} (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2 = \sigma^2(r-1)(c-1) + \sum_{i,j} (\mu_{ij} - \bar{\mu}_{i.} - \bar{\mu}_{.j} + \mu)^2.$$

будут:

$$\begin{aligned}\frac{1}{r-1} E \sum_{i,j} (\bar{x}_{i.} - \bar{x})^2 &= \sigma^2 + \frac{1}{r-1} \sum_{i,j} (\bar{\mu}_{i.} - \mu)^2, \\ \frac{1}{c-1} E \sum_{i,j} (\bar{x}_{.j} - \bar{x})^2 &= \sigma^2 + \frac{1}{c-1} \sum_{i,j} (\bar{\mu}_{.j} - \mu)^2, \\ \frac{1}{(r-1)(c-1)} E \sum_{i,j} (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2 &= \sigma^2.\end{aligned}$$

Эти результаты можно включить в табл. 7.8 (см. табл. 7.9)

Таблица 7.9

Двухфакторный дисперсионный анализ; модель I

Тип дисперсии	Сумма квадратов	Число степеней свободы	Оценка дисперсии	Математическое ожидание оценок
Между строками	$c \sum_i (\bar{x}_{i.} - \bar{x})^2 = S_r^2$	$r-1$	$S_r^2/(r-1)$	$\sigma^2 + \frac{\sum_{i,j} (\bar{\mu}_{i.} - \mu)^2}{r-1}$
Между столбцами	$r \sum_j (\bar{x}_{.j} - \bar{x})^2 = S_c^2$	$c-1$	$S_c^2/(c-1)$	$\sigma^2 + \frac{\sum_{i,j} (\bar{\mu}_{.j} - \mu)^2}{c-1}$
Остаточная дисперсия	$\sum_i \sum_j (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2 = S_{rc}^2$	$(r-1)(c-1)$	$\frac{S_{rc}^2}{(r-1)(c-1)}$	σ^2
Общая дисперсия	$\sum_i \sum_j (x_{ij} - \bar{x})^2 = S_T^2$	$rc-1$	$S_T^2/(rc-1)$	—

Модель II. Статистические выводы о компонентах дисперсии

Модель I двухфакторного дисперсионного анализа, как уже отмечалось, во многом похожа на модель II. В описываемой ниже модели II любой результат наблюдения x_{ij} представляет собой сумму константы μ и значений y_i , z_j и ε_{ij} случайных величин:

$$x_{ij} = \mu + y_i + z_j + \varepsilon_{ij}. \quad (7.30)$$

В качестве примера рассмотрим результаты измерения силы звуковых волн, проходящих через воду и обнаруживаемых с помощью приборов, расположенных на дне океана. Допустим, что звуковые волны и, следовательно, эффект их обнаружения зависят от случайных изменений как скорости ветра, так и скорости течения. В данном случае y_i можно рассматривать

в качестве эффекта случайного воздействия, обусловленного изменениями скорости ветра, на интенсивность обнаруженного звука, а z_j — как результат действия на эту интенсивность случайных флуктуаций течения. Таким образом, y_i и z_j рассматриваются как значения случайных величин, причем y_i представляет собой отклонение выборочного среднего значения интенсивности звука при i -й скорости течения от истинного среднего μ . Величина z_j есть отклонение оценки средней интенсивности звука при j -м значении скорости ветра от генеральной средней μ . По определению, математические ожидания как для y_i , так и для z_j равны 0, а их дисперсии $Ey_i^2 = \sigma_r^2$ и $Ez_j^2 = \sigma_c^2$. Дисперсия случайной величины ϵ_{ij} , равная σ^2 , характеризует неоднородность в столбцах по результатам строк. Формула (7.30) выражает известное *условие аддитивности*. В результате величина x_{ij} , образованная значениями случайных величин, может также рассматриваться как значение случайной переменной, математическое ожидание которой

$$Ex_{ij} = \mu,$$

а дисперсия

$$\text{var } x_{ij} = \text{var } y_i + \text{var } z_j + \text{var } \epsilon_{ij} = \sigma_r^2 + \sigma_c^2 + \sigma^2. \quad (7.31)$$

С помощью (7.30) и (7.31) можно получить следующие равенства:

$$(1) \quad \bar{x}_{i.} = \frac{1}{c} \sum_{j=1}^c x_{ij} = \mu + y_i + \frac{1}{c} \sum_{j=1}^c z_j + \frac{1}{c} \sum_{j=1}^c \epsilon_{ij},$$

откуда

$$E\bar{x}_{i.} = \mu$$

и

$$\text{var } \bar{x}_{i.} = \sigma_r^2 + \frac{\sigma_c^2}{c} + \frac{\sigma^2}{c}. \quad (7.32)$$

$$(2) \quad \bar{x}_{.j} = \frac{1}{r} \sum_{i=1}^r x_{ij} = \mu + \frac{1}{r} \sum_{i=1}^r y_i + z_j + \frac{1}{r} \sum_{i=1}^r \epsilon_{ij};$$

$$E\bar{x}_{.j} = \mu;$$

$$\text{var } \bar{x}_{.j} = \frac{\sigma_r^2}{r} + \sigma_c^2 + \frac{\sigma^2}{r}. \quad (7.33)$$

Наконец,

$$(3) \quad \bar{\bar{x}}_{..} = \bar{\bar{x}} = \frac{1}{rc} \sum_{i,j} x_{ij} = \\ = \mu + \frac{1}{rc} \sum_{i,j} (y_i - \mu) + \frac{1}{rc} \sum_{i,j} (z_j - \mu) + \frac{1}{rc} \sum_{i,j} \epsilon_{ij}.$$

Таким образом,

$$E(\bar{x}) = \mu, \\ \text{var } \bar{x} = \frac{\sigma_r^2}{r} + \frac{\sigma_c^2}{c} + \frac{\sigma_r^2}{rc}. \quad (7.34)$$

Разложение общей суммы квадратов

Как и раньше, определим общую сумму квадратов как

$$\sum_i \sum_j (x_{ij} - \bar{x})^2.$$

Эту сумму квадратов можно разложить на следующие три слагаемых:

$$\begin{aligned} & \sum_i \sum_j (\bar{x}_{i.} - \bar{x})^2, \\ & \sum_i \sum_j (\bar{x}_{.j} - \bar{x})^2, \\ & \sum_i \sum_j (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2. \end{aligned}$$

Математические ожидания сумм квадратов двухфакторного дисперсионного анализа; модель II

$$\begin{aligned} E \sum_{i,j} (\bar{x}_{i.} - \bar{x})^2 &= E \sum_{i,j} [(\bar{x}_{i.} - \mu) - (\bar{x} - \mu)]^2 = \\ &= E \sum_{i,j} (\bar{x}_{i.} - \mu)^2 + E \sum_{i,j} (\bar{x} - \mu)^2 - 2E \sum_{i,j} (\bar{x}_{i.} - \mu)(\bar{x} - \mu) = \\ &= rc \text{var } \bar{x}_{i.} + rc \text{var } \bar{x} - 2E \sum_{i,j} (\bar{x}_{i.} \bar{x} - \bar{x}_{i.} \mu - \mu \bar{x} + \mu^2) = \\ &= rc \text{var } \bar{x}_{i.} + rc \text{var } \bar{x} - 2E(\bar{x} - \mu)^2 = rc [\text{var } \bar{x}_{i.} - \text{var } \bar{x}] = \\ &= rc \left(\sigma_r^2 + \frac{\sigma_c^2}{c} + \frac{\sigma^2}{r} - \frac{\sigma_r^2}{r} - \frac{\sigma_c^2}{c} - \frac{\sigma^2}{rc} \right) = (r-1)(\sigma^2 + c\sigma_r^2). \quad (7.35) \end{aligned}$$

Аналогично

$$E \sum_{i,j} (\bar{x}_{.j} - \bar{x})^2 = (c-1)(\sigma^2 + r\sigma_c^2). \quad (7.36)$$

И наконец, рассмотрим выражение

$$E \sum_{i,j} (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2,$$

которое при одновременном добавлении и вычитании 2μ равно

$$E \sum_{i,j} [(x_{ij} - \mu) - (\bar{x}_{i.} - \mu) - (\bar{x}_{.j} - \mu) + (\bar{x} - \mu)]^2.$$

Возведя в квадрат выражение в квадратных скобках под знаком суммы, получим $rc \text{ var } x_{ij} + rc \text{ var } \bar{x}_{i.} + rc \text{ var } \bar{x}_{.j} + + rc \text{ var } \bar{x} +$ (шесть смешанных произведений).

Математические ожидания смешанных произведений

$$\begin{aligned} 1. E \sum_i \sum_j (x_{ij} - \mu)(\bar{x}_{i.} - \mu) &= E \sum_i \sum_j (x_{ij}\bar{x}_{i.} - x_{ij}\mu - \bar{x}_{i.}\mu + \mu^2) = \\ &= c \sum_i E(\bar{x}_{i.}^2 - 2\bar{x}_{i.}\mu + \mu^2) = cr \text{ var } \bar{x}_{i.}. \end{aligned} \quad (7.37)$$

$$\begin{aligned} 2. E \sum_i \sum_j (x_{ij} - \mu)(\bar{x}_{.j} - \mu) &= E \sum_i \sum_j (x_{ij}\bar{x}_{.j} - x_{ij}\mu - \mu\bar{x}_{.j} + \mu^2) = \\ &= r \sum_j E(\bar{x}_{.j}^2 - 2\bar{x}_{.j}\mu + \mu^2) = \\ &= cr \text{ var } \bar{x}_{.j}. \end{aligned} \quad (7.38)$$

$$\begin{aligned} 3. E \sum_i \sum_j (x_{ij} - \mu)(\bar{x} - \mu) &= E \sum_i \sum_j (x_{ij}\bar{x} - x_{ij}\mu - \mu\bar{x} + \mu^2) = \\ &= rcE(\bar{x}^2 - 2\bar{x}\mu + \mu^2) = \\ &= rc \text{ var } \bar{x}. \end{aligned} \quad (7.39)$$

$$\begin{aligned} 4. E \sum_i \sum_j (\bar{x}_{.j} - \mu)(\bar{x}_{i.} - \mu) &= E \sum_i \sum_j (\bar{x}_{.j}\bar{x}_{i.} - \bar{x}_{i.}\mu - \bar{x}_{.j}\mu + \mu^2) = \\ &= rcE(\bar{x}^2 - 2\mu\bar{x} + \mu^2) = \\ &= rc \text{ var } \bar{x}. \end{aligned} \quad (7.40)$$

$$\begin{aligned} 5. E \sum_i \sum_j (\bar{x}_{i.} - \mu)(\bar{x} - \mu) &= E \sum_i \sum_j (\bar{x}_{i.}\bar{x} - \bar{x}_{i.}\mu - \bar{x}\mu + \mu^2) = \\ &= rcE(\bar{x}^2 - 2\bar{x}\mu + \mu^2) = \\ &= rc \text{ var } \bar{x}. \end{aligned} \quad (7.41)$$

$$\begin{aligned} 6. E \sum_i \sum_j (\bar{x}_{.j} - \mu)(\bar{x} - \mu) &= E \sum_i \sum_j (\bar{x}_{.j}\bar{x} - \bar{x}_{.j}\mu - \mu\bar{x} + \mu^2) = \\ &= rcE(\bar{x}^2 - 2\bar{x}\mu + \mu^2) = \\ &= rc \text{ var } \bar{x}. \end{aligned} \quad (7.42)$$

Учитывая эти формулы, для математических ожиданий смешанных произведений получим

$$\begin{aligned} E \sum_{i,j} (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2 &= \\ &= rc(\text{var } x_{ij} + \text{var } \bar{x}_{i.} + \text{var } \bar{x}_{.j} + \text{var } \bar{x} - 2 \text{ var } \bar{x}_{i.} - 2 \text{ var } \bar{x}_{.j} + \\ &+ 2 \text{ var } \bar{x} + 2 \text{ var } \bar{x} - 2 \text{ var } \bar{x} - 2 \text{ var } \bar{x}) = \\ 166 \quad &= rc(\text{var } x_{ij} - \text{var } \bar{x}_{i.} - \text{var } \bar{x}_{.j} + \text{var } \bar{x}), \end{aligned}$$

что можно с помощью (7.31) — (7.34) записать в более кратком виде:

$$rc\sigma^2\left(1 - \frac{1}{c} - \frac{1}{r} - \frac{1}{rc}\right) = (r-1)(c-1)\sigma^2.$$

Разложение общей суммы квадратов с соответствующими значениями математических ожиданий приведено в табл. 7.10. Используя данные табл. 7.10, легко найти оценки для μ , σ_r^2 , σ_c^2 и σ^2 . Так, оценкой для μ будет $\bar{x}$, для σ_r^2 — величина

$$\frac{S_r^2}{c(r-1)} - \frac{S_{rc}^2}{c(r-1)(c-1)},$$

для σ_c^2

$$\frac{S_c^2}{r(c-1)} - \frac{S_{rc}^2}{r(r-1)(c-1)}$$

и для σ^2

$$\frac{S_{rc}^2}{(r-1)(c-1)}.$$

Т а б л и ц а 7.10

Двухфакторный дисперсионный анализ; модель II

Тип дисперсии	Сумма квадратов	Число степеней свободы	Оценка дисперсии	Математическое ожидание оценок
Между строками	$\sum_{i,j} (\bar{x}_{i.} - \bar{x})^2 = S_r^2$	$r-1$	$S_r^2/(r-1)$	$\sigma^2 + c\sigma_r^2$
Между столбцами	$\sum_{i,j} (\bar{x}_{.j} - \bar{x})^2 = S_c^2$	$c-1$	$S_c^2/(c-1)$	$\sigma^2 + r\sigma_c^2$
Остаточная дисперсия	$\sum_{i,j} (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2 = S_{rc}^2$	$(r-1)(c-1)$	$\frac{S_{rc}^2}{(r-1)(c-1)}$	σ^2
Общая дисперсия	$\sum_{i,j} (x_{ij} - \bar{x})^2 = S_T^2$	$rc-1$	$S_T^2/(rc-1)$	

Модель III. Смешанная

На практике нередко приходится рассматривать как *случайные*, так и *фиксированные компоненты эксперимента*.

В качестве примера рассмотрим коллекцию образцов, отобранных из разных рудных жил. Допустим, что нам нужно изучить содержание нерастворимого остатка в этих образцах, которые переданы *c* и только *c* лабораториям, проводящим такие исследования. В данном случае содержание нерастворимого

остатка x_{ij} (в процентах) может быть представлено в виде следующего выражения:

$$x_{ij} = \mu + (\bar{y}_{i.} - \mu) + (\bar{\mu}_{.j} - \mu) + \varepsilon_{ij} = \mu + a_i + \xi_j + \varepsilon_{ij}, \quad (7.43)$$

где μ — константа. Величина $(\bar{\mu}_{.j} - \mu) = \xi_j$ является также константой, выражающей величину расхождения между истинным значением среднего содержания нерастворимого остатка по данным j -й лаборатории и генеральной средней μ . Случайная величина $(\bar{y}_{i.} - \mu) = a_i$ характеризует случайные отклонения выборочного среднего значения $\bar{y}_{i.}$ содержания нерастворимого остатка в i -м образце по результатам определения всех лабораторий от истинной генеральной средней μ . В данном случае $Ea_i = 0$ и $\text{var } a_i = \sigma_r^2$. Как обычно, $E\xi_j = 0$ и $\text{var } \xi_j = \sigma^2$. Таким образом, $Ex_{ij} = \bar{\mu}_{.j}$ и

$$\text{var } x_{ij} = \sigma_r^2 + \sigma^2. \quad (7.44)$$

Разложение суммы квадратов

Так как $\bar{x}_{i.} = \mu + a_i + \varepsilon_{i.}$,
то

$$E\bar{x}_{i.} = \mu$$

и

$$\text{var } \bar{x}_{i.} = \sigma_r^2 + \frac{\sigma^2}{c}, \quad (7.45)$$

и так как

$$\bar{x}_{.j} = \mu + \frac{1}{r} \sum_{i=1}^r a_i + \xi_j + \bar{\varepsilon}_{.j},$$

то

$$E\bar{x}_{.j} = \mu + \xi_j = \bar{\mu}_{.j}$$

и

$$\text{var } \bar{x}_{.j} = \frac{\sigma_r^2}{r} + \frac{\sigma^2}{r}. \quad (7.46)$$

Наконец, так как

$$\bar{x} = \bar{x}_{..} = \mu + \frac{1}{r} \sum_{i=1}^r a_i + \bar{\varepsilon}_{..},$$

то

$$E\bar{x} = \mu$$

и

$$\text{var } \bar{x} = \frac{\sigma_r^2}{r} + \frac{\sigma^2}{rc}. \quad (7.47)$$

Общая сумма квадратов, определенная как

$$\sum_{i=1}^r \sum_{j=1}^c (x_{ij} - \bar{x})^2,$$

как и в предыдущем рассмотрении, равна

$$\sum_i \sum_j (\bar{x}_{.j} - \bar{x})^2 + \sum_i \sum_j (\bar{x}_{i.} - \bar{x})^2 + \sum_i \sum_j (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2.$$

Первая сумма квадратов характеризует в этом примере ту долю общей изменчивости, которая обусловлена различиями в определении содержания нерастворимого остатка отдельными лабораториями. Вторая сумма квадратов выражает изменчивость, вызванную различиями в содержании нерастворимого остатка в разных образцах. Последняя сумма квадратов, или остаточный член, характеризует погрешности определения содержаний нерастворимого остатка в лабораториях.

Математические ожидания сумм квадратов

$$\text{A.} \quad E \sum_{i=1}^r \sum_{j=1}^c (\bar{x}_{i.} - \bar{x})^2 = E \sum_i \sum_j [(\bar{x}_{i.} - \mu) - (\bar{x} - \mu)]^2.$$

Используя результаты (7.45) и (7.47), получим

$$\begin{aligned} rc \operatorname{var} \bar{x}_{i.} + rc \operatorname{var} \bar{x} - 2 \sum_i \sum_j E (\bar{x}_{i.} \bar{x} - \mu \bar{x}_{i.} - \mu \bar{x} + \mu^2) &= \\ &= rc (\operatorname{var} \bar{x}_{i.} - \operatorname{var} \bar{x}) = rc \left(\sigma_r^2 + \frac{\sigma^2}{c} - \frac{\sigma_r^2}{r} - \frac{\sigma^2}{rc} \right) = \\ &= rc \frac{r-1}{r} \left(\sigma_r^2 + \frac{\sigma^2}{c} \right) = (r-1)(\sigma^2 + c\sigma_r^2). \end{aligned} \quad (7.48)$$

$$\text{B.} \quad E \sum_i \sum_j (\bar{x}_{.j} - \bar{x})^2 = E \sum_i \sum_j [(\bar{x}_{.j} - \bar{\mu}_{.j}) - (\bar{x} - \mu) + (\bar{\mu}_{.j} - \mu)]^2.$$

С помощью формул (7.46) и (7.47) это выражение можно представить как

$$\begin{aligned} rc \operatorname{var} \bar{x}_{.j} + rc \operatorname{var} \bar{x} + \sum_i \sum_j (\bar{\mu}_{.j} - \mu)^2 + \\ + E \sum_i \sum_j (\text{трех смешанных произведений}). \end{aligned}$$

Рассмотрим, не принимая во внимание знак, три следующих смешанных произведения:

$$1. \quad E \sum_i \sum_j (\bar{x}_{.j} \bar{x} - \mu \bar{x}_{.j} - \bar{\mu}_{.j} \bar{x} + \mu \bar{\mu}_{.j}).$$

Вначале просуммируем по j . Затем, получив математические ожидания и проведя суммирование по i , найдем

$$\begin{aligned} &rc \operatorname{var} \bar{x}. \\ 2. \quad &E \sum_i \sum_j (\bar{x}_{.j} - \bar{\mu}_{.j})(\bar{\mu}_{.j} - \mu). \end{aligned}$$

Записав это выражение как

$$\sum_i \sum_j (\bar{\mu}_{.j} - \mu) E(\bar{x}_{.j} - \mu_{.j})$$

и получив математическое ожидание, найдем, что оно равно 0. Наконец,

$$3. E \sum_i \sum_j (\bar{x} - \mu)(\bar{\mu}_{.j} - \mu) = \sum_i \sum_j (\bar{\mu}_{.j} - \mu) E(\bar{x} - \mu) = 0.$$

Таким образом,

$$E \sum_i \sum_j (\bar{x}_{.j} - \bar{x})^2 = rc \text{ var } \bar{x}_{.j} + rc \text{ var } \bar{x} + \\ + \sum_i \sum_j (\bar{\mu}_{.j} - \mu)^2 - 2rc \text{ var } \bar{x}.$$

На основе (7.46) и (7.47) это выражение можно представить в следующем виде:

$$rc \left(\frac{\sigma_r^2}{r} + \frac{\sigma^2}{r} - \frac{\sigma_r^2}{r} + \frac{\sigma^2}{rc} \right) + \sum_i \sum_j (\bar{\mu}_{.j} - \mu)^2 = \\ = (c-1) \sigma^2 + \sum_i \sum_j (\bar{\mu}_{.j} - \mu)^2. \quad (7.49)$$

$$\text{С.} \quad E \sum_i \sum_j (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2 = \\ = E \sum_i \sum_j [(x_{ij} - \bar{\mu}_{.j}) - (\bar{x}_{i.} - \mu) - (\bar{x}_{.j} - \bar{\mu}_{.j}) + (\bar{x} - \mu)]^2 = \\ = rc [\text{var } x_{ij} + \text{var } \bar{x}_{i.} + \text{var } \bar{x}_{.j} + \text{var } \bar{x}] + \\ + E \sum_i \sum_j (\text{шести смешанных произведений}).$$

$$1. \quad E \sum_i \sum_j (x_{ij} - \bar{\mu}_{.j})(\bar{x}_{i.} - \mu).$$

Это выражение после суммирования по j , определения математических ожиданий и сложения по i примет следующий вид:

$$rc \text{ var } \bar{x}_{i.}.$$

$$2. \quad E \sum_i \sum_j (x_{ij} - \bar{\mu}_{.j})(\bar{x}_{.j} - \bar{\mu}_{.j}).$$

Это выражение после суммирования по i , определения математических ожиданий и сложения по j будет равно

$$rc \text{ var } \bar{x}_{.j}.$$

$$3. \quad E \sum_i \sum_j (x_{ij} - \bar{\mu}_{.j})(\bar{x} - \mu) = rc \text{ var } \bar{x}.$$

Этот результат будет получен при последовательном суммировании по i и j и вычислении математических ожиданий.

$$4. \quad E \sum_i \sum_j (x_{ij} - \bar{\mu}_{.j})(\bar{x}_{.j} - \bar{\mu}_{.j}) = rc \operatorname{var} \bar{x}.$$

$$5. \quad E \sum_i \sum_j (\bar{x}_{i.} - \mu)(\bar{x} - \mu) = rc \operatorname{var} \bar{x}.$$

$$6. \quad E \sum_i \sum_j (\bar{x}_{.j} - \bar{\mu}_{.j})(\bar{x} - \mu) = rc \operatorname{var} \bar{x}.$$

В результате выражения (7.49) можно представить как

$$\begin{aligned} &rc(\operatorname{var} x_{ij} + \operatorname{var} \bar{x}_{i.} + \operatorname{var} \bar{x}_{.j} + \operatorname{var} \bar{x} - 2 \operatorname{var} \bar{x}_{i.} - 2 \operatorname{var} \bar{x}_{.j} + \\ &+ 2 \operatorname{var} \bar{x} + 2 \operatorname{var} \bar{x} - 2 \operatorname{var} \bar{x} - 2 \operatorname{var} \bar{x}) = \\ &= rc\sigma^2 \left(1 - \frac{1}{c} - \frac{1}{r} - \frac{1}{rc}\right) = (r-1)(c-1)\sigma^2. \quad (7.50) \end{aligned}$$

Суммы квадратов, типы дисперсий, выборочные средние квадраты и соответствующие математические ожидания приведены в табл. 7.11.

Т а б л и ц а 7.11

Тип дисперсии	Число степеней свободы	Сумма квадратов	Оценка дисперсии	Дисперсия
Между строками	$r-1$	$\sum_{i,j} (\bar{x}_{i.} - \bar{x})^2 = S_r^2$	$S_r^2/(r-1)$	$\sigma^2 + c\sigma_r^2$
Между столбцами	$c-1$	$\sum_{i,j} (\bar{x}_{.j} - \bar{x})^2 = S_c^2$	$S_c^2/(c-1)$	$\sigma^2 + \frac{\sum_{i,j} (\bar{\mu}_{.j} - \mu)^2}{c-1}$
Остаточная дисперсия	$(r-1)(c-1)$	$\sum_{i,j} (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2 = S_e^2$	$\frac{S_e^2}{(r-1)(c-1)}$	σ^2
Общая дисперсия	$rc-1$	$\sum_{i,j} (x_{ij} - \bar{x})^2 = S_T^2$	$S_T^2/(rc-1)$	

ПРОВЕРКА ГИПОТЕЗ С ПОМОЩЬЮ ДИСПЕРСИОННОГО АНАЛИЗА

Получение оценок параметров и обобщение эмпирических данных в виде таблиц дисперсионного анализа — лишь малая доля того, что может дать исследователю техника дисперсионного анализа. Статистические выводы, основанные на проверке гипотез, являются наиболее важным этапом статистического анализа, который может быть осуществлен с помощью матема-

тических моделей, изложенных в предыдущих разделах, где рассматривалась техника обобщения данных и получения оценок параметров при дисперсионном анализе. Исследование математических моделей позволит выделить те гипотезы, которые могут быть легко проверены и изучены. Рассмотрим систематические, или неслучайные, факторы, соответствующие ранее упомянутым планам эксперимента. Помимо генеральной средней μ , к ним относятся, например, такие величины, как $(\mu_i - \mu)$, ξ_i и η_j .

Рассматривая модели, содержащие такие факторы, мы имеем целью определить их значение в степени изменчивости, присущей изучаемой величине. Таким образом, в каждой из рассмотренных математических моделей изучаемая случайная величина разлагается на составляющие ее компоненты. Вопрос ставится следующим образом: оказывают ли систематические (неслучайные) факторы эффективное воздействие на изменчивость полученных результатов наблюдений? Проверяемая гипотеза могла бы быть сформулирована в терминах значений этих факторов: например, равны ли все они нулю, равны ли они все между собой и т. д.? В случае изучения эффективности действия систематических факторов рассматриваемая нулевая гипотеза формулируется в следующем виде: $H_1: \mu_1 = \mu_2 = \dots = \mu_r$ при альтернативе H_2 , заключающейся в том, что, хотя бы для одной пары μ_i и μ_j при $i \neq j$, имеет место неравенство $\mu_i \neq \mu_j$. Гипотезы стандартного параметрического типа рассмотрены в гл. 5 и 6, где было показано, что при проверке любой гипотезы нужно располагать набором заранее известных фактов; кроме того, следует предположить наличие некоторых свойств критерия. Например, выбираем подходящую статистику для проверки данной гипотезы, после того как задано значение уровня значимости α . Затем с целью выбора некоторого критерия, или областей принятия и отклонения гипотезы, рассматриваем множество всех возможных исходов эксперимента и на основе данного рассмотрения выбираем решающее правило. Для этого необходимо знать распределения выбранного статистического критерия при условии, что проверяемая нулевая гипотеза верна. Однако, чтобы определить распределение статистики (задающей критерий), мы должны располагать некоторыми сведениями о распределениях изучаемых случайных величин, значения которых наблюдаются в процессе эксперимента; по этим величинам, собственно, и вычисляется значение критерия. Если распределение рассматриваемых случайных величин известно, то по нему можно вычислить распределение критерия и определить области принятия и отклонения проверяемой гипотезы. И только после того, как все это будет сделано, можно приступить непосредственно к проверке гипотезы.

Таким образом, и в задаче проверки гипотезы $\mu_1 = \mu_2 = \dots = \mu_r$ мы должны двигаться вдоль намеченной выше линии, начав с перечисления известных факторов и принимаемых допущений. Очевидно, что нужно или ввести предположение о виде распределения изучаемой случайной величины x , или каким-то другим способом определить его и с его помощью получить распределение критерия, каким бы оно ни было. Итак, эвристически объяснена необходимость последнего допущения, чтобы процедура дисперсионного анализа носила законченный характер. Полезно отметить еще раз, что предположение о форме распределения изучаемых величин не играет существенной роли в той стадии дисперсионного анализа, когда производится простое обобщение данных и получение оценок параметров. До сих пор техника дисперсионного анализа развивалась без учета каких-либо данных о распределении изучаемой величины — факт, который часто остается незамеченным в статистических приложениях. В подобных случаях сделанные допущения ограничиваются предположениями о случайности, аддитивности и однородности дисперсий. Вопрос о специфике форм распределения случайной величины рассмотрен в следующих разделах.

Проверка гипотез

Фиксированные компоненты

В качестве примера рассмотрим модель I однофакторного дисперсионного анализа. В данном случае представляющая интерес гипотеза была бы $H_1: \mu_1 = \mu_2 = \dots = \mu_r$. Естественна постановка вопроса: какая величина будет служить статистическим критерием для проверки этой гипотезы? Изучение табл. 7.3 поможет ответить на этот вопрос. Если верна нулевая гипотеза, т. е. все 14 машин производят пружины одного и того же качества, то $\sum_{i=1}^{14} (\mu_i - \mu)^2 = \sum_{i=1}^r \xi_i^2$ будет равна нулю, а математическое

ожидание среднего квадрата между машинами будет равно математическому ожиданию среднего квадрата для пружин, производимых одной и той же машиной, а именно σ^2 . Это значит, что если гипотеза H_1 верна, то

$$E\left(\frac{S_m^2}{r-1}\right) = E\left[\frac{S_r^2}{r(c-1)}\right] = \sigma^2.$$

Следовательно, отношение

$$\frac{S_m^2/(r-1)}{S_r^2/r(c-1)}$$

должно быть в среднем близко к 1. Статистический критерий такого типа был бы удовлетворительным, если бы было известно его распределение, при условии, что гипотеза H_1 верна. Чтобы выяснить распределение этого отношения, добавим к ранее сформулированным допущениям об аддитивности, случайности и однородности дисперсий предположение о том, что случайная величина ε_{ij} распределена нормально с математическим ожиданием $E(\varepsilon_{ij}) = 0$ и дисперсией $\text{var } \varepsilon_{ij} = \sigma^2$. Из этого следует, что x_{ij} также нормально распределены со средним μ и дисперсией σ^2 . Напомним, что в условиях сделанного предположения и числитель и знаменатель критерия будут распределены как величина $\sigma^2 \chi^2/n$ с n степенями свободы (гл. 4), а отношение двух таким образом распределенных случайных величин будет распределено по закону Фишера (F -распределение) с n_1 и n_2 степенями свободы. Теперь, зная распределение этого дисперсионного отношения, можно выбрать те области, в которых будет принята или отклонена гипотеза, и, следовательно, проверить ее в нашей задаче.

Случайные компоненты

Рассмотрим модель II, содержащую случайные компоненты, построенную для однофакторного дисперсионного анализа, в которой представляют интерес не средние значения компонент, а скорее их дисперсии. Дисперсия случайной компоненты это параметр, характеризующий ее вклад в изменчивость случайной величины x , значения которой наблюдаются в результате эксперимента. В данном случае вопрос ставится так: существует ли (т. е. будет ли отличной от нуля) дисперсия рассматриваемых компонент? Так, например, однородность состава шлифов, взятых из гранита (табл. 7.5), будет отражать гипотеза $H_1: \sigma_s^2 = 0$. Табл. 7.5 облегчает выбор статистики для проверки этой гипотезы. Если бы нулевая гипотеза была верна, то математическое ожидание среднего квадрата внутри шлифов было бы равным математическому ожиданию среднего квадрата различий между шлифами. Тогда отношение

$$\frac{S_s^2/(r-1)}{S_i^2/r(c-1)} \quad (7.51)$$

в среднем стало бы близким к 1. Если допустить, что ε_{ij} и $(\bar{y}_i - \mu)$ распределены нормально с математическим ожиданием, равным 0, и дисперсиями σ^2 и σ_s^2 соответственно, то x_{ij} также будет нормально распределенной величиной со средним μ и дисперсией $\sigma_s^2 + \sigma^2$. Следовательно, величина, определяемая выражением (7.51), будет распределена по закону Фишера с $(r-1)$ и $r(c-1)$ степенями свободы. Область принятия нуле-

вой гипотезы и область ее отклонения могут быть теперь установлены, и гипотеза может быть проверена.

Эту процедуру проверки можно распространить на другие математические модели, рассматриваемые в данной книге, а также на некоторые из них, здесь не рассматриваемые. В каждом случае статистика, с помощью которой проверяется гипотеза, может быть получена на основе математических моделей, которые описывают данную ситуацию; далее, предположение о виде распределения случайных компонент позволяет считать распределение изучаемых случайных величин известным, что в свою очередь делает возможным определить распределение статистики. При рассмотрении как случайных факторов, так и систематических статистика, используемая в качестве критерия, будет распределена по закону Фишера. Процедуры дисперсионного анализа для всех задач и моделей, рассмотренных в этой главе, обобщены в табл. 7.12—7.16.

ВЫЯВЛЕНИЕ ЗНАЧИМЫХ РАЗЛИЧИЙ

Если при проверке гипотезы о равенстве более двух средних она отвергается, возникает задача: найти различающиеся между собой средние. При процедуре дисперсионного анализа выявляется лишь факт существования различий, но мы не получаем ответа на вопрос, какие именно средние различаются. Единой процедуры для определения существенных различий вообще нет. Ниже изложены два метода, которые представляются вполне удовлетворительными и легко применимыми для решения поставленной задачи. Это так называемый *T*-метод Тьюки, или *метод множественного сравнения*, и *S*-метод Шеффэ, называемый *методом отыскания всех контрастов*.

T-метод множественного сравнения¹

Описываемый ниже метод основан на рассмотрении верхних процентных точек $q_{f, n}$ студентизированного размаха, или *q*-распределения. В данном случае это позволяет при построении метода сравнения установить доверительные пределы. Так, если s^2 представляет собой имеющую f степеней свободы оценку дисперсии, обусловленной ошибками эксперимента, и если истинное значение одного из всех $[n(n-1)]/2$ возможных различий есть $\mu_i - \mu_j$ и $i \neq j$, $i, j = 1, 2, \dots, n$, то

$$P\{a_{ij} - q_{f, n}ks \leq \mu_i - \mu_j \leq a_{ij} + q_{f, n}ks\} = 1 - \alpha,$$

¹ H. Scheffé, A method for judging all contrasts in the analysis of variance, *Biometrika*, 40, 1953; кроме того, см. русский перевод: Г. Шеффэ, Дисперсионный анализ, Физматгиз, М., 1963,

Таблица 7.12

Полный однофакторный дисперсионный анализ; модель I

Тип дисперсии	Число степеней свободы	Сумма квадратов	Средний квадрат	Математическое ожидание среднего квадрата	Проверяемая гипотеза	Статистика
Между машинами	$r - 1$	$\sum_i \sum_j (\bar{x}_{i.} - \bar{x})^2 = S_m^2$	$S_m^2 / (r - 1)$	$\sigma^2 + \frac{c}{r-1} \sum_{i=1}^r \xi_i^2$	$\mu_1 = \mu_2 = \dots = \mu_r$ $\left(\sum_{i=1}^r \xi_i^2 = 0 \right)$	$\frac{S_m^2 / (r - 1)}{S_T^2 / r (c - 1)}$
По отдельным машинам	$r(c - 1)$	$\sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2 = S_r^2$	$S_r^2 / r(c - 1)$	σ^2	—	—
Общая дисперсия	$rc - 1$	$\sum_i \sum_j (x_{ij} - \bar{x})^2 = S_T^2$	$S_T^2 / (rc - 1)$		—	—

Модель: $X = x_{ij} = \mu + (\mu_i - \mu) + \varepsilon_{ij}$.

Предполагаемое распределение: ε_{ij} распределена как $N(0, \sigma^2)$.

Распределение X : $N(\mu, \sigma^2)$.

Таблица 7.13

Полный однофакторный дисперсионный анализ; модель II

Тип дисперсии	Число степеней свободы	Сумма квадратов	Средний квадрат	Математическое ожидание среднего квадрата	Проверяемая гипотеза	Статистика
Между шлифами	$r - 1$	$\sum_i \sum_j (\bar{x}_{i.} - \bar{x})^2 = S_s^2$	$S_s^2 / (r - 1)$	$\sigma^2 + c\sigma_s^2$	$\sigma_s^2 = 0$	$\frac{S_s^2 / (r - 1)}{S_t^2 / r (c - 1)}$
Между пересечениями в шлифах	$r(c - 1)$	$\sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2 = S_t^2$	$S_t^2 / r(c - 1)$	σ^2	—	—
Общая дисперсия	$rc - 1$	$\sum_i \sum_j (x_{ij} - \bar{x})^2 = S_T^2$	$S_T^2 / (rc - 1)$	—	—	—

Модель: $X = x_{ij} = \mu + (\bar{y}_i - \mu) + e_{ij}$.

Предполагаемые распределения: e_{ij} распределена как $N(0, \sigma^2)$,

$(\bar{y}_i - \mu)$ — как $N(0, \sigma_s^2)$.

Распределение X : $N(\mu, \sigma_s^2 + \sigma^2)$.

Таблица 7.14

Полный двухфакторный дисперсионный анализ; модель I

Тип дисперсии	Число степеней свободы	Сумма квадратов	Средний квадрат	Математическое ожидание среднего квадрата	Проверяемая гипотеза	Статистика
Между строками	$r-1$	$c \sum_i (\bar{x}_{i.} - \bar{x})^2 = S_r^2$	$S_r^2/(r-1)$	$\sigma^2 + \frac{\sum_{i,j} (\bar{\mu}_{i.} - \mu)^2}{r-1}$	$\bar{\mu}_{.1} = \bar{\mu}_{.2} = \dots$ $\dots = \bar{\mu}_{.r.}$ $\left(\sum_{i,j} \xi_i^2 = 0 \right)$	$\frac{S_r^2/(r-1)}{S_{rc}^2/(r-1)(c-1)}$
Между столбцами	$c-1$	$r \sum_j (\bar{x}_{.j} - \bar{x})^2 = S_c^2$	$S_c^2/(c-1)$	$\sigma^2 + \frac{\sum_{i,j} (\bar{\mu}_{.j} - \mu)^2}{c-1}$	$\bar{\mu}_{.1} = \bar{\mu}_{.2} = \dots$ $\dots = \bar{\mu}_{.c}$ $\left(\sum_{i,j} \eta_j^2 = 0 \right)$	$\frac{S_c^2/(c-1)}{S_{rc}^2/(r-1)(c-1)}$
Остаточная дисперсия	$(r-1)(c-1)$	$\sum_i \sum_j (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2 = S_{rc}^2$	$\frac{S_{rc}^2}{(r-1)(c-1)}$	σ^2	—	—
Общая дисперсия	$rc-1$	$\sum_i \sum_j (x_{ij} - \bar{x})^2 = S_T^2$	$S_T^2/(rc-1)$	—	—	—

Модель: $X = x_{ij} = \mu + \xi_i + \eta_j + \varepsilon_{ij}$.

Предполагаемое распределение: ε_{ij} распределена как $N(0, \sigma^2)$.

Распределение X : $N(\mu + \xi_i + \eta_j, \sigma^2)$.

Полный двухфакторный дисперсионный анализ; модель II

Тип дисперсии	Число степеней свободы	Сумма квадратов	Средний квадрат	Математическое ожидание среднего квадрата	Проверяемая гипотеза	Статистика
Между строками	$r - 1$	$\sum_{i,j} (\bar{x}_{i.} - \bar{x})^2 = S_r^2$	$S_r^2 / (r - 1)$	$\sigma^2 + c\sigma_r^2$	$\sigma_r^2 = 0$	$\frac{S_r^2 / (r - 1)}{S_{rc}^2 / (r - 1)(c - 1)}$
Между столбцами	$c - 1$	$\sum_{i,j} (\bar{x}_{.j} - \bar{x})^2 = S_c^2$	$S_c^2 / (c - 1)$	$\sigma^2 + r\sigma_c^2$	$\sigma_c^2 = 0$	$\frac{S_c^2 / (c - 1)}{S_{rc}^2 / (r - 1)(c - 1)}$
Остаточная дисперсия	$(r - 1)(c - 1)$	$\sum_{i,j} (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2 = S_{rc}^2$	$\frac{S_{rc}^2}{(r - 1)(c - 1)}$	σ^2	—	—
Общая дисперсия	$rc - 1$	$\sum_{i,j} (x_{ij} - \bar{x})^2 = S_T^2$	$S_T^2 / (rc - 1)$	—	—	—

Модель: $X = x_{ij} = \mu + y_i + z_j + \varepsilon_{ij}$.

Предполагаемые распределения: ε_{ij} распределена как $N(0, \sigma^2)$,

y_i — как $N(0, \sigma_r^2)$,

z_j — как $N(0, \sigma_c^2)$.

Распределение X : $N(\mu, \sigma_r^2 + \sigma_c^2 + \sigma^2)$.

Таблица 7.16

Полный двухфакторный дисперсионный анализ; модель III

Тип дисперсии	Число степеней свободы	Сумма квадратов	Средний квадрат	Математическое ожидание среднего квадрата	Проверяемая гипотеза	Статистика
Между строками	$r-1$	$\sum_i \sum_j (\bar{x}_{i.} - \bar{x})^2 = S_r^2$	$S_r^2/(r-1)$	$\sigma^2 + c\sigma_r^2$	$\sigma_r^2 = 0$	$\frac{S_r^2/(r-1)}{S_{rc}^2/(r-1)(c-1)}$
Между столбцами	$c-1$	$\sum_i \sum_j (\bar{x}_{.j} - \bar{x})^2 = S_c^2$	$S_c^2/(c-1)$	$\sigma^2 + \frac{\sum_{i,j} (\bar{\mu}_{.j} - \mu)^2}{c-1}$	$\bar{\mu}_{.1} = \bar{\mu}_{.2} = \dots = \bar{\mu}_{.c}$ $\left(\sum_{i,j} (\bar{\mu}_{.j} - \mu)^2 = 0 \right)$	$\frac{S_c^2/(c-1)}{S_{rc}^2/(r-1)(c-1)}$
Остаточная дисперсия	$(r-1)(c-1)$	$\sum_i \sum_j (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2 = S_{rc}^2$	$\frac{S_{rc}^2}{(r-1)(c-1)}$	σ^2	—	—
Общая дисперсия	$rc-1$	$\sum_i \sum_j (x_{ij} - \bar{x})^2 = S_T^2$	$S_T^2/(rc-1)$	—	—	—

Модель: $X = x_{ij} = \mu + \alpha_i + \xi_j + e_{ij}$.

Предполагаемые распределения: e_{ij} распределена как $N(0, \sigma^2)$, α_i — как $N(0, \sigma_r^2)$.

Распределение X : $N(\mu + \xi_j, \sigma_r^2 + \sigma^2)$.

Таблица 7.17

Полный двухфакторный дисперсионный анализ с повторениями

Тип дисперсии	Число степеней свободы	Сумма квадратов	Средний квадрат	Математическое ожидание среднего квадрата	Проверяемая гипотеза	Статистика
Между строками	$r-1$	$c \sum_i (\bar{x}_{i..} - \bar{x})^2 = S_r^2$	$S_r^2/(r-1)$	$\sigma^2 + p\sigma_{rc}^2 + cp\sigma_r^2$	$\sigma_r^2 = 0$	$\frac{S_r^2/(r-1)}{S_{rc}^2/(r-1)(c-1)}$
Между столбцами	$c-1$	$rp \sum_j (\bar{x}_{.j} - \bar{x})^2 = S_c^2$	$S_c^2/(c-1)$	$\sigma^2 + p\sigma_{rc}^2 + pc\sigma_c^2$	$\sigma_c^2 = 0$	$\frac{S_c^2/(c-1)}{S_{rc}^2/(r-1)(c-1)}$
Промежуточная дисперсия	$(r-1)(c-1)$	$p \sum_{i,j} (\bar{x}_{ij} - \bar{x}_{i..} - \bar{x}_{.j} + \bar{x})^2 = S_{rc}^2$	$\frac{S_{rc}^2}{(r-1)(c-1)}$	$\sigma^2 + p\sigma_{rc}^2$	$\sigma_{rc}^2 = 0$	$\frac{S_{rc}^2/(r-1)(c-1)}{S_e^2/rc(p-1)}$
Остаточная дисперсия	$rc(p-1)$	$\sum_{i,j,k} (x_{ijk} - \bar{x}_{ij.})^2 = S_e^2$	$S_e^2/rc(p-1)$	σ^2	—	—
Общая дисперсия	$rcp-1$	$\sum_{i,j,k} (x_{ijk} - \bar{x})^2 = S_T^2$	$S_T^2/(rcp-1)$	—	—	—

Модель: $X = x_{ijk} = \mu + y_i + z_j + w_{ij} + \varepsilon_{ijk}$.

Предполагаемые распределения: y_i, z_j, w_{ij} и ε_{ijk} распределены как $N(0, \sigma_r^2)$, $N(0, \sigma_c^2)$, $N(0, \sigma_{rc}^2)$, $N(0, \sigma^2)$.

Распределение X : $(\mu, \sigma_r^2 + \sigma_c^2 + \sigma_{rc}^2 + \sigma^2)$

¹ Если принята $H_1: \sigma_{rc}^2 = 0$, то $S_{rc}^2/(r-1)(c-1)$ или $S_e^2/rc(p-1)$ могут быть использованы как знаменатель при проверке гипотез $\sigma_r^2 = 0$ и $\sigma_c^2 = 0$. Если $H_1: \sigma_{rc}^2 = 0$ принята, то $S_{rc}^2/(r-1)(c-1)$ и $S_e^2/rc(p-1)$ являются оценками для σ^2 . Следовательно, эти оценки можно использовать для получения новой, более точной оценки с большим числом степеней свободы. Этой оценкой будет величина $(S_{rc}^2 + S_e^2)/[(r-1)(c-1) + rc(p-1)]$.

где $q_{f,n}$ — верхнее значение α в q -распределении¹

$$Ea_{ij} = \mu_i - \mu_j.$$

Изменив обозначения Шеффе, можно выразить все *возможные сравнения* как

$$\Psi = \sum_{i=1}^n c_i \mu_i,$$

где $\sum_{i=1}^n c_i = 0$. В этом случае

$$P \left\{ b_{ij} - q_{f,n,\alpha} k s \left(\frac{1}{2} \sum_{i=1}^n |c_i| \right) \leq \Psi \leq b_{ij} + \right. \\ \left. + q_{f,n,\alpha} k s \left(\frac{1}{2} \sum_{i=1}^n |c_i| \right) \right\} = 1 - \alpha,$$

где

$$E[b_{ij}] = \Psi.$$

В основу этого критерия положены следующие предположения:

1. Оценки сравниваемых параметров статистически независимы.

2. Оценки распределены нормально с параметрами μ_i и $a^2 \sigma^2$, где $i=1, 2, \dots, n$; a — известная положительная константа.

3. Оценка s^2 распределена как случайная величина $\sigma^2 \chi^2 / j$ и независима от набора оценок сравниваемых параметров.

Примеры

Если в результате дисперсионного анализа гипотеза $H_1: \mu_1 = \mu_2 = \dots = \mu_n$ отвергается, то при сравнении любой пары из $n(n-1)/2$ возможных пар параметров значения разностей между различными μ_i будут удовлетворять соотношению

$$P \left\{ (\bar{x}_i - \bar{x}_j) - q_{f,n,\alpha} \frac{s}{\sqrt{n}} \leq \mu_i - \mu_j \leq \right. \\ \left. \leq (\bar{x}_i - \bar{x}_j) + q_{f,n,\alpha} \frac{s}{\sqrt{n}} \right\} = 1 - \alpha,$$

где $\bar{x}_i$ и $\bar{x}_j$ — оценки для μ_i и μ_j , $k = \frac{1}{\sqrt{n}}$, $a_{ij} = \bar{x}_i - \bar{x}_j$.

Допустимое расхождение между парами средних определяется величиной $q_{f,n,\alpha} \frac{s}{\sqrt{n}}$. Если разность между парой оце-

¹ См. сноску на стр. 175 и E. S. Pearson, H. D. Hartley, Biometrika Tables for Statisticals, Cambridge Univ. Press., 1954.

нок средних больше допустимой, параметры следует считать существенно различными. И наоборот, если расхождение между оценками меньше допустимой величины, параметры нужно рассматривать как несущественно отличающиеся друг от друга.

В качестве примера можно привести результаты исследования Кана [139], который изучал содержание обломочного материала в ориентированном кубическом образце песчаника из Потсвилла. Этот автор изучал образец в трех взаимно перпендикулярных сечениях, в каждом из которых он выбирал шесть направлений. По каждому направлению Кан получил оценку среднего содержания из результатов подсчета по пяти перпендикулярным к ним пересечениям. Эти оценки приведены в табл. 7.18. Кан показал, что набор оценок средних содержит

Таблица 7.18

Средние значения содержания обломочного материала
по всем пересечениям, %

	Направление	Сечение		
		XY	XZ	YZ
	0°	61,86	59,24	62,40
	30°	61,86	64,52	57,92
	60°	60,01	65,66	63,87
	90°	62,42	58,42	54,34
	120°	62,76	59,17	64,40
	150°	57,85	49,99	55,66
Среднее арифметическое по шлифам		61,13	59,50	59,77

существенно отличающиеся друг от друга элементы. В данном случае нужно решить, какие направления и в каких сечениях существенно отличаются друг от друга по средним значениям содержания обломочного материала. Применяя *T*-метод, мы поступаем следующим образом:

$$s^2 = 22,3866, \quad \frac{s}{\sqrt{n}} = \frac{4,73}{4,23} = 1,12,$$

$$f = 72,$$

$$n = 18,$$

$$q_{f, n, \alpha} = q_{72, 18, 0,05} = 5,1,$$

$$5,1 \cdot 1,12 = 5,71 \text{ (допустимое расхождение).}$$

Так, например, в сечении XZ оценка среднего в направлении 150° , равная 49,99, существенно отличается от всех остальных, так как последние превышают допустимую величину $49,99 + 5,71 = 55,70$. Это направление обладает минимальным значением среднего, тогда как максимум соответствует направлению 60° .

С помощью T -метода при анализе сечения YZ мы узнаем, что делать выводы о положении единственного максимума или минимума в этом сечении нельзя.

S-метод Шеффэ для нахождения всех «контрастов»

Если гипотеза $H_1: \mu_1 = \mu_2 = \dots \mu_k$, где μ_i — средние значения нормально распределенных случайных величин, отвергнута, то спрашивается, как поступить дальше, чтобы исключить те значения μ_i , которые нарушают однородность рассматриваемого ряда? Один из способов решения этой задачи был предложен Шеффэ¹.

Его метод позволяет производить сравнение любой пары или, что является главным, производить итоговое сравнение оценок всех параметров μ_i . Шеффэ показал, что если величина, выражающая результат сопоставления («контраст»), определена как

$$\theta = \sum_{i=1}^k c_i \mu_i,$$

где

$$\sum_{i=1}^k c_i = 0,$$

и если $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$ и s^2 — оценки для $\mu_1, \mu_2, \dots, \mu_k$ и σ^2 по k выборкам объемом $n_1, n_2, \dots, n_k$, таким, что $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$ совместно нормально распределены и не зависят от s^2 , а также если

$$E\bar{x}_i = \mu_i$$

и ковариация двух средних $\text{cov}(\bar{x}_i, \bar{x}_j)$ постоянна, тогда

$$P\{\hat{\theta} - S\hat{\sigma}_{\hat{\theta}} \leq \theta \leq \hat{\theta} + S\hat{\sigma}_{\hat{\theta}}\} = 1 - \alpha.$$

В этом выражении

$$\hat{\theta} = \sum_{i=1}^k c_i \bar{x}_i,$$

$$\hat{\sigma}_{\hat{\theta}}^2 = \text{var } \hat{\theta} = \text{var } \sum_{i=1}^k c_i \bar{x}_i$$

$$S = \left\{ (k-1) F_{\alpha} \left[\frac{k-1}{n_k-1} \right] \right\}^{1/2}.$$

Таким образом,

1. Если $\{-S\hat{\sigma}_{\hat{\theta}} < \hat{\theta} < S\hat{\sigma}_{\hat{\theta}}\}$, величина θ несущественно отличается от нуля.

2. Если $\{\hat{\theta} \geq S\hat{\sigma}_{\hat{\theta}}\}$, величина θ существенно отличается от нуля в большую сторону.

3. Если $\{\hat{\theta} \leq -S\hat{\sigma}_{\hat{\theta}}\}$, величина θ существенно отличается от нуля в меньшую сторону.

ОЦЕНКИ КОМПОНЕНТ ДИСПЕРСИИ

Модели II и III дисперсионного анализа полностью или частично связаны с проверкой гипотез о равенстве нулю тех или иных компонент дисперсии. Если эта гипотеза отвергается, то нужно каким-либо способом оценить величину этих компонент дисперсии.

В однофакторном дисперсионном анализе модели II (табл. 7.13) сумма квадратов между группами является оценкой для $c\sigma_s^2 + \sigma^2$. Точно так же сумма квадратов внутри групп представляет собой оценку дисперсии погрешностей эксперимента σ^2 . Таким образом, оценками для σ^2 и σ_s^2 будут $S_t^2/r(c-1)$ и $S_s^2/c(r-1) - S_t^2/rc(c-1)$ соответственно.

В случае полностью рандомизированных блоков¹ при p результатах наблюдения в каждом блоке оценками для σ^2 , σ_r^2 , σ_c^2 и σ_{rc}^2 являются $\hat{\sigma}^2$, $\hat{\sigma}_r^2$, $\hat{\sigma}_c^2$ и $\hat{\sigma}_{rc}^2$ соответственно, где

$$\begin{aligned}\hat{\sigma}^2 &= \frac{S_e^2}{rc(p-1)}, \\ \hat{\sigma}_c^2 &= \frac{S_c^2/(c-1) - \hat{\sigma}^2 - p\hat{\sigma}_{rc}^2}{rp}, \\ \hat{\sigma}_{rc}^2 &= \frac{S_{rc}^2/(r-1)(c-1) - \hat{\sigma}^2}{p}, \\ \hat{\sigma}_r^2 &= \frac{S_r^2/(r-1) - \hat{\sigma}^2 - p\hat{\sigma}_{rc}^2}{cp}.\end{aligned}$$

¹ См. определение дисперсионного анализа со вполне случайными блоками (стр. 153). — Прим. ред.

Если $p=1$, табл. 7.17 превратится в табл. 7.15 и оценками для σ^2 , σ_r^2 , σ_c^2 будут

$$\begin{aligned}\hat{\sigma}^2 &= \frac{S_{rc}^2}{(r-1)(c-1)}, \\ \hat{\sigma}_r^2 &= \frac{S_r^2}{c(r-1)} - \frac{\hat{\sigma}^2}{c}, \\ \hat{\sigma}_c^2 &= \frac{S_c^2}{r(c-1)} - \frac{\hat{\sigma}^2}{r}.\end{aligned}$$

В смешанной модели, описанной в табл. 7.16, σ^2 оценивается так же, как и в предыдущем случае, а оценкой для σ_r^2 будет величина

$$\frac{S_r^2}{c(r-1)} - \frac{\hat{\sigma}^2}{c}.$$

Компоненты дисперсии в других моделях оцениваются аналогично.

Стабилизация дисперсии и преобразование переменных

Задачу «подгонки» конкретных примеров под главные предположения, лежащие в основе техники дисперсионного анализа, как средство проверки гипотез нельзя считать неразрешимой. Наиболее неприятные отклонения от теоретических предположений следующие: крайняя асимметрия, заметное отклонение от аддитивности, изменения в дисперсии погрешностей эксперимента и наличие больших ошибок¹. Аномальные результаты наблюдения лучше исключать из рассмотрения. С прагматической точки зрения асимметрия и неоднородность дисперсий наиболее вредны; однако их действие можно устранить путем подходящего преобразования исходных данных. Нередко неоднородность дисперсий бывает обусловлена функциональной зависимостью между дисперсией и средним значением результатов измерения. Если эта зависимость может быть установлена *a priori* или эмпирическим путем, можно подобрать и подходящее преобразование, которое делает эти величины независимыми. Бартлетт² рассмотрел эту задачу и дал сводку преобразований, пригодных для некоторых отдельных типов зависимости между средними и дисперсиями набора выборок. При этом учитывается и распределение данных. Эти преобразования приведены в табл. 7.19. Так, например, если дисперсии набора выборок про-

¹ W. G. Cochran, Some consequences when the assumptions for the analysis of variance are not satisfied, *Biometrics*, 3, 1947.

² M. S. Bartlett, The use of transformations, *Biometrics*, 3, 1947.

порциональны выборочным значениям средних, т. е. бóльшим дисперсиям соответствуют бóльшие значения средних, то подходящим в этом случае преобразованием будет корень квадратный из первоначальных результатов наблюдений. Возможны случаи, когда функциональные зависимости между средними и дисперсиями в наборе независимых выборок не будут соответствовать ни одной связи из перечисленных в табл. 7.19. В подобной ситуации следует определить прямую или кривую, наиболее подходящую для описания зависимости от имеющихся данных. Эту зависимость легко установить графически, рассматривая дисперсию как функцию от среднего. Если полученная функция имеет вид $\sigma^2 = f(m)$, то нужное преобразование $g(m)$ может быть найдено путем интегрирования¹, а именно

$$g(m) = \int \frac{c \, dm}{\sqrt{f(m)}}. \quad (7.51a)$$

В качестве примера применения преобразования исходных данных для стабилизации дисперсии, а также для приведения эмпирического распределения к нормальному виду рассмотрим результаты работы Розенфельда и Гриффитса [236]. Цель их работы первоначально состояла в разработке приемов визуального сравнения степени сферичности и окатанности кварцевых зерен в осадочных образованиях. Авторы сравнивали эмпирические распределения и, помимо прочих выводов, пришли к заключению, что распределения существенно отличаются от нормального. Результаты были нанесены на график оценок среднего показателя сферичности и дисперсий по данным 21 выборки; таким образом эти авторы показали отсутствие функциональной зависимости между средними и дисперсиями. Так как область значений показателя сферичности заключена между 0 и 1, т. е. она является процентной величиной, или пропорциональной формой, Розенфельд и Гриффитс решили вместо исходных данных рассматривать их арксинусы.

Данные изучения окатанности в наборе независимых выборок свидетельствуют о том, что между средними и дисперсиями существует функциональная связь. Следует отметить, что значения показателя окатанности также заключены между 0 и 1. Рассмотрение вместо исходных данных их арксинусов не привело ни к нормальному распределению, ни к функциональной зависимости между средними и дисперсиями. Тогда был использован описанный выше метод Бартлетта (7.51a).

Линия, наилучшим образом отражающая зависимость между средними и дисперсиями (гл. 8), получена в виде

¹ Более точное описание см. Э. Леман, Проверка статистических гипотез, Наука, М., 1964, стр. 374. — *Прим. ред.*

Обобщение преобразований

Дисперсия в единицах среднего m	Преобразование	Приближенная дисперсия новой величины	Подходящее распределение
m	$\left\{ \sqrt{x} \text{ (или } \sqrt{x + \frac{1}{2}} \right.$	0,25	Пуассоновское
$\lambda^2 m$	$\left. \begin{array}{l} \text{для малых значений} \\ \ln x \end{array} \right\}$	$0,25\lambda^2$	Эмпирическое
$2m^2/(n-1)$		$2/(n-1)$	Выборочные дисперсии
$\lambda^2 m^2$	$\left\{ \begin{array}{l} \ln x, \ln(x+1) \\ \lg x, \lg(x+1) \end{array} \right\}$	$\left\{ \begin{array}{l} \lambda^2 \\ 0,189\lambda^2 \end{array} \right\}$	Эмпирическое
$m(1-m)/n$	$\left\{ \begin{array}{l} \arcsin \sqrt{x} \text{ (радианы)} \\ \arcsin \sqrt{x} \text{ (градусы)} \end{array} \right\}$	$\left\{ \begin{array}{l} 0,25/n \\ 821/n \end{array} \right\}$	Биномиальное
$m(1-m)/n$	Пробит	Не постоянны	—
$m(1-m)/n$	$\ln [x/(1-x)]$	$1/[mn(1-m)]$	
$\lambda^2 m^2 (1-m)^2$	$\ln [x/(1-x)]$	λ^2	Эмпирическое
$(1-m^2)/(n-1)$	$\frac{1}{2} \ln [(1+x)/(1-x)]$	$1/(n-3)$	Выборочная корреляция
$m + \lambda^2 m^2$	$\left\{ \begin{array}{l} \lambda^{-1} \text{ Arch } [\lambda \sqrt{x}] \text{ или} \\ \lambda^{-1} \text{ Arch } \left[\lambda \sqrt{x + \frac{1}{2}} \right] \end{array} \right\}$	0,25	Отрицательно биномиальное
$\mu(m + \lambda^2 m^2)$	$\left\{ \begin{array}{l} \text{для малых значений} \\ \text{К ожидаемому нормаль-} \\ \text{ному распределению} \end{array} \right\}$	$0,25\mu^2$	Эмпирическое
—		1 для больших n	Упорядоченные данные

Таблица взята из работы Бартлетта, 1947 (см. сноску 2 на стр. 186).

выражения

$$s^2 = 0,01107 - 0,01208\bar{x} = f(m).$$

Следуя изложенным выше рекомендациям, был вычислен интеграл

$$c \int \frac{dm}{f(m)} = c \int \frac{d\bar{x}}{(0,01107 - 0,01208\bar{x})^{1/2}},$$

и преобразование приняло форму

$$P_A = (0,01107 - 0,0121P)^{1/2}.$$

где P — измеренное значение окатанности, а P_A — преобразованное значение.

В итоге с помощью относительно простых преобразований выборочных данных были устранены некоторые из трудностей, связанные с проверкой гипотез.

Двухфакторный дисперсионный анализ с повторениями

При рассмотрении двухфакторного анализа предполагалось, что одной ячейке, т. е. пересечению одной строки и одного столбца в исходной таблице, соответствует одно наблюдение. Чтобы уменьшить ошибку сравнения в строках или столбцах, лучше много раз повторить наблюдения в каждой ячейке. Если в каждой ячейке сделано по p повторных наблюдений, единичный результат можно записать как x_{ijk} , где $i=1, 2, \dots, r$; $j=1, 2, \dots, c$; $k=1, 2, \dots, p$: $x_{ijk} = \mu + a_i + b_j + w_{ij} + \epsilon_{ijk}$. Кроме того, $E(x_{ijk}) = \mu$, а $\text{var}(x_{ijk}) = \sigma_a^2 + \sigma_b^2 + \sigma_{ab}^2 + \sigma^2$. В табл. 7.17 наряду с гипотезами, которые могут быть проверены, приведена вся схема дисперсионного анализа данной модели.

Вычислительная техника дисперсионного анализа

Суммы квадратов, приведенные в табл. 7.2—7.5 и 7.8—7.17, удобны в работе с реальными задачами, когда число наблюдений невелико и получающиеся числа не очень громоздки. В большей части работы, кроме той, что запрограммирована для высокоскоростной вычислительной техники, приходится пользоваться тем, что обычно называют *вычислительными формулами дисперсионного анализа*. Эти формулы позволяют использовать

Таблица 7.20

Обобщенные формулы для вычисления сумм квадратов при однофакторном дисперсионном анализе для всех моделей

Тип дисперсии	Сумма квадратов
Между машинами или шлифами (как пример)	$\sum_{i=1}^m \left(\sum_{j=1}^r x_{ij} \right)^2 / r - \left(\sum_{i,j} x_{ij} \right)^2 / mr$
Внутри машин, шлифов, или, что равносильно, пружин, пересечений	Разность между нижней и верхней формулами
Общая дисперсия	$\sum_{i,j} x_{ij}^2 - \left(\sum_{i,j} x_{ij} \right)^2 / mr$

данные в их первоначальной форме (без вычисления средних значений) для отыскания различных типов дисперсий. Формулы для процедур дисперсионного анализа, описанных в этой главе, приведены в табл. 7.20—7.22.

Таблица 7.21

Обобщенные формулы для вычисления сумм квадратов при двухфакторном дисперсионном анализе без повторения для всех моделей

Тип дисперсии	Сумма квадратов
Между строками	$\sum_{i=1}^r \left(\sum_{j=1}^c x_{ij} \right)^2 / c - \left(\sum_{i,j} x_{ij} \right)^2 / rc = S_r^2$
Между столбцами	$\sum_{j=1}^c \left(\sum_{i=1}^r x_{ij} \right)^2 / r - \left(\sum_{i,j} x_{ij} \right)^2 / rc = S_c^2$
Остаточная дисперсия	Разность $S_T^2 - S_r^2 - S_c^2$
Общая дисперсия	$\sum_{i=1}^r \sum_{j=1}^c x_{ij}^2 - \left(\sum_{i=1}^r \sum_{j=1}^c x_{ij} \right)^2 / rc = S_T^2$

Таблица 7.22

Обобщенные формулы для вычисления сумм квадратов при двухфакторном дисперсионном анализе с p повторениями в каждой ячейке для всех моделей

Тип дисперсии	Сумма квадратов
Между строками	$\sum_{i=1}^r \left(\sum_{j,k} x_{ijk} \right)^2 / cp - \left(\sum_{i,j,k} x_{ijk} \right)^2 / rcp = S_r^2$
Между столбцами	$\sum_{j=1}^c \left(\sum_{i,k} x_{ijk} \right)^2 / pr - \left(\sum_{i,j,k} x_{ijk} \right)^2 / rcp = S_c^2$
Остаточная дисперсия	$\sum_{i,j} \left(\sum_{k=1}^p x_{ijk} \right)^2 / p - S_r^2 - S_c^2 - \left(\sum_{i,j,k} x_{ijk} \right)^2 / rcp = S_{rc}^2$
Внутри ячеек	$S_T^2 - S_r^2 - S_c^2 - S_{rc}^2$
Общая дисперсия	$\sum_{i,j,k} x_{ijk}^2 - \left(\sum_{i,j,k} x_{ijk} \right)^2 / rcp = S_T^2$

НЕПАРАМЕТРИЧЕСКИЙ ДИСПЕРСИОННЫЙ АНАЛИЗ; ОДНОФАКТОРНЫЙ РАНГОВЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ КРАСКЛА — УЭЛЛИСА

Краскл и Уэллис¹ рассматривали в качестве проверяемой нулевой гипотезы утверждение, согласно которому выборочные средние, полученные по данным K независимых выборок, несущественно отличаются одно от другого. Предполагалось, что о распределении изучаемых случайных величин ничего не известно, кроме того лишь, что они непрерывны. В этом предположении заключается основное отличие от параметрического анализа.

В качестве проверочной статистики использовалась величина

$$H = \frac{12}{N(N+1)} \sum_{i=1}^K \frac{R_i^2}{n_i} - 3(N+1), \quad (7.52)$$

где K — число выборок объемом $n_1, n_2, \dots, n_k$; N — общее число наблюдений по всем выборкам; R_i — сумма рангов в i -й выборке. Под рангом имеется в виду номер, приписанный каждому члену данной выборки в общей упорядоченной возрастающей последовательности результатов наблюдения по всем выборкам. Так, ранг наименьшего значения равен 1, ранг следующего по величине значения равен 2, а наибольшего — N . Иногда необходима специальная обработка данных, которая описана Красклом и Уэллисом, а также Сигалом². Если каждая выборка содержит более пяти наблюдений ($n_i > 5$), то статистика, выраженная формулой (7.52), будет распределена как χ^2 с $K - 1$ степенями свободы. Точные значения вероятностей для случая, когда $n_i \leq 5$ и число выборок равно 3, были получены М. Фридманом в 1937 г.; этот вопрос рассматривал Сигал², который и приводит соответствующие таблицы Фридмана. Уэллис и Робертс³ показали, что для приближенного определения вероятностей, соответствующих значениям H , можно также пользоваться таблицами нормального распределения.

В качестве примера рассмотрим табл. 7.23, построенную по данным следующего опыта Вейсера (личное сообщение, 1952 г.). В четырех резервуарах при постоянной температуре в каждом, но различной в отдельных резервуарах выращивались пресноводные гастроподы *Limnea*. В конце периода их роста в рако-

¹ W. H. Kruskal, A nonparametric test for the several sample problem, *Annals, Math. Stats.*, 23, 1952; W. H. Kruskal, W. A. Wallis, Use of ranks in one criterion variance analysis, *Journ. Am. Stat. Assoc.*, 47, 1952.

² S. Siegal, *Nonparametric statistics for the behavioral sciences*, New York, 1956.

³ W. A. Wallis, H. V. Roberts, *Statistics, a new approach*, The Free Press, Glencoe, Illinois, 1956.

винах определили некоторые редкие элементы, накапливаемые гастроподами из окружающей среды. В табл. 7.23 приведены результаты определения содержаний бария, выраженные в частях на миллион (при этом предполагается, что точность анализа достаточна для надежного установления реального количественного различия).

Таблица 7.23

Резервуар I		Резервуар II		Резервуар III		Резервуар IV	
барий	ранг	барий	ранг	барий	ранг	барий	ранг
20	2	21	3	18	1	27	6
30	8	25	5	24	4	28	7
44	13	33	9	35	10	37	11
51	14	39	12	62	17	55	15
65	19	60	16	75	23	64	18
72	21	80	24	85	25	68	20
100	26					73	22
$R_1 = 103$		$R_2 = 69$		$R_3 = 80$		$R_4 = 99$	

$$N = 26, n_1 = 7, n_2 = 6, n_3 = 6, n_4 = 7, K = 4,$$

$$H = \frac{12}{26 \cdot (26 + 1)} \left(\frac{103^2}{7} + \frac{69^2}{6} + \frac{80^2}{6} + \frac{99^2}{7} \right) - 3 \cdot (26 + 1) = 0,19.$$

Нулевая гипотеза заключается в том, что между средними содержаниями бария в раковинах особей, выращенных в четырех разных резервуарах, нет разницы. Положим уровень значимости $\alpha = 0,07$. Тогда критическая область будет содержать все те значения H , которые столь велики, что соответствуют вероятности отклонения нулевой гипотезы (если она верна), меньшей или равной 0,05. Для числа степеней свободы $K - 1 = 4 - 1 = 3$ допустимое значение χ^2 при $\alpha = 0,05$ равно 7,82. Так как $H = 0,19$, т. е. значительно меньше 7,82, нулевую гипотезу следует принять и средние содержания бария рассматривать как равные. Если бы вычисленное значение H превысило 7,82, гипотезу о равенстве средних следовало бы отвергнуть.

Изложенный метод полезен в тех случаях, когда исследователь не уверен в выполнении допущений, лежащих в основе формального дисперсионного анализа, в данной конкретной ситуации (например, в выполнении предположений о нормальности и однородности дисперсий).

Нередко исследователь не ограничивается решением задачи только с одной случайной величиной, а решает задачу с двумя, тремя и с большим числом переменных величин, изучаемых совместно. Примерами подобных исследований могут служить работы Стрейлера¹, Милтона² и Гриффитса³. Наблюдения над парами случайных величин (x, y) или (y, z) или же их группами (x, y, z) и (p, q, r, s, t) не являются чем-либо необычным.

При рассмотрении двух случайных величин может возникнуть вопрос о существовании между ними какой-либо зависимости. Если она существует, то какова ее форма? Можно ли, располагая значениями одной величины, предсказать значения другой? Один из подходов к решению этой задачи состоит в следующем: вводится предположение о форме зависимости между изучаемыми величинами. Таким образом, в качестве функции можно рассматривать y , а x — как ее аргумент, а затем ввести допущение, согласно которому математическое ожидание y при данном значении x является константой. Например, если положить $y = f(x)$, то $f(x)$ может быть

$$f(x) = k_1 x + k_2, \quad (8.1)$$

или

$$f(x) = k_3 x^{k_4} + k_5 x^3 + \log 2x, \quad (8.2)$$

или

$$f(x) = \log x, \quad (8.3)$$

или

$$f(x) = \log(1 + x),$$

¹ A. Strahler, Dimensional analyses applied to fluvially eroded landforms, Bull. Geol. Soc. Am., 69, 1958.

² M. Melton, An analysis of the relations among elements of climate, surface properties, and geomorphology, Tech. Rept. No. 11, Dept. of Geol., Columbia Univ., New York, 1957.

³ J. C. Griffiths, Modal analysis of sediments, Revue de Geogr. phys et de Geol. Dynamique (2), III, f. 1, Paris, 1960.

где $k_1, k_2, \dots, k_5$ — константы или параметры. После того как введены предположения о форме зависимости, нужно выбрать подходящие оценки для неизвестных параметров, например для k_1 и k_2 в выражении (8.1) и аналогично для k_3, k_4 и k_5 в (8.2). Получив оценки параметров и установив их распределения, можно перейти к построению критериев для проверки гипотез об этих параметрах.

ЛИНЕЙНАЯ РЕГРЕССИЯ ПО ДАННЫМ ВЫБОРКИ ОБЪЕМОМ n ПРИ ОДНОМ НАБЛЮДЕНИИ ДЛЯ КАЖДОГО ЗНАЧЕНИЯ АРГУМЕНТА

Допустим, что $f(x)$ имеет следующий вид:

$$f(x) = \text{const} + \text{const}(x), \quad (8.4)$$

т. е. зависимость является линейной. Следовательно, мы предполагаем, что зависимость линейная, но с неизвестными параметрами. Выражение (8.4) можно записать в виде

$$f(x) = \alpha + \beta x,$$

или

$$f(x) = \alpha + \beta(x - \bar{x}). \quad (8.5)$$

В дальнейшем мы будем предполагать зависимость в виде (8.5).

Из вводных замечаний следует, что

$$E(y|x) = \mu = \alpha + \beta(x - \bar{x}) \quad (8.6)$$

и что

$$\text{var}(y|x) = \sigma^2.$$

В частности,

$$E(y_i|x_i) = \mu_i$$

и

$$\text{var}(y_i|x_i) = \sigma_i^2 = \sigma^2.$$

Оценки параметров α , β и σ^2 и предположение о виде распределения y

Чтобы построить критерии для проверки гипотез о неизвестных параметрах, нужно знать распределения оценок, которые сами являются функциями от y , что в свою очередь требует введения предположения о распределении случайной величины y . Таким образом, *мы предполагаем, что каждому значению x соответствует нормально распределенная величина y со средним $\alpha + \beta(x - \bar{x}) = \mu$ и дисперсией σ^2* . Используя метод максимального правдоподобия (гл. 9), мы можем получить оценки пара-

Функция правдоподобия L дается выражением

$$L = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n [y_i - \alpha - \beta(x_i - \bar{x})]^2 \right\}$$

и

$$\ln L = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n [y_i - \alpha - \beta(x_i - \bar{x})]^2.$$

Тогда

$$\frac{\partial \ln L}{\partial \alpha} = \frac{1}{\sigma^2} \sum_{i=1}^n [y_i - \alpha - \beta(x_i - \bar{x})] = 0, \quad (8.7)$$

$$\frac{\partial \ln L}{\partial \beta} = \frac{1}{\sigma^2} \sum_{i=1}^n [y_i - \alpha - \beta(x_i - \bar{x})] (x_i - \bar{x}) = 0 \quad (8.8)$$

и

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n [y_i - \alpha - \beta(x_i - \bar{x})]^2 = 0. \quad (8.9)$$

Обозначив через a , b и $\hat{\sigma}^2$ оценки для α , β и σ^2 соответственно и решив систему из трех уравнений (8.7) — (8.9), получим следующие оценки:

$$a = \bar{y}, \quad (8.10)$$

$$b = \frac{\sum_{i=1}^n y_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad (8.11)$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n [y_i - a - b(x_i - \bar{x})]^2}{n}. \quad (8.12)$$

Из выражения (8.11) видно, почему необходимо располагать по крайней мере двумя значениями x . Если будет использовано только одно значение x , то (8.11) превратится в неопределенность и оценки для β и σ^2 потеряют смысл.

Свойства оценок a и b

$$E(a) = E(\bar{y}) = \frac{E \sum_{i=1}^n [\alpha + \beta(x_i - \bar{x})]}{n} = \alpha, \quad (8.13)$$

$$\text{var}(a) = \text{var}(\bar{y}) = \frac{\sigma^2}{n}; \quad (8.14)$$

$$\begin{aligned}
 E(b) &= E \frac{\sum_{i=1}^n y_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{E \sum_{i=1}^n (x_i - \bar{x}) [a + \beta (x_i - \bar{x})]}{\sum_{i=1}^n (x_i - \bar{x})^2} = \\
 &= \beta \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta,
 \end{aligned} \tag{8.15}$$

$$\begin{aligned}
 \text{var}(b) &= \text{var} \frac{\sum_{i=1}^n y_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \text{var } y_i}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} = \\
 &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.
 \end{aligned} \tag{8.16}$$

РАЗДЕЛЕНИЕ ПРИЧИН ИЗМЕНЧИВОСТИ

Уклонение результата любого наблюдения от его ожидаемого значения можно записать как

$$[y_i - a - \beta(x_i - \bar{x})] = y_i - \mu_i,$$

что показано на фиг. 8.1. Это уклонение можно разделить на две части:

$$[y_i - a - b(x_i - \bar{x})] \quad \text{и} \quad [a + b(x_i - \bar{x}) - a - \beta(x_i - \bar{x})],$$

которые представляют собой соответственно уклонение любого результата наблюдения от вычисленного на основе эмпирического уравнения регрессии и уклонение последнего от истинного значения. Таким образом,

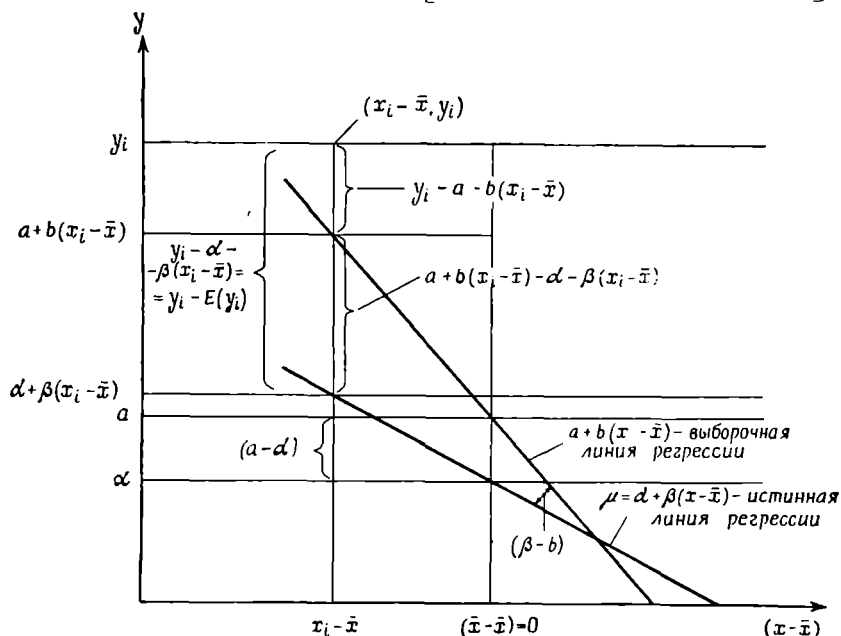
$$y_i - \mu_i = [y_i - a - b(x_i - \bar{x}) + (a - \alpha) + (b - \beta)(x_i - \bar{x})].$$

Сумма квадратов уклонений результатов наблюдения от истинной линии регрессии можно представить в виде следующего разложения:

$$\begin{aligned}
 \sum_{i=1}^n (y_i - \mu_i)^2 &= \sum_{i=1}^n [y_i - a - b(x_i - \bar{x})]^2 + n(a - \alpha)^2 + \\
 &\quad + (b - \beta)^2 \sum_{i=1}^n (x_i - \bar{x})^2
 \end{aligned}$$

плюс набор смешанных произведений. Смешанные произведения равны нулю, так как

$$\begin{aligned} n(a - \alpha) \sum_{i=1}^n [y_i - a - b(x_i - \bar{x})] &= \\ &= n(a - \alpha) [n\bar{y} - n\bar{y} - b(n\bar{x} - n\bar{x})] = 0, \\ (b - \beta) \sum_{i=1}^n [y_i - a - b(x_i - \bar{x})](x_i - \bar{x}) &= \\ &= (b - \beta) \left[\sum_{i=1}^n y_i(x_i - \bar{x}) - b \sum_{i=1}^n (x_i - \bar{x})^2 \right]. \end{aligned}$$



Фиг. 8.1. Графическое изображение типов изменчивости при линейной регрессии с одной независимой переменной внутри группы при n наблюдениях в группе.

$[y_i - a - b(x_i - \bar{x})]$ — отклонение любого наблюдения от выборочной линии регрессии;

$[y_i - \dots] = y_i - \mu_i$ — отклонение любого наблюдения от истинной линии регрессии;

$[a + b(x_i - \bar{x}) - \alpha - \beta(x_i - \bar{x})]$ — отклонение истинной линии регрессии от выборочной.

В силу формулы (8.11) последнее выражение равно нулю. Наконец,

$$n(a - \alpha)(b - \beta) \sum_{i=1}^n (x_i - \bar{x}) = 0.$$

Таким образом, общая сумма квадратов, характеризующая уклонения значений случайной величины y от ожидаемого значения, разложена на три части:

$$\sum_{i=1}^n (y_i - \mu_i)^2 = \sum_{i=1}^n [y_i - a - b(x_i - \bar{x})]^2 + n(a - \alpha)^2 + (b - \beta)^2 \sum_{i=1}^n (x_i - \bar{x})^2; \quad (8.17)$$

эти три части соответственно представляют собой отклонения результатов наблюдения от эмпирической линии регрессии, различия между оценкой для α и истинным значением α и аналогичную разницу между β и ее оценкой.

Распределение величин a и b

Так как $\sum_{i=1}^n (y_i - \mu_i)^2$ распределена как $\sigma^2 \chi^2$ (гл. 4) с n степенями свободы и представлена как сумма трех квадратичных форм, ранги которых равны соответственно $(n-2)$, 1 и 1, то, согласно теореме Кохрена, величины $\sum_{i=1}^n [y_i - a - b(x_i - \bar{x})]^2$, $n(a - \alpha)^2$ и $(b - \beta)^2 \sum_{i=1}^n (x_i - \bar{x})^2$ стохастически независимы и распределены как $\sigma^2 \chi^2$ соответственно с $(n-2)$, 1 и 1 степенями свободы. Таким образом, a , b и $\hat{\sigma}^2$ независимы. Далее, так как $n(a - \alpha)^2 = \sigma^2 \chi^2$, следовательно,

$$\frac{(a - \alpha)^2}{\sigma^2/n} = \chi^2.$$

Используя результаты (8.13) и (8.14), получим

$$\frac{[a - E(a)]^2}{\text{var } a} = \chi^2,$$

из чего следует, что a распределена как $N(\alpha, \sigma^2/n)$. Точно так же величина b распределена как $N\left[\beta, \sigma^2 / \sum_{i=1}^n (x_i - \bar{x})^2\right]$, так как

$$\sum_{i=1}^n (x_i - \bar{x})^2 (b - \beta)^2 = \sigma^2 \chi^2.$$

Следовательно,

$$\frac{(b - \beta)^2}{\sigma^2 \left| \sum_{i=1}^n (x_i - \bar{x})^2 \right|} = \chi^2,$$

но последнее выражение идентично $(b - Eb)^2/\text{var } b^1$.

Проверка гипотез $\alpha=0$ и $\beta=0$

Так как величины a и b нормально распределены, то гипотезы об их математических ожиданиях, а именно α и β , могут быть проверены способом, аналогичным изложенному в гл. 6, в которой рассматривались гипотезы об оценках для μ нормальной совокупности. В случае когда дисперсия σ^2 неизвестна, можно воспользоваться ее оценкой. Тогда обе статистики

$$\frac{a}{\sum_{i=1}^n [y_i - a - b(x_i - \bar{x})]^2 / n(n-2)}$$

и

$$\frac{b \sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n [y_i - a - b(x_i - \bar{x})]^2 / (n-2)}$$

будут распределены по закону Стьюдента с $n-2$ степенями свободы, если гипотезы $\alpha=0$ и $\beta=0$ верны.

Связь между дисперсионным анализом и регрессионным; обоснование выбора формы уравнения регрессии как $Y = a + b(x - \bar{x})$, а не $Y = a + bx$

Причины изменчивости, связанные с регрессией в выборке объемом n , сведены в табл. 8.1. Сравнение сумм квадратов в табл. 8.1 и критериев для проверки гипотез $\alpha=0$ и $\beta=0$ наводит на мысль о зависимости между ними. Это действительно так. Например, критерий для проверки гипотезы $\alpha=0$ есть

$$t_{(n-2)} = \sqrt{\frac{s_a^2}{s^2}},$$

а для проверки гипотезы $\beta=0$

$$t_{(n-2)} = \sqrt{\frac{s_b^2}{s^2}}.$$

¹ Нормальность распределений a и b вытекает из того факта, что эти оценки суть линейные комбинации нормальных величин; см. (8.10) и (8.11).—
Прим. ред.

Далее, отношение s_a^2/s^2 распределено как $F\left[\frac{1}{n-2}\right]$, так как величина s_a^2 распределена как $\sigma^2\chi^2$ с одной степенью свободы, а s^2 — как $\sigma^2\chi^2$ с $n-2$ степенями свободы. Все это верно независимо от значений α и β . Если регрессия дана в форме $y=\alpha+\beta(x-\bar{x})$, а общая сумма квадратов разложена на три части [см. уравнение (8.17)], то a и b независимы. Если же уравнение регрессии дано как $\alpha+\beta$, то a и b становятся зависимыми. Наличие зависимости между суммами квадратов и критериями для проверки гипотез об α и β , показанное выше, не является неожиданным, так как существует связь между t -распределением Стьюдента и F -распределением Фишера, что будет показано ниже (стр. 203—204).

Т а б л и ц а 8.1

Обобщение причин изменчивости для регрессионного анализа в выборке объемом n

Причина изменчивости	Число степеней свободы	Сумма квадратов	Средний квадрат
Уклонение от эмпирической регрессии	$n-2$	$\sum_{i=1}^n [y_i - a - b(x_i - \bar{x})]^2 = S^2$	$S^2/(n-2) = s^2$
a от α	1	$n(a-\alpha)^2 = S_a^2$	$S_a^2/1 = s_a^2$
b от β	1	$(b-\beta) \sum_{i=1}^n (x_i - \bar{x})^2 = S_b^2$	$S_b^2/1 = s_b^2$
Общая дисперсия	n	$\sum_{i=1}^n (y_i - \mu_i)^2$	—

ЛИНЕЙНАЯ РЕГРЕССИЯ В СЛУЧАЕ n_i НАБЛЮДЕНИЙ ДЛЯ КАЖДОГО ЗНАЧЕНИЯ АРГУМЕНТА x_i

Допустим, что каждому значению x_i соответствует n_i наблюдений над случайной величиной y ($i=1, 2, \dots, c$). Как и в предыдущих случаях, $E[y|x]$ представляют собой $\alpha+\beta(x-\bar{x})$ и, в частности,

$$E[y_{ij}|x_i] = \mu_i = \alpha + \beta(x_i - \bar{x}),$$

где

$$i = 1, 2, \dots, c,$$

$$j = 1, 2, \dots, n_i$$

и

$$\text{var}[y_{ij}|x_i] = \sigma_i^2 = \sigma^2 \text{ для всех } i, j.$$

Оценки параметров

В данном случае функция максимального правдоподобия дается выражением

$$L = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^c \sum_{j=1}^{n_i} [y_{ij} - a - b(x_i - \bar{x})]^2 \right\},$$

где

$$n = \sum_{i=1}^c n_i;$$

тогда уравнения для нахождения оценок параметров a , b и σ^2 примут вид

$$\frac{1}{\sigma^2} \sum_{i=1}^c \sum_{j=1}^{n_i} [y_{ij} - a - b(x_i - \bar{x})] = 0,$$

$$\frac{1}{\sigma^2} \sum_{i=1}^c \sum_{j=1}^{n_i} [y_{ij} - a - b(x_i - \bar{x})] (x_i - \bar{x}) = 0$$

и

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^c \sum_{j=1}^{n_i} [y_{ij} - a - b(x_i - \bar{x})]^2 = 0.$$

Таким образом,

$$a = \frac{1}{n} \sum_{i=1}^c \sum_{j=1}^{n_i} y_{ij} = \frac{1}{n} \sum_{i=1}^c n_i \bar{y}_i = \bar{y},$$

$$b = \frac{1}{\sum_{i=1}^c n_i (x_i - \bar{x})^2} \sum_{i=1}^c \sum_{j=1}^{n_i} y_{ij} (x_i - \bar{x}).$$

Это можно записать как

$$\frac{\sum_{i=1}^c y_{i.} (x_i - \bar{x})}{\sum_{i=1}^c n_i (x_i - \bar{x})^2} = \frac{\sum_{i=1}^c n_i \bar{y}_i (x_i - \bar{x})}{\sum_{i=1}^c n_i (x_i - \bar{x})^2},$$

и

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^c \sum_{j=1}^{n_i} [y_{ij} - a - b(x_i - \bar{x})]^2.$$

$$\begin{aligned} Ea &= \alpha, \\ Eb &= \beta, \\ \text{var } a &= \frac{\sigma^2}{n}, \\ \text{var } b &= \frac{\sigma^2}{\sum_{i=1}^c n_i (x_i - \bar{x})^2}. \end{aligned}$$

Разложение суммы квадратов

Сумма квадратов, характеризующая отклонения значений случайной величины y от ожидаемых значений, дается выражением

$$\sum_{i=1}^c \sum_{j=1}^{n_i} [y_{ij} - \alpha - \beta(x_i - \bar{x})]^2 = \sum_i \sum_j (y_{ij} - \mu_i)^2.$$

Эта величина распределена как $\sigma^2 \chi^2$ с n степенями свободы; ее можно разложить на следующие суммы квадратов, которые независимы и распределены как $\sigma^2 \chi^2$ с $(\sum_{i=1}^c n_i) - c$, $c - 2$, 1 и 1 степенями свободы, а именно:

$$\begin{aligned} & \sum_i \sum_j (y_{ij} - \bar{y}_i)^2, \\ & \sum_i \sum_j [\bar{y}_i - a - b(x_i - \bar{x})]^2 = \sum_i n_i [\bar{y}_i - a - b(x_i - \bar{x})]^2, \\ & \sum_i n_i (a - \alpha)^2 \quad \text{и} \quad (b - \beta)^2 \sum_{i=1}^c n_i (x_i - \bar{x})^2. \end{aligned}$$

Это показано на фиг. 8.2 и обобщено в табл. 8.2. Каждая из сумм квадратов в табл. 8.2 распределена как $\sigma^2 \chi^2$ с соответствующим числом степеней свободы. Таким образом,

$$\frac{(a - \alpha)^2 \sum_{i=1}^c n_i}{\sigma^2} = \chi^2,$$

а величина a распределена как $N[\alpha, \sigma^2/n]$. Точно так же величина b распределена как

$$N\left[\beta, \frac{\sigma^2}{\sum_{i=1}^c n_i (x_i - \bar{x})^2}\right].$$

Обобщенная таблица причин изменчивости

Причина изменчивости	Число степеней свободы	Сумма квадратов	Средний квадрат
Внутри групп	$\sum_{i=1}^c (n_i - 1)$	$\sum_{i=1}^c \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 = S_g^2$	$\frac{S_g^2}{\sum_{i=1}^c (n_i - 1)} = s_g^2$
Средние отклонения от эмпирического уравнения регрессии	$c - 2$	$\sum_{i=1}^c n_i [\bar{y}_{i.} - a - b(x_i - \bar{x})]^2 = S^2$	$S^2/(c - 2) = s^2$
a от α	1	$(a - \alpha)^2 \sum_{i=1}^c n_i = n(a - \alpha)^2 = S_a^2$	$S_a^2/1 = s_a^2$
b от β	1	$(b - \beta)^2 \sum_{i=1}^c n_i (x_i - \bar{x})^2 = S_b^2$	$S_b^2/1 = s_b^2$
Общая дисперсия	n	$\sum_{i,j} (y_{ij} - \mu_i)^2$	—

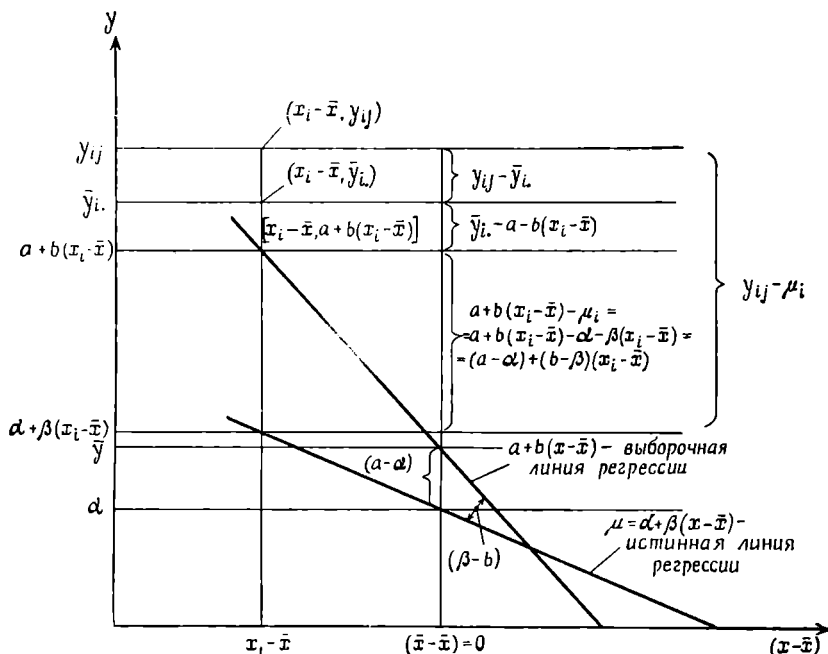
Следовательно, оба критерия для проверки гипотез $\alpha=0$ и $\beta=0$ соответственно распределены по закону Стьюдента с $c - 2$ степенями свободы. Эти критерии даются выражениями

$$\sqrt{\frac{a^2 n}{\sum_{i=1}^c n_i [\bar{y}_{i.} - a - b(x_i - \bar{x})]^2 / (c - 2)}} = \frac{s_a}{s}; \quad (8.18)$$

$$\sqrt{\frac{b \sum_{i=1}^c n_i (x_i - \bar{x})^2}{\sum_{i=1}^c n_i [\bar{y}_{i.} - a - b(x_i - \bar{x})]^2 / (c - 2)}} = \frac{s_b}{s}. \quad (8.19)$$

Известно, что корень квадратный из случайной величины, распределенной по закону Фишера (F -распределение), распределен как модуль величины, подчиненной распределению Стьюдента с 1 и $c - 2$ степенями свободы.

Таким образом, выражения (8.18) и (8.19) свидетельствуют о связи между критериями для проверки гипотез $\alpha=0$ и $\beta=0$ и дисперсионным анализом, так как отношения s_a^2/s^2 и s_b^2/s^2



Фиг. 8.2. Графическое изображение типов изменчивости при линейной регрессии с одной независимой переменной и n наблюдениями величины $x_i - \bar{x}$.

$y_{ij} - \mu_i = [\dots]$ — отклонение любого наблюдения от соответствующей средней или математического ожидания;

$y_{ij} - \bar{y}_i$ — отклонение любого наблюдения от среднего по группе;

$[\bar{y}_i - a - b(x_i - \bar{x})]$ — отклонение среднего по группе от выборочной линии регрессии;

$(a - \alpha) + (b - \beta)(x_i - \bar{x})$ — отклонение выборочной регрессии от истинной линии регрессии;

$\bar{y}_i - \mu$ — отклонение среднего по группе от истинной регрессии.

F -распределены с 1 и $s - 2$ степенями свободы, что также может быть использовано для проверки этих гипотез.

Критерии для проверки предположений о линейной регрессии

Успех применения линейного регрессионного анализа в значительной степени зависит от того, сколь легко могут быть

проверены и использованы введенные теоретические предположения. Они сводятся к следующему:

- 1) y_{ij} распределены нормально со средним μ_i и дисперсией σ_i^2 ;
- 2) μ_i дается в форме $\alpha + \beta(x_i - \bar{x})$;
- 3) $\sigma_i^2 = \sigma^2$ для всех значений i .

Прежде чем приступить к весьма трудоемкому анализу, рекомендуется проверить справедливость сделанных предположений, так как без этой проверки впоследствии может оказаться, что статистический вывод неверен из-за невыполнения предположений, лежащих в основе анализа.

Нормальность

Каждому значению x (допустим, x_i) соответствует n_i наблюдаемых значений y . Таким образом, каждому значению x_i соответствует распределение случайной величины y независимо от значения n_i . Если n_i достаточно велико, для условного распределения $f(y|x)$ может быть проверена гипотеза о согласованности с нормальным законом (гл. 6). Однако если n_i мало, то проверку гипотезы о нормальности провести нельзя и придется основываться на непроверенном предположении о нормальности¹.

Равенство дисперсий

Так как каждому значению x_i соответствует распределение случайной величины y , то при $n_i > 1$ всегда можно получить оценку для σ_i^2 . Пусть

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$$

будет оценкой для σ_i^2 . Тогда отношение любой пары выборочных дисперсий будет распределено по закону Фишера при условии, что гипотеза о нормальности проверена и принята или что просто она введена в качестве предположения. Если значений x больше 2, нулевая гипотеза H_0 будет $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_c^2 = \sigma^2$, и ее можно проверить при помощи критерия Бартлетта (гл. 6).

Линейность

Допустим $\mu_i \neq \alpha + \beta(x_i - \bar{x})$. Рассмотрим выражение

$$\bar{y}_{i.} - E\bar{y}_{i.} = \bar{y}_{i.} - \mu_i.$$

¹ Трудности, связанные с проверкой гипотезы о нормальности по малым выборкам, описаны в статье А. А. Петрова (см. «Теория вероятностей и ее применение», 1, 2, 1956). — *Прим. ред.*

Запишем его в следующем виде:

$$\begin{aligned} \bar{y}_{i.} - a - b(x_i - \bar{x}) - \mu_i + \alpha + \beta(x_i - \bar{x}) + a + b(x_i - \bar{x}) - a - \\ - \beta(x_i - \bar{x}) = \{[\bar{y}_{i.} - a - b(x_i - \bar{x})] - [\mu_i - \alpha - \beta(x_i - \bar{x})]\} + \\ + [(a - \alpha) + (b - \beta)(x_i - \bar{x})]. \end{aligned} \quad (8.20)$$

Так как величина $\sum_{i=1}^c n_i (\bar{y}_{i.} - \mu_i)^2$ распределена как $\sigma^2 \chi^2$ с c степенями свободы, то, следовательно,

$$E \sum_{i=1}^c n_i (\bar{y}_{i.} - \mu_i)^2 = \sigma^2 c,$$

а, возведя в квадрат (8.20) и затем умножив результат на n_i , после суммирования по i получим

$$\begin{aligned} \sum_{i=1}^c n_i [\bar{y}_{i.} - a - b(x_i - \bar{x}) - \mu_i + \alpha + \beta(x_i - \bar{x})]^2 + \\ + (a - \alpha)^2 \sum_{i=1}^c n_i + (b - \beta)^2 \sum_{i=1}^c n_i (x_i - \bar{x})^2 = \sum_{i=1}^c n_i (\bar{y}_{i.} - \mu_i)^2. \end{aligned}$$

Каждое из трех полученных слагаемых распределено как $\sigma^2 \chi^2$ с $c - 2$, 1 и 1 степенями свободы соответственно.

Таким образом,

$$\begin{aligned} E \sum_i n_i [\bar{y}_{i.} - a - b(x_i - \bar{x}) - \mu_i + \alpha + \beta(x_i - \bar{x})]^2 = \sigma^2 (c - 2), \\ E \sum_i n_i [\bar{y}_{i.} - a - b(x_i - \bar{x})]^2 + E \sum_i n_i [\mu_i - \alpha - \beta(x_i - \bar{x})]^2 - \\ - 2 \sum_i n_i [\mu_i - \alpha - \beta(x_i - \bar{x})] E [\bar{y}_{i.} - a - b(x_i - \bar{x})] = \sigma^2 (c - 2); \end{aligned}$$

тогда

$$E [\bar{y}_{i.} - a - b(x_i - \bar{x})] = \mu_i - \alpha - \beta(x_i - \bar{x}).$$

Следовательно,

$$\begin{aligned} \sigma^2 (c - 2) = E \sum_i n_i [\bar{y}_{i.} - a - b(x_i - \bar{x})]^2 - \\ - \sum_i n_i [\mu_i - \alpha - \beta(x_i - \bar{x})]^2 \end{aligned}$$

и

$$\frac{E \sum_{i=1}^c n_i [\bar{y}_{i.} - a - b(x_i - \bar{x})]^2}{c - 2} = \sigma^2 + \frac{\sum_{i=1}^c n_i [\mu_i - \alpha - \beta(x_i - \bar{x})]^2}{c - 2}.$$

Итак, если $\mu_i \neq \alpha + \beta(x_i - \bar{x})$, математические ожидания средних квадратов уклонений от эмпирического уравнения регрессии $E(s^2)$ будут равными не σ^2 , а σ^2 плюс уклонение μ_i от линии $\alpha + \beta(x_i - \bar{x})$. С другой стороны, если $H_1: \mu_i + \beta(x_i - \bar{x})$, то

$$Es^2 = \sigma^2 + \frac{1}{c-2} \sum_{i=1}^c n_i (\mu_i - \mu_i) = \sigma^2.$$

Поэтому, чтобы проверить гипотезу $\mu_i = \alpha + \beta(x_i - \bar{x})$, представляющую собой предположение о линейности, можно воспользоваться отношением s_g^2/s (табл. 8.2), которое, если верна гипотеза H_1 , распределено как

$$F \left[\begin{array}{c} \sum_{i=1}^c (n_i - 1) \\ c - 2 \end{array} \right].$$

С тех пор как была опубликована работа Хаксли [128] «Проблемы относительного роста», появилось множество статей, в которых рассматривается этот вопрос в различных аспектах. Некоторые из этих работ были посвящены изучению ископаемых организмов: можно назвать, например, работы Робба [233], Плиджера и Патнэма [212] и Грея [101]. Авторы более ранних работ предполагали, что отношение размеров двух и более частей скелета можно использовать как характеристику вида или более крупной таксономической единицы. Однако до появления работ Рива [229, 230] более строгую обработку данных никто не проводил. Формализацией такой обработки занимались Тиссьер [271], Кермак и Халдейн [143], а также Олсон и Миллер [198]. Позднее этот подход использовали Кермак [144], Пэстилс [206] и Имбри [130].

Многие работы, посвященные соотношениям размеров частей скелета животных, основаны на предположении, согласно которому изменение отношения любых двух частей можно выразить уравнением

$$y = bx^k.$$

Если $k=1$, то данная часть организма растет с одной и той же скоростью, что видно из уравнения, которое становится линейным: $y = a + bx$ ¹. Оно определяет *изогонический* рост. Если же одна из частей скелета растет со скоростью, отличающейся от скорости роста другой части, коэффициент k будет отличаться от единицы и уравнение примет вид $y = a + bx^k$. Оно называется уравнением *гетерогонического* роста.

¹ Подойдем формально к аллометрическому росту по Хаксли [128] и Кермаку [144]. Рассмотрим два измеренных морфологических признака x и y . Скорость роста x есть изменение x за определенное время (возраст) и выражается как dx/dt . Аналогично и для y скорость роста dy/dt . Относительные скорости роста можно представить в виде $(1/x)(dx/dt)$ и $(1/y)(dy/dt)$.

Это уравнение было детально изучено Хаксли [128], который первым использовал его для изучения величины отношения роста какой-нибудь части скелета (например, конечности) к росту всего скелета как целого.

Аналогично уравнение относительного роста можно применить к весу скелета или его длине. Однако в более поздних работах Хаксли применял его к любым двум частям организма. Этот вопрос освещен, например, в работах Ричардса [228], Рива [230], Феллера [92] и Ламера и др. [170]. Кроме того, Олсон и Миллер [198], исследовавшие ископаемые рептилии и амфибии, показали, что уравнение $y = a + bx$ во многих случаях оказалось удовлетворительным. Даже в таких случаях, когда k превышает единицу ($k = 1,25$) или, наоборот, меньше ее ($k = 0,75$), гетерогоническое уравнение, по-видимому, дает не лучшее приближение.

Таким образом, необходим критерий для проверки линейности уравнения наилучшего приближения пар измерений. Ниже мы рассмотрим критерии для проверки предположения о линейности линии регрессии, а также методы сравнения коэффициентов регрессии по данным различных выборок.

Изучение относительного роста любых двух частей скелета ископаемых животных может быть разделено на две части:

А. Таксономическое изучение, в котором модели роста рассматриваются как групповые характеристики.

Б. Интерпретация моделей относительного роста.

а) Два или более морфологических признака могут быть измерены в выборках индивидуумов одного и того же вида, но

Согласно Хаксли, отношение двух относительных скоростей роста должно быть равно константе, которую обозначим через k . Тогда

$$\frac{\frac{1}{y} \left(\frac{dy}{dt} \right)}{\frac{1}{x} \left(\frac{dx}{dt} \right)} = k, \quad \text{или} \quad \frac{dy}{y dt} = k \left(\frac{dx}{x dt} \right).$$

Но, так как

$$\frac{\frac{dy}{dt}}{\frac{dx}{dt}} = \frac{dy}{dx},$$

получим

$$\frac{x}{y} \left(\frac{dy}{dx} \right) = k_1.$$

Решая

$$\int \frac{dy}{y} = k \int \frac{dx}{x},$$

получим, что

$$\ln y = k \ln x + \ln c, \quad \text{или} \quad y = cx^k,$$

определяющее c как символ b из аллометрического уравнения роста.

существовавших в различных условиях, таких, например, как условия илистого, песчаного или каменистого дна. Влияние этих факторов составляет основной объект исследования (см., например, [144, 75]).

б) При палеонтологических исследованиях может возникнуть необходимость в выяснении влияния пространственного распределения организмов на их морфологию.

Так, например, может потребоваться изучить морфологию отдельно живущих индивидуумов и особей, живущих колониями. Возможно, что анализ относительного роста окажет помощь при решении задач подобного типа. Например, отношение размеров различных морфологических признаков может меняться в результате близкого физического контакта с другими индивидуумами, как это бывает у *Ostrea*. Хотя такая реакция и не имеет ничего общего с наследственностью, ее влияние отражается на индивидуумах. Возможно, что изменения в организме, возникшие после его рождения, под влиянием окружающих условий, как это предполагается согласно пунктам «а» и «б», будут неправильно интерпретированы как доказательство проявления наследственности и, следовательно, эволюционного развития.

в) При исследовании относительного роста основное внимание должно быть сосредоточено на функционировании организма и его развитии. Вот один из примеров возможного подхода к данному типу задач: если произошли небольшие изменения в комплексе генов, то 1) исходное состояние организма будет отличаться от исходного состояния организмов его предков; 2) на ранних стадиях роста результаты генетических изменений проявятся в дифференциации частей организма, и влияние этих изменений будет увеличиваться по мере того, как ростом будет усложняться организм; 3) во взрослой форме проявится результат относительных изменений, происшедших в процессе развития.

Относительное значение исходного онтогенетического изменения может быть косвенно измерено с помощью фиксации результирующего изменения в морфологии взрослой особи [198]. Считая все изложенное как обоснование постановки задачи, рассмотрим различные типы линий регрессии, а также статистические критерии, которые могут быть использованы при анализе и интерпретации этих линий.

РЕГРЕССИОННЫЕ МЕТОДЫ

Детальный анализ относительного роста частей ископаемых организмов принадлежит Олсону и Миллеру [198], которые в своей работе рассматривают критерий для проверки гипотезы

о линейности, а также методы изучения углов наклона, точек пересечения и сравнения двух линий регрессии. Мы не касаемся этих методов, так как им обычно предпочитают метод сокращенной главной оси, который изложен ниже.

ЕДИНАЯ ЛИНИЯ ОРГАНИЧЕСКОЙ КОРРЕЛЯЦИИ (СОКРАЩЕННАЯ ГЛАВНАЯ ОСЬ)

Согласно Кермаку [144], функцию и аргумент при изучении морфологических признаков нередко выбирают без реального обоснования, что сильно осложняет использование условной линии регрессии. Относительно такого, например, утверждения: увеличение длины черепа зависит от увеличения длины тела — нельзя иметь отчетливого суждения; более реально рассматривать изменения длины черепа и длины тела как результат действия совокупности общих факторов.

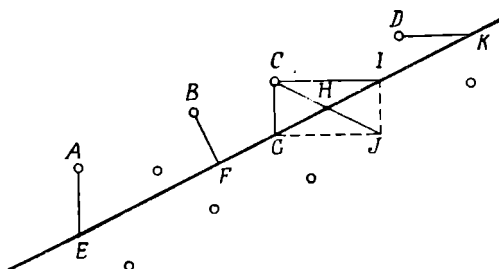
В случаях применения обычных методов регрессии наклон регрессионной линии непосредственно зависит от того, какой из двух морфологических признаков выбран в качестве аргумента. Так, например, результаты изучения скоростей роста будут зависеть от выбора независимой переменной. В случаях слабой корреляционной зависимости сложность увеличивается из-за большого угла между возможными линиями регрессии. Пирсон¹ вместо обычных линий регрессии использовал единую линию, которая является главной осью корреляционной поверхности. Продолжая мысль Пирсона, Тиссер [271] предложил единую линию, названную сокращенной главной осью и обладающую свойствами, более подходящими для анализа пар морфологических признаков, чем главная ось Пирсона.

Свойства сокращенной главной оси, называемой также единой линией органической корреляции, изучены Кермаком и Халдейном [143] и Красклом [160]. Этот вопрос детально рассмотрел Имбри [130], который привел примеры применения как регрессионных методов, так и методов сокращенной главной оси. На фиг. 9.1, взятой из указанной работы, показана зависимость между различными методами построения линии на точечной диаграмме.

Если при просмотре точечной диаграммы окажется, что прямая линия типа $y = b + Kx$ неудовлетворительно описывает нанесенные точки, то следует построить новую диаграмму, но на логарифмической бумаге (с логарифмическим масштабом на обеих осях). Если в этом случае полученная прямая окажется удовлетворительной, ее можно выразить уравнением вида

¹ K. Pearson, On lines and planes of closets fit to systemes of points. in space, Phil. Mag., 6, № 2, 1901.

$\lg y = \lg b + k \lg x$, которое эквивалентно аллометрическому уравнению $y = bx^k$. Называемые ниже статистики могут быть точно так же использованы в новом логарифмическом уравнении. В случае когда уравнение сокращенной главной оси дается



Ф и г. 9.1. Диаграмма, иллюстрирующая различные методы подбора линии для рассеянных точек. Регрессия y на x минимизирует сумму квадратов отклонений, измеряемых как расстояние AE . Линия регрессии x на y минимизирует соответствующую сумму квадратов отклонений, измеряемых как DK . Главная ось минимизирует сумму квадратов отклонений, измеряемых как BF . Сокращенная главная ось обеспечивает минимум суммы площадей треугольников GCI [130].

в форме прямой $y = b + Kx$ (изогоническое уравнение роста), желательно применение следующих статистик.

Статистики

Тангенс угла наклона

$$K = \frac{s_y}{s_x},$$

где s_y — оценка стандартного отклонения случайной величины y , а s_x — аналогичная оценка для x . Это выражение может быть представлено в более удобной для счета форме:

$$K = \sqrt{\frac{\sum y^2 - \bar{y} \sum y}{\sum x^2 - \bar{x} \sum x}}.$$

Стандартная ошибка s_K этой оценки может быть выражена в виде

$$s_K = \frac{s_y}{s_x} \sqrt{\frac{1 - r^2}{n}},$$

где r^2 — квадрат оценки коэффициента корреляции, а n — объем выборки.

Точка пересечения

$$b = \bar{y} - \bar{x}K.$$

Стандартную ошибку s_b оценки точки пересечения можно определить из выражения

$$s_b = s_y \sqrt{\frac{1-r^2}{n} \left(1 + \frac{\bar{x}^2}{s_x^2}\right)}.$$

Дисперсия относительно сокращенной главной оси s_d

$$s_d = \sqrt{2(1-r)(s_x^2 + s_y^2)}.$$

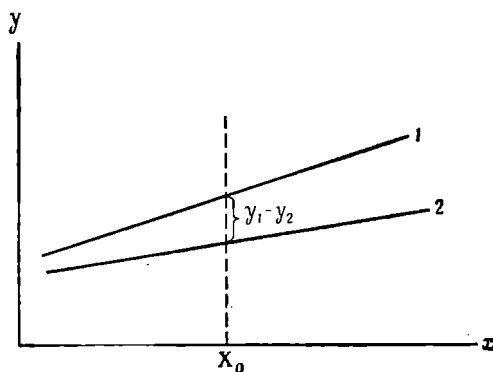
Сравнение наклона двух сокращенных главных осей. Проверяемая гипотеза заключается в том, что расхождение между K_1 и K_2 можно принять равным нулю.

Для проверки этой гипотезы

1. Вычисляем статистику

$$Z = \frac{K_1 - K_2}{\sqrt{s_{K_1}^2 + s_{K_2}^2}}.$$

2. Критическое значение Z при соответствующей альтернативе и заданном уровне значимости находим из таблицы значений нормального распределения.



Ф и г. 9.2.

Проверка гипотезы, заключающейся в том, что две сокращенные главные оси построены по выборкам, взятым из одной популяции, т. е. что различия между двумя выборочными линиями роста относятся к расхождению между выборочными характеристиками, а не к существующим на самом деле различиям. Из всех описанных методов проверки этой гипотезы мы приведем здесь только метод Имбри [130] как относительно простой в работе и последующей интерпретации. Сущность критерия

заклучается в выборе на оси x точки X_0 ¹ (в соответствии с биологическими предпосылками) и оценке разницы между линиями в этой точке при помощи отношения

$$\frac{\text{Различие по вертикали между линиями в точке } X_0}{\text{Стандартная ошибка этого различия}}.$$

В частности, чтобы проверить гипотезу, согласно которой в точке X_0 истинные линии роста совпадают, нужно вычислить следующую статистику:

$$Z = \frac{X_0(K_1 - K_2) + (b_1 - b_2)}{\sqrt{s_{K_1}^2(X_0 - \bar{x}_1)^2 + s_{K_2}^2(X_0 - \bar{x}_2)^2}}.$$

Критическое значение для Z , соответствующее заданному уровню значимости, находят из таблиц нормального распределения.

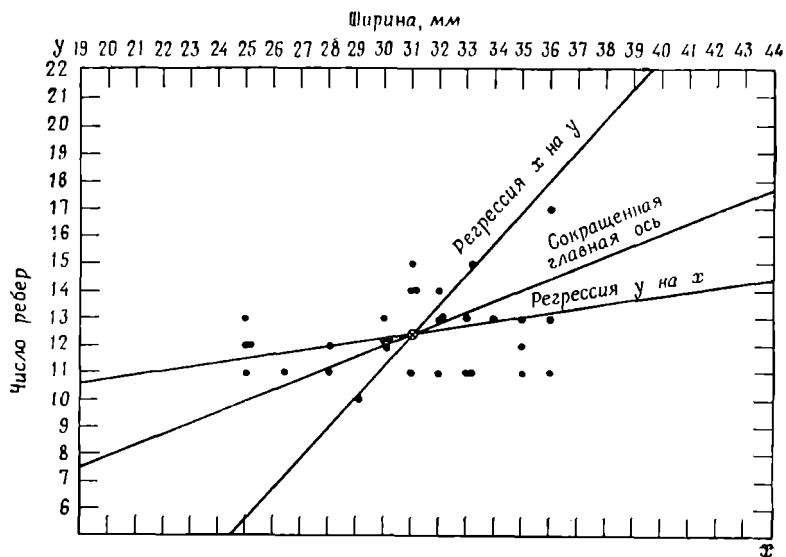
Пример

В качестве примера, иллюстрирующего применение сокращенной главной оси, и проверки существенности расхождения между осями можно рассмотреть случай, когда из коллекции брахиопод *Spirifer pennatus* в количестве 1000 экземпляров, отобранных из одного района, взяты две выборки, содержащие по 40 особей каждая. В качестве признаков рассматривались максимальная ширина раковины и число ребер спинной створки.

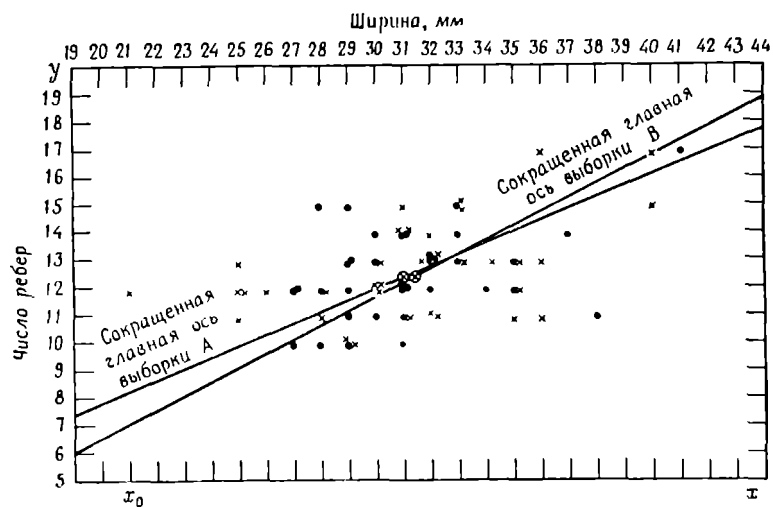
Пусть x — максимальная ширина, y — число ребер. Необходимые данные приведены в табл. 9.1, по которой вычислены характеристики сокращенных главных осей для каждой выборки, а для выборки A , кроме того, построены две линии регрессии y на x и x на y (табл. 9.2).

На фиг. 9.3 показана сокращенная главная ось, построенная по данным выборки A , а также обычные линии регрессии y на x и x на y , которые значительно различаются углами наклона. С этим различием и связаны главные трудности применения обычного регрессионного анализа, когда нет достаточных оснований для объявления одной переменной аргументом, а другой функцией от нее.

¹ Выбор значения X_0 зависит от природы задачи. Например, исследователь, пытающийся показать, что между двумя линиями существует значимое различие, выбирает в качестве X_0 ту из точек пересечения прямых с осью x , в которой вертикальное различие между линиями минимально [130]. В связи с тем что выбор точки X_0 носит субъективный характер, а применяемая статистика основана на предположении, что X_0 выбирается заранее, в применении этого метода существует опасность «порочного круга».



Фиг. 9.3. Сокращенная главная ось A ($y = 0,2821 + 0,4091x$).



Фиг. 9.4. Сокращенная главная ось B ($y = -3,8034 + 0,5165x$).

Таблица 9.1

	Выборка А	Выборка В
$\sum_{i=1}^n x_i$	1 240,00	1 257,00
$\sum_{i=1}^n y_i$	496,00	497,00
$\sum x_i^2$	39 011,00	39 861,00
$\sum y_i^2$	6 246,00	6 273,00
$\sum xy$	15 461,00	15 677,00
$\bar{x}$	31,00	31,42
$\bar{y}$	12,40	12,425
n	40	40
s_x	1,566	1,582
s_y	3,828	3,063

Таблица 9.2

	Выборка А	Выборка В
r_{xy}	0,364	0,312
K	0,4091	0,5165
s_K	0,0603	0,0776
b	-0,2821	-3,8034
s_b	4,9047	6,8034
Регрессия ¹		
у на x . . .	0,1489	
Пересечение у		
на x	7,80	
Регрессия x		
на y	0,8891	
Пересечение x		
на y	19,98	

¹ Имеется в виду также угол наклона линии регрессии y на x (и ее пересечение с осью $y=b$). — *Прим. ред.*

На фиг. 9.4 показаны сокращенные главные оси обеих выборок (А и В), для которых сравниваются углы наклона.

Сравнение характеристик наклона в двух выборках (А и В)

$$Z = \frac{K_A - K_B}{\sqrt{s_{K_A}^2 + s_{K_B}^2}} = \frac{0,4091 - 0,5165}{\sqrt{0,0603^2 + 0,0776^2}} = -1,0926.$$

Абсолютное значение Z равно 1,0926. Примерно в 28% случаев будет наблюдаться такое же или большее значение $|Z|$. Поэтому нет оснований рассматривать выборочные тангенсы K_A и K_B как существенно отличающиеся друг от друга, что позволяет считать обе линии обладающими одним и тем же углом наклона. Естественно, что в данном случае не следовало бы ожидать существенных расхождений, так как обе выборки взяты из большой совокупности особей одного вида и из одного района.

Сравнение двух сокращенных главных осей по выборкам А и В

Следуя рекомендации Имбри [130], выберем наибольшее различие между значениями ординат двух сравниваемых выборочных осей, соответствующее фиксированному значению X_0 . Это

различие выбирается внутри области значений результатов измерения. По мнению авторов, экстраполяция линий в точке их пересечения с осями в рассматриваемом случае является неоправданной. Дело в том, что в выборке представлены только поздние онтогенетические стадии, и мы не можем судить по имеющимся данным о распределении X и Y на ранних стадиях развития. Предлагаемый метод позволяет сравнивать наклоны осей более точно, чем изложенный выше.

Рассмотрев фиг. 9.4, выберем $X_0=21$, т. е. то значение ширины раковины, которое является наименьшим в объединении двух выборок. При $X_0=21$ $Y_A - Y_B = 1,2$ (фиг. 9.2)

$$Z = \frac{21 \cdot (0,4091 - 0,5165) + [(-0,2821) - (-3,8034)]}{\sqrt{(0,0603)^2 (21 - 31)^2 + (0,0776)^2 (21 - 31,42)^2}},$$

$$Z = 1,2558.$$

При условии что гипотеза об однородности верна, такое же или большее значение $|Z|$ будет наблюдаться примерно в 21% случаев. Поэтому отвергать гипотезу нет оснований, и выборки A и B следует рассматривать как взятые из одной гомогенной популяции. Имбри¹ применял изложенный метод при изучении земноводных триаса.

ВЫВОДЫ. РЕГРЕССИОННЫЙ АНАЛИЗ В ПАЛЕОБИОМЕТРИИ

Точечные диаграммы, на которых каждая точка представляет собой результат измерения двух морфологических признаков у одной особи, можно проанализировать как на субъективной основе, так и с помощью одного из формальных методов. Гипотеза относительного роста позволяет рассматривать подбор линий, «наилучшим образом» описывающих точечные диаграммы, как объективный шаг на пути содержательной интерпретации данных. Эта линия может быть выражена, например, уравнением $y = b + Kx$ или уравнением $y = bx^k$. В случае любого из уравнений параметры имеют биологический смысл. Так, например, b представляет собой исходное значение показателя роста, а k — скорость роста [128, 144].

Модель, развитая Олсоном и Миллером [198], иллюстрирует зависимость между скоростями роста отдельных морфологических признаков и точечными диаграммами. При рассмотрении этой модели нетрудно видеть, что результаты анализа точеч-

¹ E. H. Colbert, J. Imbrie, Triassic Metoposaurid amphibians, Bull. Am. Mus. Nat. Hist., 110, 1956.

ных диаграмм могут вызвать сомнение в случае изучения ископаемых, особенно когда выборки малы, а изменение скоростей роста индивидуумов значительно. Это сомнение большей частью бывает следствием того, что палеонтолог обычно не знает (или знает лишь в очень широких пределах) возраст ископаемой особи.

Следует отметить, что формальный анализ точечных диаграмм, проведенный или с помощью обычных методов регрессии, или методом сокращенной главной оси, является широко распространенным и важным приемом исследования палеонтологических данных.

В следующей главе будет показано, что регрессионный и корреляционный анализы тесно связаны с дисперсионным анализом.

АНАЛИЗ КОВАРИАЦИИ

Постановка задачи

Наблюдения в геологических науках нередко бывают представлены в виде данных об изменении более чем одного признака в одном элементе выборки. Таким образом, результат одного наблюдения может состоять из k значений, и такое k -мерное наблюдение может быть повторено n раз (гл. 5).

Так, Милтон¹ (Колумбийский университет) с целью выяснить зависимость между характеристиками климата, свойствами поверхности и геоморфологическими условиями рассматривал совместно шесть переменных.

При изучении шума моря в бухте Нарраганзетт-Бей (Род-Айленд) Дитц, Кан и Берч² вели наблюдения над 19 звуковыми частотами, направлением и скоростью ветра, высотой волн, фазами прилива и количеством выпадаемых осадков. Всего они одновременно рассматривали 24 величины.

В любом из приведенных выше экспериментов одна из k переменных может действительно быть функцией (в математическом смысле) от другой переменной. Естественно, что для выбора функции и аргумента нужны некоторые сведения об изучаемых величинах, до получения которых от принятия решения лучше воздержаться.

Каждая из изучаемых величин может быть проанализирована отдельно от других и различными способами. Например, обработку результатов наблюдения значений непрерывной случайной величины при выполнении соответствующих предположений можно провести с помощью дисперсионного анализа (гл. 7) как одного из возможных методов. Однако возможно, что зависимости между упомянутыми выше k -переменными

¹ M. A. Melton, An analysis of the relations among elements of climate surface properties, and geomorphology, Techn. Rept. № 11, Dept. of Geol., Columbia Univ., New York, 1957.

² E. T. Dietz, J. S. Kahn, W. B. Birch, Effect of wind on shallow water ambient noise, Abst. of paper pres. at A. S. meet. at Brown Univ., June 9, 1960.

окажутся такими, что они будут влиять на изменчивость отдельных случайных величин, образующих k -мерный комплекс.

Зависимости между парами случайных величин могут быть выражены при помощи линий регрессии (гл. 8), а также коэффициентов корреляции (гл. 6). Влияние одной переменной на наблюдаемые значения другой может вызвать существенные различия в результатах наблюдения над одной из случайных величин, тогда как на самом деле они вызваны не действительно имеющимися различиями в значениях этой величины, а расхождениями в значениях коррелируемой переменной. Следовательно, было бы весьма желательным прежде всего устранить различия между величинами, которые возникают за счет зависимости одной переменной от другой. Исправленные таким образом данные следовало бы проанализировать снова, чтобы можно было судить об эффекте устранения этих различий.

Эти задачи могут быть решены путем применения комбинации регрессионного и дисперсионного анализов, которая называется *анализом ковариации*.

Модели

Так как ковариационный анализ является *комбинацией линейного регрессионного и дисперсионного анализов*, то математическая модель, соответствующая анализу ковариации, также представляет собой комбинацию моделей, связанных с двумя указанными типами исследования. В частности, при рассмотрении любой модели дисперсионного анализа мы просто добавляем слагаемое $\beta(x - \bar{x})$, где β — коэффициент регрессии y на x , а $\bar{x}$ — оценка генерального среднего. Используя ту же форму записи, что и в разделе о дисперсионном анализе, и модели I и II, построим математические модели ковариационного анализа.

Модель I. Все факторы фиксированы

1. Полностью рандомизированная схема для однофакторного анализа:

$$y_{ijkl\dots} = \mu + \beta(x_{ijkl\dots} - \bar{x}) + \alpha_i + \gamma_{ij} + \delta_{ijk} + \dots + \epsilon_{ijkl\dots}$$

2. Полностью рандомизированные блоки при двухфакторном анализе:

$$y_{ij} = \mu + \beta(x_{ij} - \bar{x}) + \alpha_i + \gamma_j + \epsilon_{ij}.$$

3. Латинский квадрат:

$$y_{ijk} = \mu + \beta(x_{ijk} - \bar{x}) + \alpha_i + \gamma_j + \delta_k + \epsilon_{ijk}.$$

Модель II. Все компоненты случайные

1. Полная рандомизация:

$$y_{ijkl\dots} = \mu + \beta(x_{ijkl\dots} - \bar{x}) + a_i + b_{ij} + c_{ijk} + \dots + \varepsilon_{ijkl\dots}$$

2. Полная рандомизация блока:

$$y_{ij} = \mu + \beta(x_{ij} - \bar{x}) + a_i + c_j + \varepsilon_{ij}$$

3. Латинский квадрат:

$$y_{ijk} = \mu + \beta(x_{ijk} - \bar{x}) + a_i + c_j + d_k + \varepsilon_{ijk}$$

Модель III

Эта модель охватывает группу математических построений, структура которых меняется в зависимости от характера задачи; следовательно, описать все возможные ситуации невозможно. Естественно, что условия каждого конкретного эксперимента будут определять случайные и фиксированные факторы; однако, каким бы образом ни разделялись факторы на фиксированные и случайные, добавляется слагаемое типа $\beta(x - \bar{x})$.

Иллюстрирующие примеры

Модель

В качестве примера рассмотрим следующую смешанную модель, представляющую собой комбинацию регрессионного и дисперсионного анализов:

$$Y = y_{ijkl} = \mu + \rho_i + f_{ij} + s_{ijk} + t_{ikl} + \beta(x_{ijkl} - \bar{x}),$$

где

$$E[Y|X] = \mu + \rho_i + \beta(x_{ijkl} - \bar{x})$$

при

$$X = x_{ijkl},$$

$$\text{var}[Y|X] = \sigma_f^2 + \sigma_s^2 + \sigma_t^2;$$

ρ_i — константа, характеризующая отклонение среднего i -й группы от генеральной средней; f_{ij} — случайная величина, характеризующая отклонения j -й подгруппы от средней в i -й группе; s_{ijk} — случайная величина, описывающая отклонения k -й подподгруппы от среднего в j -й подгруппе; t_{ikl} — случайные погрешности; β — коэффициент регрессии Y на X ; x_{ijkl} — фиксированные значения X , при которых производятся наблюдения над Y .

Таблица 10.1

Обобщенная таблица ковариационного анализа

Тип изменчивости	Число степеней свободы	Сумма квадратов	Сумма смешанных произведений	Коэффициент корреляции	Коэффициент регрессии
Между группами r	$r - 1$	$\sum_{ijkl} (\bar{\bar{y}}_{i...} - \bar{y})^2 = S_r^2(y)$	$\sum_{ijkl} (\bar{\bar{y}}_{i...} - \bar{y}) (\bar{\bar{x}}_{i...} - \bar{x}) = S_r(xy)$	$r_r = \frac{S_r(xy)}{\sqrt{S_r^2(y) S_r^2(x)}}$	$b_r = \frac{S_r(xy)}{S_r^2(x)}$
Между подгруппами внутри групп f	$r(f - 1)$	$\sum_{ijkl} (\bar{\bar{y}}_{ij...} - \bar{\bar{y}}_{i...})^2 = S_f^2(y)$	$\sum_{ijkl} (\bar{\bar{y}}_{ij...} - \bar{\bar{y}}_{i...}) (\bar{\bar{x}}_{ij...} - \bar{\bar{x}}_{i...}) = S_f(xy)$	$r_f = \frac{S_f(xy)}{\sqrt{S_f^2(y) S_f^2(x)}}$	$b_f = \frac{S_f(xy)}{S_f^2(x)}$
Между подподгруппами внутри подгрупп и групп s	$rf(s - 1)$	$\sum_{ijkl} (\bar{\bar{y}}_{ijk.} - \bar{\bar{y}}_{ij..})^2 = S_s^2(y)$	$\sum_{ijkl} (\bar{\bar{y}}_{ijk.} - \bar{\bar{y}}_{ij..}) (\bar{\bar{x}}_{ijk.} - \bar{\bar{x}}_{ij..}) = S_s(xy)$	$r_s = \frac{S_s(xy)}{\sqrt{S_s^2(y) S_s^2(x)}}$	$b_s = \frac{S_s(xy)}{S_s^2(x)}$
Ошибка	$rfst - 1$	$\sum_{ijkl} (y_{ijkl} - \bar{\bar{y}}_{ijk.})^2 = S_t^2(y)$	$\sum_{ijkl} (y_{ijkl} - \bar{\bar{y}}_{ijk.}) (x_{ijkl} - \bar{\bar{x}}_{ijk.}) = S_t(xy)$	$r_t = \frac{S_t(xy)}{\sqrt{S_t^2(y) S_t^2(x)}}$	$b_t = \frac{S_t(xy)}{S_t^2(x)}$
Общая дисперсия	$rfst - 1$	$\sum_{ijkl} (y_{ijkl} - \bar{y})^2 = S_T^2(y)$	$\sum_{ijkl} (y_{ijkl} - \bar{y}) (x_{ijkl} - \bar{x}) = S_T(xy)$	$r_T = \frac{S_T(xy)}{\sqrt{S_T^2(y) S_T^2(x)}}$	$b_T = \frac{S_T(xy)}{S_T^2(x)}$

Каждую переменную (как x , так и y) можно проанализировать отдельно. Общая сумма квадратов, соответствующая каждой из этих величин, может быть разделена на ряд слагаемых, характеризующих различные причины изменчивости, как это сделано в гл. 7. Далее, оценки для характеристик регрессии и корреляционных зависимостей можно получить из сумм смешанных произведений x и y .

Все сказанное обобщено в удобной для пользования табл. 10.1, в которую включены суммы квадратов для x и y , суммы смешанных произведений, оценки коэффициентов регрессии, соответствующие каждой причине изменчивости, и аналогичные значения выборочных коэффициентов корреляции.

Предполагая, что зависимость между x и y может быть описана функцией регрессии, мы должны признать, что успех анализа величины y зависит от наших возможностей устранить любое влияние на величину y , связанное с регрессией. Такой подход уменьшит ошибку при проверке значимости изучаемых различий. В табл. 10.1 перечислены типы изменчивости величины y , обусловленные различными причинами. Кроме того, для каждого типа изменчивости в этой таблице приведены оценки коэффициентов регрессии в подгруппах. Ковариационный анализ позволяет исключить любое влияние на величину y изменений величины x , обусловленных регрессией по всем типам изменчивости.

Сумма квадратов, соответствующая общей изменчивости величины y , без поправки на регрессию дается выражением $\sum_{ijkl} (y_{ijkl} - \bar{y})^2$, тогда как величина $\sum_{ijkl} [y_{ijkl} - \bar{y} - b_T(x_{ijkl} - \bar{x})]^2$, где величина b_T определена как отношение

$$b_T = \frac{\text{Общая сумма смешанных произведений}}{\text{Общая сумма квадратов величины } x},$$

представляет собой исправленную общую сумму квадратов величины y после устранения регрессионного эффекта.

Исправленная сумма квадратов Проверка значимости уменьшения ошибки

Исправленные суммы квадратов можно свести в табл. 10.2. Заметим, что каждая сумма квадратов после исправления теряет одну степень свободы за счет введения в сумму величины, зависящей от оценки $\bar{x}$. Любую из исправленных сумм квадратов теперь можно сравнить с соответствующей неисправленной суммой. Схематически это показано в табл. 10.3, в которой неисправленные суммы квадратов разбиты на две части, т. е. на

ошибку, вносимую регрессией, и исправленную сумму квадратов.

Сокращение влияния этой ошибки обеспечивает увеличение точности эксперимента.

Т а б л и ц а 10.2

Исправление суммы квадратов

Тип ошибки	Число степеней свободы	Исправленная сумма квадратов
Между подгруппами внутри групп	$rf - r - 1$	$\sum_{ijkl} [\bar{y}_{ij..} - \bar{y}_{i...} - b_f (\bar{x}_{ij..} - \bar{x}_{i...})]^2$
Между подподгруппами внутри подгрупп	$rfs - rf - 1$	$\sum_{ijkl} [\bar{y}_{ijk.} - \bar{y}_{ij..} - b_s (\bar{x}_{ijk.} - \bar{x}_{ij..})]^2$
Ошибка эксперимента	$rfst - rfs - 1$	$\sum_{ijkl} [y_{ijkl} - \bar{y}_{ijk.} - b_t (x_{ijkl} - \bar{x}_{ijk.})]^2$

Т а б л и ц а 10.3

Проверка значимости поправки для сумм квадратов

Тип изменчивости	Число степеней свободы	Сумма квадратов	Средний квадрат
Неисправленная ошибка (между подгруппами)	$r(f - 1)$	$S_{f(y)}^2$	$S_{f(y)}^2 / r(f - 1)$
Регрессия	1	$b_f S_{f(xy)}$	$b_f S_{f(xy)}$
Исправленное значение	$r(f - 1) - 1$	$S_{f(y)}^2 - b_f S_{f(xy)}$	$\frac{S_{f(y)}^2 - b_f S_{f(xy)}}{r(f - 1) - 1}$

Критерий для проверки гипотезы, согласно которой коэффициент регрессии равен нулю, есть

$$F_{[r(f-1)-1]} = \frac{b_f S_{f(xy)}}{(S_{f(y)}^2 - b_f S_{f(xy)}) / r(f-1)}.$$

Так как отношение среднего квадрата, обусловленного влиянием регрессии, к исправленному среднему квадрату распределено по закону Фишера с 1 и $r(f - 1) - 1$ степенями свободы, если влияние регрессии незначительно, то этим можно воспользоваться для проверки гипотезы о равенстве коэффициента регрессии нулю.

Процедура исправления значений

Предположив, что коэффициент регрессии существенно отличается от нуля, мы можем исправить значения y , которые соответствуют изменениям x . Исправления, соответствующие данным любого типа изменчивости, могут быть внесены при помощи коэффициента регрессии, т. е. путем применения коэффициентов регрессии к неисправленным значениям. Процедура проста: она заключается в исправлении значений y , соответствующих значениям x , при помощи величины скорости изменения y на единицу изменения x (коэффициент регрессии y на x для данного уровня изменчивости). Например, если среднее значение x равно $\bar{x}$, то $x - \bar{x}$ есть число, на которое x отличается от $\bar{x}$ в большую или меньшую сторону, и если b — коэффициент регрессии y на x , то $b(x - \bar{x})$ — поправка для y . Таким образом, значение выражения $y - b(x - \bar{x})$ является исправленным. Конечно, если величина b несущественно отличается от нуля, эта процедура исправления неприменима.

Таблица 10.4

Сравнение исправленных сумм квадратов

Тип изменчивости	Число степеней свободы	Сумма квадратов и смешанных произведений		
		S_y^2	S_{xy}	S_x^2
Частный уровень изменчивости, например в группах	$r - 1$	$S_{r(y)}^2$	$S_{r(xy)}$	$S_r^2(x)$
Ошибка в группах (подгруппах)	$r(f - 1)$	$S_{f(y)}^2$	$S_{f(xy)}$	$S_f^2(x)$
Неисправленная сумма квадратов в группах	$rf - 1$	$S_{r(y)}^2 + S_{f(y)}^2$		
Регрессия	1	$\frac{[S_{r(xy)} + S_{f(xy)}]^2}{S_r^2(x) + S_f^2(x)}$		
Исправленная сумма квадратов в группах	$rf - 2$	$S_{r(y)}^2 + S_{f(y)}^2 - \frac{[S_{r(xy)} + S_{f(xy)}]^2}{S_r^2(x) + S_f^2(x)}$		
Исправленная сумма квадратов	$rf - r - 1$	$S_{f(y)}^2 - \frac{[S_{f(xy)}]^2}{S_f^2(x)} = S_{f(y_c)}^2$		
Приведенная сумма квадратов по группам	$r - 1$	$S_{r(y)}^2 - \frac{[S_{r(xy)} + S_{f(xy)}]^2}{S_r^2(x) + S_f^2(x)} + \frac{S_{f(xy)}}{S_f^2(x)} = S_r^2(y_c)$		

Суммы квадратов для приведенных средних

Если ошибка, исключенная путем введения коэффициента регрессии, была большой (табл. 10.3), по исправленным значениям y можно проверить следующие гипотезы:

1) Параметры, оцененные при помощи приведенных средних в различных группах изменчивости, различаются несущественно.

2) Дисперсионные компоненты равны нулю в предположении, что они содержатся в рассматриваемой математической модели. Чтобы проверить эти гипотезы, нужно сначала подсчитать приведенные к общему значению x суммы квадратов y , которые соответствуют различным уровням изменчивости. Это можно сделать, исправив суммы квадратов, соответствующих этим уровням. Схематически это показано в табл. 10.4.

Проверка гипотез $\mu_1 = \mu_2 = \dots = \mu_k$ и $\sigma_j^2 = 0$

Метод проверки этих двух гипотез одинаков. Критерием служит частное от деления исправленного среднего квадрата для изучаемых приведенных средних на исправленный средний квадрат. Полученная величина, если гипотеза справедлива, распределена с тем же числом степеней свободы, что и F -отношение, приведенное в начале раздела о ковариационном анализе (табл. 10.4).

МНОЖЕСТВЕННАЯ РЕГРЕССИЯ И МНОЖЕСТВЕННЫЙ КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Множественная регрессия

Функциональная зависимость нескольких переменных нам известна. Примером может служить уравнение $pV = nRT$, выражающее зависимость между давлением p , объемом V и температурой T идеального газа; R — газовая постоянная. Задача, которая встает перед нами при многих экспериментах, состоит в предсказании, основанном на эмпирических данных. Простая линейная зависимость между двумя переменными, рассмотренная выше, может быть распространена на случай любого числа рассматриваемых величин. Таким образом, общее выражение можно записать в следующей форме:

$$E(y) = \mu = \alpha + \beta_1(x_1 - \bar{x}_1) + \beta_2(x_2 - \bar{x}_2) + \dots + \beta_k(x_k - \bar{x}_k).$$

В частности,

$$E[y_{ij} | (x_{i1}, x_{i2}, \dots, x_{ik}) = X_i] = \mu_i =$$

$$= \alpha + \beta_1(x_{i1} - \bar{x}_{.1}) + \beta_2(x_{i2} - \bar{x}_{.2}) + \dots + \beta_k(x_{ik} - \bar{x}_{.k}),$$

где

$$i = 1, 2, \dots, c,$$

$$j = 1, 2, \dots, k,$$

и

$$\text{var}[y_{ij}|x_i] = \sigma^2 \quad \text{для всех } i, j.$$

Предполагается, что каждому фиксированному значению x соответствует нормально распределенная величина y со средним μ_j и дисперсией σ^2 . Через β_j обозначены частные коэффициенты регрессии, а соответствующие им оценки будем записывать как $b_{yxj} \cdot x_1, x_2, \dots, x_{j-1}, x_{j+1}, \dots, x_k$, т. е. регрессия y на x_j не зависит от $x_1, x_2, \dots, x_{j-1}, x_{j+1}, \dots, x_k$. Оценивающее уравнение запишется в виде

$$y = a + b_1(x_1 - \bar{x}_1) + b_2(x_2 - \bar{x}_2) + \dots + b_k(x_k - \bar{x}_k),$$

где $a, b_1, b_2, \dots, b_k$ — оценки для $\alpha, \beta_1, \beta_2, \dots, \beta_k$ соответственно.

Затем следует найти оценки неизвестных параметров. Применение метода максимального правдоподобия приводит к $k+1$ уравнениям, содержащим k оценок коэффициентов регрессии и оценку для α . Эти уравнения с $k+1$ неизвестными могут быть решены совместно в предположении, что

$$\Delta = \begin{vmatrix} S_{x_1}^2 & S_{x_1 x_2} & \dots & S_{x_1 x_k} \\ S_{x_2 x_1} & S_{x_2}^2 & \dots & S_{x_2 x_k} \\ \dots & \dots & \dots & \dots \\ S_{x_k x_1} & S_{x_k x_2} & \dots & S_{x_k}^2 \end{vmatrix} \neq 0.$$

Существуют различные методы нахождения этих коэффициентов регрессии. Без сомнения, можно было бы выбрать один или два метода и шаг за шагом проследить весь процесс решения. Однако здесь этого не сделано, так как читатель, интересующийся этими вопросами, может найти соответствующий материал в работах Хальда [19], Снедекора¹, Эзекила² и Остли³.

Благодаря внедрению вычислительной техники в практику этот трудоемкий вид анализа стал вполне осуществим; теперь не надо затрачивать дни и даже недели на вычисления, проверку и перепроверку результатов. Для соответствующей счетной машины составляется стандартная программа. Затем исследователь обращается со своими данными в вычислительный центр, переносит их на перфоленту или перфокарты и проводит дальнейшие счетные операции самостоятельно или с помощью работника центра.

¹ G. W. Snedecor, Statistical Method, Iowa, 1953.

² M. Ezekiel, Method of correlation analysis, New York, 1930.

³ B. Ostle, Statistics in research, Iowa, 1954.

После того как получены оценки параметров и характеристики дисперсий этих оценок, обычными методами осуществляется проверка гипотез о неизвестных параметрах. Например, если $s_{b_j}^2$ — оценка дисперсии величины b_j , проверка гипотезы

$$H_1: E(b_j) = \beta_j$$

при альтернативе

$$H_2: E(b_j) \neq \beta_j$$

осуществляется с помощью двустороннего критерия, основанного на величине

$$t = \frac{b_j - \beta_j}{s_{b_j}}.$$

Аналогично, пусть a — оценка для α , которая распределена нормально со средним α и дисперсией σ^2/ck . Тогда для проверки нулевой гипотезы

$$H_1: \alpha = 0$$

при альтернативе

$$H_2: \alpha \leq 0$$

применяется односторонний критерий

$$t = \frac{a}{\hat{\sigma}/\sqrt{ck}},$$

где $\hat{\sigma}$ — оценка для σ , равная корню квадратному из суммы квадратов отклонений эмпирических значений y от теоретических (полученных из уравнения регрессии), деленному на корень квадратный из числа наблюдений ck , уменьшенного на число линейных зависимостей для y , которое в данном случае равно $k+1$. Величину $\hat{\sigma}$ называют также *стандартной ошибкой оценки*; ее используют для построения области случайных отклонений около уравнения регрессии.

Корреляционный анализ трех и более переменных

Эмпирического выражения зависимости между переменными y и k для целей исследования иногда бывает недостаточно; тогда необходимо охарактеризовать силу связи между y и любой величиной x_j . Эта процедура описана в гл. 6.

В случае когда y зависит более чем от одной переменной, можно получить ложное представление о зависимости между y и x_j , если исключить зависимости между y , x и $x_1, x_2, \dots, x_{j-1}, x_{j+1}, \dots, x_k$. Таким образом, нам нужна не просто оценка линейного коэффициента корреляции r_{xy} , а мера корреляционной зависимости между y и x_j , обеспечивающая устранение влияния

Такой мерой служит частный коэффициент корреляции, или, как его еще называют, *условный корреляционный коэффициент*¹ ρ_{yxj} , оценкой которого будет

$$r[yx_j | x_1, x_2, \dots, x_{j-1}, x_{j+1}, \dots, x_k]$$

ОДНОРОДНОСТЬ ОЦЕНОК КОЭФФИЦИЕНТОВ КОРРЕЛЯЦИИ

Интерпретация существенных расхождений между оценками коэффициентов корреляции в применении к палеонтологии

Успех применения коэффициентов корреляции при сравнении популяций зависит от выполнения двух предположений:

1. Популяция, в которой устанавливаются пары коррелируемых признаков, является наименьшей таксономической единицей.

2. Большие расхождения в значениях оценок коэффициентов корреляции пары признаков в сравниваемых популяциях не связаны с выборочными флуктуациями; они выражают реальные, биологически интерпретируемые различия в силе связи.

Таблица 10.5

Интерпретация коэффициента корреляции

№ выборки	Объем выборки n_i	Коэффициент корреляции r_i	$\arctg r_i = Z_i$	$n_i - 3$	$Z_i (n_i - 3)$	$Z_i^2 (n_i - 3)$
1	50	0,364	0,38	47	17,86	6,7868
2	50	0,312	0,32	47	15,04	4,8128
3	30	0,405	0,43	27	11,61	4,9923
4	25	0,327	0,34	22	7,48	2,5432
$n = 143 \quad T_1 = 51,99 \quad T_2 = 19,1351$						

Критерий для проверки гипотезы об однородности набора выборочных коэффициентов корреляции описан Рао [226]. Этот критерий удобен для сравнения $k \geq 2$ оценок коэффициентов корреляции, вычисленных по данным выборок разного объема. Критерий основан на статистике

$$\chi^2 = T_2 - \frac{T_1^2}{N},$$

¹ Т. Андерсон, Введение в многомерный статистический анализ, Физматгиз, гл. 4, раздел 3, 1963.

где

$$T_1 = \sum_{i=1}^k (n_i - 3) Z_i,$$

$$Z_i = \arctg r_i,$$

n_i — объем i -й выборки,

$$T_2 = \sum_{i=1}^k (n_i - 3) Z_i^2,$$

$$N = \sum_{i=1}^k (n_i - 3).$$

Случайная величина $\chi^2 = T_2 - T_1^2/N$, если верна гипотеза о равенстве всех k -коэффициентов корреляции, распределена по закону χ^2 с $k - 1$ степенями свободы.

В качестве иллюстрации критерия приведем следующий пример.

Из коллекции *Spirifer pennatus* взяты наудачу две выборки объемом по 50 и две дополнительные объемом 30 и 25, по данным которых были вычислены четыре оценки коэффициентов корреляции между двумя признаками: 0,364, 0,312, 0,405 и 0,327. Спрашивается, можно ли рассматривать набор этих оценок как однородный или, наоборот, различия между ними столь велики, что они превышают выборочные погрешности в одной популяции?

$$\chi^2 = T_2 - \frac{T_1^2}{N} = 19,1351 - \frac{2702,96}{143} = 0,2333,$$

число степеней свободы $k - 1 = 3$, $\chi_{3, 0,05}^2 = 7,81$.

Так как вычисленное значение χ^2 оказалось меньше 7,81, ряд выборочных коэффициентов корреляции следует рассматривать как однородный.

В качестве второго примера можно привести данные [202], связанные с проверкой следующих предположений:

1. Значения r , по данным выборок из размножающейся популяции, однородны. Из размножающейся популяции голубей взято пять выборок объемом 20 каждая. По данным этих выборок были вычислены оценки коэффициентов корреляции для всех возможных пар десяти морфологических признаков и проверены 45 гипотез об однородности. Оказалось, что нулевая гипотеза не отвергается в 44 случаях из 45 при уровне значимости 0,05.

2. Значения r , по данным выборок, взятых в эволюционной последовательности, неоднородны. Выборки, представляющие собой результаты измерения девяти морфологических признаков, взяты в различное время из двух популяций *Pentremites*.

Второе предположение заключается в том, что значения величины r для любой пары признаков существенно отличаются в выборках, соответствующих различному времени в одной популяции. В случае первой популяции изучались четыре выборки, соответствующие разному времени. Всего было проведено 36 проверок гипотезы об однородности, из которых 29 дали отрицательный результат, что служит подтверждением неоднородности коэффициентов корреляции. Во втором случае рассматривались три периода времени и оказалось, что все 36 проверенных гипотез об однородности были отвергнуты. Не следует ожидать, что *все* коэффициенты корреляции должны быть обязательно различными. В действительности приходится рассматривать коэффициенты корреляции, которые, будучи стабильными в эволюционном смысле, не меняются, а также которые меняются, отражая тем самым изменения в развивающейся популяции.

3. *Значения r однородны в онтогенических сериях.* Исследовали белых крыс. В выборке из 20 индивидуумов были измерены десять морфологических признаков последовательно в пяти онтогенических стадиях (1, 10, 40, 250 дней и взрослые). Были также проверены гипотезы об однородности значений r в пяти изученных онтогенических стадиях. В результате оказалось, что в половине случаев проверки гипотезу пришлось отвергнуть, а в половине — принять. Однако после того, как были исключены результаты измерения первых двух стадий (1 и 10 дней), для оставшихся коэффициентов корреляции гипотеза об однородности была принята в 43 случаях из 45 при уровне значимости 0,05. Это позволяет предположить, что большие отклонения значений коэффициентов корреляции по выборкам, взятым из единой совокупности, связаны прежде всего с очень ранними онтогеническими стадиями, представители которых редко сохраняются в ископаемых остатках.

Рассмотренные выше способы обоснования статистических выводов, заключающиеся в нахождении оценок неизвестных параметров и в проверке гипотез, были представлены в виде общего описания некоторых приемов и методов, применяемых в геологии. Поставив перед собой задачу показать эти методы в их совокупности как единое целое, авторы в данной главе рассматривают результаты исследований, проведенных Каном¹. Хотя они по природе своей и являются петрографическими, применяемые в них статистические приемы нельзя ограничивать петрографической спецификой. Авторы об этом говорят лишь с целью продемонстрировать комплекс всех способов статистической обработки результатов геологических исследований.

ПОСТАНОВКА ВОПРОСА

Петрографическая задача исследования состояла в изучении особенностей структурного параметра — заполнения обломочным материалом терригенных пород. При решении этой задачи, предложенной Гриффитсом, не располагали достаточной информацией о природе рассматриваемой величины. Была разработана техника измерения указанного параметра. При решении задачи рассматривались две величины: мера части пространства, заполненной зернами, называемая *плотностью заполнения* или *объемным содержанием зерен* P_d , и мера близости зерен друг к другу, называемая *близостью заполнения* P_p .

¹ J. S. Kahn, The analysis and distribution of the properties of packing in sandstones, Journ. Geol., 64, № 4, 1956; The analysis and distribution of the properties of packing in sand-size sediments; The distribution of packing measurements and an example of packing analysis, Journ. Geol., 64, № 6, 1956.

МЕТОД ВЫБОРА

Было решено изучить три типа осадочных пород, определенных Крайнином¹, по данным, отобранным следующим способом.

Образцы каждого типа пород отобрали из трех различных формаций, каждая формация характеризовалась двумя шлифами, а в каждом шлифе были изучены пять пересечений.

Выборочный метод, принятый в этой работе, схематически представлен в табл. 11.1. Формации, соответствующие одному

Таблица 11.1

Схема плана выборки, применявшегося при изучении плотности осадочных пород

Тип породы	Формация	Шлиф	Пересечение
Кварциты	Тускарора	{ LM742 — В Р — 72	Пять пересечений в каждом шлифе
	Орискани	{ O7A — 3A O7A — 3B	
	Кьюч	{ GR 1200 ⁶ — 96 471 AGH3676	
Граувакки	Поконо	{ 6225 6285	
	Освего	{ 6307 6301	
	Бредфорд	{ 5820—3 Р (4)—(6)	
Аркозы	Портленд	{ Mid 5—2 Mid 8—8B	
	Рослин	{ ab — 3a ef — 2a	
	Стоктон	{ 1 2	
	Всего	{ 3 типа пород 9 формаций 18 шлифов 90 пересечений	

типу пород, рассматриваются как взятая наудачу выборка из множества формаций этого типа; шлифы — как случайная выборка из множества всех возможных шлифов, которые могли

¹ P. D. Kryniue, The Megascopic Study and field classification of sedimentary rocks, Mineral Industries, Exper. Station Techn. Paper 130, College of Min. Industries, The Pennsylvania State Univ., State College Penn., 1947.

бы быть выбраны; пересечения также выбраны наудачу из бесчисленного множества пересечений, которые можно было бы изучить.

При этом были поставлены следующие вопросы:

1. Каковы функции распределения случайных величин P_d и P_p ?
2. Какими параметрами характеризуются эти функции?
3. Каковы оценки средних значений P_d и P_p для различных типов пород, формаций, шлифов и пересечений?
4. Существуют ли значимые различия в распределениях рассматриваемых величин в трех типах пород?
5. Каковы оценки различных дисперсионных компонент?
6. Есть ли линейная зависимость между P_d и P_p ?
7. Все ли изменения значений P_p обусловлены зависимостью между P_p и P_d или между P_p и другими переменными?

План эксперимента

Поставленные вопросы требовали построения модели, которая бы схематически отражала характер сформулированных выше задач. Для этой цели лучше всего было бы воспользоваться смешанной моделью однофакторного дисперсионного анализа, в которой любой результат наблюдения над величиной y представляет собой сумму

$$\mu + \rho_i + f_{ij} + s_{ijk} + t_{ijkl}.$$

В этом случае величина ρ_i — константа, представляющая собой отклонение среднего значения рассматриваемой характеристики в данном типе пород от генеральной средней μ ; величина f_{ij} рассматривается как случайная, характеризующая отклонения средних значений признака в формации от соответствующего среднего по типу породы. Предполагается, что f_{ij} распределена нормально со средним 0 и дисперсией σ_f^2 . Случайные отклонения средних значений по шлифам от соответствующего им значения средней по формации представлены величинами s_{ijk} , распределенными как $N(0, \sigma_s^2)$. Величины t_{ijkl} рассматриваются как погрешности эксперимента с распределением $N(0, \sigma_t^2)$. Все сказанное сведено в табл. 11.2.

Распределение рассматриваемых случайных величин

В гл. 6 было показано, что если известен вид функций распределения данной случайной величины, то известно и число параметров, от которых зависит эта функция. Далее, если можно определить точные значения этих параметров или по крайней мере оценить их, то это означает что исследователь распо-

Математическая модель гнездового дисперсионного анализа

Тип изменчивости	Число степеней свободы	Сумма квадратов	Средний квадрат	Дисперсионные компоненты
Типы пород r	$r-1$	$\left[\sum_i^r \left(\sum_{j,k,l}^{fst} X \right)^2 / fst \right] - T = S_r^2$	$S_r^2 / (r-1)$	$\sigma_t^2 + t\sigma_s^2 + st\sigma_f^2 + \frac{fst}{r-1} \sum_i^r \rho_i^2$
Формации f (внутри типов пород)	$r(f-1)$	$\left[\sum_{i,j}^{rf} \left(\sum_{k,l}^{st} X \right)^2 / st \right] - T - S_r^2 = S_f^2$	$S_f^2 / r(f-1)$	$\sigma_t^2 + t\sigma_s^2 + st\sigma_f^2$
Шлифы s (внутри формаций и типов пород)	$rf(s-1)$	$\left[\sum_{i,j,k}^{rfs} \left(\sum_l^t X \right)^2 / t \right] - T - S_r^2 - S_f^2 = S_s^2$	$S_s^2 / rfs(s-1)$	$\sigma_t^2 + t\sigma_s^2$
Пересечения t (внутри шлифов, формаций и типов пород)	$rfst(t-1)$	$S_{общая}^2 - S_r^2 - S_f^2 - S_s^2 = S_t^2$	$S_t^2 / rfst(t-1)$	σ_t^2
Общая дисперсия	$rfst-1$	$\sum_{i,j,k,l}^{rfst} X^2 - T = S_{общая}^2$	$S_{общая}^2 / (rfst-1)$	—

$$T = \left(\sum_{i,j,k,l}^{rfst} X \right)^2 / rfst; \quad i=1, 2, \dots, r; \quad j=1, 2, \dots, f; \quad k=1, 2, \dots, s; \quad l=1, 2, \dots, t.$$

$X = x_{ijkl} = \mu + \rho_i + f_{ij} + s_{ijk} + t_{ijkl}$, где ρ_i — константа, f_{ij} — случайная величина, распределенная как $N(0, \sigma_f^2)$; s_{ijk} — распределенная как $N(0, \sigma_s^2)$, а t_{ijkl} — как $N(0, \sigma_t^2)$.

лагает достаточно полными сведениями об изучаемой случайной величине. Например, если известно, что случайная величина y распределена нормально, значит известен вид функции распределения с параметрами μ и σ^2 . Чтобы можно было указать конкретную функцию распределения, нужно определить точные значения μ и σ .

Функции распределения случайных величин P_d и P_p а priori неизвестны. Один из способов определения близости наблюдаемого эмпирического распределения к теоретическому заключается в сравнении частот, полученных в результате эксперимента, с теоретическими частотами. Процедура этого сравнения описана в гл. 6. Имеющиеся в нашем распоряжении данные собраны согласно выборочной методике, которая описана выше. Эти эмпирические данные приведены в виде частот в табл. 11.3.

Таблица 11.3

Распределение плотности заполнения и близости заполнения в кварцитах, граувакках, аркозах в их совокупности, %

	Класс	Кварциты	Граувакки	Аркозы	Все типы пород
Плотность заполнения	90—100	4	1	6	11
	80—90	5	6	12	23
	70—80	1	13	9	23
	60—70	9	9	3	21
	50—60	8	1	—	9
	40—50	3	—	—	3
	N	30	30	30	90
	$\bar{P}_d$	68,00	74,00	82,00	74,67
	s	15,75	8,72	9,00	12,95
Мера близости	80—90	—	—	3	3
	70—80	3	2	4	9
	60—70	4	1	9	14
	50—60	4	3	8	15
	40—50	4	11	2	17
	30—40	6	9	4	19
	20—30	6	4	—	10
	10—20	3	—	—	3
	N	30	30	30	90
	$\bar{P}_p$	43,00	43,00	60,33	48,78
	s	18,50	12,77	14,32	17,42

Гипотеза о согласованности эмпирического распределения P_d с нормальным законом была проверена в гл. 6. В резуль-

тате значение статистики

$$\chi_{(2)}^2 = \sum_{i=1}^5 \frac{(f_i - np_i)^2}{np_i}$$

оказалось меньше допустимого при уровне значимости $\alpha=0,05$ и двух степенях свободы, что позволило принять гипотезу о нормальном распределении с параметрами, которые равны соответствующим оценкам.

С другой стороны, величина P_p представляет собой долю контактов, наблюдаемых в пересечении, от общего числа n всех возможных контактов. Так, в крайнем случае, если всего в пересечении $n+1$ зерен, то при условии, что они все находятся на прямой и соприкасаются друг с другом, между ними будет n соприкосновений.

Таким образом,

$$P_p = \left(\frac{1}{n} \sum_{i=1}^n c_i \right) \cdot 100\%,$$

где $\sum_{i=1}^n c_i$ — число контактов между зернами в данном пересечении. Величины c_i принимают значения 1 и 0, что означает наличие или отсутствие контакта между двумя соседними зернами. Если вероятность события $c_i=1$ есть p , то при n возможных контактах случайная величина $\sum_{i=1}^n c_i$ будет биномиально распределена с параметром p (гл. 2).

Следовательно, с увеличением n распределение случайной величины $\frac{1}{n} \sum_{i=1}^n c_i$ будет приближаться к нормальному со средним $\mu=p$ и дисперсией $\sigma^2=p(1-p)/n$. Проверка гипотезы о согласованности распределения случайной величины P_p с нормальным законом при помощи критерия χ^2 показала, что эту гипотезу можно принять как подтвердившуюся, так как вычисленное значение $\chi^2=2,82$ меньше допустимого $\chi_{3,0,05}^2=7,82$ при трех степенях свободы и уровне значимости $\alpha=0,05$.

ОЦЕНКИ ПАРАМЕТРОВ

В связи с тем что изучаемые величины распределены нормально, нужно в каждом случае оценить два параметра — μ и σ . Этими оценками служат среднее арифметическое $\bar{x}$ и величина s , вычисляемая обычным способом. Значения этих оценок приведены в табл. 11.3—11.5.

Оценки средних значений для P_d и P_p по типам пород и формациям, %

Формация	P_d	P_p
Тускарора	87,45	63,28
Орискани	60,93	27,93
Кьюч	55,73	36,85
Среднее по кварцитам	68,04	42,68
Поконо	72,59	39,85
Освего	82,47	56,98
Бредфорд	70,77	33,61
Среднее по грауваккам	75,28	43,48
Портленд	76,64	46,86
Рослин	90,54	73,52
Стоктон	78,52	58,54
Среднее по аркозам	81,90	59,64
Генеральные средние	75,07	48,60

ПРОВЕРКА ГИПОТЕЗ О μ И σ

На основании описанной модели эксперимента можно предложить метод проверки главных линейных гипотез относительно генеральных значений μ и σ .

Читатель, очевидно, помнит изложенные в гл. 7 главные предположения, взятые за основу дисперсионного анализа. Относящееся к ним предположение о нормальности было проверено и принято. Требование случайности выбора было выполнено при проведении эксперимента. Следующее важное условие — предположение об однородности дисперсий. Как уже отмечалось выше, это последнее предположение сильно ограничивает возможность применения дисперсионного анализа. По имеющимся данным были проверены гипотезы об однородности дисперсии с использованием метода, изложенного в гл. 6. В результате проверки этих гипотез были приняты табл. 11.7 и 11.8.

Теперь с помощью дисперсионного анализа можно проверить следующие гипотезы, относящиеся как к P_d , так и к P_p .

1. H_1 : $\mu_1 = \mu_2 = \mu_3 = \mu$, где μ_1 — среднее значение для аркозов, μ_2 — для граувакк, μ_3 — для кварцитов. Таким образом, согласно нулевой гипотезе, между средними значениями в типах

Т а б л и ц а 11.5

Результаты подсчетов по пересечениям в шлифах, %

№ шлифа Пресечение	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
<i>Плотность заполнения</i>																		
1	91,01	89,25	66,92	62,03	49,11	63,47	78,93	73,68	88,21	75,13	68,88	68,09	80,97	65,44	85,35	91,72	82,32	85,91
2	78,33	86,06	56,31	58,68	46,92	64,77	79,30	63,85	89,55	76,60	68,63	78,63	84,73	77,77	93,59	90,83	82,30	77,30
3	88,86	90,54	58,51	58,24	40,34	65,59	79,28	57,83	88,51	77,91	64,43	76,77	81,31	79,25	90,01	92,75	79,86	73,26
4	87,25	90,00	65,76	64,57	50,25	63,39	85,05	69,93	87,12	77,97	68,91	72,97	76,27	68,22	89,00	95,33	75,48	76,99
5	82,60	90,59	60,16	58,14	54,67	58,74	81,79	62,30	92,96	76,70	64,52	75,85	80,53	71,88	88,34	88,50	65,60	86,20
$\overline{P_d}$	85,61	89,29	61,53	60,33	48,26	63,19	80,87	64,32	89,27	75,66	67,07	74,46	80,76	72,51	89,26	91,83	77,11	70,93
s^2	26,11	3,55	20,53	8,19	27,58	7,05	6,77	33,55	5,01	5,23	5,64	16,86	9,08	35,47	8,88	6,30	49,20	33,78
<i>Мера близости</i>																		
1	64,47	68,33	35,63	29,10	19,81	45,90	41,29	40,27	68,54	40,28	32,68	36,99	61,29	37,65	71,29	80,77	64,58	59,26
2	57,61	70,21	28,57	18,02	36,11	41,10	38,61	42,41	71,76	55,35	29,67	40,09	53,95	55,79	82,11	66,97	54,35	58,93
3	60,17	70,59	36,00	23,48	18,87	55,00	37,91	35,76	57,30	43,95	23,66	48,50	45,24	31,31	60,87	80,00	60,53	38,64
4	46,67	66,04	30,19	30,77	27,35	52,00	44,67	41,10	58,89	49,71	29,38	36,63	55,21	30,19	60,42	76,85	60,53	59,18
5	50,53	78,13	24,76	22,79	40,40	31,94	38,93	37,50	78,57	45,40	23,13	35,37	55,81	42,16	79,78	76,15	60,00	69,39
$\overline{P_p}$	55,89	70,66	31,03	24,83	28,51	45,19	40,28	39,41	67,01	46,94	27,70	39,52	54,30	39,42	70,89	76,15	60,00	57,08
s^2	52,16	20,70	22,99	26,47	92,28	83,85	7,63	7,39	79,69	33,52	17,18	28,23	33,54	107,40	103,78	30,23	13,37	126,04

пород нет существенных различий при альтернативе, что есть хотя бы одно ρ_i , для которого $\rho_i \neq \rho$.

2. H_2 : $\sigma_f^2 = 0$ при альтернативе $\sigma_f^2 > 0$.

3. H_3 : $\sigma_s^2 = 0$ при альтернативе $\sigma_s^2 > 0$.

Результаты анализа приведены в табл. 11.9 и 11.11. Эти результаты одинаковы для обеих величин P_d и P_p , т. е. если отвергается гипотеза H_1 и делается вывод о существенном различии в средних для типов пород, это одинаково относится к P_d и P_p .

Аналогично если отвергаются гипотезы H_2 и H_3 и делается вывод, согласно которому как шлифы, так и формации вносят свою долю в общую изменчивость, то этот вывод относится к P_d и P_p . Оценки средних для типов пород приведены в табл. 11.3. Методом Тьюки можно установить, какие именно средние нарушают однородность ряда (табл. 11.10 и 11.12). Например, среднее значение для P_p формации Рослин сильно отличается от средних значений формаций Кьюч, Бредфорд и Орискани. Формация Тускарора отличается только от формации Орискани. Средние значения других формаций не сильно отличаются друг от друга. Оценки для σ_i^2 , σ_s^2 и σ_f^2 приведены в табл. 11.13.

Линейные зависимости между P_d и P_p

Линейную зависимость между двумя любыми случайными величинами можно выявить двумя способами. Во-первых, при помощи линейного регрессионного анализа (гл. 8); в данном случае следует изучить возможность зависимости вида $y = \hat{f}(x) = \alpha + \beta(x - \bar{x})$. Во-вторых, путем проверки гипотезы о равенстве нулю коэффициента корреляции (гл. 6).

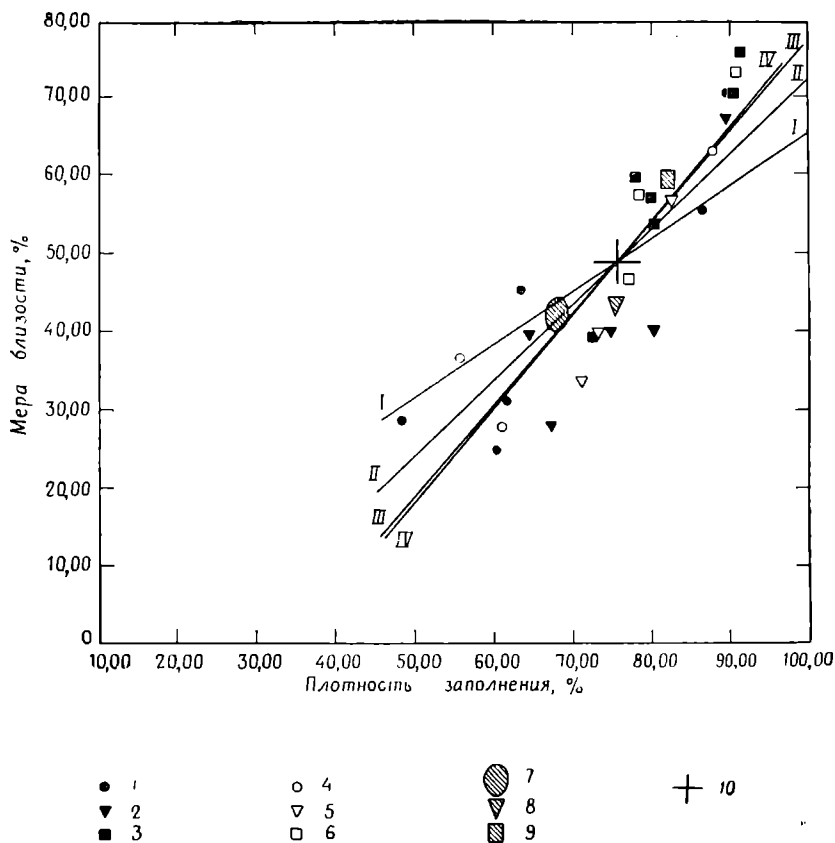
Линейный регрессионный анализ

Задача весьма сходна с общей задачей, рассмотренной в гл. 8. Оценки для α и β по данным пересечений, шлифов, формаций и типов пород приведены в табл. 11.6 для различных уровней изменчивости.

На основе результатов проверки этих гипотез приходим к выводу, что между рассматриваемыми случайными величинами резко выражена линейная зависимость. Это показано на фиг. 11.1, на которой приведены графики уравнений регрессии для различных причин изменчивости.

Корреляционный анализ

Оценки коэффициентов корреляции, соответствующие различным источникам изменчивости, приведены в табл. 11.6.



Фиг. 11.1. Усредненные зависимости между P_p и P_d .

I — регрессия по пересечениям
 $y = 0,6901x - 3,20$
 $r = 0,4079$
 $r^2 = 16,64\%$

III — регрессия по формациям
 $y = 1,1900x - 40,73$
 $r = 0,9081$
 $r^2 = 82,46\%$

II — регрессия по шлифам
 $y = 0,9754x - 24,62$
 $r = 0,7709$
 $r^2 = 59,43\%$

IV — регрессия по типам пород
 $y = 1,2507x - 45,29$
 $r = 0,8739$
 $r^2 = 76,37\%$

Для шлифов: 1 — кварциты; 2 — граувакки; 3 — аркозы. Для формаций: 4 — кварциты; 5 — граувакки; 6 — аркозы. Для типов пород: 7 — кварциты; 8 — граувакки; 9 — аркозы; 10 — генеральная средняя.

Таблица 11.6

**Уравнения регрессии, коэффициенты регрессии и корреляции
для пересечений, шлифов, формаций и типов пород**

Тип изменчивости	Уравнение регрессии	b	r	$r_{0,05}$	$r_{0,01}$
По пересечениям (внутри шлифов)	$y = 0,6901x - 3,20$ $n = 90$, ст. св. ¹ = 88	0,6901	0,4079 ²	0,205	0,267
По шлифам (внут- ри формаций)	$y = 0,9754x - 24,62$ $n = 18$, ст. св. ¹ = 16	0,9754	0,7709 ²	0,468	0,590
По формациям (внутри типов пород)	$y = 1,1900x - 40,73$ $n = 9$, ст. св. ¹ = 7	1,1900	0,9081 ²	0,666	0,798
По типам пород	$y = 1,2507x - 45,29$ $n = 3$, ст. св. ¹ = 1	1,2507	0,8739 NS ³	0,997	1,000

¹ ст. св. — число степеней свободы.² Отклонение от нуля значимо при уровне 0,01.³ NS — гипотеза $\rho=0$ не отвергается.

Таблица 11.7

**Проверка гипотезы об однородности дисперсии P_p
по всем типам пород**

Тип породы	Число степеней свободы	χ^2	$\chi_{0,05}^2$
Кварциты	5	4,37	11,1
Граувакки	5	8,34	11,1
Аркозы	5	6,96	11,1
Все типы пород	17	24,07	27,6

Таблица 11.8

**Проверка гипотезы об однородности дисперсии P_d
по всем типам пород**

Тип породы	Число степеней свободы	χ^2	$\chi_{0,05}^2$
Кварциты	5	6,14	11,1
Граувакки	5	7,01	11,1
Аркозы	5	7,19	11,1
Все типы пород	17	23,16	27,6

Результаты дисперсионного анализа P_p по трем типам осадочных пород

Причины изменчивости	Число степеней свободы	Сумма квадратов	Средний квадрат	F
Типы пород	2	5 493,64	2746,82	1,24
Формации (внутри типов пород)	6	13 256,67	2209,45	5,95 ¹
Шлифы (внутри формаций и типов пород)	9	3 338,94	370,99	7,53 ²
Пересечения (внутри шлифов, формаций и типов пород)	72	3 545,69	49,25	—
Общая изменчивость (n = 90)	89	25 634,94		

¹ Различия существенно при уровне значимости 0,01.

² То же при 0,001.

Таблица 11.10

Значения P_p в формациях, %

Формация	P_p	Допустимое значение P_p^1
Рослин	73,52	39,41
Гускарора	63,28	29,17
Стоктон	58,54	—
Освего	56,98	24,43
Портленд	46,86	
Поконо	39,85	
Кьюч	36,85	
Бредфорд	33,61	
Орискани	27,93	

¹ Допустимое значение при уровне значимости 0,05 равно 34,11%.

Результаты дисперсионного анализа P_d по трем типам осадочных пород

Причина изменчивости	Число степеней свободы	Сумма квадратов	Средний квадрат	F
Типы пород	2	2 885,46	1442,73	1,12
Формации (внутри типов пород)	6	7 718,99	1286,50	5,55 ¹
Шлифы (внутри формаций и типов пород)	9	2 085,83	231,76	13,47 ²
Пересечения (внутри шлифов, формаций и типов пород)	72	1 238,11	17,96	
Общая изменчивость ($n = 90$)	89	13 928,39		

¹ Различие существенно при уровне значимости 0,05.

² То же при 0,001.

Таблица 11.12

Значения P_d в формациях, %

Формация	P_d	Допустимое значение P_d^1
Рослин	90,54	63,58
Тускарора	87,45	60,49
Освего	82,47	55,51
Стоктон	78,52	
Портленд	76,64	
Поконо	72,59	
Бредфорд	70,77	
Орисканн	60,93	
Кьюч	55,73	

¹ Допустимое значение при уровне значимости 0,05 равно 56,96%.

Таблица 11.13
Дисперсионные компоненты для результатов измерения

Дисперсионная компонента	Оценка дисперсионной компоненты	
	для P_p	для P_d
σ_t^2	49,25	17,96
σ_s^2	64,35	42,76
σ_f^2	183,85	105,47

УПРАЖНЕНИЕ

Проверить гипотезы $\alpha=0$ и $\beta=0$ по данным, приведенным в табл. 11.6.

КОВАРИАЦИОННЫЙ АНАЛИЗ И ИЗУЧЕНИЕ ЗАВИСИМОСТИ МЕЖДУ МЕРОЙ БЛИЗОСТИ И ПЛОТНОСТЬЮ ЗАПОЛНЕНИЯ

Схема ковариационного анализа для выявления зависимости между P_p и P_d приведена в табл. 11.14. Эффект, возникающий в результате действия P_d , описываемый величиной $b_{yx}S_{xy}$, исключен из суммы квадратов S_y величины P_p по всем источникам изменчивости. Результат будет выражать величину погрешности эксперимента для P_p . Полученные таким образом новые выражения ошибки будут использованы для изучения различий.

В случае простого дисперсионного анализа для P_p (табл. 11.9) были отвергнуты гипотезы $\sigma_f^2=0$ и $\sigma_s^2=0$. Различия между типами пород оказались несущественными. Спрашивается, сколь значимы эти результаты, относящиеся исключительно к P_p , действительно ли P_p меняется в зависимости от изменений P_d ? Путем дисперсионного анализа выявлены те различия для P_p , которые обусловлены только изменчивостью P_p , тогда как ковариационный анализ позволяет выявить такие различия, которые обусловлены влиянием изменчивости величины P_d .

В результате ковариационного анализа путем проведения преобразований, показанных в табл. 11.14, гипотеза $\sigma_s^2=0$ была отвергнута и, кроме того, было установлено существенное различие между типами пород. Гипотеза $\sigma_f^2=0$ на этот раз была принята как подтвердившаяся. Изменчивость величины P_p по формациям в действительности обусловлена изменчивостью величины P_d , от которой зависит P_p . Следовательно, величина P_p , по крайней мере частично, зависит от P_d , т. е. число

Таблица 11.14

Результаты анализа величины P_p с учетом зависимости между P_p и P_d ($x = P_p$, $y = P_d$)

Причина изменчивости	Число степеней свободы	S_x^2	S_{xy}^2	S_y^2	r	b	Погрешность оценки		
							сумма квадратов	число степеней свободы	средний квадрат
Типы пород	2	2 885,46	3 478,92	5 493,64	0,8739	1,2057			
Формации	6	7 718,99	9 185,97	13 256,67	0,9081	1,1900	2325,36	5	465,07
Шлифы	9	2 085,83	2 034,46	3 338,94	0,7709	0,9754	1354,53	8	169,32
Пересечения	72	1 238,11	854,47	3 545,69	0,4079	0,6901	2955,99	71	41,63
Общая изменчивость .	89	13 928,49	15 553,82	25 634,94			8265,99	88	
Типы пород и формации	8	10 604,45	12 664,89	18 750,31			3624,63	7	
Формации и шлифы . .	15	9 804,82	11 220,44	16 595,61			3754,95	14	
Шлифы и пересечения .	81	3 323,94	2 888,93	6 884,63			4373,86	80	
Расхождения между приведенными средними							1299,27	2	64 963
То же по формациям							2400,42	6	400,07
» » по шлифам							1417,86	9	157,54

Таблица 11.15

Результаты анализа неисправленного среднего квадрата

Тип изменчивости	Число степеней свободы	Сумма квадратов	Средний квадрат	Приведенное значение F	
				для проверки	приведенное значение F
E_T — неисправленная ошибка . .	6	13 256,67	2 209,44	По типам пород По формациям По шлифам	1,396 NS ¹
Влияние регрессии	1	10 931,31	10 931,31		2,360 NS ¹
Приведенная ошибка	5	2 325,36	465,07		3,784 ²
E_f — неисправленная ошибка . .	9	3 338,94	370,99		11,720 ²
Влияние регрессии	1	1 984,41	1 984,41		
Приведенная ошибка	8	1 354,53	169,32		
E_s — неисправленная ошибка . .	72	3 545,69	46,25		14,175 ³
Влияние регрессии	1	589,69	589,69		
Приведенная ошибка	71	2 956,00	41,63		

¹ NS — различие несущественно.² Различие существенно при уровне значимости 0,01.³ То же при 0,001.

контактов между зернами зависит от степени заполнения ими пространства.

В табл. 11.14 приведены необходимые данные для сравнения сумм квадратов. Исходная сумма квадратов для P_p по формациям равна 13 256,67; при введении поправки, связанной с влиянием P_d , эта сумма уменьшается до 2400,42, что в свою очередь приводит к изменению значения дисперсионного отношения F , меняющего решение при проверке гипотезы $\sigma_f^2 = 0$. Результаты сравнения неисправленных сумм квадратов, характеристика погрешностей эксперимента, а следовательно, и точности приведены в табл. 11.15. Уменьшение первоначального абсолютного значения ошибки свидетельствует о некоторой изменчивости, ранее рассматривавшейся как «ошибка эксперимента», тогда как на самом деле изменчивость обусловлена влиянием P_d на P_p . Эта изменчивость выявлена и устранена. В результате точность сравнения повышается.

Общая сумма квадратов, соответствующая каждой причине изменчивости, разлагается на две части (табл. 11.15). Одна из этих частей, обладающая одной степенью свободы, характеризует изменчивость P_p , вызванную зависимостью от P_d , а другая часть с оставшимся числом степеней свободы представляет собой меру погрешностей эксперимента. Если эффект исключения влияния регрессии несуществен, отношение этих двух средних квадратов будет распределено по закону Фишера, что может быть использовано в качестве критерия при проверке гипотезы об отсутствии влияния одной переменной на другую. Так, общий средний квадрат для формаций, равный 370,99, после исключения влияния регрессии сильно изменился — до значения 169,32, что свидетельствует о большом влиянии величины P_d на значения P_p .

Аналогичный ковариационный анализ был проведен с целью изучения зависимости между P_p и размером зерен, слагающих породу. Кроме того, была изучена зависимость между размером зерен и P_d . В результате выяснилось, что величина пространства, занятого зернами, и число контактов между зернами P_p в значительной степени зависят от размеров зерен. Таким образом, взаимную зависимость между этими тремя величинами можно считать установленной. Зависимость между величиной P_p , выбранной в качестве аргумента, величиной P_d и размером зерен, рассматриваемых как функции, можно выразить, произведя множественный регрессионный анализ.

МНОЖЕСТВЕННАЯ РЕГРЕССИЯ

Точно так же, как после устранения эффекта влияния величины P_d на P_p , остается эффект влияния размеров зерен, а при устранении влияния размеров зерен остается эффект влияния P_d .

С целью выразить строгую зависимость между этими переменными нужно изменения функции, обусловленные изменениями аргумента, измерять вне зависимости от других величин. Следовательно, изменение величины P_p , связанное с единичным изменением P_d , должно быть выражено независимо от размеров зерен. Точно так же изменение величины P_p , обусловленное единичным изменением размера зерен, следует выразить независимо от P_d . Таким образом, изменчивость величины P_p можно выразить в виде двух составляющих: во-первых, в виде изменчивости P_p , связанной с P_d , независимо от размеров зерен и, во-вторых, в виде изменчивости P_p , обусловленной размерами зерен, независимо от P_d . Мерой изменения величины, выбранной в качестве функции в зависимости от единичного изменения аргумента, служит множественный коэффициент регрессии

$$b_{yx_1, x_2, x_3, \dots, x_{j-1}, x_{j+1}, \dots, x_k}$$

Множественные коэффициенты регрессии для P_p при исключении влияния размеров зерен и P_d для формаций равны соответственно 1,5850 и -143,91. Эта множественная регрессионная зависимость для формаций может быть выражена уравнением множественной регрессии

$$\hat{P}_p = y = c + b_{yx_1, x_2} x_1 + b_{yx_2, x_1} x_2,$$

где $\hat{P}_p$ — предсказанное значение P_p ; $x_1 = P_d$; x_2 — размер зерен. Это выражение позволяет с определенной точностью предсказывать значение P_p любой формации, если для нее известны значения P_d и размер зерен. В данном конкретном случае это уравнение имеет вид

$$y = 1,585x_1 - 143,9x_2 - 51,31.$$

Это выражение соответствует случаю, когда существуют средние по формациям для x_1 и x_2 . Нередко бывает необходимо предсказать значения P_p в шлифах в пределах одной формации, что можно сделать также с помощью уравнения множественной регрессии; в нашем случае это уравнение, составленное по данным шлифов, принимает следующий вид:

$$y = 0,827x_1 - 101,79x_2 - 0,0081,$$

где x_1 — средние значения P_d в шлифах, x_2 — средние значения размеров зерен в шлифах.

В гл. 12 и 13 рассмотрены общие вопросы многомерного, или мультивариантного, анализа и даны примеры его применения. Термин *многомерный анализ* буквально следует понимать как статистическое исследование задач, содержащих две и более случайные величины. Однако по поводу обозначения этого термина иногда возникают разногласия.

Андерсон¹ делит многомерный анализ на пять основных групп задач, к которым мы и переходим.

КОРРЕЛЯЦИЯ

Методы измерения степени зависимости двух случайных величин можно распространить на случай, когда одна величина рассматривается как функция многих переменных. Но нас сейчас интересует пример, когда мерой зависимости служит множественный коэффициент корреляции. Кроме того, эти методы можно использовать при измерении зависимости путем исключения эффекта, обусловленного корреляцией между переменными. В разделе, посвященном картированию и тренду, рассматривается множественная регрессия и ее применение. В списке литературы приведены работы, посвященные применению этих методов в геологических науках. В гл. 13 рассматриваются факторный анализ и связанные с ним вопросы.

АНАЛОГИ ОДНОМЕРНЫХ СТАТИСТИЧЕСКИХ МЕТОДОВ

Одномерные статистические задачи нередко имеют свои многомерные аналоги; например, 1) аналогом проверки гипотез

¹ Т. Андерсон, Введение в многомерный статистический анализ, Физматгиз, 1963.

тезы, согласно которой среднее равно нулю, будет проверка гипотезы, постулирующей, что несколько средних равны нулю; 2) многомерным аналогом t -критерия Стьюдента является обобщенный критерий T^2 . В многочисленных задачах, связанных с применением статистического анализа, рассматривается проверка гипотез о средних при неизвестном значении дисперсии. Если изучаемая случайная величина распределена нормально со средним μ и неизвестной дисперсией σ^2 , то в качестве статистики, характеризующей отклонение выборочной средней $\bar{x}$ от генеральной средней μ , используется величина

$$t = \frac{(\bar{x} - \mu)\sqrt{n}}{s},$$

где n — объем выборки, s — оценка стандартного отклонения (корень квадратный из выборочной дисперсии). Если $E(\bar{x}) = \mu$, то величина t распределена по закону Стьюдента с $n - 1$ степенями свободы (гл. 6). С помощью этого критерия можно проверить гипотезу $\mu = \mu_0$: неизвестное значение μ находится в заданном интервале при данном уровне значимости.

Статистика T^2 , рассмотренная в этой главе, представляет собой многомерный аналог статистики t ; 3) метод наименьших квадратов обобщается на случай более двух переменных; 4) дисперсионный анализ одномерной величины приводит к своему многомерному аналогу. Последний охватывает задачи, связанные с проверкой гипотез о равенстве средних многомерных случайных величин, имеющих общую ковариационную матрицу. Этот тип анализа детально рассматривается в настоящей главе.

ПРОБЛЕМЫ СИСТЕМЫ КООРДИНАТ

Задачи этой категории связаны с отысканием нормализованных линейных комбинаций случайных величин, соответствующих максимуму или минимуму дисперсии (нахождение главных компонент). Основная особенность этой группы задач заключается в нахождении путем поворота такого положения системы осей координат, которое приводит ковариационную матрицу к диагональной форме. Поворот системы координат, который приводит эту систему в новое положение, обеспечивающее нужные статистические свойства, ведет к следующей группе задач, называемой *факторным анализом*. Андерсон дает следующее обоснование этим интересным для практики задачам:

«Во многих исследовательских работах число рассматриваемых случайных величин, которые требуется обработать, слишком велико. Поскольку в этих работах интерес представляют именно отклонения, применяемый метод сокращения числа

случайных величин состоит в том, что линейные комбинации, имеющие малые дисперсии, отбрасываются, а рассматриваются лишь линейные комбинации с большими дисперсиями. Например, антрополог-физиологист может проделать десятки измерений, таких, как длина уха, ширина уха, длина лица, ширина лица и т. д., у большого числа индивидуумов. Его может интересовать описание и анализ различий индивидуумов по такого рода физиологическим характеристикам. Возможно, он «объяснит» эти различия, но вначале он хотел бы знать, какие измерения или комбинации измерений дают значительные отклонения; другими словами, что следует изучать дальше. Главные компоненты представляют собой новое множество измерений, полученных в результате линейной комбинации исходных измерений. Может оказаться, что различия между индивидуумами сводятся к трем линейным комбинациям. Тогда антрополог может обратиться к изучению этих трех величин. Другие линейные комбинации так мало изменяются от одного человека к другому, что изучение их дает мало новых знаний о различиях между индивидуумами».

БОЛЕЕ ДЕТАЛИЗИРОВАННЫЕ ЗАДАЧИ, СВЯЗАННЫЕ С РАЗДЕЛЕНИЕМ НАБОРА СЛУЧАЙНЫХ ВЕЛИЧИН НА ПОДГРУППЫ; НАПРИМЕР, ПРОВЕРКА ГИПОТЕЗЫ О НЕЗАВИСИМОСТИ ПОДГРУПП

В этой главе рассматривается так называемое обобщенное расстояние D^2 . Факторный анализ изложен в гл. 13.

ЗАВИСИМЫЕ НАБЛЮДЕНИЯ

Согласно Андерсону, пятая группа задач включает статистический анализ коррелированных последовательностей и стохастических разностных уравнений. Первые кратко рассмотрены в гл. 15 (изучение временных рядов), тогда как вторые в этой книге не рассматриваются.

В основе многомерного анализа в большинстве случаев лежит предположение о нормальном распределении. Андерсон отмечает, что «одной из причин, побуждающих к изучению многомерных нормальных распределений, является то, что частные и условные распределения, полученные из многомерных нормальных распределений, также нормальны, более того, линейные комбинации нормально распределенных величин также распределены нормально» (см. гл. 4 настоящей книги).

В качестве дополнения можно упомянуть также полиномиальное распределение (обобщение биномиального распределения на многомерный случай) (гл. 2).

В гл. 12 и 13 излагаются лишь отдельные задачи, связанные с применением многомерного анализа, но не сам анализ в целом. Наиболее интересные вопросы изложены в примерах, заимствованных из литературных источников. Авторы надеются, что сущность методов многомерного анализа будет более понятна читателям из приведенных примеров, из которых авторы иногда пытались выбрать метод, наилучший в данной ситуации. Как и при чтении других разделов этой книги, авторы советуют читателю обращаться к соответствующим учебникам по статистике, а также консультироваться у математиков.

В настоящей главе рассматривается ряд связанных между собой вопросов. Большое внимание уделяется проверке гипотезы о равенстве средних двух многомерных случайных величин. Вместо того чтобы сравнивать средние в последовательности одномерных случайных величин, рассмотрим их как одну многомерную случайную величину, так как это дает возможность учесть не только дисперсии, но и ковариации. При таком подходе принимаются во внимание также и зависимости между случайными величинами, образующими многомерную величину.

Первые четыре рассмотренные задачи включают критерий Хотеллинга T^2 , многомерный дисперсионный анализ, содержащий Λ -критерий, обобщенное расстояние D^2 и коэффициент расового сходства (C.R.L.). Все четыре задачи внутренне связаны между собой. (В современной практике вместо коэффициента расового сходства обычно используется обобщенное расстояние D^2 . Однако в связи с тем, что этот коэффициент широко использовался в более ранней литературе, он включен в рассмотрение.)

Так, например, критерий T^2 отличается от коэффициента расового сходства главным образом тем, что, применяя его, мы можем вместо точных значений параметров использовать их оценки. Кроме того, коэффициент расового сходства эквивалентен (с точностью до некоторого множителя) обобщенному расстоянию D^2 . Вопрос о связи критериев T^2 будет рассмотрен ниже, когда мы перейдем к дисперсионному анализу.

КРИТЕРИЙ ГОТЕЛИНГА T^2

Критерий Хотеллинга T^2 представляет собой многомерный аналог критерия Стьюдента, при помощи которого проверяется гипотеза о равенстве средних по данным двух выборок. В отличие от критерия Стьюдента, относящегося к одномерным случайным величинам, статистика T^2 позволяет рассматривать многомерные случайные величины¹.

¹ H. Hotelling, The generalization of Student's ratio, Ann. Math. Stat., 2, 1931; Relations between two sets of variates, Biometrika, 28, 1936.

В данном случае нулевая гипотеза заключается в том, что две p -мерные выборки обладают одинаковыми p -мерными средними ($H_0: \mu_1 = \mu_2$); при этом предполагаем, что точные значения дисперсий и ковариаций равны. Следует помнить, что на практике очень часто эти важные предположения сами по себе являются объектами проверки, что рассматривается в других местах этой книги.

Рассмотрим подробнее статистику T^2 . Пусть X_{aij} представляет собой значение i -й характеристики j -го индивидуума в выборке a , где $a=1, 2$; $i=1, 2, \dots, p$ и $j=1, 2, \dots, N_a$. Пусть также $\bar{X}_{1i}$ и $\bar{X}_{2i}$ — средние арифметические, полученные по результатам наблюдения i -й характеристики в первой и второй выборках соответственно. Определим величину d_i как

$$d_i = \frac{\bar{X}_{1i} - \bar{X}_{2i}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

а для i -й и k -й характеристик как

$$ns_{ik} = \sum_{j=1}^{n_1} (X_{1lj} - \bar{X}_{1l})(X_{1kj} - \bar{X}_{1k}) + \sum_{j=1}^{n_2} (X_{2lj} - \bar{X}_{2l})(X_{2kj} - \bar{X}_{2k}),$$

где

$$n = n_1 + n_2 - 2.$$

Величины s_{ik} образуют матрицу A

$$A = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \dots & \dots & \dots & \dots \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{pmatrix}.$$

Обозначим элементы матрицы A^{-1} , обратной матрице A , через s^{ik} . Следует отметить, что при больших значениях p нахождение величин s^{ik} — весьма трудоемкая операция. Статистика T^2 дается выражением

$$T^2 = \sum_{i=1}^p \sum_{k=1}^p s^{ik} d_i d_k. \quad (12.1)$$

Величина

$$\frac{(n+1-p) T^2}{n-p}$$

при условии, что гипотеза о равенстве двух p -мерных средних верна, распределена по закону Фишера с p и $n+1-p$ степенями свободы. Нулевая гипотеза отвергается, если вычисленное значение F превышает допустимое при заданном уровне значи-

В этой главе авторы не приводят примеров применения критерия Готелинга; вместо этого они предлагают рассмотреть другой подход к проверке той же гипотезы на примере приложения ее к палеонтологии.

ДИСПЕРСИОННЫЙ АНАЛИЗ

В олигоценовых отложениях найдены два близких вида ископаемых млекопитающих. Они очень похожи друг на друга по морфологическому облику. Предполагается, что существуют различия в длине лицевой части черепа, его ширине и характеристиках черепной коробки¹.

Таблица 12.1

Результаты измерения четырех признаков у 14 особей *Merychiododon culbertsoni* и у 12 особей *Prodesmatochoerus meeki*

<i>Merychiododon culbertsoni</i> (по Олсону)				<i>Prodesmatochoerus meeki</i>			
X	Y	Z	Q	X	Y	Z	Q
45	91	16	7,5	37	88	17	3,9
46	93	17	6,5	43	79	14	4,0
48	92	19	5,0	43	84	19	4,2
46	91	19	6,0	42	80	17	5,2
45	86	15	6,5	39	83	12	4,5
51	93	19	7,5	39	87	15	4,5
47	92	16	5,0	40	86	18	4,5
48	89	18	6,5	34	77	16	4,8
47	91	17,5	6,0	35	82	15	4,6
50	91	17	7,2	45	88	17	4,9
48	91	19	7,0	33	80	15	3,9
49	93	17,5	7,0	42	85	13	4,0
49	87	17,0	6,5				
49	91	19	7,7				

X — ширина черепной коробки в области теменного шва;

Y — максимальная длина ряда коренных зубов;

Z — максимальная длина черепной коробки;

Q — максимальная глубина черепной коробки, измеренная от заднего основания на уровне затылочного бугра.

По имеющимся эмпирическим данным нужно проверить предположение, согласно которому два вида различаются по

¹ Олсен, личное сообщение; C. B. Schultz, C. H. Falkenbach, Desmatochoerinae, a new subfamily of Oreodonts, Bull. Am. Nat. Hist., 105, art. 2, 1954.

совокупности признаков. По мнению авторов, этот пример наиболее характерен для иллюстрации задач такого типа, так как две рассматриваемые группы особей весьма сходны и наблюдаемые различия между ними основаны на качественном анализе, а не на каком-либо подсчете. При совместном рассмотрении средних значений всех признаков можно построить более чувствительный критерий, чем при сравнении средних значений каждого признака в отдельности. Халльд [19] показал, что даже в том случае, когда средние значения отдельных признаков различаются несущественно, их совместное рассмотрение за счет учета корреляционных зависимостей может привести к отрицанию гипотезы о равенстве многомерных средних.

Мы воспользуемся результатами измерения четырех признаков двух видов *Merychoidodon culbertsoni* и *Prodesmatochoerus meeki*, между которыми установлено качественное различие. С помощью многомерного дисперсионного анализа попытаемся выявить различия между многомерными средними значениями признаков. Результаты измерения этих признаков, любезно представленные Олсоном, приведены в табл. 12.1. В табл. 12.2

Таблица 12.2

Результаты сравнения средних арифметических отдельных признаков двух видов при помощи критерия Стьюдента в предположении, что распределения нормальны и дисперсии i -й характеристики равны ($\sigma_{1i}^2 = \sigma_{2i}^2$)

Признак	Среднее арифметическое		Результат применения критерия Стьюдента; различие
	<i>Merychoidodon culbertsoni</i>	<i>Prodesmatochoerus meeki</i>	
X	47,71	39,33	Существенное
Y	90,79	83,25	»
Z	17,57	15,67	Несущественное
Q	6,61	4,42	Существенное

$N_1=14$; $N_2=12$; уровень значимости $\alpha=0,05$.

приведены результаты применения одномерного критерия Стьюдента для сравнения средних значений отдельных признаков.

Необходимо отметить, что результаты любого подобного исследования значительно улучшатся, если все имеющиеся характеристики будут рассматриваться совместно. В этом случае основное внимание следует обратить на то, что суммарный эффект различия между средними всех четырех характеристик

существен для разделения двух видов. Из табл. 12.2 видно, что

расхождения между оценками средних для трех характеристик из четырех являются существенными при их рассмотрении порознь и без учета корреляционных зависимостей. Этот пример следует считать вспомогательным, иллюстрирующим подход к многомерному анализу.

Казалось бы, что лучший результат можно получить при рассмотрении как можно большего числа анатомических признаков двух изучаемых видов с последующим применением быстродействующих счетных машин для сокращения времени, которое требуется на вычисления. Однако при этом возникает серьезное затруднение, заключающееся в возможном превышении числа признаков над объемом выборки. Например, рассмотрим предположение о равенстве ковариационных матриц и предположение о p -мерном нормальном распределении. Очевидно, что очень большое число признаков приведет к ослаблению этих предположений. Насколько известно, этот вопрос специально не исследовался. В тех случаях, когда число признаков превышает число наблюдений, можно рекомендовать подход к этому типу задач, изложенный в главе о факторном анализе. Однако в любом случае следует консультироваться у математиков-специалистов.

Многомерный анализ требует очень большой затраты труда, поэтому в каждом конкретном случае нужны веские основания для его применения. Авторы считают, что многомерный анализ предпочтительнее в ситуации, аналогичной рассмотренному примеру, когда в основе таксономических различий лежит степень близости скорее видовых признаков, чем родовых. Это прежде всего относится к сравнению сортов, подвидов и близких видов. В подобных обстоятельствах применение статистического критерия служит лишь началом количественного анализа. Например, если при статистическом изучении двух видов выявлены существенные различия в комплексах признаков, то следующий этап исследования будет заключаться в отыскании тех признаков, на которых можно определить различие, или же в разработке количественного правила, позволяющего ученому отнести новую особь к тому или иному виду. Такое правило может быть найдено путем построения линейной дискриминантной функции, которая будет рассмотрена в следующем разделе.

Дисперсионный анализ; уравнения и пример

Методы обобщенного дисперсионного анализа, являющегося многомерным аналогом обычного одномерного дисперсионного анализа, подробно рассмотрены Рао [226]. В этих методах дисперсии и ковариации используются для проверки предположения о существенном различии средних значений многомерных

случайных величин¹. Обозначим матрицу, образованную несмещенными оценками дисперсий и ковариаций, буквой W , а ее детерминант $|W|$; вторую матрицу, элементы которой представляют собой несмещенные оценки дисперсий и ковариаций, вычисленные в предположении, что нулевая гипотеза верна, обозначим Q , а ее детерминант $-|Q|$.

Отношение $|W|$ к $|W+Q|$ представляет собой критерий для проверки гипотезы о равенстве многомерных средних, который был построен Уилксом [285] и применялся Рао [226]. В табл. 12.4 даны формулы для вычисления элементов матриц W , Q и $W+Q$ на клавишных электрических машинах. Матрица W — та же, что и матрица A в рассмотренном критерии Хотеллинга T^2 . Поэтому

$$\Lambda = \frac{1}{1 + \frac{T^2}{|W|}},$$

из чего следует наличие зависимости между критериями Λ и T^2 . Исходные данные для вычисления элементов матриц W и Q приведены в табл. 12.3 в виде сумм первых степеней, сумм квадратов и сумм смешанных произведений. Эти элементы (табл. 12.5) удобно вычислять, применяя формулы табл. 12.4. Элементы W_{ij} образуют матрицу W :

$$W = \begin{pmatrix} W_{11} & W_{12} & \dots & W_{1p} \\ W_{21} & W_{22} & \dots & W_{2p} \\ \dots & \dots & \dots & \dots \\ W_{p1} & W_{p2} & \dots & W_{pp} \end{pmatrix}.$$

Первая строка этой матрицы в нашем примере будет $W_{xx} W_{xy} W_{xz} W_{xq}$, вторая — $W_{yx} W_{yy} W_{yz} W_{yq}$ и т. д. Детерминант этой матрицы

$$|W| = \begin{vmatrix} 21,0 & 72,2 & 31,6 & 9,5 \\ 72,2 & 208,0 & 34,8 & -2,0 \\ 31,6 & 34,8 & 70,0 & 4,5 \\ 9,5 & -2,0 & 4,5 & 13,0 \end{vmatrix} = 29\,132\,813,3.$$

Элементы матрицы $W+Q$ обозначим через S_{ij} . Эти величины представляют собой суммы, соответствующие элементам Q_{ij} и W_{ij} матриц Q и W , приведенных в табл. 12.4.

¹ Хотя в данном примере рассматриваются только две выборки, дисперсионный анализ можно легко распространить на случай, когда их число больше двух. При двух выборках рекомендуем пользоваться критерием T^2 .

В данном примере

$$|S| = |W + Q| = \begin{vmatrix} 663,0 & 480,2 & 134,8 & 128,2 \\ 480,2 & 576,0 & 127,5 & 104,7 \\ 134,8 & 127,5 & 94,0 & 31,5 \\ 128,2 & 104,7 & 31,5 & 44,0 \end{vmatrix} = 174\,082\,293,9;$$

$$\Lambda = \frac{|W|}{|W + Q|} = \frac{29\,132\,813}{174\,082\,294} = 0,10735.$$

Конечная статистика

$$V = -m \ln \Lambda$$

при условии, что гипотеза о равенстве средних значений сравниваемых p -мерных случайных величин верна, распределена как χ^2 с pq степенями свободы. При этом в данном примере использованы следующие обозначения:

k — число выборок	$k = 2$
$q = k - 1$	$q = 2 - 1 = 1$
p — число признаков	$p = 4$
$n = N_1 + N_2 - 1$	$n = 12 + 14 - 1 = 25$
$m = n - \frac{p+q+1}{2}$	$m = 25 - \frac{4+1+1}{2} = 22$
pq — число степеней свободы	$pq = 4$

Таблица 12.3

Этапы вычислений

		<i>Merychoidodon</i> <i>culbertsoni</i> (выборка I)	<i>Prodesmatochoerus</i> <i>meeki</i> (выборка II)	Сумма
Сумма	$\sum X$	668	472	1 140
	$\sum Y$	1 271	999	2 270
	$\sum Z$	246	188	434
	$\sum Q$	92,5	53,0	145,5
	n	14	12	26
Сумма квадратов	$\sum X^2$	31 916	18 732	50 648
	$\sum Y^2$	115 447	83 317	198 764
	$\sum Z^2$	4 345,5	2 992,0	7 337,5
	$\sum Q^2$	621,4	236,1	857,5

		<i>Merychoidodon culbertsoni</i> (выборка I)	<i>Prodesmatochoerus meeki</i> (выборка II)	Сумма
Сумма смешанных произведений	$\sum XY$	60 658	39 353	100 011
	$\sum XZ$	11 754	7 410	19 164
	$\sum XQ$	4 420,1	2 087,7	6 507,8
	$\sum YZ$	22 349	15 670	38 019
	$\sum YQ$	8 398	4 410,0	12 808
	$\sum ZQ$	1 627,6	832,6	2 460,2

$$N = n_I + n_{II} \quad \sum_{I, II} X^2 = \sum_I X^2 + \sum_{II} X^2; \text{аналогично } \sum_{I, II} Y^2 \text{ и т. д.; } \sum_{I, II} XY = \\ = \sum_I XY + \sum_{II} XY; \text{аналогично } \sum_{I, II} XZ \text{ и т. д.}$$

Таким образом, $V = -m \ln \Lambda = 22 \cdot 1,791 = 39,402$. В связи с тем что вычисленное значение V превысило допустимое значение χ^2 при уровне значимости 0,05 и 4 степенях свободы ($\chi^2_{0,05} = 9,49$), нулевую гипотезу следует забраковать и средние значения совокупности четырех признаков у *M. culbertsoni* и *P. meeki* считать различными.

ОБОБЩЕННОЕ РАССТОЯНИЕ

В 1927 г. Махаланобис¹ ввел статистическое понятие *кастового различия*, или *обобщенного расстояния*. Простейшая форма обобщенного расстояния (предполагается, что число наблюдений для каждого признака в одной выборке одинаково, т. е. $n_{1i} = n_{1j}$, $n_{2i} = n_{2j}$ и т. д.) эквивалентна коэффициенту расового сходства Пирсона; при этом следует помнить об одной важной особенности: коэффициент $n_1 n_2 / (n_1 + n_2)$ устранен. Это позволяет избежать некоторых трудностей, связанных с объемом выборки, возникающих при применении коэффициента расового сходства в качестве меры расхождения. Дело в том, что коэффициент расового сходства возрастает с увеличением объема выборки, тогда как обобщенное расстояние D^2 свободно от этого недостатка.

¹ P. C. Mahalanobis, Analysis of race-mixture in Bengal, Journ. As. Soc. Beng., 23, 1927; On test and measures of group divergence, Journ. As. Soc. Beng., 26, 1930.

Таблица 12.4

Рабочая схема многомерного дисперсионного анализа двух выборок (I и II) на примере трех переменных X, Y, Z . Такую форму записи можно легко распространить на случай любого числа признаков. В данном примере в целях упрощения величина Q устранена

Тип изменчивости	Число степеней свободы	Суммы произведений				
		X^2	Y^2	Z^2	XY	XZ YZ
Между выборками	Число выборок минус единица	$Q_{ij} = \left[\frac{(\sum X_I)^2}{n_I} + \frac{(\sum X_{II})^2}{n_{II}} \right] - \frac{(\sum X_I + \sum X_{II})^2}{n_I + n_{II}}$	То же для Y и Z		$\left(\frac{\sum X_I \sum Y_I}{n_I} + \frac{\sum X_{II} \sum Y_{II}}{n_{II}} \right) - \left[\frac{(\sum X_I + \sum X_{II})(\sum Y_I + \sum Y_{II})}{n_I + n_{II}} \right]$	То же для XZ и YZ
Внутри выборок	Число индивидуумов минус число выборок	$W_{ij} = (\sum X_I^2 + \sum X_{II}^2) - \left[\frac{(\sum X_I)^2}{n_I} + \frac{(\sum X_{II})^2}{n_{II}} \right]$	То же для Y и Z		$\left(\sum_I XY + \sum_{II} XY \right) - \left[\frac{\sum_I X \sum_I Y}{n_I} + \frac{\sum_{II} X \sum_{II} Y}{n_{II}} \right]$	То же для XZ и YZ
Общая изменчивость		$S_{ij} = (\sum X_I^2 + \sum X_{II}^2) - \frac{(\sum X_I + \sum X_{II})^2}{n_I + n_{II}}$	То же для Y и Z		$\left(\sum_I XY + \sum_{II} XY \right) - \left[\frac{(\sum_I X + \sum_{II} X)(\sum_{II} Y + \sum_{II} Y)}{n_I + n_{II}} \right]$	То же для XZ и YZ

Число выборок — 2, $n_1 = 14$, $n_{11} = 12$, четыре
случайные величины X, Y, Z, Q

Сравниваются два вида: *Merychoidodon culbertsoni* и *Prodesmatoherus meeki*

Тип изменчивости	Число степеней свободы	Компоненты изменчивости	Суммы квадратов и смешанные произведения									
			X ²	Y ²	Z ²	Q ²	XY	XZ	XQ	YZ	YQ	ZQ
Между выборками . .	1	Q _{ij}	(453)	(368)	(24)	(31)	408,0	103,2	118,7	92,7	106,7	27
Внутри выборок . . .	27	W _{ij}	(210)	(208)	(70)	(13)	72,2	31,6	9,5	34,8	—2,0	4,5
Общая изменчивость	25	S _{ij}	(663)	(576)	(94)	(44)	480,2	134,8	128,2	127,5	104,7	31,5

Рассмотрим следующую типичную задачу. Допустим, что нужно сравнить в довольно большой коллекции близкие виды, чтобы можно было охарактеризовать «таксономическое расстояние» каждого вида от всех остальных. Таксономическое расстояние должно быть некоторым числовым показателем, основанным на результатах измерения признаков в выборках из совокупности особей каждого вида. Значения признаков следует рассматривать совместно, что приведет к анализу ковариационной матрицы. Одним из возможных решений предложенной задачи будет следующее: в качестве показателя надо применить обобщенное расстояние D^2 , вычисленное для всех пар сравниваемых видов.

Предложенное Рао [226] выражение для вычисления D^2 имеет следующий вид:

$$D^2 = \sum_i \sum_j \lambda^{ij} d_i d_j,$$

где $\lambda^{ij} = \sigma^{ij}$, $d_i = \mu_{1i} - \mu_{2i}$. Если значения неизвестных параметров заменить оценками, то выборочное обобщенное расстояние $\hat{D}^2$ будет представлять выражение

$$\hat{D}^2 = \sum_i \sum_j \hat{\lambda}^{ij} \hat{d}_i \hat{d}_j,$$

которое в развернутом виде можно записать как

$$\hat{D}^2 = \frac{1}{|R|} \sum_{i=1}^p \sum_{j=1}^p R_{ij} \frac{\bar{x}_{1i} - \bar{x}_{2i}}{\sigma_i} \frac{\bar{x}_{1j} - \bar{x}_{2j}}{\sigma_j},$$

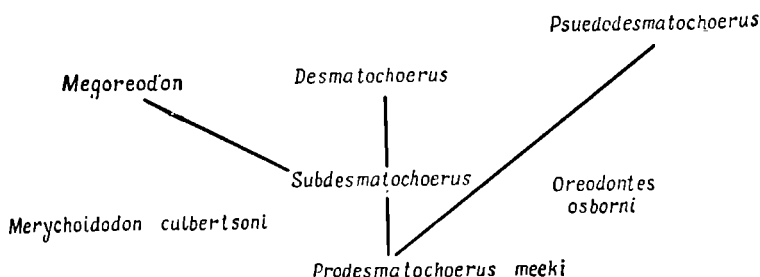
где R — матрица, образованная оценками коэффициентов корреляции:

$$R = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \dots & \dots & \dots & \dots \\ r_{p1} & r_{p2} & \dots & 1 \end{pmatrix},$$

а R_{ij} — алгебраическое дополнение элемента, стоящего на пересечении i -й строки и j -го столбца. Признаки обозначены через $1, 2, \dots, i, \dots, j, \dots, p$, а средние арифметические результатов измерения i -го признака в первой и второй выборках — соответственно через x_{1i} и x_{2i} . Дисперсия σ^2 i -го признака предполагается одинаковой для обеих выборок, т. е. $\sigma_{1i}^2 = \sigma_{2i}^2 = \sigma_i^2$.

В ситуациях, требующих рассмотрения свыше трех и четырех признаков у нескольких видов, приведенная выше форма D^2 становится непригодной при обычных вычислительных средствах в связи с резким возрастанием объема вычислений. Рао [226] предложил преобразование, в результате которого получается новый набор характеристик признаков, которые уже не коррелированы. В итоге работа становится менее трудоемкой, так как коэффициенты корреляции нового набора случайных величин равны нулю.

Рассмотрим реальный пример. С целью характеристики таксономического расстояния при помощи D^2 были проанализированы семь разных видов ореодонтов. Зависимости в стратиграфическом смысле для семи видов этого рода описаны в работе Шульца и Фалькенбаха (1954 г.), из которой заимствована фиг. 12.1. Линии на этой фигуре характеризуют степень таксономического родства.



Фиг. 12.1. Таксономические связи ореодонтов.

При изучении количественных различий между семью упомянутыми видами были использованы те же четыре признака, что и в предыдущем примере, касающемся *Merychoidodon culbertsoni* и *Prodesmatochoerus meeki*. В табл. 12.6 приведены

Subdesmatcoherus sp.

X	Y	Z	Q
47,0	99	26,0	15,0
42,0	93	26,0	16,0
40,0	90	22,0	13,0
46,0	100	22,0	11,0
46,0	96	24,0	16,0
42,0	88	26,0	15,0
43,0	89	23,0	14,0
44,0	78	23,0	13,0
44,0	90	25,0	11,0
47,0	99	27,0	15,0
47,0	92	27,0	13,0

$n = 11$

$\sum X = 488$	$\sum Y = 1\,014$	$\sum Z = 271$	$\sum Q = 152$
$\sum X^2 = 21\,708$	$\sum Y^2 = 93\,880$	$\sum Z^2 = 6713$	$\sum Q^2 = 2132$
$\sum XY = 45\,067$	$\bar{X} = 44,364$		
$\sum XZ = 12\,041$	$\bar{Y} = 92,182$		
$\sum XQ = 6\,743$	$\bar{Z} = 24,636$		
$\sum YZ = 25\,012$	$\bar{Q} = 13,818$		
$\sum YQ = 14\,030$			
$\sum ZQ = 3\,760$			

Megoreodon gigas loomisi

X	Y	Z	Q
78,0	165,0	35,0	18,0
77,0	165,0	37,0	19,0
65,0	148,0	30,0	20,0
74,0	163,0	31,0	15,0
65,0	169,0	31,0	16,0
70,0	176,0	34,0	23,0
69,0	161,0	28,0	13,0
67,0	178,0	31,0	14,0

X	Y	Z	Q
65,0	174,0	34,0	18,0
64,0	168,0	28,0	13,0
68,0	166,0	32,0	15,0

$n = 11$

$$\begin{aligned} \sum X &= 762 & \sum Y &= 1\,833 & \sum Z &= 351 & \sum Q &= 184 \\ \sum X^2 &= 53\,034 & \sum Y^2 &= 306\,121 & \sum Z^2 &= 11\,281 & \sum Q^2 &= 3178 \\ \bar{X} &= 69,273 & \bar{Y} &= 166,636 & \bar{Z} &= 31,909 & \bar{Q} &= 16,727 \end{aligned}$$

$$\sum XY = 126\,947$$

$$\sum XZ = 24\,405$$

$$\sum XQ = 12\,784$$

$$\sum YZ = 58\,554$$

$$\sum YQ = 30\,653$$

$$\sum ZQ = 5\,930$$

Oreodontes osborni

x	y	z	q
42,0	81,0	15,0	8,0
48,0	83,0	18,0	8,6
45,0	87,0	18,0	9,0
48,0	83,0	17,0	8,0
46,0	84,0	16,0	6,1
51,0	87,0	21,0	7,9
46,0	80,0	17,0	7,0
50,0	90,0	18,0	8,1
46,0	85,0	16,0	6,5
48,0	85,0	15,0	7,2
47,0	85,0	17,0	8,0
49,0	83,0	18,0	7,7

	<i>x</i>	<i>y</i>	<i>z</i>	<i>q</i>
<i>n</i> = 15	43,0	79,0	15,0	7,1
	47,0	87,0	19,0	7,5
	46,0	87,0	18,0	8,0

$$\begin{aligned}
 \sum X &= 702,00 & \sum Y &= 1266 & \sum Z &= 258 & \sum Q &= 114,7 \\
 \sum X^2 &= 32\,934 & \sum Y^2 &= 106\,976 & \sum Z^2 &= 4476 & \sum Q^2 &= 885,23 \\
 \bar{X} &= 46,800 & \bar{Y} &= 84,400 & \bar{Z} &= 17,200 & \bar{Q} &= 7,647 \\
 \sum XY &= 59\,306,00 \\
 \sum XZ &= 12\,112,00 \\
 \sum XQ &= 5\,372,00 \\
 \sum YZ &= 21\,815,00 \\
 \sum YQ &= 9\,690,50 \\
 \sum ZQ &= 1\,980,70
 \end{aligned}$$

Desmatochoerus hatcheri

	<i>x</i>	<i>y</i>	<i>z</i>	<i>q</i>
<i>n</i> = 11	58,0	129,0	26,0	16,0
	52,0	126,0	27,0	18,0
	50,0	122,0	28,0	22,0
	52,0	123,0	29,0	18,0
	60,0	133,0	33,0	17,0
	61,0	122,0	28,0	17,0
	54,0	132,0	30,0	17,0
	65,0	131,0	32,0	18,0
	55,0	130,0	32,0	17,0
	64,0	125,0	26,0	16,0
	56,0	124,0	28,0	16,0

$$\begin{aligned}
 \sum X &= 627 & \sum Y &= 1402 & \sum Z &= 319 & \sum Q &= 192 \\
 \sum X^2 &= 35\,991 & \sum Y^2 &= 178\,944 & \sum Z^2 &= 9311 & \sum Q^2 &= 3380 \\
 \bar{X} &= 57,000 & \bar{Y} &= 127,455 & \bar{Z} &= 29,000 & \bar{Q} &= 17,455 \\
 \sum XY &= 79\,989 \\
 \sum XZ &= 18\,200 \\
 \sum XQ &= 10\,900 \\
 \sum YZ &= 40\,743 \\
 \sum YQ &= 24\,446 \\
 \sum ZQ &= 5\,571
 \end{aligned}$$

Psuedodesmatochoerus

<i>x</i>	<i>y</i>	<i>z</i>	<i>q</i>
60,0	114,0	27,0	20,0
60,0	118,0	31,0	19,0
60,0	111,0	31,0	21,0
58,0	102,0	30,0	20,0
55,0	116,0	28,0	20,0
59,0	117,0	29,0	17,0
59,0	114,0	24,0	17,0
60,0	121,0	25,0	19,0

n = 8

$$\begin{aligned}
 \sum X &= 471 & \sum Y &= 913 & \sum Z &= 225 & \sum Q &= 153 \\
 \sum X^2 &= 27\,751 & \sum Y^2 &= 104\,427 & \sum Z^2 &= 6377 & \sum Q^2 &= 2941 \\
 \bar{X} &= 58,875 & \bar{Y} &= 114,125 & \bar{Z} &= 28,125 & \bar{Q} &= 19,125 \\
 \sum XY &= 53\,765 \\
 \sum XZ &= 13\,247 \\
 \sum XQ &= 9\,006 \\
 \sum YZ &= 25\,639 \\
 \sum YQ &= 17\,439 \\
 \sum ZQ &= 4\,316
 \end{aligned}$$

Merychoidodon culbertsoni

<i>x</i>	<i>y</i>	<i>z</i>	<i>q</i>
45,0	91,0	16,0	7,5
46,0	93,0	17,0	6,5
48,0	92,0	19,0	5,0
46,0	91,0	19,0	6,0
45,0	86,0	15,0	6,5
51,0	93,0	19,0	7,5
47,0	92,0	16,0	5,0
48,0	89,0	18,0	6,5
47,0	91,0	17,5	6,0
50,0	91,0	17,0	7,2
48,0	91,0	19,0	7,6

	<i>x</i>	<i>y</i>	<i>z</i>	<i>q</i>
<i>n</i> = 14	49,0	93,0	17,5	7,0
	49,0	87,0	17,0	6,5
	49,0	91,0	19,0	7,7

$\sum X = 668$	$\sum Y = 1271$	$\sum Z = 246,0$	$\sum Q = 92,5$
$\sum X^2 = 31\,916$	$\sum Y^2 = 115\,447$	$\sum Z^2 = 4345,5$	$\sum Q^2 = 621,39$
$\bar{X} = 47,714$	$\bar{Y} = 90,786$	$\bar{Z} = 17,571$	$\bar{Q} = 6,607$
$\sum XY = 60\,658$			
$\sum XZ = 11\,754,0$			
$\sum XQ = 4\,420,1$			
$\sum YZ = 22\,349,0$			
$\sum YQ = 8\,398,0$			
$\sum ZQ = 1\,627,6$			

Prodesmatochoerus meeki

	<i>x</i>	<i>y</i>	<i>z</i>	<i>q</i>
<i>n</i> = 12	37,0	88,0	17,0	3,9
	43,0	79,0	14,0	4,0
	43,0	84,0	19,0	4,2
	42,0	80,0	17,0	5,2
	39,0	83,0	12,0	4,5
	39,0	87,0	15,0	4,5
	40,0	86,0	18,0	4,5
	34,0	77,0	16,0	4,8
	35,0	82,0	15,0	4,6
	45,0	88,0	17,0	4,9
	33,0	80,0	15,0	3,9
	42,0	85,0	13,0	4,0

$\sum X = 472$	$\sum Y = 999$	$\sum Z = 188$	$\sum Q = 53,0$
$\sum X^2 = 18\,732$	$\sum Y^2 = 83\,317$	$\sum Z^2 = 2992$	$\sum Q^2 = 236,06$
$\bar{X} = 39,333$	$\bar{Y} = 83,250$	$\bar{Z} = 15,666$	$\bar{Q} = 4,417$
$\sum XY = 39\,353$			
$\sum XZ = 7\,410$			
$\sum XQ = 2\,087,7$			
$\sum YZ = 15\,670$			
$\sum YQ = 4\,410,0$			
$\sum ZQ = 832,6$			

названия видов и результаты измерения признаков, для которых введены обозначения, данные в табл. 12.1.

Ниже мы продолжим рассмотрение вычислительных операций, необходимых для решения задач.

В табл. 12.7 приведены оценки средних значений всех четырех признаков для семи видов.

Табл. 12.8 представляет собой ковариационную матрицу, элементы которой являются обобщенными по всем семи видам оценками дисперсий и ковариаций. Например, для первого признака оценка дисперсии вычисляется по формуле

$$\lambda_{xx} = \frac{\left(\sum_I X^2 + \dots + \sum_{VII} X^2 \right) - \left[\frac{\left(\sum_I X \right)^2}{n_I} + \dots + \frac{\left(\sum_{VII} X \right)^2}{n_{VII}} \right]}{n_I + n_{II} + \dots + n_{VII} - 7},$$

где n_I — число наблюдений для *Merychoidodon culbertsoni*; ...; n_{VII} — число наблюдений для *Megoreodon*.

Аналогично вычисляются величины λ_{yy} , λ_{zz} , λ_{qq} .

Оценки ковариаций λ_{xy} , λ_{xz} , λ_{xq} , λ_{yz} , λ_{yq} , λ_{zq} вычисляются по формуле, аналогичной

$$\lambda_{xy} = \frac{\left(\sum_I XY + \dots + \sum_{VII} XY \right) - \left[\frac{\sum_I X \sum_I Y}{n_I} + \dots + \frac{\sum_{VII} X \sum_{VII} Y}{n_{VII}} \right]}{n_I + n_{II} + \dots + n_{VII} - 7}.$$

Таблица 12.7

Средние значения признаков в группах

Группа	x	y	z	q
<i>Merychoidodon culbertsoni</i>	47,714	90,786	17,571	6,607
<i>Prodesmatochoerus</i>	39,333	83,250	15,666	4,417
<i>Subdesmatochoerus</i>	44,364	92,182	24,636	13,818
<i>Pseudodesmatochoerus</i>	58,875	114,125	28,125	19,125
<i>Desmatochoerus</i>	57,000	127,455	29,000	17,455
<i>Oreodontes osborni</i>	46,800	84,400	17,200	7,647
<i>Megoreodon</i>	69,273	166,636	31,909	16,727
Сумма	363,359	758,834	164,107	85,796
Генеральные средние	51,908	108,405	23,444	12,257

Обобщенная ковариационная матрица

	x	y	z	q
X	11,6	3,587	2,600	0,071
Y		25,360	2,880	-0,389
Z			4,467	1,363
Q				2,610

Чтобы получить нормированные величины x' , y' , z' и q' , преобразуем данные табл. 12.7 следующим образом: из каждого значения среднего, приведенного в таблице, вычтем соответствующее значение генеральной средней, а затем разделим полученную разность на соответствующую оценку стандартного отклонения. Например, чтобы вычислить первый элемент табл. 12.10, нужно сделать следующее:

$$\frac{47,714 - 51,908}{3,406} = -\frac{4,194}{3,406} = -1,23140.$$

Нормированную случайную величину можно записать как $(\bar{x} - \bar{x})/\sigma\bar{x} = x'$. Если коэффициент корреляции ρ_{ij} представить в форме $\lambda_{ij}/\sqrt{\lambda_{ii}\lambda_{jj}}$, матрица оценок коэффициентов корреляции величин X , Y , Z , Q , приведенная в табл. 12.9, будет служить ковариационной матрицей для величин x' , y' , z' и q' .

Таблица 12.9

Матрица оценок коэффициентов корреляции

	x	y	z	q	
X	1	0,209	0,361	0,013	Например: $\rho_{xy} = \frac{\lambda_{xy}}{\sqrt{\lambda_{xx}\lambda_{yy}}} =$ $= \frac{3,587}{\sqrt{11,6 \cdot 25,36}} = 0,209.$
Y		1	0,271	-0,047	
Z			1	0,767	
Q				1	
Оценка стандартных отклонений	3,406	5,036	2,113	1,616	Например: $\sqrt{11,6} = 3,406$

Теперь нам нужно преобразовать x' , y' , z' и q' так, чтобы они стали независимыми и имели средние равными нулю, а дисперсии равными единице. Согласно Рао (1955 г.), новые величины x , y , z и q , полученные из уравнений

$$x = x',$$

$$y = y' - \rho_{xy}x,$$

$$z = z' - a'_{32}y - a'_{31}x = z' - a'_{32}y - \rho_{xz}x,$$

$$q = q' - a'_{43}z - a'_{42}y - a'_{41}x = q' - a'_{43}z - a'_{42}y - \rho_{xq}x,$$

хотя и являются некоррелированными, имеют дисперсии, не равные единице. Однако, если x , y , z и q разделить на соответствующие стандартные отклонения, их дисперсия будет равна единице.

Таблица 12.10

Нормированные средние значения признаков

Группа	x'	y'	z'	q'
<i>Merychiododon culbertsoni</i>	-1,231	-3,499	-2,779	-3,497
<i>Prodesmatochoerus</i>	-3,692	-4,995	-3,680	-4,853
<i>Subdesmatochoerus</i>	-2,215	-3,221	0,564	0,967
<i>Pseudodesmatochoerus</i>	2,046	1,136	2,215	4,251
<i>Desmatochoerus</i>	1,495	3,783	2,629	3,218
<i>Oreodontes osborni</i>	-1,500	-4,767	-2,954	-2,853
<i>Megoreodon</i>	5,098	11,563	4,006	2,767

Теперь нужно вычислить коэффициенты a'_{32} , a'_{42} , a'_{43} и оценки дисперсий $V(y)$, $V(z)$ и $V(q)$:

$$a'_{21} = \rho_{xy} = 0,209;$$

$$V(y) = 1 - \rho_{xy}^2 = 1 - 0,209^2 = 0,956; \quad \sqrt{V(y)} = 0,978;$$

$$\rho_{xz} = a'_{31} = 0,3612;$$

$$b'_{32} = \rho_{yz} - \rho_{xy}\rho_{xz} = 0,271 - 0,209 \cdot 0,361 = 0,195;$$

$$a'_{32} = \frac{b'_{32}}{\sqrt{V(y)}} = \frac{0,195}{0,956} = 0,204;$$

$$V(z) = 1 - \rho_{xz}^2 - b'_{32}a'_{32} = 1 - 0,361^2 - \frac{0,195^2}{0,956} = 0,830;$$

$$\sqrt{V(z)} = 0,911;$$

$$b'_{41} = \rho_{xq} = 0,013 = a'_{41};$$

$$b'_{42} = \rho_{YQ} - \rho_{XY}\rho_{XQ} = -0,047 - 0,209 \cdot 0,013 = -0,050;$$

$$a'_{42} = \frac{b'_{42}}{V(y)} = -\frac{0,050}{0,956} = -0,052;$$

$$\begin{aligned} b'_{43} &= \rho_{ZQ} - a'_{32}b'_{42} - \rho_{XZ}\rho_{XQ} = \\ &= 0,768 + 0,204 \cdot 0,050 - 0,361 \cdot 0,013 = 0,773; \end{aligned}$$

$$a'_{43} = \frac{b'_{43}}{V(z)} = \frac{0,773}{0,830} = 0,931;$$

$$\begin{aligned} V(q) &= 1 - \rho_{XQ}^2 - b'_{42}a'_{42} - b'_{43}a'_{43} = \\ &= 1 - 0,013^2 - 0,050 \cdot 0,052 - 0,773 \cdot 0,931 = 0,278; \end{aligned}$$

$$\sqrt{V(q)} = 0,527.$$

Таким образом, следуя рекомендациям Рао (1955 г.), получим:

$$x^* = x = x',$$

$$0,978y^* = y = y' - 0,209x,$$

$$0,911z^* = z = z' - 0,204y - 0,361x,$$

$$0,527q^* = q = q' - 0,931z + 0,052y - 0,013x.$$

Каждое из этих четырех уравнений применяется последовательно ко всем семи видам. Значения x , y , z и q в табл. 12.11 получены путем соответствующих преобразований данных табл. 12.10, а последующее их деление на соответствующие оценки стандартных отклонений дает в результате x^* , y^* , z^* , q^* , приведенные в табл. 12.12. Значения x^* получаются простым перенесением значений x' и x в табл. 12.12. Вычисление y^* можно рассмотреть на примере для *Merychoidodon culbertsoni*: $0,978 y^* = -3,499 + 0,209 \cdot 1,23 = -3,241$.

Следовательно, $y^* = -3,314$.

Т а б л и ц а 12.11

Группа	x	y	z	q
<i>Merychoidodon culbertsoni</i>	-1,231	-3,241	-1,673	-2,092
<i>Prodesmatochoerus</i>	-3,692	-4,223	-1,485	-3,643
<i>Subdesmatochoerus</i>	-2,215	-2,758	1,927	-0,943
<i>Pseudodesmatochoerus</i>	2,046	0,708	1,331	3,022
<i>Desmatochoerus</i>	1,495	3,470	1,381	2,094
<i>Oreodontes osborni</i>	-1,500	-4,453	-1,504	-1,666
<i>Megoreodon</i>	5,098	10,497	0,023	3,227

Оценки средних, преобразованных значений признаков

Группа	x^*	y^*	z^*	q^*
<i>Merychiododon culbertsoni</i>	-1,231	-3,314	-1,837	-3,970
<i>Prodesmatochoerus</i>	-3,692	-4,319	-1,630	-6,911
<i>Subdesmatochoerus</i>	-2,215	-2,820	2,115	-1,788
<i>Pseudodesmatochoerus</i>	2,046	0,724	1,462	5,733
<i>Desmatochoerus</i>	1,495	3,549	1,516	3,973
<i>Oreodontes osborni</i>	-1,500	-4,554	-1,651	-3,160
<i>Megoreodon</i>	5,098	10,735	0,025	6,123

Для *Prodesmatochoerus*: $0,978y^* = -4,995 + 0,209 \cdot 3,692 = -4,223$.

Следовательно, $y^* = -4,319$.

Для *Megoreodon*: $0,978y^* = 11,563 - 0,209 \cdot 5,098 = 10,497$.

Следовательно, $y^* = 10,735$.

Значения z^* , например, для *Merychiododon culbertsoni* получают из уравнения $0,911z^* = -2,7792 + 0,204 \cdot 3,241 + 0,361 \times \times 1,231 = -1,673$.

Следовательно, $z^* = -1,837$.

Для *Megoreodon*: $0,911z^* = 4,006 - 0,204 \cdot 10,497 - 0,361 \times \times 5,098 = 0,023$.

Следовательно, $z^* = 0,025$.

Величины q^* находят из выражений для *Merychiododon culbertsoni*: $0,527q^* = -3,497 + 0,0931 \cdot 1,673 - 0,052 \cdot 3,241 + 0,013 \times \times 1,234 = -2,092$.

Следовательно, $q^* = -3,970$.

Чтобы по возможности избежать ошибок при вычислениях, следует сначала узнать значения x , y , z и q , а затем разделить их на соответствующие оценки стандартных отклонений $\sqrt{V(x)}$, ..., $\sqrt{V(q)}$, что дает результаты, представленные в табл. 12.13.

Обобщенное расстояние D^2 для двух видов (A и B) находится из соотношения

$$D_{AB}^2 = (x_A^* - x_B^*)^2 + (y_A^* - y_B^*)^2 + \dots + (q_A^* - q_B^*)^2;$$

например, D^2 для *M. culbertsoni* и *Prodesmatochoerus* будет $D^2 = (-1,231 + 3,692)^2 + (-3,314 + 4,319)^2 + (-1,837 + 1,630)^2 + + (-3,970 + 6,911)^2 = 15,761$; для *Subdesmatochoerus* и *Pseudodesmatochoerus* $D^2 = (-2,215 - 2,046)^2 + \dots + (-1,788 - 5,733)^2 = 87,823$.

Т а б л и ц а 12.13

Mc		Pr		Su		Oo		Ps		De		Me	
Oo	2,297	Mc	15,761	Oo	19,586	Mc	2,297	De	11,385	Ps	11,385	De	71,467
Pr	15,761	Oo	18,929	Mc	21,585	Pr	18,929	Su	87,723	Me	71,467	Ps	111,746
Su	21,585	Su	44,698	Pr	44,698	Su	19,586	Me	111,746	Su	87,882	Su	304,185
De	128,847	De	217,154	Ps	87,823	Ps	129,217	Oo	129,217	Mc	128,847	Mc	342,752
Ps	132,066	Ps	227,793	De	87,882	De	135,531	Mc	132,066	Oo	135,531	Oo	366,258
Me	342,752	Me	476,487	Me	304,185	Me	366,258	Pr	227,793	Pr	217,154	Pr	476,487

Значения D^2 для всех пар видов расположены в порядке их возрастания, что позволяет отчетливее видеть степень близости. Например, в первом столбце показано, что по морфологическим признакам *Merychiododon culbertsoni* и *Oreodonts osborni* наиболее близкие, тогда как *Megoreodon* наиболее удаленный от *Merychiododon culbertsoni*.

Mc — *Merychiododon culbertsoni*; Pr — *Prodesmatochoerus*; Su — *Subdesmatochoerus*; Oo — *Oreodontes osborni*; Ps — *Pseudodesmatochoerus*; De — *Desmatochoerus*; Me — *Megoreodon* (см. фиг. 12.2).

Наиболее раннее применение количественных методов в биологии принадлежит Пирсону, который разработал статистический критерий, применявшийся им для сравнения человеческих черепов, оказавшийся эффективным, особенно в условиях малых выборок. Этот критерий, известный под названием *коэффициента расового сходства*, был опубликован Пирсоном в работе Тилдесли¹. При этом малый объем выборок сочетался с измерением большого числа признаков.

Коэффициент расового сходства (C.R.L.) для двух p -мерных выборок объемом n_1 и n_2 соответственно дается выражением

$$C. R. L. = \frac{1}{p} \sum_{l=1}^p \frac{n_{1l}n_{2l}}{n_{1l} + n_{2l}} \left(\frac{\bar{x}_{1l} - \bar{x}_{2l}}{s_l} \right)^2, \quad (12.2)$$

где

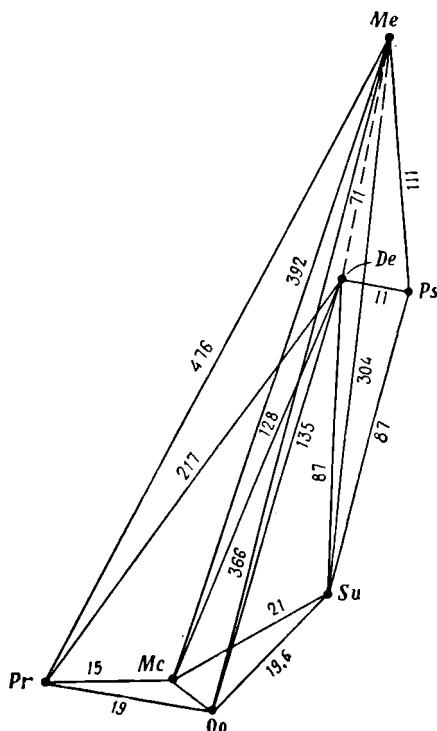
$$s_l = \sqrt{\frac{n_{1l}s_{1l}^2 + n_{2l}s_{2l}^2}{n_{1l} + n_{2l} - 2}},$$

в предположении, что признаки независимы и обе выборки имеют одинаковые дисперсии. Предположение о равенстве дисперсий введено в связи с тем, что в

рассматриваемых малых выборках не обязательно $n_{1i} = n_{2i}$. При сравнении значений для двух пар выборок используется выражение

$$C. R. L. = \frac{\bar{n}_1 + \bar{n}_2}{n_1 n_2} \frac{1}{p} \left\{ \sum \frac{n_{1l}n_{2l}}{n_{1l} + n_{2l}} \left(\frac{\bar{x}_{1l} - \bar{x}_{2l}}{s_l} \right)^2 - 1 \right\}, \quad (12.3)$$

где $\bar{n}_1 = \sum_{l=1}^p (n_{1l}/p)$, а $\bar{n}_2$ — определяется точно так же.



Фиг. 12.2. Графическое изображение данных табл. 12.13, характеризующее связи семи видов, выраженные величинами D^2 для всех возможных пар. Формальный метод выделения подгрупп близко связанных членов (плеяды) см. [226].

Условные обозначения см. табл. 12.13.

¹ M. Tildesley, A first study of the Burmese skull, Biometrika, 13, 1921. **275**

Необходимо отметить следующее:

1. Если две выборки взяты из одной и той же популяции, то для i -го признака математическое ожидание величины $\bar{x}_{1i} - \bar{x}_{2i}$ равно нулю.

2. В любом случае дисперсия величины $\bar{x}_{1i} - \bar{x}_{2i}$ оценивается как $s_{1i}^2/n_1 + s_{2i}^2/n_2$.

3. Проведенные эмпирические исследования показали, что распределения результатов измерений биологических признаков очень часто можно рассматривать как приблизительно нормальные. Из этого следует, что распределения средних арифметических будут приближаться к нормальному закону. Поэтому величину $\frac{\bar{x}_{1i} - \bar{x}_{2i}}{\sqrt{s_{1i}^2/n_1 + s_{2i}^2/n_2}}$ можно рассматривать как распределен-

ную нормально со средним, равным 0, и дисперсией, равной 1¹.

Согласно положениям, развитым в гл. 4 и касающимся χ^2 -распределения, нетрудно видеть следующее: 1) для каждого i существует приведенная выше случайная величина, которая распределена нормально со средним, равным 0, и дисперсией, равной 1; 2) каждая такая величина соответствует одному из p независимых признаков; 3) следовательно, сумма квадратов таких случайных величин $\left(\sum_{i=1}^p x_i^2\right)$ будет распределена как χ^2 с p степенями свободы. Вывод С.Р.Л. основан главным образом на предположении о независимости измеряемых характеристик. Очевидно, Пирсон принимал это предположение как обоснованное по крайней мере для измеряемых признаков черепов *Homo sapiens*. Ходжес, цитируя Тилдесли, отмечал, что между некоторыми измеряемыми признаками черепов корреляция очень незначительная и часто ею можно пренебречь. Такие же исследования были проведены авторами [202] на обширных группах как позвоночных, так и беспозвоночных животных. Оказалось, что корреляционные зависимости обычно бывают значительными. Следует отметить, что, кроме высокой корреляционной зависимости, нет никаких ограничений, упоминаемых Тилдесли. Пирсон² отмечал, что нетрудно разработать теорию для случая зависимых признаков, но вычисления, которые пришлось бы производить при ее применении, были бы практически невыполнимы даже при умеренном числе признаков. Поэтому он предложил при применении С.Р.Л. выбирать такие характеристики, которые или независимы, или связаны слабой корреляционной зависимостью. Пирсон в той же работе рассматривал предположение о независимости, лежащее в основе С.Р.Л.; он предло-

¹ В действительности, как показано в гл. 6, это выполняется достаточно точно только при больших выборках [уравнение (6.5)].

² К. Pearson, On the coefficient of racial likeness, *Biometrika*, 18, 1926.

жил вместо него другую форму, пригодную для случая, когда характеристики нельзя рассматривать как почти независимые и когда их не очень много.

Предполагается, что сравниваемые выборки имеют равные ковариационные матрицы и, следовательно, для i -й и k -й характеристик $r_{1ik}=r_{2ik}=r_{ik}$. Однако автор считает, что для живых организмов, за исключением *Homo sapiens*, это предположение не всегда оправдано и должно быть проверено в каждом конкретном случае. Рассмотренные выше результаты исследования многочисленных коэффициентов корреляции [202] характеристик широкой группы животных показали, что коэффициенты видов и даже подвидов сильно отличаются друг от друга; причем это касается не только одновременно существующих групп животных, но и видов и подвидов в их эволюционной последовательности.

Однако критерий, заменяющий С.Р.Л., в этих условиях может быть построен на основании следующего предположения: положим

$$y_l = \sqrt{\frac{\bar{n}_1 \bar{n}_2}{n_1 + n_2}} \left(\frac{\bar{x}_{1l} - \bar{x}_{2l}}{\sigma_l} \right)$$

для i -характеристики, и пусть R будет корреляционной матрицей p характеристик

$$R = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \dots & \dots & \dots & \dots \\ r_{p1} & r_{p2} & \dots & 1 \end{pmatrix},$$

Обозначим через R_{ik} алгебраическое дополнение элемента r_{ik} . Тогда величина

$$\frac{1}{|R|} \sum_{l=1}^p \sum_{k=1}^p R_{lk} y_l y_k \quad (12.4)$$

будет распределена как χ^2 с p степенями свободы при условии, что матрица точно известна, а выборки взяты из одной и той же популяции¹.

Если все признаки независимы, то $r_{ik}=0$ при $i \neq k$ и матрица R будет иметь вид:

$$\begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix},$$

¹ Обоснование этого построения заключается в следующем: если величины x_i распределены как $N(0, 1)$, то $\sum_{i=1}^n x_i^2$ будет распределена как χ^2 ; см. гл. 4 и приложения D, E и F.

а ее детерминант $|R|$ будет равен 1. Величины $R_{ih}=1$ при $i=k$ и равны 0 во всех остальных случаях, так что (12.4) можно будет записать как

$$\sum_{i=1}^p \left(\sum_{i=1}^p y_i^2 \right).$$

Если предположить, что $\sigma_{ii}=\sigma_{2i}=\sigma_i$, то выражению $s_{ii}^2/n_1 + s_{2i}^2/n_2$ будет эквивалентно $\frac{n_1 n_2}{(n_1 + n_2) s_i^2}$ и тогда (12.4) можно записать в форме (12.2).

Вычисленное эмпирическое значение С.Р.Л. не применяется в качестве меры таксономической близости или таксономического расстояния. Этот критерий предназначен для проверки гипотезы, согласно которой две выборки взяты из одной популяции. Гипотеза отвергается или принимается в зависимости от того, превысит или нет С.Р.Л. значение, соответствующее заданному уровню значимости, и, кроме этого вывода, никаких других делать нельзя. Например, если С.Р.Л. для двух выборок A и B больше, чем для выборок A и C , и в обоих случаях является существенным, то это не будет служить основанием для вывода о большем таксономическом удалении A от B , чем от C . Этот вопрос следует рассматривать с тех же позиций, с которых рассматривалась задача о таксономическом расстоянии в предыдущем разделе.

ЛИНЕЙНАЯ ДИСКРИМИНАНТНАЯ ФУНКЦИЯ

Приложения

Линейная дискриминантная функция используется при построении многомерного критерия или набора критериев, позволяющих относить данный индивидуум к одной из нескольких популяций. Этот метод применяется в различных областях, представляющих интерес для геологов¹.

Допустим, что матрица коэффициентов корреляции соответствует девяти случайным величинам, которые представляют удельный вес породы, присутствие в ней обломочного материала и различных минералов. Предварительно для каждого элемента выборочной корреляционной матрицы проверяется гипотеза о его несущественном отличии от нуля. Затем матрица сокращается путем устранения имеющих значение элементов [226] и наконец строится линейная дискриминантная функция. Далее дискриминантный показатель z может быть использован

¹ J. C. Griffiths, Petrographical investigation of the salt wasch sediments, Final Report, RME 3151, Off. Techn. Serv., 1957.

при сравнении безрудных и рудоносных образований, но сначала нужно исключить из рассмотрения те переменные, которые не влияют на различия. К наиболее важным признакам отличия рудных и безрудных остаточных образований относятся присутствие кварца, цемента, карбонатов и процент окремнения; такие же признаки, как обломочный материал, размер зерен, степень их отсортированности и их ориентировка, могут оказаться несущественными.

Этот способ позволяет проследить все этапы линейного дискриминантного анализа, включая и устраненные признаки как несущественные.

Вопрос о применении линейных дискриминантных функций рассматривали, хотя и не столь детально, в более ранней работе Гриффитс и др.¹

Вистелиус² в своей работе, посвященной составам тяжелых фракций песков, с помощью дискриминантных функций сравнивает три минеральные ассоциации.

Обсуждение

Понятие о линейной дискриминантной функции было введено Фишером [94], а впервые этот метод анализа применил Бернард³ в своем знаменитом исследовании черепов, найденных при раскопках в Египте. Фишер детально рассмотрел линейный дискриминантный анализ. Задачу, которая встает перед нами при изучении черепов, можно сформулировать следующим образом: из двух популяций, различия между которыми известны и обоснованы а priori, взяты выборки. В выборке из популяции A множество признаков измерялось на каждом из N_A индивидуумов, и те же характеристики измерялись на каждом из N_B элементов выборки B . Теперь нужно найти такой «показатель», который бы позволил относить любой новый образец к популяции A или B . Следуя интуиции, для решения этой задачи следовало бы сначала подсчитать оценки средних значений для некоторого признака как в выборке A , так и в выборке B . Затем для решения вопроса, к какой выборке ближе подходит новый индивидуум, следовало бы сравнить его характеристики с полученными двумя средними. Как отмечает Ходжес⁴ это по

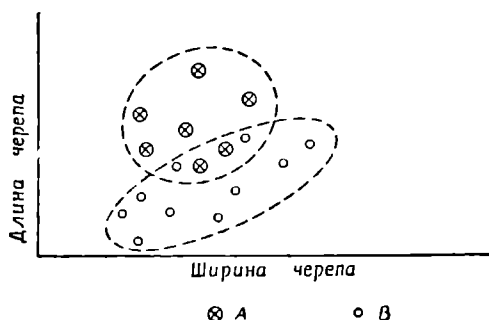
¹ J. C. Griffiths, J. A. Cochran, J. J. Hulta, R. Steinmetz, Petrographical investigation of salt wash sediments, Ann. Techn. Rep. RME 3122 (Pts. I, II), off. Techn. Serv. Washington, 1954.

² А. Б. Вистелиус, О минералогическом составе тяжелой фракции песков продуктивной толщи дельты Волги, ДАН СССР, 71, № 2, 1950.

³ M. M. Barnard, The secular variations of skull characters in four series of Egyptian skulls, Ann. of Eugenics, 6, 1935.

⁴ J. L. Hodges, Discriminatory analysis, USAF School of Aviation Medicine, Proj., № 21-49-004, Report, № 1, 1950.

существо и есть процедура, использованная Фишером. Допустим, что мы имеем дело с ситуацией, когда в выборках А и В измерялись два признака и результаты измерения нанесены на точечную диаграмму (фиг. 12.3). На этой фигуре показана основная область перекрытия результатов измерения. Эту область



Фиг. 12.3. Точечная диаграмма зависимости длины черепа от ширины.

можно рассматривать в терминах различия между группами и в терминах рассеяния внутри каждой группы. Фишер при этом обратил основное внимание на рассмотрение отношения

$$\frac{\text{Различие между выборочными средними}}{\text{Стандартная ошибка внутри двух выборок}}.$$

Хоул [122] подошел к этому вопросу с позиций геометрического решения (фиг. 12.4), которое мы рекомендуем вниманию читателя. Его основные положения сводятся к следующему. Если точечную диаграмму рассматривать как плоскость и ввести третью координату R , легко заметить, что, спроектировав точки, которые представляют собой результаты измерения длины черепа l и его ширины w , на плоскость, показанную на фиг. 12.4, можно получить более отчетливое разделение двух групп. Если бы мы нашли такую плоскость, которая обеспечила бы максимальное различие между группами и свела к минимуму рассеяние внутри групп, то мы бы получили оптимальный вариант. Пусть уравнение плоскости будет иметь вид

$$R = \lambda_1 x_1 + \lambda_2 x_2,$$

где x_1 и x_2 — значения двух измеряемых признаков. Задача заключается в отыскании таких значений коэффициентов λ_1 и λ_2 , которые бы обеспечивали требуемые условия разделения. Как отмечает Хоул, случай представления результатов измерения двух признаков в трехмерном пространстве может быть обобщен для любого числа измерений.

[illegible]
$$S_{ij} = \sum_{k=1}^{n_a} (x_{akl} - \bar{x}_{ai})(x_{akj} - \bar{x}_{aj}) + \sum_{k=1}^{n_b} (x_{bkl} - \bar{x}_{bl})(x_{bkj} - \bar{x}_{bj});$$
$$\begin{aligned} i &= 1, 2, \dots, i, \\ j &= 1, 2, \dots, p; \end{aligned}$$
$$d_l = \bar{x}_{al} - \bar{x}_{bl}.$$

Предлагаемая форма решения представляет собой систему уравнений с p неизвестными, которыми являются $\lambda_1, \dots, \lambda_p$.

Примеры

Теперь мы перейдем к применению линейной дискриминантной функции. Два близких вида *Merychoidodon culbertsoni* и *Prodesmatochoerus* будут служить основой для ее построения с использованием тех же результатов измерения признаков, которые уже использовались для сравнения многомерных средних. Для удобства приводим часть табл. 12.7 (табл. 12.14).

Весьма важно, что с помощью многомерного сравнения средних было установлено, что эти два вида сильно отличаются друг от друга. Как отмечалось выше, мы должны быть а priori

Таблица 12.14

Признак	Средние по видам	
	<i>Merychoidodon culbertsoni</i>	<i>Prodesmatochoerus meeki</i>
X	47,71	39,33
Y	90,78	83,25
Z	17,57	15,67
Q	6,61	4,42

уверены в различии популяций, иначе построение линейной дискриминантной функции будет бесполезным. На практике для предварительного обоснования разделения совокупности обычно до построения линейной дискриминантной функции используют критерий Хотеллинга T^2 .

Коэффициенты S_{ii} и S_{ij} при неизвестных $\lambda_1, \dots, \lambda_p$ представляют собой

элементы ковариационной матрицы. Для их вычисления можно воспользоваться теми же формулами сумм квадратов и промежуточных произведений W_{ij} , которые использовались при многомерном дисперсионном анализе. Исходные значения для вычисления детерминантов приведены в табл. 12.5.

В данном случае для сравнения двух четырехмерных выборок дискриминантная матрица образуется элементами системы уравнений:

$$S_{xx}\lambda_x + S_{xy}\lambda_y + S_{xz}\lambda_z + S_{xq}\lambda_q = d_x,$$

$$S_{yx}\lambda_x + S_{yy}\lambda_y + S_{yz}\lambda_z + S_{yq}\lambda_q = d_y,$$

$$S_{zx}\lambda_x + S_{zy}\lambda_y + S_{zz}\lambda_z + S_{zq}\lambda_q = d_z,$$

$$S_{qx}\lambda_x + S_{qy}\lambda_y + S_{qz}\lambda_z + S_{qq}\lambda_q = d_q;$$

$$d_x = \bar{X}_1 - \bar{X}_2 = \frac{668}{14} - \frac{472}{12} = 47,71 - 39,33 = 8,38,$$

$$d_y = \bar{Y}_1 - \bar{Y}_2 = \frac{1271}{14} - \frac{999}{12} = 90,78 - 83,25 = 7,53,$$

$$d_z = \bar{Z}_1 - \bar{Z}_2 = \frac{246}{14} - \frac{188}{12} = 17,57 - 15,67 = 1,90,$$

$$d_q = \bar{Q}_1 - \bar{Q}_2 = \frac{92,5}{14} - \frac{53}{12} = 6,61 - 4,42 = 2,19,$$

где средние значения взяты из табл. 12.14. Подставляя в систему уравнений значения из табл. 12.5, после округления получим¹:

$$\begin{aligned}210,0\lambda_X + 72,2\lambda_Y + 31,6\lambda_Z + 9,5\lambda_Q &= 8,4, \\72,2\lambda_X + 208,0\lambda_Y + 34,8\lambda_Z - 2,0\lambda_Q &= 7,5, \\31,6\lambda_X + 34,8\lambda_Y + 70,0\lambda_Z + 4,5\lambda_Q &= 1,9, \\9,5\lambda_X - 2,0\lambda_Y + 4,5\lambda_Z + 13,0\lambda_Q &= 2,2.\end{aligned}$$

Решая эту систему с помощью детерминантов, найдем корни:

$$\begin{aligned}\lambda_X &= 0,02349; \\ \lambda_Y &= 0,03214; \\ \lambda_Z &= -0,01107; \\ \lambda_Q &= 0,15991.\end{aligned}$$

Подставляя найденные значения в линейную дискриминантную функцию

$$R = \lambda_X X + \lambda_Y Y + \lambda_Z Z + \lambda_Q Q,$$

получим окончательный результат:

$$R = 0,02349X + 0,03214Y - 0,01107Z + 0,15991Q.$$

Применяя эту „оптимальную“ линейную комбинацию к первоначальным данным, приведенным в табл. 12.2 и 12.5, получим четкое разграничение между популяциями, как показано в табл. 12.15.

Таблица 12.15

Значения R для отдельных образцов

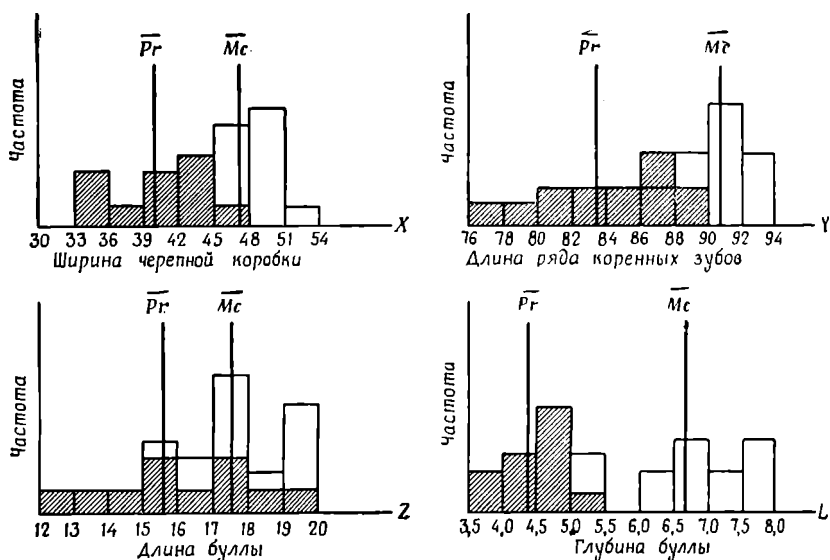
<i>M. culbertsoni</i>	5,004	4,921	4,674	4,754	4,694	5,176	4,683
<i>P. meeki</i>	4,133	4,034	4,171	4,201	4,170	2,466	4,224
<i>M. culbertsoni</i>	4,828	4,795	5,062	5,057	5,066	4,798	5,097
<i>P. meeki</i>	3,864	4,027	4,481	3,804	4,214	—	—

В данном случае перекрытие отсутствует. Наибольшее значение для *P. meeki* равно 4,481, что значительно меньше наименьшего значения для *M. culbertsoni*.

Новый образец можно теперь отнести к тому или иному виду на основе вычисления соответствующего ему значения дискриминантной функции, по близости которого в верхнем или

¹ Во всех вычислениях следует сохранять одно и то же число десятичных знаков.

нижнем ряду значений табл. 12.15 будет проводиться классификация. На фиг. 12.5, построенной эмпирически, показано распределение значений четырех признаков для обоих рассмотренных видов. На приведенных графиках видно перекрытие значений в той или иной степени всех четырех признаков. В подобном случае, применяя линейную дискриминантную функцию, можно было бы выявить существенное различие этих видов¹.



Фиг. 12.5. Гистограммы, иллюстрирующие перекрытие значений характеристик видов *Merychoidodon culbertsoni* и *Prodesmatochoerus meeki*. Обозначения $\overline{Pr}$ и $\overline{Mc}$ показывают положение средних арифметических.

В качестве второго примера применения дискриминантного анализа можно привести результаты сравнения двух групп миссисипских бластонидей (морских бутонов), которые названы *пириформными* и *годониформными*. Они были собраны в Голконде (южный Иллинойс). Измерялись два признака — длина и ширина чашечки.

	Объем выборки	Средние	
Пириформные	$N_p = 31$	$\overline{L}_p = 17,5882$	$\overline{W}_p = 11,3412$
Годониформные	$N_g = 32$	$\overline{L}_g = 9,3278$	$\overline{W}_g = 9,2500$

¹ Полученный результат можно проверить при помощи статистического сравнения средних для двух видов или однофакторного дисперсионного анализа (М. А. Кендалл, The advanced theory of statistics, II, New York, 1948).

Общие для обеих групп оценки дисперсий и ковариаций следующие:

$$S_{LL} = 1074,444,$$

$$S_{WL} = 443,528,$$

$$S_{WW} = 252,904.$$

Значения d :

$$d_L = \bar{L}_p - \bar{L}_g = 8,2604,$$

$$d_W = \bar{W}_p - \bar{W}_g = 2,0912.$$

Теперь нужно найти значения коэффициентов λ_1 и λ_2 из системы уравнений

$$\lambda_1 S_{LL} + \lambda_2 S_{WL} = C d_L$$

$$\lambda_1 S_{WL} + \lambda_2 S_{WW} = C d_W.$$

Величина C введена с целью более наглядной иллюстрации предположения Хоула. Вычисление значения R значительно упрощается, если C выбрать так, чтобы одна из величин λ_1 или λ_2 была равна 1.

Решая систему линейных уравнений методом детерминантов, обычно полагают $D_1/D = \lambda_1$, где для приведенного выше случая величина D_1 получена путем подстановки d_L и d_W в первую строку D . Если умножить D_1/D на D/D_1 , то величина λ_1 будет равна 1. В данном примере:

$$1074,444\lambda_1 + 443,528\lambda_2 = 8,2604C,$$

$$443,528\lambda_1 + 252,904\lambda_2 = 2,0912C.$$

Пусть

$$C = \frac{\begin{vmatrix} 1074,444 & 443,528 \\ 443,528 & 252,904 \end{vmatrix}}{\begin{vmatrix} 8,2604 & 443,528 \\ 2,0912 & 252,904 \end{vmatrix}} = \frac{D}{D_1} = 64,58.$$

Тогда $\lambda_1 = 1$ и

$$\lambda_2 = \frac{\begin{vmatrix} 1074,444 & 8,2604 \\ 443,528 & 2,0912 \end{vmatrix}}{\begin{vmatrix} 1074,444 & 443,528 \\ 443,528 & 252,904 \end{vmatrix}} = -1,220.$$

Таким образом, для $R = \lambda_1 L + \lambda_2 W$ мы получим $R = L - 1,22W$.

Подставляя, как и в предыдущем примере, в эту линейную комбинацию первичные данные, получим хорошее разделение двух групп.

Значения R для некоторых образцов следующие:

Голони- формные	-0,86	-1,54	-1,68	-1,35	-1,56	-1,20	-0,24	-3,22	-2,56	+0,04	-3,18 ч т. д.
Пириформ- ные . . .	+3,50	+8,15	+1,18	+0,61	+6,52	+3,76	+3,18	+4,25	+3,38	+5,25	+0,88 и т. д.

Применяя дискриминантную функцию, не следует забывать, что популяции A и B , рассмотренные в начале этого раздела, должны быть заведомо различными. Поэтому сначала к выборкам из популяций A и B следует применить критерий T^2 для установления реальности различий между ними. Только после этого может быть построена дискриминантная функция, которая в дальнейшем используется для классификации вновь получаемых данных.

ПРИМЕНЕНИЕ МНОГОМЕРНОГО АНАЛИЗА ПРИ РЕШЕНИИ
ГЕОМОРФОЛОГИЧЕСКИХ ЗАДАЧ

Ниже рассмотрены две работы, в которых излагаются данные исследования бассейнов рек, характеризующихся набором признаков.

Первая из них¹ посвящена геометрическим свойствам сформировавшихся дренажных систем. Описание строения рельефа производится с помощью двух параметров: D (модуль стока) и F (частота речной сети).

Путем регрессионного анализа найдено, что зависимость между величинами D и F может быть выражена с помощью уравнения

$$F = 0,694D^2,$$

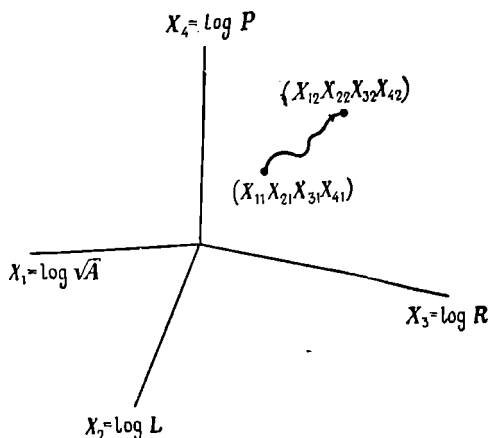
которое при нанесении на логарифмическую бумагу представляется прямой линией. Разброс 156 точек, соответствующих отдельным бассейнам, вокруг этой линии незначителен. Кроме того, с помощью корреляционного анализа установлено, что величина F/D^2 «не зависит» от периметра бассейна.

Милтон использовал для изображения морфометрических характеристик сформировавшихся речных систем в единой системе координат геометрию четырехмерного пространства. Как видно из заглавия его статьи, эти построения он проводил в евклидовом четырехмерном пространстве (E_4).

Милтон сосредоточил основное внимание на четырех характеристиках бассейнов: A — площадь, L — общая длина русел, R — рельеф бассейна и P — периметр бассейна. Пространство E_4 определяется четырьмя взаимно перпендикулярными координатами: $X_1 = \lg \sqrt{A}$, $X_2 = \lg L$, $X_3 = \lg R$ и $X_4 = \lg D$ (фиг. 13.1). Все другие характеристики бассейнов рассматриваются как функции A , L , R и P . Бассейн отдельной реки теперь удобно представить как точку в пространстве E_4 . Эту точку можно также рассматривать как вектор с компонентами (X_{11} , X_{21} , X_{31} ,

¹ M. A. M e l t o n, Geometric properties of nature drainage sistems and their representation in E_4 phase space, Journ. Geol., 66, № 1, 1958.

X_{41}), где, согласно Милтону, X_{ij} — i -я координата j -й точки. Общая длина русел, величина площади бассейна, его морфологические показатели и периметр могут меняться во времени, так что точка в пространстве E_4 описывает некоторую кривую, переходя в последующий момент времени в положение 2 с координатами $(X_{12}, X_{22}, X_{32}, X_{42})$, как показано на фиг. 13.1. Милтон предположил, что пространство E_4 можно использовать для представления во всех остальных количественных характеристиках



Фиг. 13.1. Схематическое изображение фазового пространства E_4 .

бассейна. Другие характеристики речных бассейнов являются функциями одного (или более) из четырех упомянутых параметров. Например, модуль стока $D=f(A, L)=X_2-2X_1$ и набор всех значений D может быть представлен как поле в фазовом пространстве.

Применение пространства E_4 весьма показательно при рассмотрении моделей роста речных бассейнов. Путь, который описывает точка, соответствующая бассейну, в пространстве E_4 (фиг. 13.1), отражает сущность модели роста. Однако Милтон отмечает, что изменения, наблюдаемые в характеристиках бассейна в вышеупомянутом пространстве, это пространственные изменения в данный момент времени, тогда как модели роста позволяют рассматривать эти изменения во времени в данном месте пространства. В той же работе Милтон сравнивает рассматриваемую задачу с эргодической задачей кинетической теории газов, ставя вопрос о возможности замены закона изменения в пространстве законами изменения во времени¹. Множе-

¹ Применимость аналогов эргодической теории в геологии довольно типична. См., например, R. L. Miller, Symposium on Statistic in Geology, Abst. G. S. A. meet. Los Angeles, 1954.

ственный регрессионный анализ Милтон также использовал, например, при нахождении оценки для величины θ (угол наклона борта длины), выраженной через значения четырех координат пространства E_4 . Если $\lg \theta$ обозначить через Y , то гиперплоскостью, удовлетворяющей требованию наименьших квадратов, будет

$$\hat{Y} = 1,3831 - 0,9612X_1 + 0,2038X_2 + 0,5672X_3 + 0,2180X_4.$$

Анализ коэффициента регрессии $b_{Y 4, 123}$ (тангенс угла наклона линии регрессии Y на X_4 при фиксированных значениях X_1 , X_2 и X_3) показал, что полученное значение этого коэффициента несущественно отличается от нуля. Поэтому величину X_4 из рассмотрения исключили и построили новое уравнение: $\hat{Y} = 1,5253 - 0,7401 X_1 + 0,2079 X_2 + 0,5637 X_3$, или $\lg \theta = 1,5253 - 0,7401 \lg \sqrt{A} + 0,2079 \lg L + 0,5637 \lg R$. Взяв антилогарифм этого выражения, получим

$$\theta = 33,52 \frac{L^{0,2079} R^{0,5637}}{(\sqrt{A})^{0,7401}}.$$

В дополнение к работе Милтона следует сказать об особенности вычислительных операций при отыскании этого выражения.

Приближенное число Стралера $HD = LR/A$ является хорошей оценкой для θ (H — характеристика рельефа, D — модуль стока)¹. Отсюда ясно, что произведение характеристики рельефа на модуль стока приблизительно равно θ .

Рассмотрим вторую работу Милтона², в которой при изучении морфометрических свойств речных систем применен многомерный анализ. В этой работе использованы корреляционные зависимости между значениями 15 признаков (табл. 13.1), расположенных в порядке их значения при изучении речных систем. Сущность исследования заключается в вычислении оценок коэффициентов корреляции для всех возможных пар признаков.

Таким образом, первый этап работы заключается в построении выборочной корреляционной матрицы, элементами которой являются оценки коэффициентов корреляции для всех возможных пар признаков, с последующим выделением «групп» и «основных пар» с помощью методов, предложенных Миллером и Веллером³ и Олсоном и Миллером⁴.

¹ A. N. Strahler, Dimensional analysis applied to fluvially eroded land-forms, Geol. Soc. Am. Bull., **69**, 1958.

² M. A. Melton, Correlation structure of morfometric properties of drainage systems and their controlling agents, Journ. Geol., **66**, № 4, 1958.

³ R. L. Miller, J. Weller, Significant comparisons in Paleontology, Journ. Paleont., **26**, № 6, 1952.

⁴ См. сноску 2 на стр. 287.

Матрица значений τ , сильно отличающихся от 0 (Милтон, 1958)

[illegible]

Особенности задачи сводятся к следующему:

1. Некоторое природное явление описывается большим числом признаков, представляющих интерес в исследовании и имеющих большое значение для анализа.

2. Возможно, что значения признаков не могут изменяться вне зависимости один от другого.

3. Зависимость между любой парой рассматриваемых признаков нередко бывает нелинейной. Это проявляется на точечных диаграммах, когда точки располагаются не по прямой линии.

4. Распределения изучаемых величин в генеральной совокупности большей частью неизвестны.

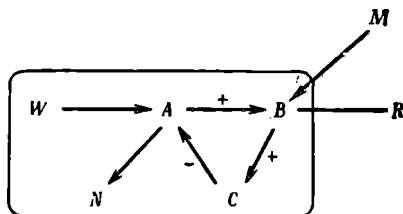
5. Имеются сведения, полученные *a priori* и в различной степени обоснованные, о причинных зависимостях между изучаемыми величинами. Учитывая эти положения, Милтон ввел понятие *системы переменных*.

Система переменных V представляет собой абстрактную последовательность случайных величин, таких, что 1) каждая из этих величин весьма сильно коррелируется со всеми другими величинами; 2) причинная зависимость между каждой парой переменных установлена; 3) одна или более переменных в V может быть коррелируема с величинами, не содержащимися в V .

Мы проиллюстрируем вышеизложенное на детальном разборе примеров, взятых из работы Милтона. Однако сначала необходимо рассмотреть структурные элементы этих систем.

Система с обратной связью

Выше было отмечено, что в условиях рассматриваемой задачи существуют сведения, полученные *a priori* о причинных



Фиг. 13.2. Система с обратной связью (Милтон, 1958).

зависимостях между случайными величинами. Пример подобных сведений схематически показан на фиг. 13.2, на которой причинные зависимости или их отсутствие показаны стрелками, линиями и знаками плюс и минус. Если изменения переменной A являются «причиной» изменений величины B , это показано

стрелкой из A в B . Если увеличение значений A влечет за собой увеличение значений B , это выражено знаком плюс. С другой стороны, если увеличение значений A приводит к уменьшению значений B , это обозначено знаком минус. Если величины B и R коррелируются, но между ними нет причинной зависимости (внутри системы!), они просто соединены прямой линией. Исключительно ярким примером такой зависимости служит сильная корреляционная зависимость между коэффициентом рождаемости людей и численностью популяции аистов в Швеции. Подобные явления могут быть включены в рассматриваемую систему или, наоборот, исключены из нее в зависимости от того, представляют ли они интерес сами по себе; однако сведения *a priori* в подобных случаях не позволяют утверждать наличие причинной зависимости.

Совокупность случайных величин и их зависимость могут быть различными:

1. Случайная величина M , находящаяся вне системы, изменяется независимо от элементов, образующих систему. Однако изменения M являются «причиной» изменений B . Поэтому M и B соединены стрелкой.

2. Точно так же величина W не зависит от изменений, происходящих внутри системы, хотя она и является элементом системы. Однако изменения величины W влекут за собой изменения величины A .

3. Величина N зависит от A , что показано стрелкой, направленной из A в N .

4. Система с обратной связью показана в виде триады ABC . Изменения A , являющиеся причиной изменений B , связаны с последними положительной корреляционной зависимостью. В свою очередь изменения B связаны положительной зависимостью с изменениями C , которые характеризуются отрицательной корреляционной зависимостью с изменениями A . Последовательность причинных зависимостей, показанная стрелками, образует замкнутую «петлю», начиная от A к B , от B к C и от C снова к A . Такая зависимость называется *обратной связью*; в данном случае перед нами отрицательная обратная связь.

5. Так как для M и W $\rho \approx 0$, только одна величина W включена в систему.

Корреляционная матрица

Пусть, как отмечалось выше, несколько случайных величин связаны линейной корреляционной зависимостью. Кроме того, пусть их распределения, как обычно, неизвестны [265]. В связи с этим Милтон предлагает использовать непараметрическую меру степени корреляционной зависимости величину τ Кендалла.

Ниже приведены этапы вычислений. Здесь мы рассмотрим корреляционные зависимости между L/P (отношение общей длины русел к периметру бассейна) и F/D^2 (относительная густота речной сети).

Первый этап

Выборочные значения L/P располагаются в порядке возрастания. Каждому из 59 рассматриваемых значений L/P , приписанных отдельным бассейнам, соответствует значение F/D^2 . В связи с тем что значения L/P упорядочены, остается изучить полученную последовательность соответствующих значений F/D^2 . Ясно, что в том случае, если F/D^2 также окажутся все упорядоченными, зависимость будет полной.

Второй этап

Рассмотрим полученную в результате упорядочения значений L/P последовательность значений F/D^2 :

43, 13, 51, 38, 58, 49, 57, 31, 40, 46, 28, 39, 45, 48, 47, 56,
12, 32, 59, 10, 16, 55, 30, 52, 2, 19, 50, 34, 33, 54, 37, 26, 41,
6, 53, 9, 1, 7, 14, 35, 4, 29, 15, 44, 8, 18, 23, 42, 22, 21, 24,
25, 5, 17, 36, 20, 11, 27, 3.

Составим таблицу величин $N_H - N_I$, соответствующих каждому из приведенных значений F/D^2 , где N_H — число рангов (элементов), больших, чем рассматриваемое значение, и расположенных справа от него, а N_I — число элементов, меньших, чем рассматриваемый, также расположенных от него справа. Первому элементу последовательности, равному 43, соответствуют справа от него 16 элементов, имеющих большие значения, и 42 — меньшие. Таким образом, величина $N_H - N_I$, соответствующая значению 43, будет равна $16 - 42 = -26$. Для предпоследнего элемента последовательности, равного 27, $N_H - N_I = 0 - 1 = -1$.

Третий этап

Теперь мы можем вычислить значение рангового коэффициента корреляции Кендалла:

$$\tau = \frac{S}{\frac{1}{2} N(N-1)},$$

где S — сумма всех значений $N_H - N_I$, которая в данном случае равна -551 . Величина N представляет собой число элементов

рассматриваемой последовательности, или число рангов. В этом примере $N=59$. Следовательно,

$$\tau = \frac{-551}{\frac{1}{2} \cdot 59 \cdot 58} = -\frac{551}{1711} = -0,322.$$

Четвертый этап

Если в последовательности значений F/D^2 нет тенденции к упорядочиванию, то при $N>10$ статистика τ будет распределена приблизительно нормально со средним $\mu_\tau=0$ и стандартным отклонением

$$\sigma_\tau = \sqrt{\frac{4(2N+5)}{9N(N-1)}}.$$

Таким образом, для проверки гипотезы об отсутствии упорядоченности можно воспользоваться величиной

$$Z = \frac{\tau}{\sigma_\tau},$$

которая, если гипотеза верна, будет распределена нормально со средним 0 и дисперсией, равной 1.

В данном примере

$$\sigma_\tau = \frac{2 \cdot (2 \cdot 59 + 5)}{9 \cdot 59 \cdot 58} = 0,0894,$$

а значение $|Z|=0,322/0,0894=3,60$. В связи с тем что вычисленное значение Z значительно превышает допустимое при уровне значимости 0,01, последовательность значений F/D^2 следует рассматривать как имеющую тенденцию к уменьшению при возрастании значений L/P , а значение $\tau=-0,322$ — существенно отличающимся от нуля.

Проверку гипотезы об отсутствии упорядоченности при $N \leq 10$ и общее рассмотрение критерия τ можно найти в работе Сигала¹.

Табл. 13.1, взятая из работы Милтона, представляет собой матрицу тех значений τ , которые существенно отличаются от нуля. На этой матрице основано построение «коррелированных множеств» и «групповых диаграмм». «Основные пары» этой матрицы образованы такими членами, для которых корреляционная зависимость является более сильной, чем та же зависимость с любой другой рассматриваемой величиной (о понятии «основные пары» см. стр. 308—309).

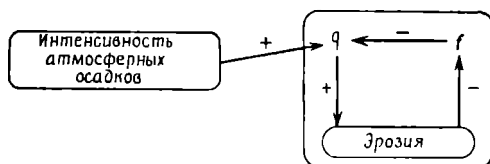
После того как из корреляционной матрицы были устранены элементы, несущественно отличающиеся от нуля, Милтон возвратился к «системе переменных» с целью интерпретации.

¹ S. Siegal, Nonparametric statistics for the behavioral sciences, New York, 1956.

Таким образом, систему переменных можно рассматривать как гипотезу, сформулированную не только относительно элементов корреляционной матрицы, но и характеристик, не входящих в эту матрицу. В качестве примера рассмотрим систему, названную *системой с положительной обратной связью инфильтрационной емкости и проникающей способности*.

Эта система построена следующим образом.

Случайные величины q (водопроницаемость) и f (инфильтрационная емкость) связаны сильной отрицательной корреляционной зависимостью (фиг. 13.3). Причинная связь, основанная на предпосылках *a priori*, показана стрелкой, направленной от f к q . Система переменных образована путем введения гипотетического элемента; в данном случае — степени эрозии.



Фиг. 13.3. Система переменных (Милтон, 1958).

Допустим, что мы начали рассмотрение с f и следуем по направлению стрелок. Снижение инфильтрационной емкости приводит к возрастанию водопроницаемости [отрицательная корреляционная зависимость (табл. 13.1)], что в свою очередь ведет к усилению эрозии. Таким образом, связь q и эрозия следует охарактеризовать знаком плюс. Усиление эрозии приводит к снижению инфильтрационной емкости в связи с устранением тех фракций, которые характеризуются наибольшей проницаемостью. Таким образом, система представляет собой замкнутое кольцо обратной связи. В качестве дополнительного элемента введена интенсивность атмосферных осадков, которая не полностью включена в систему как величина, не зависящая от этой последней. Однако этот элемент положительно воздействует на величину q , т. е. увеличение интенсивности атмосферных осадков приводит к возрастанию значений q . Обратная связь является положительной, так как изменения, начавшиеся в произвольно выбранной точке f , приводят снова к этой же точке с тем же знаком, который был вначале.

Этот метод хорош тем, что при его применении особое внимание уделяется противоречиям в знаках и направлениях стрелок, что приводит к пересмотру сведений, полученных *a priori*, или корреляционной структуры.

Интересный метод проверки обоснованности соединения стрелкой двух величин заключается в применении частных

коэффициентов корреляции. Например, Милтон рассматривает случай, когда неясно, нужно ли соединять стрелкой величину H (показатель расчленения рельефа) с θ (угол склона долины): частный коэффициент корреляции между H и θ (при постоянном рельефе) оказывается таким же, как коэффициент корреляции между D и θ , так как H по определению равно D , а величина r общая для всего бассейна. Из табл. 13.1 видно, что выборочный коэффициент корреляции между D и θ несущественно отличается от нуля, из чего следует, что соединять стрелкой величины H и θ не нужно.

Авторы полагают, что к дополнительному вопросу о проверке гипотезы относительно внутренней непротиворечивости системы переменных можно подойти более формально, используя теорию частной корреляции для случая нормального распределения. Возможно, что с этой целью придется рассмотреть частный коэффициент корреляции сначала для пары случайных величин, предположив, что другие величины фиксированы, затем для всех других возможных пар. Подход, аналогичный общей задаче, рассмотренной Милтоном, отчасти используется в генетике¹.

ФАКТОРНЫЙ АНАЛИЗ

В наиболее ограниченном понимании можно рассматривать факторный анализ как изучение степени зависимости между парами случайных величин. Сила зависимости измеряется с помощью коэффициента корреляции r для всех возможных пар². Однако если требуется, можно использовать непараметрические оценки для ρ или другие характеристики, такие, как ранговый коэффициент корреляции Спирмена [13]. В недавних исследованиях подобного типа применялась величина χ^2 .

Однако при изучении матрицы (табл. 13.2) может потребоваться переход ко второму основному этапу факторного анализа, заключающемуся в выделении из матрицы групп случайных величин с помощью тех или иных средств. В табл. 13.3 сведены два метода, наиболее часто применяемых для оценки силы зависимости. Это так называемые R - и Q -методы.

R -метод основан на результатах измерения двух величин, где каждая пара эмпирических значений величин, например x и y , принадлежат одному индивидууму; этот метод требует измерений большого числа индивидуумов.

С помощью Q -метода сравниваются два индивидуума на основе данных измерения многих признаков, что также требует большого числа характеристик.

¹ S. Wright, The method of path coefficients, Ann. Math. Stats., 5, 1934.

² См. сноску 3 на стр. 289; L. L. Thurston, Edge-marking method of analyzing data, Journ. Am. Stat. Assoc., 43, 1948.

Таблица 13.2

Матрица выборочных коэффициентов корреляции

	1	2	3	4...	N
1		r_{12}	r_{13}	$r_{14} \dots$	r_{1N}
2			r_{23}	$r_{24} \dots$	r_{2N}
3				$r_{34} \dots$	r_{3N}
4					r_{4N}
$\vdots$					$\vdots$
$\vdots$					$\vdots$
$\vdots$					$r_{(N-1)N}$
N					

Таблица 13.3

Связь между R - и Q -методами

	Случайная величина X	Случайная величина Y	...	Случайная величина Z
Индивидуум 1	x_1	y_1	...	z_1
Индивидуум 2	x_2	y_2	...	z_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
Индивидуум N	x_N	y_N	...	z_N
	Корреляция с помощью R -метода			

Кроме того, существуют так называемые P - и Q -методы, предназначенные для применения факторного анализа к отдельным индивидуумам [49]¹.

Мы сосредоточим внимание на R -методе, так как именно с его помощью решаются геологические задачи, а затем рассмотрим Q -метод. Собственно уже вычисления оценок коэффи-

¹ См. также R. B. Cattell, Growing points in factor analysis, Austr. Journ. Psych., 6, 1954.

циентов корреляции для всех возможных пар случайных величин можно считать началом R -метода факторного анализа. Так, для случайных величин будет вычислено $\frac{1}{2} N(N-1)$ выборочных коэффициентов корреляции. На основе полученной матрицы коэффициентов корреляции (табл. 13.2) выделяют факторы. Рассмотрим различные способы этого выделения.

Многофакторный анализ

При многофакторном анализе выделение факторов осуществляется путем представления случайных величин как векторов и с помощью некоторых действий над матрицами. Следует отметить, что эти факторы образуют независимые координаты пространства, в котором корреляционные зависимости представлены как векторы.

Литература, посвященная этим методам, приведена в конце книги.

Анализ групп

Методы выделения групп с внутренней корреляцией из общей корреляционной матрицы рассмотрены в работах Кателла [48] и Трайона [276]. Эти методы объединены под общим названием *анализ групп*.

Мы будем называть группой такую совокупность случайных величин, в которой корреляционная зависимость между всеми возможными парами принадлежащих ей переменных равна или превышает некоторый произвольно заданный уровень.

МЕТОДЫ ВЫДЕЛЕНИЯ ГРУПП

Метод ветвящихся связей [48]

Скажем, что «связь» двух случайных величин существует, если значение коэффициента корреляции превысит некоторый произвольно заданный уровень. Таким образом, две случайные величины являются связанными, если соответствующее им значение коэффициента корреляции больше или равно величине, заданной исследователем.

В некоторых случаях процедура выделения групп может включать требование, согласно которому истинное значение коэффициента корреляции ρ должно быть больше (или равно) некоторого произвольно заданного значения. При этом требовании обычно обращаются к таблицам, составленным Дэви-

дом¹; в этих таблицах дана зависимость объема выборки и вероятности того, что значение коэффициента корреляции больше заданного (табл. 13.4).

Шаг 1. Упорядочим случайные величины. Например, примем порядок верхней строки матрицы.

Шаг 2. Из установленного таким образом порядка последовательно выпишем на карточку каждую случайную величину со всеми величинами, с которыми она связана.

Шаг 3. Начнем с первой карточки и выберем из записанных случайных величин одну, следующую за первой справа. Рассмотрим следующую карточку. Выберем на ней первую случайную величину и следующую за ней справа, которая есть и на первой карточке. Каждая новая случайная величина должна быть и на предшествующих карточках в этой последовательности. Надо продолжать эту процедуру до тех пор, пока все карточки не будут исчерпаны. В результате выбранные таким образом карточки образуют группу.

Шаг 4. Вернемся к первой карточке и начнем построение новой последовательности; это построение следует производить до тех пор, пока все возможности не будут исчерпаны.

Шаг 5. Пропустим первую карточку и начнем со второй (т. е. второй случайной величины). Произведя описанную выше процедуру, зафиксируем как группу все вновь выбранные карточки. Эта процедура повторяется для всех случайных величин.

В результате будут полностью выделены все максимальные группы. Для образования групп, состоящих из двух элементов, не обязательно прибегать к этому методу, так как их можно взять непосредственно из матрицы.

Эта процедура показана в табл. 13.5 для случая матрицы порядка 6×6 . Связанные пары, т. е. те пары случайных величин, для которых коэффициент корреляции больше или равен 0,9, обозначены буквой X. Пустые квадраты соответствуют случаям, когда связи нет, т. е. $|r| \leq 0,9$:

(1)	(2)	(3)	(4)	(5)	(6)
(1), 2, 3, 6	1, (2), 3, 4, 5	1, 2, (3), 5, 6	2, (4)	2, 3, (5), 6	1, 3, 5, (6)

Карточка 1 связана с карточками 2, 3, 6. Переходя к карточке 2, видим, что случайная величина 3 первая справа от случайной величины 2, которая содержится и на карточке 1. Далее, на карточке 3 первая случайная величина справа от 2 есть 5. Однако 5 нет на карточке 1. Закончив на этом процедуру, фиксируем группу, состоящую из случайных величин 1,

¹ F. N. David, Tables of the ordinates and probability integral of the distribution of the correlation coefficient in small samples, Biometrika, Office, London, 1948.

Матрица оценок коэффициентов корреляции

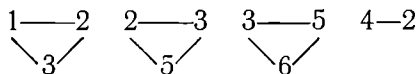
[illegible]

Корреляционная матрица шестого порядка

	1	2	3	4	5	6
1		X	X			X
2			X	X	X	
3					X	X
4						
5						X
6						

2, 3. Исключив карточку 1, начинаем отбор с карточки 2 в новой последовательности. В результате получаем группу 2, 3, 5. Аналогично, начав с карточки 3, фиксируем группу 3, 5, 6. Для карточки 4 группа будет состоять из одной пары 4, 2, которую можно получить непосредственно из матрицы. Точно так же для карточки 5 группа будет состоять из двух элементов — 5 и 6.

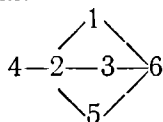
Для выражения результатов весьма удачна следующая форма:



Вернувшись к первой карточке (шаг 4), начнем построение новой последовательности. На первой карточке первая непроверенная справа случайная величина есть 3. Переходя к карточке 3, получим триаду 1, 3, 6.

Выше было принято, что линия, соединяющая пару случайных величин, означает наличие между ними связи. В результате будут получены максимальные отделенные группы. Под «отделенными» мы понимаем такие группы, которые не содержатся полностью ни в какой другой, например пара 1, 6. Наоборот, пара 5, 6 является группой, содержащейся в большей группе 3, 5, 6. Заметим, что некоторые члены бывают общими для нескольких групп. Вопрос об этом перекрытии групп и способы его устранения рассмотрены при описании p - F -группового анализа. Иногда графическое изображение группы можно представить так, что перекрытие будет отчетливо видно. В рассмо-

тренном примере этого можно достигнуть путем следующего комплексного изображения:



Q-МЕТОД

Как отмечалось в начале раздела о факторном анализе, Q-метод включает рассмотрение корреляционной зависимости между парой индивидуумов по множеству измерений. Если степень зависимости (выраженная, например, с помощью коэффициента корреляции r) высока, то можно сделать вывод, что пропорции измеряемых признаков сходны. Таким образом, если значение каждого признака, измеренного у образца A , будет, например, почти точно вдвое больше значения соответствующего признака у образца B , то корреляционная зависимость будет высокой. Аналогично если результат каждого измерения образца A почти такой же, как и результат измерения образца B , то корреляционная зависимость также будет высокой. Таким образом, в Q-методе учтена только корреляционная зависимость, и можно сказать, что этот метод заставляет концентрировать внимание на различии или сходстве пропорций независимо от различий в размерах.

Рассмотрим применение этого метода в палеонтологии.

Пусть имеется коллекция особей, которую мы разделим на несколько групп, используя различия в коэффициентах корреляции. Эти группы выбираются так, чтобы между особями одной и той же группы была высокая корреляционная зависимость, а между каждым из ее индивидуумов и любым не принадлежащим этой группе элементом была бы слабая зависимость. Q-метод был применен при изучении трех ископаемых видов рода *Hyopsodus* из нижнего и среднего эоцена. Каждый из видов — *H. lysitensis*, *H. powellianus* и *H. paulus* — был охарактеризован четырьмя особями. В анализе были использованы результаты 27 различных измерений серии верхних зубов. Значения коэффициентов корреляции для всех возможных пар отдельных особей приведены в табл. 13.4. Согласно результатам исследования, независимо от размеров зубов различия, существующие между их пропорциями, не столь существенны, что бы можно было совокупности имеющихся 12 особей разбить на более дробные категории.

Однако, прежде чем делать какие-либо заключения, нужны были дополнительные данные. Для сравнения выбрали единственный индивидуум — обезьяну *Aotus trivirgatus*, на котором были произведены те же измерения. Коэффициенты корреляции ме-

жду этой особью и каждым из 12 индивидуумов рода *Hyopsodus* даны в табл. 13.4. Так как зубы обезьяны *Aotus trivirigatus* полностью отличаются по внешнему виду от зубов ископаемых представителей рода *Hyopsodus*, следует ожидать низкую корреляционную зависимость. Данные табл. 13.4 свидетельствуют о справедливости этого предположения.

Далее были измерены особи, отобранные из серии ископаемых ореодонтов олигоценового возраста. При этом произвели измерения 44 признаков черепа каждого индивидуума нескольких отличающихся один от другого видов. Коэффициент корреляции для двух особей одного и того же вида (*Prodesmatochoerus meeki*) оказался равным 0,994. Корреляционная зависимость между представителем *Desmatochoerus* и индивидуумом, принадлежащим к совершенно другому виду — *Megoreodon gigas*, также оказалась значительной — 0,983. Наконец, коэффициент корреляции между особями двух близких видов — *Prodesmatochoerus meeki* и *Merychiodon culbertsoni* — равен 0,993. Таким образом, как и в случае рода *Hyopsodi*, здесь нельзя выделить какие-либо категории, так как значения коэффициентов корреляции очень высокие. Следует отметить, что при сравнении многомерных средних оказалось, что различия между видами *Merychiodon culbertsoni* и *Prodesmatochoerus* (гл. 12) существенны.

Попытаемся сделать из сказанного вывод. Там, где линейные зависимости резко выражены, Q-метод бывает чувствителен только в тех случаях, когда виды сильно отличаются один от другого; например, *Hyopsodus* и *Aotos trivirigatus*. Там, где подобный подход имел бы практическое значение, т. е. при изучении близко связанных серий, результаты оказываются неприемлемыми; например, в сериях *Huopsodus* и *Oreodont*.

Однако Сокэл и Мичинер (1957 г.), применяя этот метод при изучении таксономии пчел, получили превосходные предварительные результаты. Но следует отметить, что эти авторы рассмотрели множество характеристик, основанных на подсчете и произвольном обозначении целыми числами качественных признаков, таких, например, как цвет. Но, так как эти данные они использовали в совокупности с непрерывными линейными измерениями, изменчивость в рассеянии точек увеличилась; следовательно, увеличились и коэффициенты корреляции.

Здесь уместно заметить, что органы или части близко связанных видов больше отличаются одни от других по своим размерам, чем по пропорциям. Таким образом, между парой особей следовало бы ожидать более высокую корреляционную зависимость при измерениях только зубов (как в случае *Hyopsodi*) или только черепов (*Oreodont*), а не всего организма.

Авторы считают, что при использовании линейных измерений (при изучении ископаемых остатков) чувствительность

Q-метода, основанного на коэффициентах корреляции, недостаточна для обеспечения удовлетворительных результатов по близким видам. Тем не менее дальнейшие исследования в этом направлении представляются весьма ценными. Возможно, что будет найден какой-либо способ приписывать «веса» измеряемым характеристикам или другой путь изучения степени близости.

Вообще проблема разделения коллекции индивидуумов на категории на основе сходства или различия между особями в случае, когда эти категории не определены заранее, является одной из главных проблем палеонтологии.

Успех Q-метода при изучении современных пчел свидетельствует о необходимости продолжать разрабатывать этот метод в применении его к палеонтологии¹.

МЕТОД КОРРЕЛЯЦИОННЫХ ПРОФИЛЕЙ; ПОКАЗАТЕЛЬ ЕДИНСТВА И ρ -F-АНАЛИЗ

Метод корреляционных профилей, предложенный Трайном [276], основан на весьма своеобразном подходе к задаче выделения групп в корреляционной матрице. Преимущество этого метода заключается в устранении перекрытия групп или по крайней мере сокращения этого перекрытия. Есть два способа применения метода корреляционных профилей. Кроме них, имеется третий, развитый одним из авторов, который будет описан ниже.

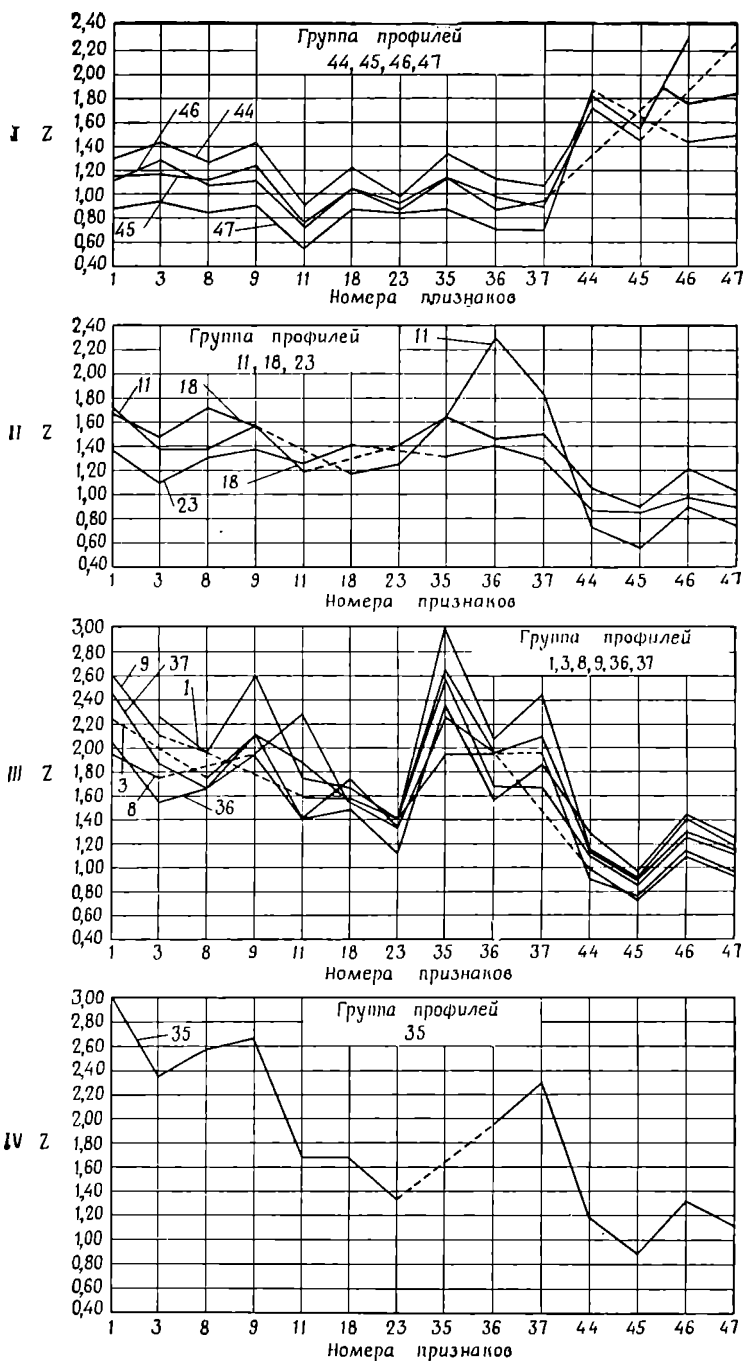
Первый способ называется *методом пробных графиков*, второй — *количественным методом*; он основан на применении коэффициента Холзингера² при построении предварительных групп.

После образования предварительных групп описываются окончательные варианты этих групп с помощью метода пробных графиков. Хотя второй метод не столь субъективен и более практичен в случае больших матриц, он иногда бывает очень запутанным. Заинтересованный читатель может найти соответствующие пояснения в работе Трайона. Мы закончим наш обзор описанием метода пробных графиков.

Процедура этого метода крайне проста. Значения коэффициентов корреляции r для данной случайной величины откладываются по ординате, а по оси абсцисс — номера случайных величин для удобства в том же порядке, что и в матрице.

¹ R. R. Sokal, H. D. Michener, A statistical method for evaluating systematic relationships, Contr. 945, Dept. of Entom., Univ. of Kansas, 1956.

² K. J. Holzinger, Student manual of factor analysis, Dept. of Educ., Univ. of Chicago, 1937.



Фиг. 13.4. Корреляционные профили, разделенные в результате исследования на четыре группы (I — IV).

Совокупность точек, соединенных на графике, образует корреляционный профиль (фиг. 13.4). В процессе исследования на график наносят корреляционные профили для других случайных величин, которые согласуются с первым профилем и всеми остальными профилями в группе. Таким способом группы случайных величин будут представлены группами согласующихся профилей. Анализируя матрицу коэффициентов корреляции (табл. 13.3), можно выбрать те группы профилей, которые представляются согласующимися один с другим. Трайон рекомендует строить сначала профиль для той случайной величины, для которой Σr принимает наибольшее значение. Следовательно, нужно сложить в столбце, соответствующем данной случайной величине, все коэффициенты корреляции (за исключением элементов главной диагонали). После построения корреляционного профиля первой случайной величины выбирают ту, которая обладает наибольшей корреляционной зависимостью с первой, а затем рассматривают согласованность полученных профилей.

Характеристика согласованности профилей, необходимая для включения случайной величины в группу или, наоборот, исключения, является субъективной. Ясно, что произвольное упорядочивание случайных величин будет влиять на согласованность. Ниже мы рассмотрим интерпретацию результатов этого анализа на реальном примере.

Факторный анализ *Equis hemonius hemonius* из Монголии

По каждому из 17 представителей диких монгольских ослов *Equis hemonius hemonius* были сделаны измерения 14 признаков. Эти данные были заимствованы из работы Мотохаши [187] (результаты вычисления коэффициентов корреляции были любезно предоставлены профессором Олсоном из Чикагского университета).

Все измерения были произведены на элементах черепа и нижней челюсти, а рассматриваемые здесь признаки выбраны из очень обширного списка самых разнообразных характеристик. Ниже приводится перечень рассматриваемых признаков под соответствующими номерами, имеющими значение кода.

Измерения длины элементов черепа

1. Базальная длина, измеренная как расстояние от центрального резца до горлового отверстия.

3. Длина носовой кости, измеренная от начала до конца внутриносового шва.

8. Горизонтальный диаметр орбиты.

9. Длина нёба, измеренная как расстояние от центрального резца до заднего конца среднего нёбного шва на горизонтальной части нёбной кости.

11. Длина диастемы, измеренная как расстояние от переднего конца альвеолы P^2 до заднего конца альвеолы I^3 .

Измерения ширины элементов черепа

18. Ширина носовых костей, измеренная в их передней части в точке сочленения с верхней челюстью.

23. Максимальная ширина затылочных костей.

Углы

35. Лицевой угол.

Измерения элементов нижней челюсти

36. Длина нижней челюсти, измеренная как расстояние от центрального резца до середины заднего края сочленения с черепом.

37. Длина диастемы, измеренная как расстояние от переднего конца альвеолы P_2 до заднего конца альвеолы I_3 .

44. Высота короноидного отростка, измеренная в средней точке его верхнего конца.

Измерения серии верхних зубов

45. Диаметр резца I^1 , представляющий наибольший диаметр поверхности зуба.

46. I^2 , как и предыдущий.

47. Высота резца I^1 .

Корреляционная матрица для этих признаков приведена в табл. 13.6. Теперь с помощью метода ветвящихся связей и метода корреляционных профилей выделим группы признаков.

АНАЛИЗ ГРУПП С ПОМОЩЬЮ МЕТОДА ВЕТВЯЩИХСЯ СВЯЗЕЙ; ПРИМЕНЕНИЕ ρ -F-МЕТОДА [198] *

Рассмотрим пару случайных величин как «связанную», если соответствующее им истинное значение коэффициента корреляции r больше или равно 0,95. В таблицах Дэвида этому условию соответствуют при объеме выборки 17 и уровне надежно-

* См. также E. C. Olson, R. L. Miller, Morphological integration, Univ. of Chicago Press, Chicago, 1958.

Корреляционная матрица 14 характеристик ($\rho \geq 0,90$, $r \geq 0,3887$) при $n = 17$

[illegible]

сти 0,95 значения выборочного коэффициента корреляции r , превышающие 0,86. При анализе матрицы, приведенной в табл. 13.4, составляются карточки, которые затем сортируются, как было описано выше. В результате будут получены следующие «отделенные r -группы»:

- I. 1—3—8—9—11—35—36—37;
- II. 1—3—8—9—18—23—35—36;
- III. 3—46;
- IV. 9—46;
- V. 44—45—46—47.

Элементы каждой группы коррелируются со всеми остальными членами этой группы так, что для любой пары признаков $|\rho| \geq 0,95$. Ясно, что группы характеризуются значительным перекрытием. Так, признаки с номерами 1, 3, 9, 35 и 36 являются общими для групп I и II. Дальнейший анализ и интерпретация в такой ситуации становятся затруднительными. Олсон [201] предложил способ применения r - F -метода, позволивший свести перекрытие групп к минимуму. Относительно исключительно высокой корреляционной зависимости, характеризующей пару морфологических признаков, предполагают, что она является следствием очень сильной биологической связи, или морфологического единства. Следует ожидать, что признаки, которые группируются вокруг такой пары, также будут сильно зависимы. В этом смысле пара становится ядром группы. Таким образом, «основная пара» представляет собой пару признаков, между которыми корреляционная зависимость выше, чем для любого из них и какой-либо другой случайной величины. Из табл. 13.6 видно, что признаки 1 и 35, образующие основную пару, окружены зависящими от них другими признаками. Последовательность сокращения области пересечения групп сводится к следующему:

1. Список основных пар. В данном примере: (1)—(35), (11)—(36), (44)—(45) и (46)—(47).

2. Исключаются все группы, не содержащие основных пар. В данном примере исключаются из рассмотрения группы III и IV.

3. а) Наличие общего члена в разных группах означает, что группа, содержащая наивысшую основную пару, связана с этим общим членом.

б) Из группы исключается любой член, который имеет более сильную связь с элементом основной пары какой-либо другой группы.

На основе сказанного из группы II исключается признак (36), так как он сильнее связан с признаком (11) в группе I, чем с (1) или с (35) в группе II. Кроме того, из группы II исключаются признаки (18) и (23), так как они характеризуются

более сильной связью с признаками (11)—(36) группы I, чем с элементами пары (1)—(35) группы II. В группе II остались следующие элементы: (1), 3, 8, 9, (35). В связи с тем что все эти признаки содержатся в группе II, ее следует исключить из рассмотрения. После анализа основных пар остаются две группы:

I. (1), 3, 8, 9, (11), (35), (37);

V. (44), (45), (46), (47).

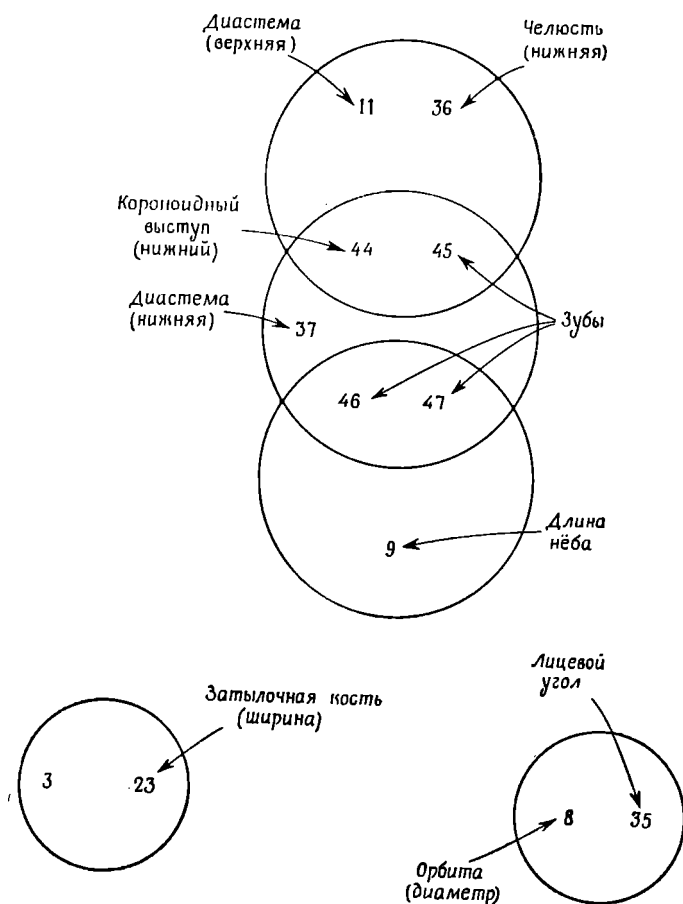
Практика показала, что фактор размера обычно играет значительную роль в остеометрических исследованиях. Рассмотрим следующий шаг ρ - F -анализа, предшествующий окончательным результатам. Меры связи между парами признаков выражены в виде частных коэффициентов корреляции, вычисленных при устранении влияния признака 1, характеризующего длину черепа, и признака 15 — ширину черепа. (Коэффициенты корреляции признака 15 со всеми остальными и частные коэффициенты корреляции r_{xy} : 1,15 были любезно предоставлены Олсоном.) В табл. 13.7 приведена матрица частных коэффициентов корреляции. В этой матрице коэффициенты корреляции основных пар подчеркнуты, а пустые ячейки соответствуют случаям, когда коэффициенты корреляции несущественно отличаются от нуля при уровне значимости 0,10. Группы, полученные в результате анализа основных пар, после устранения влияния размеров приведены на фиг. 13.5. Матрица, приведенная в табл. 13.7, значительно отличается от матрицы табл. 13.6, так как устранено влияние длины и ширины черепа. При вычислении частных коэффициентов корреляции длина черепа была выбрана в качестве «фиксированной константы». Хотя этот выбор субъективен, показано, что фактор размеров играет важную роль при изучении корреляционных зависимостей признаков черепа не только у современных животных, но и при исследовании ископаемых беспозвоночных и позвоночных животных. Теперь можно сформулировать гипотезы, которые лежат в основе ρ - F -анализа, и затем применить их к рассмотренному примеру.

Факторы, на которых основано выделение внутренне коррелируемых групп остеометрических признаков, следующие: 1 — размер, 2 — функциональные факторы, 3 — наследственные факторы, которые необязательно являются функциональными и которые мы называем *факторами развития*. После устранения влияния размера интерпретация индивидуальной корреляции и, следовательно, группы будет основана на следующем:

$$\rho = F \{ \text{функция} + \beta (\text{развитие}) + \gamma (\text{остаток}) \}.$$

Оценив коэффициенты α , β , γ , мы ищем интерпретацию группы в терминах этой функции, т. е. хотим определить влияние этих факторов, на основе сведений *a priori*. Таким образом,

смысл термина « ρ - F » становится понятным из приведенного выше уравнения, в котором ρ — истинное значение коэффициента корреляции. Сведения *a priori* получают при изучении скелета и анатомии мягких тканей современных животных (наиболее близких к вымершим видам).



Фиг. 13.5. Группы, полученные в результате анализа основных пар, после устранения влияния размеров.

Вернувшись к рассматриваемому примеру, можно в порядке опыта изолировать четыре фактора:

1. Фактор размера (рассмотрен выше).
2. Жевательный функциональный фактор. Это показано на фиг. 13.5, на которой признаки длина нёба, длина диастемы на

Т а б л и ц а 13.7

Матрица частных коэффициентов корреляции r_{xy} : 1,15, где признак 1 — длина черепа,
а 15 — максимальная ширина черепа

3	8	9	11	18	23	35	36	37	44	45	46	47
3			0,64		<u>—0,72</u>		0,46	—0,39	0,49		0,44	
8				0,45		0,52						
9											0,40	0,45
11							<u>0,82</u>		—0,73	—0,64	—0,57	
18												
23												
35												
36												
37												
44												
45												
46												
47												

Число характеристик равно 13

$n = 17$

верхней и нижней челюстях и длина нижней челюсти окружают ядро, образованное тремя характеристиками зубов и высотой короноидного выступа.

3. *Зрительный функциональный фактор.* Эта изолированная основная пара образована измерением орбиты и лицевым углом. Так как число признаков в группе мало, невозможно четко определить назначение фактора. Поэтому для этих двух факторов предлагается общая функция «видения».

4. *Основная пара без назначения.* Эта пара образована длиной носовой кости и шириной затылочной части.

Изложенным выше способом было проанализировано несколько ископаемых и ныне живущих видов беспозвоночных. Очень важно, что при этом был детализирован подход к задаче анализа эволюционных серий, когда действие факторов, изолированных с помощью ρ - F -анализа, прослеживалось во времени как внутри популяций, так и между ними¹. Для решения этой задачи Олсон и Миллер [198] предложили так называемый показатель общего морфологического единства I_ρ . При изучении степени зависимости морфологических признаков установлено, что иногда величины пропорций бывают сильно коррелированы, как, например, в случае пентремитовой цветковой чашечки. В другом примере (череп *Sciurus niger*) Олсон и Миллер [198] выделили области высокой внутренней корреляции, тогда как вне этих областей корреляция отсутствует. Кроме того, некоторые отношения величин признаков отличаются высокой корреляцией с другими характеристиками независимо от топологического родства. Для сравнения общих морфологических характеристик видов была предложена величина, выражающая морфологическое единство. Она называется *показателем единства* I_ρ :

$$I_\rho = \frac{4(B_{0;\rho})^2}{K_\rho(n^2 - n)^2} \quad \text{при } B_{0;\rho} \geq 1;$$
$$I_\rho = 0 \quad \text{при } B_{0;\rho} = 0,$$

где $B_{0;\rho}$ — число наблюдаемых связей, существенных при заданном значении ρ ; таким образом, $B_{0;0,95}$ — число значимых связей при $\rho \geq 0,95$ из всех возможных связей между парами признаков; K_ρ — число полных ρ -групп², отмеченных при заданном значении ρ ; термином *связь* в данном случае обозначают корреляционную зависимость между парой признаков: если для пары $\rho \geq 0,95$, эта пара связана; если же $\rho < 0,95$, связи нет.

¹ E. C. Olson, Integrating factors in amphibian skulls, Journ. Geol., 61, 1953.

² Полной ρ -группой называется группа таких признаков, все возможные пары которых зависимы; кроме того, полная ρ -группа не содержится в большей ρ -группе.

Подводя итог всему сказанному относительно ρ - F -метода, следует отметить, что его применение окажется эффективным при решении двух вопросов, представляющих интерес для палеонтологов:

1. Детализация исследования и конкретизация тесных связей между морфологией и развитием, что позволит предположить действие генотипических или фенотипических факторов; кроме того, при грубом анатомическом исследовании, с которым обычно приходится иметь дело палеонтологу, установление этих связей оказывается весьма эффективным.

2. Изучение общего облика животного и его изменений в процессе эволюции.

ГРУППОВОЙ АНАЛИЗ *EQUIS HEMONIUS HEMONIUS* С ПОМОЩЬЮ МЕТОДА КОРРЕЛЯЦИОННЫХ ПРОФИЛЕЙ

Применение этого метода следует начинать с построения профилей для различных признаков; при этом на график наносят коэффициенты корреляции одного признака со всеми остальными. Полученный профиль сравнивают с профилями (соответствующими всем другим признакам), полученными аналогичным путем.

Для построения этих групп профилей авторы предложили *метод существенных разностей*. Первым шагом этого метода [95] является замена элементов r матрицы, приведенной в табл. 13.6, величинами Z (табл. 13.8), где Z получают с помощью следующего преобразования: $Z = \frac{1}{2} \ln [(1+r)/(1-r)]$.

Статистика [226]
$$\frac{Z_1 - Z_2}{\sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}}$$
 распределена прибли-

тельно нормально со средним, равным 0, и дисперсией, равной 1, если два сравниваемых коэффициента корреляции ρ_1 и ρ_2 равны. Величины Z_1 и Z_2 — трансформированные оценки r_1 и r_2 , а N_1 и N_2 — число наблюдений в первой и второй выборках соответственно.

Так как рассматриваемая в нашем примере корреляционная матрица построена на основе одного и того же числа наблюдений, можно построить простой критерий, позволяющий быстро исследовать матрицу, образованную величинами Z , с точки зрения увеличения или уменьшения разностей между ними. Итак,

$$P_r \left\{ |Z_1 - Z_2| \sqrt{\frac{N-3}{2}} < c \right\} = 0,95;$$

Величины Z_{xy} (преобразованные значения табл. 13.6)
 $P_r \{ |Z_1 - Z_2| \leq 0.628 \} = 0.95$

	1	3	8	9	11	18	23	35	36	37	44	45	46	47
1		2,25	1,95	2,61	1,74	1,66	1,38	3,00	2,04	2,44	1,10	0,87	1,29	1,13
3	2,25		1,74	2,10	1,38	1,48	1,10	2,36	1,53	1,84	1,26	0,95	1,43	1,16
8	1,95	1,74		1,95	1,38	1,73	1,30	2,60	1,66	1,66	1,08	0,83	1,26	1,10
9	2,61	2,10	1,95		1,59	1,59	1,38	2,66	1,95	2,10	1,13	0,89	1,42	1,22
11	1,74	1,38	1,38	1,59		1,19	1,26	1,66	2,30	1,84	0,71	0,55	0,91	0,77
18	1,66	1,48	1,73	1,59	1,19		1,42	1,66	1,48	1,53	1,05	0,89	1,22	1,05
23	1,38	1,10	1,30	1,38	1,26	1,42		1,33	1,42	1,30	0,87	0,85	0,98	0,91
35	3,00	2,36	2,60	2,66	1,66	1,66	1,33		1,95	2,30	1,13	0,87	1,33	1,13
36	2,04	1,53	1,66	1,95	2,30	1,48	1,42	1,95		1,95	0,89	0,71	1,13	0,97
37	2,44	1,84	1,66	2,10	1,84	1,53	1,30	2,30	1,95		0,95	0,70	1,07	0,93
44	1,10	1,26	1,08	1,13	0,71	1,05	0,87	1,13	0,89	0,95		1,87	1,74	1,82
45	0,87	0,95	0,83	0,89	0,55	0,89	0,85	0,87	0,71	0,70	1,87		1,42	1,53
46	1,29	1,43	1,26	1,42	0,91	1,22	0,98	1,33	1,13	1,07	1,74	1,42		2,30
47	1,13	1,16	1,10	1,22	0,77	1,05	0,91	1,13	0,97	0,93	1,82	1,53	2,30	

величину c при вероятности 0,95 можно положить равной 1,95. Тогда

$$P_r \left\{ |Z_1 - Z_2| < \frac{1,95}{\sqrt{\frac{N-3}{2}}} \right\} = 0,95.$$

В данном примере $N=17$ и, следовательно,

$$P_r \{ |Z_1 - Z_2| < 0,628 \} = 0,95.$$

При отыскании согласующихся профилей метод существенных разностей устраняет элементы субъективизма. Последний может возникнуть в ситуациях, когда различные профили, построенные по коэффициентам корреляции, рассматриваются как несогласующиеся на основе несущественных расхождений между ними. Это серьезное осложнение можно избежать, используя критерий существенных разностей, зависящий от объема выборки.

Элементы Z матрицы, приведенной в табл. 13.8, изучены путем попарного сравнения строк по всем возможным вариантам сравнения в столбцах. Например, строки 1 и 2 (табл. 13.8) изучались путем сравнения вертикальных пар. Существенные значения разностей, превышающие по абсолютной величине 0,60, отмечаются для этой пары строк в столбцах 35 и 37. Поэтому можно сделать вывод, что профили, соответствующие признакам 1 и 3, существенно различаются за счет расхождения между коэффициентами корреляции с признаками 35 и 37. Полученные в результате таких сравнений предварительные группы согласующихся профилей приведены в табл. 13.9.

Таблица 13.9

Предварительные группы, полученные методом существенных разностей	Члены групп
I	1 — 3 — 8 — 9 — 36 — 37
II	1 — 3 — 35 — 37
III	3 — 8 — 11 — 18 — 23
IV	3 — 8 — 11 — 18 — 36 — 37
V	23 — 46
VI	44 — 45 — 46 — 47

Так как совокупность полученных групп характеризуется значительным числом перекрытий, то следующим шагом анализа будет нанесение профилей на график так, чтобы члены, общие для нескольких групп, были оставлены только в одной группе. Это делается на основании согласованности профилей. На фиг. 13.4 показаны окончательные группы профилей, в кото-

рых перекрытия устранены путем графических сравнений. Эти группы следующие:

1—3—8—9—36—37,

11—18—23,

35,

44—45—46—47.

В соответствии с изложенной выше теорией группового анализа типа ρ - F преобразуем частные коэффициенты корреляции $r_{xy; 1,15}$, вычисленные при условии устранения влияния длины и ширины черепа, в величины Z .

Это преобразование делается так же, как и раньше, с обычными коэффициентами корреляции. Так как в данном случае рассматриваются частные коэффициенты корреляции, то объем выборки N следует уменьшить на 2, что соответствует числу признаков, влияние которых устранено. Таким образом,

$$P\left\{|Z_1 - Z_2| < 1,95/\sqrt{\frac{N-2-3}{2}}\right\} = 0,95.$$

При $N=17$ правая часть неравенства в скобках равна 0,796. Путем изучения матрицы, образованной величинами Z (табл. 13.11), были найдены следующие группы, которые были установлены в последовательности согласно фиг. 13.2.

Т а б л и ц а 13.10

Предварительные группы, полученные методом существенных разностей	Члены групп
I	8 — 9 — 18 — 23 — 35
II	11 — 23 — 36 — 37
III	44 — 45 — 46 — 47

Интерпретация групп корреляционных профилей требует некоторых пояснений. В настоящем исследовании рассматриваются различные измерения черепа и челюстей, которые в той или иной степени связаны между собой. Большинство значений этих измерений должно изменяться в узких пределах, чтобы животное могло выжить. Изучая группу, образованную признаками 44, 45, 46, 47, можно видеть, что все они взаимодействуют с другими признаками как единое целое, хотя для этого не обязательно должно быть выполнено требование высокой степени корреляции. Таким образом, группу профилей можно рассматривать как доказательство наличия взаимодействия со всеми

Т а б л и ц а 13.11

 $r_{xy; 1,15}$ (преобразованная табл. 13.4) при $P_r \{ |Z_1 - Z_2| < 0,796 \} = 0,95$

	8	9	11	18	23	35	36	37	44	45	46	47
3			—0,76		—0,91		0,50	—0,41	0,54		0,48	
8				0,49		0,59					0,43	0,49
9											0,43	
11	—0,76						1,16		—0,93	—0,76	—0,65	
18	0,49											
23	—0,91											
35	0,59											
36	0,50		1,16						—0,62	—0,44		
37	—0,41								0,53	—0,74	—0,71	—0,73
44	0,54		—0,93				—0,62	0,53		1,53	1,10	1,19
45			—0,76				—0,44	—0,74	1,53		1,05	1,07
46	0,48	0,43	0,43	—0,65				—0,71	1,10	1,05		1,74
47	0,49							—0,73	1,19	1,07	1,74	

другими признаками, тогда как группа типа ρ - F свидетельствует о взаимодействии входящих в нее членов. Мы можем теперь, комбинируя два типа группового анализа, отыскать группу с большой степенью корреляции между ее членами, которая также будет представлять собой группу корреляционных профилей, взаимодействующую как одно целое с другими признаками. Примером такой двойственной группы может служить совокупность признаков жевательного аппарата — 44, 45, 46, 47. Эти признаки образуют не только сильно внутренне связанную группу, но и группу корреляционных профилей.

Авторы считают, что продолжать исследования в этом направлении необходимо; новый материал окажется весьма ценным при остеометрическом анализе, требующем выделения действующих факторов, а также в других областях, когда приходится использовать факторный анализ.

Общий интерес по данному разделу представляет статья Куртена¹. В ней автор делает обзор корреляционного анализа, излагает результаты вычисления коэффициентов корреляции при изучении молочных зубов и сравнивает эти результаты с моделями, построенными на основе генетических предпосылок. Кроме того, в указанной статье Куртен рассматривает так называемые корреляционные поля, являющиеся обобщением корреляционной матрицы морфологических характеристик, выраженным графически. Форма корреляционных полей явно зависит от порядка расположения признаков в матрице, что в свою очередь обусловлено зоологическими критериями (например анатомическое сходство). Возможно, что эти обобщения, по крайней мере в принципе, связаны с морфологическими полями Батлера² и полями показателей роста Хаксли [128].

МЕТОД «ДРОБОВОГО ВЫСТРЕЛА» (ТОЧЕЧНЫХ ДИАГРАММ)

Анализ взаимодействия количественных характеристик, когда отсутствуют сильные предположения *a priori*³

Использование современных аналогий при изучении древних фаций или геологических процессов вполне логично. Однако при этом нередко возникают опасения, вызываемые появлением ошибок при проведении аналогий. Если допустить значительное однообразие в частной задаче, то можно сосредоточить

¹ B. Kurtén, On the variation and population dynamics of fossil and recent mammal populations, Acta Zool. Fennica, 76, 1953.

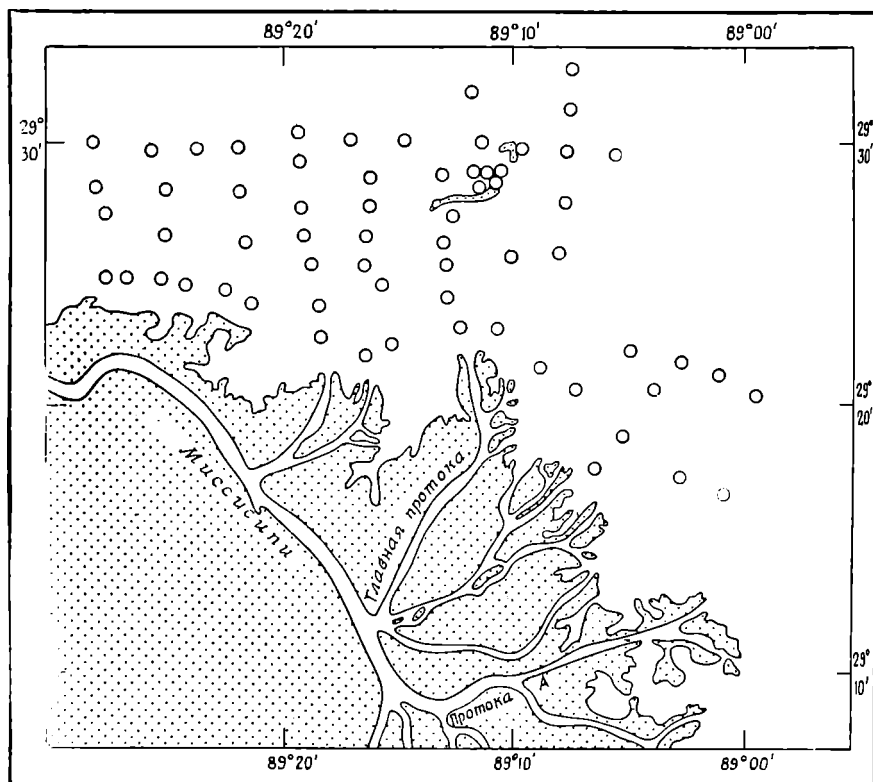
² P. M. Butler, Studies of the mammalian dentition, I, The teeth of *Centetes caudatus* and its allies, Proc. Zool. Soc. London, Ser. B., 1937.

³ R. L. Miller, An Analysis of the interaction of quantitative variables in a modern environment of sedimentation, Bull. G. S. A., 65, № 12, part 2, 1954.

внимание по крайней мере на наиболее сомнительной части сложной ситуации, когда рассматривается большое число неконтролируемых признаков. В частности, ничем не ограниченное использование одной случайной величины иногда вызывает подозрение на возможное взаимодействие с большим числом других случайных величин, между которыми установлены связи.

Постановка задачи

Задача в общем виде заключается в изучении взаимодействия большого числа случайных величин. Она будет рассмо-



Ф и г. 13.6. Расположение точек опробования.

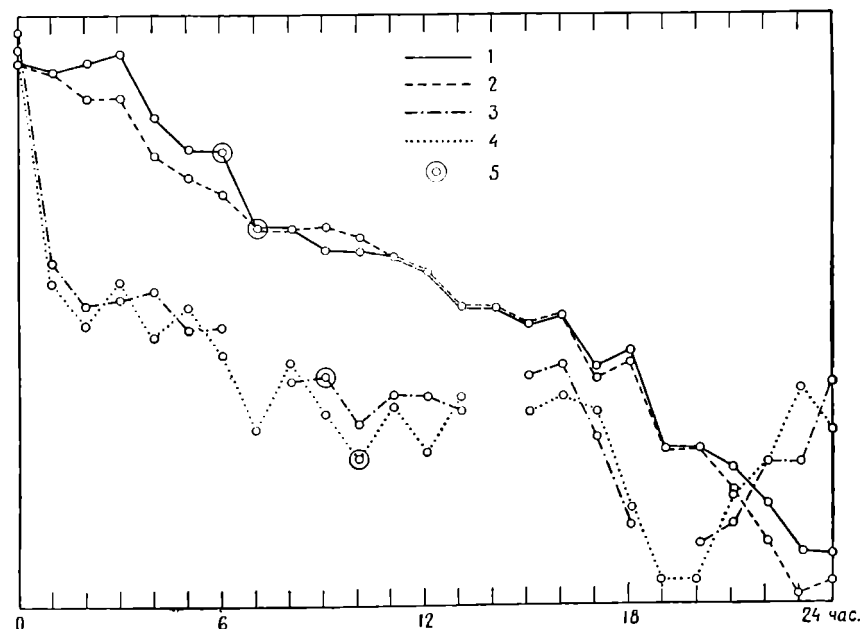
трева на примере исследования 19 биологических и физических характеристик современной морской фации.

В частности, была сделана попытка разделить характеристики по их зависимым и независимым связям. Применение термина *дробовой выстрел* обусловлено отсутствием формального

определения данного статистического метода; наоборот, процедуру построения гипотез и последовательно проводимых серий статистических проверок различных альтернатив удалось намеренно избежать.

Мы рассмотрим все возможные корреляционные связи, не касаясь сведений, полученных *a priori*, чтобы избежать предвзятости в исследовании, а выводы сделаем только после завершения анализа.

Трудности, связанные с этой процедурой, будут выявлены в процессе рассмотрения задачи.



Фиг. 13.7. Флуктуации различных признаков, замеренные на одной станции за 24 час.

1 — температура воды на поверхности; 2 — то же на дне; 3 — содержание хлора на поверхности; 4 — то же на дне; 5 — моменты наблюдения.

1. *Флуктуации, зависящие от времени, и тренд.* Результаты измерения характеристик воды в каждой точке, показанной на фиг. 13.6, были получены в разное время. Так, величина *A* могла быть измерена в данной точке в другое время, чем величина *B*. Таким образом, в изменчивость данной случайной величины вводится особое слагаемое, связанное с суточными и сезонными циклами, которое не принимается во внимание. Эта ситуация выражена на фиг. 13.7, где показаны в виде четырех серий наблюдений изменения температуры и хлористость воды

на одной из станций в течение 24 час. Предположим, что на этой станции замеры температуры воды на поверхности были начаты в 6 час., а замеры хлористости на дне — в 10 час. Значения характеристик в моменты получения результатов измерений показаны кружками. Кроме этой изменчивости, на результаты измерения влияли сезонные флуктуации, так как измерения на разных станциях производились в различные дни.

2. *Форма кривых распределения.* Скрытое предположение о нормальности распределений в этом случае является особенно вредным. Соответствующие преобразования приводят к симметричным распределениям, по крайней мере для некоторых случайных величин; исключения составляют случаи многовершинных распределений. Как правило, распределения следует рассматривать как неизвестные.

3. *Отдельные критерии для проверки предположения о независимости или зависимости* имеют не причинный, а вероятностный смысл, т. е. когда нет надежных сведений *a priori*, корреляционная зависимость не служит доказательством причинной связи. Действительно, критерии, которые будут использованы в дальнейшем, характеризуют только силу корреляционной связи. Согласно Крамеру [16], «две некоррелируемые случайные величины не обязательно независимы».

Пункты 1 и 3 будут снова рассмотрены в заключении, а пункт 2 — в разделе, посвященном статистическому анализу.

Фактический материал

Данные представляют значения 18 числовых случайных величин, собранные весной и осенью 1953 г. на нескольких станциях близ берега восточнее дельты Луизианы. Число наблюдений меняется от 54 для некоторых характеристик до 28 для остальных. Данные были представлены Американским нефтяным институтом. Этот раздел следует рассматривать как общий обзор вопросов, затронутых в книге, иллюстрирующий статистическую методологию в геологических науках.

На фиг. 13.6 показано положение станций, в которых производились измерения. Ниже приведен список характеристик, используемых в исследовании:

1. Физические факторы

1. Температура воды на поверхности.
2. Температура воды на дне.
3. Глубина.
4. Температура воздуха.
5. Прозрачность.
6. Скорость поверхностного течения.
7. Скорость донного течения.

II. Химические факторы

1. Хлористость воды на поверхности.
2. Хлористость воды на дне.
3. Величина pH в верхней части осадка.
4. Величина pH в нижней части осадка.

III. Биологические факторы

1. Число видов фораминифер.
2. Число особей фораминифер.
3. Число видов крупных беспозвоночных.
4. Число особей крупный беспозвоночных.

IV. Параметры размеров осадочных частиц

1. Md_Φ — средний размер осадочных частиц в Φ -единицах.
2. σ_Φ — стандартное отклонение размеров частиц.
3. α_Φ — степень асимметрии распределения размеров частиц.

Статистический анализ

По имеющимся эмпирическим данным, для каждой из 18 случайных величин были оценены все возможные коэффициенты корреляции. Изучение полученной в результате корреляционной матрицы имело два этапа: исследовались сначала группы с сильной корреляционной зависимостью, а потом совокупности характеристик, для которых выборочные коэффициенты корреляции несущественно отличались от нуля.

Относительно большинства характеристик предположение о нормальности распределения вызвало сомнения. Поэтому были использованы различные непараметрические способы измерения степени связи. Для выборок, объем которых меняется в интервале 25—60, ранговые методы весьма трудоемки (см. обзор ранговой корреляции, гл. 14). Для подобной ситуации Бломквист¹ построил тетрагорический критерий, основанный на изучении точечных диаграмм, который очень прост в применении. Если двумерную точечную диаграмму разделить на четыре квадранта линиями, представляющими перпендикуляры к осям x и y в точках $x = Md_x$ (медиана x) и $y = Md_y$ (медиана y), то проверяемая нулевая гипотеза о независимости и гипотеза о равном распределении точек по всем четырем квадрантам будут равносильны. Существуют различные вероятностные методы изучения степени уклонения распределения точек от теоретического (см., например, метод, описанный Диксоном и Мэсси [79],

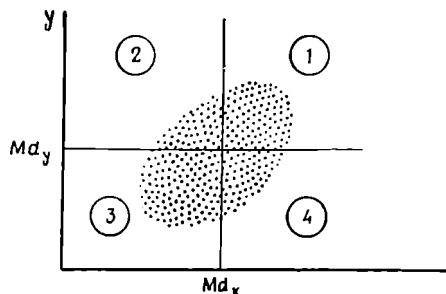
¹ N. Blomqvist, On a measure of dependence between two random variables, Ann. Math. Stats., 21, № 4, 1950.

а также метод Мостеллера¹ или метод Олмстида и Тьюки²).

Методы Бломквиста более применимы в данной ситуации, так как весьма удобны для случая малых выборок. Процедура их применения заключается в следующем:

I. *Проверка гипотезы о независимости (в вероятностном смысле)*

Шаг 1. Подсчитывается общее число n_1 точек, попавших в квадранты 1 и 3, согласно обозначениям фиг. 13.8. Обозначим через $2K$ наибольшее четное число, меньшее или равное объему выборки.



Фиг. 13.8. Точечная диаграмма, разделенная на квадранты.

Шаг 2. Для выборок, имеющих $2K$ от 4 до 50, Бломквист составил таблицы вероятности $P\{|n_1 - K| \geq v\}$ в зависимости от $2K$ и v . Если объем выборки превышает 50, можно воспользоваться нормальным приближением. Рассмотрим пример проверки гипотезы о независимости. Пусть выборка представлена 41 точкой; в первый и третий квадранты попало 18 точек, а во второй и четвертый — 23, т. е. $n_1 = 18$, $2K = 40$. По таблице находим $P\{|n_1 - K| \geq 2\} = 0,752$. Так как значение вероятности 0,752 велико, x и y нужно рассматривать как независимые.

II. *Вычисление оценки коэффициента корреляции*

Обозначим абсолютное значение неизвестного коэффициента корреляции через $|q|$, и пусть $|q_0|$ — заданное значение. Оценку для $|q|$ обозначим через $|q'|$:

$$|q'| = \frac{n_1 - n_2}{n_1 + n_2},$$

где n_1 — число точек, попавших на диаграмме в первый и третий квадранты, а n_2 — число точек во втором и четвертом квадрантах. Проверим гипотезу $|q| = |q_0|$ при альтернативе

¹ F. Mosteller, On some useful inefficient statistics, Ann. Math. Stats., 17, 1946.

² P. S. Olmstead, J. W. Tukey, A corner test for association, Ann. Math. Stats., 18, 1947.

$|q| < |q_0|$. Нулевая гипотеза должна быть отвергнута, когда $|q'| < c$.

Доверительный интервал для q' легко построить, используя следующие рассуждения Краскла. Положим, что

$$P_{|q_0|} \{ -c \leq q' \leq c \} = \alpha,$$

где α — заданное значение вероятности. Тогда

$$\begin{aligned} P_{|q_0|} \{ -c \leq q' \leq c \} &= \\ &= P_{|q_0|} \left\{ \frac{-c - |q_0|}{\sqrt{\frac{1 - q_0^2}{n}}} \leq \frac{q' - |q_0|}{\sqrt{\frac{1 - q_0^2}{n}}} \leq \frac{c - |q_0|}{\sqrt{\frac{1 - q_0^2}{n}}} \right\} \approx \\ &\approx \Phi \left(\frac{c - |q_0|}{\sqrt{\frac{1 - q_0^2}{n}}} \right) - \Phi \left(\frac{-c - |q_0|}{\sqrt{\frac{1 - q_0^2}{n}}} \right), \end{aligned}$$

где $\Phi(\cdot)$ — значение нормальной функции распределения (гл. 4). Приравняв первое слагаемое правой части приближенного равенства к α , получим

$$\Phi \left(\frac{c - |q_0|}{\sqrt{\frac{1 - |q_0|^2}{n}}} \right) = \alpha$$

и

$$\frac{c - |q_0|}{\sqrt{\frac{1 - |q_0|^2}{n}}} = \Phi^{-1}(\alpha),$$

где Φ^{-1} — функция, обратная Φ . Наконец,

$$c = \left(-\sqrt{\frac{1 - q_0^2}{n}} \right) \cdot \Phi^{-1}(\alpha) + |q_0|.$$

Например, заданное значение $|q|$ равно 0,8, а $\alpha = 0,05$. Требуется найти величину c при объеме выборки 25:

$$c = \left(-\sqrt{\frac{1 - 0,64}{25}} \right) \cdot 1,645 + 0,8 = 0,6026.$$

Таким образом, гипотеза $|q| = 0,8$ для выборок объемом 25 будет отвергаться, если $|q'| < 0,6026$.

Результаты

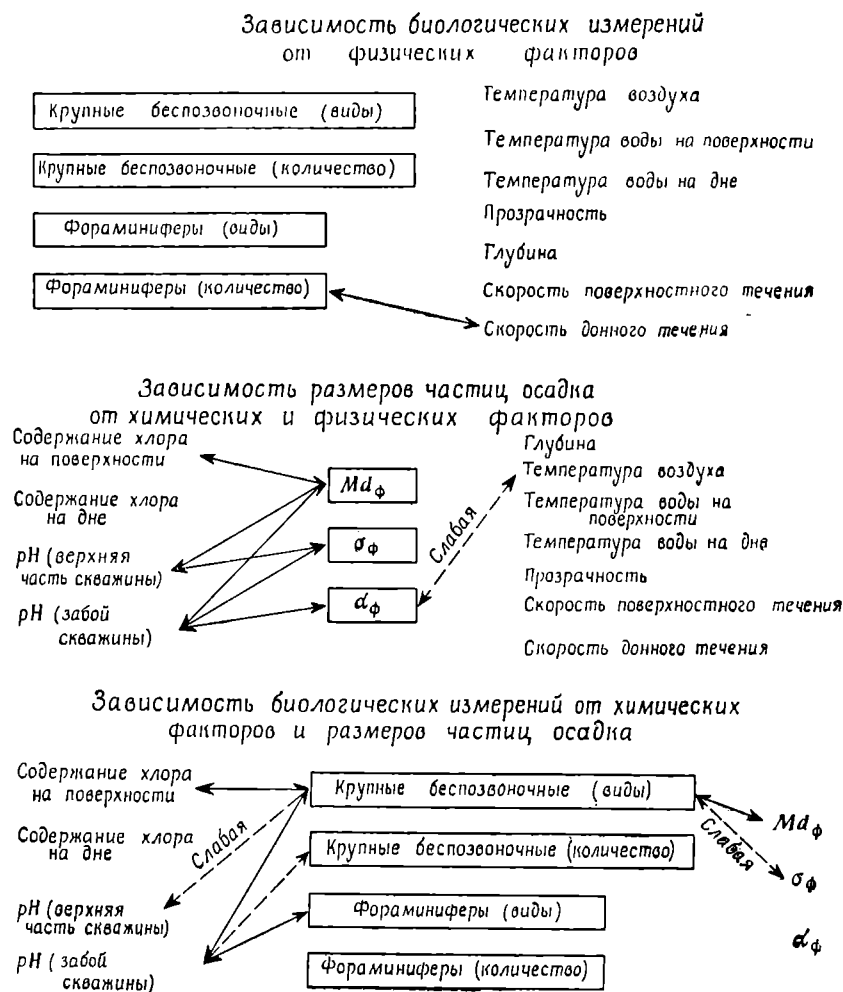
Табл. 13.12 представляет корреляционную матрицу, в которой через нуль обозначены те положительные результаты проверки гипотезы о независимости, при которых коэффициент

Корреляционные зависимости 18 случайных величин

O — независимы; H — сильно зависимы; $|r| \geq 0,70$; L — слабая связь; $0 < |r| < 0,70$

[illegible]

корреляции можно считать равным нулю. Для случаев, когда гипотеза о независимости отвергается, использованы два уровня силы связи. Первый из них соответствует случаям сильной зависимости, когда гипотеза $|r| \geq 0,70$ принимается при $\alpha = 0,003$.

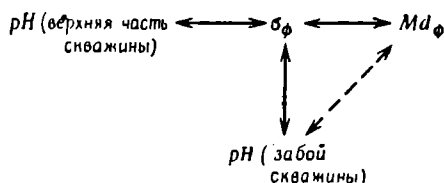


Фиг. 13.9. Схема внутренних связей.

В табл. 13.12 эти результаты обозначены через H . Второй уровень характеризует слабую зависимость, когда $0 < |r| < 0,70$, а соответствующие ему случаи обозначены через L . На фиг. 13.9

и 13.10 приведены результаты исследования, расположенные в произвольном порядке. Конечно, можно было бы выбрать другой способ расположения результатов, однако в любом случае, похожем на ситуацию в табл. 13.12, форма представления будет в основном такая же, как в этом примере.

При изучении фиг. 13.9 и 13.10 обнаруживается ряд интересных зависимостей. Некоторые характеристики обладают сильной внутренней связью. В то время как одни связи ожидалось, другие оказались неожиданными. Следует также отметить удивительно большое число независимых переменных в обстановке,



Фиг. 13.10. Сильная линейная зависимость.

Пунктир — $\rho \geq 0,90$, риск 0,003; сплошная линия — $\rho \geq 0,70$.

где можно было ожидать по крайней мере слабую связь. Так как в задачу авторов не входит интерпретация полученных результатов, они ее касаться не будут.

ВЫВОДЫ

Метод точечных диаграмм дает положительные результаты при механически планируемых выборках для свертывания информации большого объема. При таком подходе число исследуемых признаков может быть очень большим, особенно если в работе используются быстродействующие вычислительные машины (в таких случаях для программирования применяются методы более удобные, чем метод Берквиста). Корреляционная матрица — лишь один из многих способов свертки информации. Авторы считают, что доводы «за» и «против» метода точечных диаграмм заключаются в следующем. Метод дает возможность выявлять действующие «факторы» при предварительном изучении полученных групп внутренне связанных характеристик, таких, как показано на фиг. 13.9 и 13.10. При этом можно обнаружить такие связи, которые не предполагались, и иногда объяснить их, например, наличием других характеристик, связанных только с некоторыми членами группы.

Однако в исследованиях подобного типа следует скорее от-
328 носиться с осторожностью не к процессу получения результа-

тов, а к истолкованию последних. В тех случаях когда имеются надежные сведения, полученные *a priori*, и предположения (например, при изучении физических процессов, для которых можно высказать предположения о высокой или низкой корреляции), интерпретация корреляционной матрицы осуществляется без дальнейших проверок, и мы принимаем ее как надежно обоснованную.

В рассмотренном примере сведения *a priori* и предположения не все достаточно обоснованы (как для физических, так и биологических характеристик), а выборки подвержены воздействию тренда и зависимости флуктуаций от времени. В связи с этим предполагаемые зависимости следует вводить в виде гипотез с последующей статистической проверкой, предшествующей любому выводу.

В этой главе рассматривается одномерный тренд. Ниже будет дано определение этого понятия, а также описаны критерии для установления наличия тренда или его отсутствия. Например, упорядоченная последовательность случайных величин может изучаться различными способами с целью выяснения вопроса о случайном расположении элементов последовательности. Если гипотеза о случайном порядке отвергается (на вероятностной основе), то из этого следует вывод о наличии тренда. Таким образом, упорядоченные последовательности классифицируются как случайные или, наоборот, имеющие тренд.

Под трендом имеется в виду *любая закономерность в упорядоченной последовательности чисел или элементов. Любая упорядоченная последовательность чисел или элементов, отличающаяся от случайной, будет рассматриваться как имеющая тренд.*

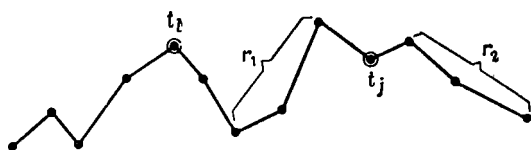
В частности, совокупность случайных флуктуаций около горизонтальной линии, например такой, как медиана, следует рассматривать как последовательность, не имеющую тренда. Наоборот, случайные отклонения от негоризонтальной прямой линии или кривой более высокого порядка требуют предположения о тренде в последовательности. Это предположение обычно основано на результатах изучения выборочной последовательности и информации *a priori*.

Критерии, описанные в этой главе, разделены на три категории.

1. К первой категории относится метод «смены знака», или, как его еще называют, метод «скачков вверх и вниз» в последовательности. Критерии, связанные с этим методом, основаны на числе точек смены знака или числе скачков вверх и вниз и на соотношении этих чисел и числа элементов в последовательности. На фиг. 14.1 дается представление об этих двух терминах. На ней изображена последовательность чисел, которые возрастают и уменьшаются. Примерами точек «смены знака» будут t_i и t_j , и их можно описать как точки максимумов и миниму-

мов. Термину *скачок вверх* соответствует на графике участок r_1 , а *скачок вниз* — r_2 . Таким образом, скачок вверх — последовательность точек, в которой значение каждой последующей точки больше значения предыдущей, а скачок вниз — последовательность точек с обратным соотношением.

2. Вторая категория критериев охватывает в основном дихотомические процессы. Примером такого процесса может служить последовательность, элементами которой будут знаки плюс и минус. Этими знаками можно обозначить положение точек относительно медианной линии (выше или ниже); но можно также воспользоваться и иными способами, приводящими к дихотомическим процессам. В результате будет получена последовательность, состоящая из двух типов элементов, которые неразличимы внутри одного типа. Пример такой последовательности: $++--++--++--++--$. Ее наиболее



Фиг. 14.1.

важными характеристиками являются общее число элементов, число элементов каждого из двух типов и число скачков. В данном случае *скачком* называется последовательность, образованная одним и более одинаковыми элементами, предшествующая элементу другого типа или следующая за ним.

На фиг. 14.1 изображена последовательность из 12 элементов, из которых 6 можно приписать знак плюс, а 6 — минус. В этой последовательности восемь скачков.

Дихотомическая процедура обычно используется там, где нужно представить последовательность чисел или результатов наблюдения как два различных класса однородных объектов при условии, что это допустимо с геологических позиций. Однако очень часто рассматриваемое явление по своей природе бывает дихотомическим и при исследовании не требуется введения искусственных приемов преобразования, таких, например, как разделение графика средней линией. В качестве примера можно привести результаты изучения окраски тонкослоистых глинистых сланцев. Введем два определения окраски — светлая L и темная D . Эти термины можно рассматривать в зависимости от количества карбоната, присутствующего в сланцах. Спрашивается, существует ли какая-либо закономерность, или тренд, в наблюдаемой последовательности $LDLDDLLDLLLLD$ окраски слоев в разрезе или же чередование индексов L и D чисто случайно.

В итоге мы проверим гипотезу о случайном расположении элементов последовательности, образованной двумя классами объектов. В конце этой главы приведен пример, в котором также рассматривается чередование трех литологических типов пород.

3. Третья категория включает вопросы применения метода ранговой корреляции Спирмена. В данном случае проверяется гипотеза о случайном расположении элементов последовательности при конкретной альтернативе. В частности, эта альтернатива будет заключаться в наличии определенного типа тренда, представляющего линейное возрастание (или уменьшение) значений.

Замечания о некоторых других критериях и указания на литературу будут даны в соответствующих местах.

Вместо рангового коэффициента корреляции ρ Спирмена можно использовать критерий Кендалла τ . Величины как ρ , так и τ при изучении одной и той же совокупности не дают одинаковых результатов, но обе они могут быть использованы в задачах, рассматриваемых в этой главе. Преимущества коэффициента ρ перед τ рассмотрены в работе Краскла¹. Коэффициент Спирмена ρ был впервые описан в 1904 г.², а величина Кендалла τ — в 1938 г.³

ОПИСАНИЕ И ОБСУЖДЕНИЕ КРИТЕРИЕВ

В этом разделе будет приведена фундаментальная теорема о распределении скачков. Прежде чем перейти к ней, рассмотрим термин *скачок*. Пусть дана последовательность элементов двух различных видов, таких, как плюс или минус, расположенная в порядке поступления результатов наблюдения: $++---+--+$. Термин *скачок* был определен выше как последовательность из одного и более одинаковых элементов. Таким образом, данная последовательность характеризуется семью скачками, из которых четыре образованы элементами плюс, а три — элементами минус. Нужно найти функцию распределения числа скачков u при данном числе элементов последовательности и заданном соотношении чисел элементов $(+)$ и $(-)$. В приведенном примере имеется восемь элементов $(+)$; обозначим это число $n_1, n_2 = 4$ элемента $(-)$. Общее число элементов равно 12. Пусть $P(u)$ — вероятность того, что в случайной последовательности из восьми элементов $(+)$ и четырех эле-

¹ W. H. Kruskal, Ordinal measures of association, Journ. Am. Stat. Assoc., 53, 1958.

² C. Spearman, The proof and measurement of association between two things, Am. Journ. Phych., 5, 1904.

³ M. G. Kendall, A new measure of rank correlation, Biometrika, 30,

ментов $(-)$ наступит ровно u скачков. Пусть $u=7$. Если значение $P(7)$ велико, значит число способов получения семи скачков в последовательности из 12 плюсов и минусов сильно зависит от общего числа всех возможных вариантов расположения 12 элементов последовательности. Если число скачков недопустимо мало, то гипотезу о случайном расположении элементов последовательности следует отвергнуть. С другой стороны, если число скачков недопустимо большое, то гипотеза о случайности также отвергается.

Теорема 14.1 [122]¹

Рассмотрим различные перестановки из n_1 результатов наблюдений одного вида и n_2 наблюдений другого вида, когда все варианты перестановок равновозможны. Пусть $P(u)$ — вероятность того, что в последовательности из $n_1 + n_2 = N$ этих наблюдений содержится точно u скачков. Используя принятые обозначения, имеем для четного u

$$P(u) = 2 \frac{(n_1 - 1)! (n_2 - 1)! n_1! n_2!}{\left(\frac{u}{2} - 1\right)! \left(\frac{u}{2} - 1\right)! \left(n_1 - \frac{u}{2}\right)! \left(n_2 - \frac{u}{2}\right)! N!}. \quad (14.1)$$

Если u — нечетное число, то

$$P(u) = \frac{(n_1 - 1)! (n_2 - 1)! n_1! n_2!}{\left(\frac{u-1}{2}\right)! \left(\frac{u-3}{2}\right)! N!} \times \\ \times \left[\frac{1}{\left(n_1 - \frac{u+1}{2}\right)! \left(n_2 - \frac{u-1}{2}\right)!} + \frac{1}{\left(n_1 - \frac{u-1}{2}\right)! \left(n_2 - \frac{u+1}{2}\right)!} \right]. \quad (14.2)$$

Доказательство

Для доказательства нужно рассмотреть процедуру счета. В последовательности из N элементов требуется подсчитать число способов расположения n_1 элементов $(+)$ и n_2 элементов $(-)$, приводящих к u скачкам. В дихотомическом процессе всегда будет или u скачков (u четное), или $u+1$. Рассмотрим случай, когда u — четное число. Если u четно, то число скачков, образованных элементами $(+)$, будет равно числу скачков, соответствующих элементам $(-)$, и равно $u/2$. Для простоты будем рассматривать скачки как участки последовательности и отделять их черточками; например, $/++++/+++/+$. Заметим, во-первых, что первая и последняя черточки фиксированы и не принимаются в счет. Так как $(u/2) + 1$ черточек обязательно образуют $u/2$ участков, то будем рассматривать $(u/2) + 1 - 2$,

¹ См. также W. Feller, An introduction to probability theory and its applications, I, New York, 1950.

или $(u/2) - 1$, нефиксированных черточек. Во-вторых, наибольшее число черточек, которыми можно разделить n_1 элементов (+), равно $n_1 - 1$. Спрашивается, сколько существует способов расположения $(u/2) - 1$ черточек на $n_1 - 1$ мест ($n_1 \geq \frac{u}{2}$, т. е. не более чем одну черточку на одно место, так как пустые отрезки не рассматриваются)? Искомое число способов можно выразить комбинаторной формулой $\binom{n_1-1}{(u/2)-1}$. Аналогичная процедура для элементов (—) приведет к выражению $\binom{n_2-1}{(u/2)-1}$. Используя эти результаты, найдем числитель для $P(u)$.

Каждому из вариантов расположения элементов (+) соответствует $\binom{n_2-1}{(u/2)-1}$ способов расположения элементов (—). Общее число способов как для (+), так и для (—) будет

$$\binom{n_1-1}{(u/2)-1} \binom{n_2-1}{(u/2)-1}.$$

Наконец, скачки должны чередоваться, так что любой знак может соответствовать первому скачку. Это приводит к удвоению общего числа скачков. В результате числитель примет следующий вид:

$$2 \binom{n_1-1}{(u/2)-1} \binom{n_2-1}{(u/2)-1}.$$

Выражение для $P(u)$ будет представлять отношение числа способов расположения n_1 объектов одного вида и n_2 другого, приводящего к u скачкам, к общему числу способов расположения двух видов объектов. Знаменатель этого отношения можно подсчитать, рассмотрев число способов, которыми можно выбрать n_1 объектов из совокупности, состоящей из $n_1 + n_2 = N$ объектов. Это число дается комбинаторной формулой $\binom{N}{n_1}$. Искомое отношение¹ можно записать как

$$\begin{aligned} P(u) &= \frac{2 \binom{n_1-1}{(u/2)-1} \binom{n_2-1}{(u/2)-1}}{\binom{N}{n_1}} = \\ &= 2 \left[\frac{(n_1-1)!}{\left(\frac{u}{2}-1\right)! \left(n_1 - \frac{u}{2} - 1 + 1\right)!} \frac{(n_2-1)!}{\left(\frac{u}{2}-1\right)! \left(n_2 - \frac{u}{2} - 1 + 1\right)!} \frac{n_1! n_2!}{N!} \right] = \\ &= \frac{2 (n_1-1)! (n_2-1)! n_1! n_2!}{\left(\frac{u}{2}-1\right)! \left(\frac{u}{2}-1\right)! \left(n_1 - \frac{u}{2}\right)! \left(n_2 - \frac{u}{2}\right)! N!}. \end{aligned}$$

¹ S. S. Wilks, Mathematical statistics, Princeton Univ. Press, Princeton, New Jersey, Chap. 10, 1943.

Это выражение совпадает с формулой (14.1) теоремы 14.1. В случае нечетного u доказательство проводится таким же способом.

Нормальное приближение для функции $P(u)$

в последовательности, состоящей из n_1 элементов одного вида и n_2 другого вида

Вычисление значений $P(u)$ становится весьма трудоемким, если или n_1 , или n_2 — большое число. Это частично связано с наличием сомножителей, представленных факториалами. Однако, если число $N = (n_1 + n_2)$ стремится к бесконечности, функция $P(u)$ приближается к нормальной. Практически приближением к нормальному распределению можно уже пользоваться, когда каждое из чисел n_1 и n_2 превышает 10. Математическое ожидание $E(u)$ и дисперсия $V(u)$ случайной величины u будут

$$E(u) = \frac{2n_1n_2}{n_1 + n_2} + 1;$$

$$V(u) = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)},$$

а для функции $P(u)$ будет справедливо приближенное равенство

$$P(u) \approx N \left[\frac{2n_1n_2}{n_1 + n_2} + 1, \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)} \right]. \quad (14.3)$$

Объясним на примере, как получается формула для параметра $E(u)$. Пусть дана последовательность: $-+ -+ +$. Точки между соседними скачками будем называть *переходными точками*. Эта последовательность характеризуется четырьмя скачками, четырьмя промежутками между элементами и тремя переходными точками. Обозначим через $E(p_i)$ вероятность того, что i -й промежуток между элементами последовательности окажется переходной точкой. Для определения $E(p_i)$ рассмотрим процедуру подсчета. Вероятность того, что в последовательности из пяти элементов, содержащей три (+) и два (—), третий элемент есть (+), равна $3/5$, а вероятность того, что четвертый элемент будет (—) при условии, что третий есть (+), равна $2/4$, так как одно из всех пяти мест в последовательности уже занято. С другой стороны, вероятность того, что третий элемент в этой последовательности есть (—), равна $2/5$, а вероятность того, что четвертый элемент будет (+) при условии, что третий (—), равна $3/4$. Так как случайные расположения элементов в последовательности независимы, то вероятность совместного наступления двух событий [третий элемент (+) и четвертый (—)] равна $3/5 \cdot 2/4$. Аналогично вероятность совместного наступления двух событий [третий элемент (—) и четвертый

(+)] равна $2/5 \cdot 3/4$. Таким образом, $E(p_i)$ в нашем примере представляет сумму $3/5 \cdot 2/4 + 2/5 \cdot 3/4$.

Теперь обобщим этот пример. Допустим, что последовательность образована n_1 элементами (+) и n_2 (-). Тогда

$$E(p_i) = \frac{n_1}{n_1 + n_2} \frac{n_2}{n_1 + n_2 - 1} + \frac{n_2}{n_1 + n_2} \frac{n_1}{n_1 + n_2 - 1},$$

или

$$E(p_i) = \frac{2n_1 n_2}{(n_1 + n_2)(n_1 + n_2 - 1)};$$

обозначим ожидаемое значение числа всех переходных точек через $E(p)$, где $E(p) = \sum E(p_i)$. Так как число промежутков между элементами последовательности на единицу меньше, чем число элементов, и так как нужно провести подсчет по всем возможным переходным точкам, то $\sum E(p_i) = (n_1 + n_2 - 1)E(p_i)$. Подставив в это выражение значение $E(p_i)$, получим

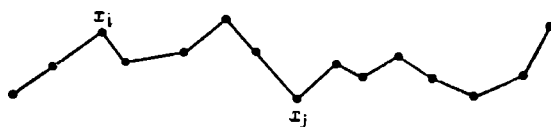
$$E(p) = \frac{2n_1 n_2}{n_1 + n_2}.$$

Число возможных скачков на единицу превышает число переходных точек. Следовательно,

$$E(u) = \frac{2n_1 n_2}{n_1 + n_2} + 1.$$

Точки смены знака и скачки вверх и вниз

Если число N случайных расположений неравных чисел стремится к бесконечности, распределение числа скачков вверх или вниз r приближается к нормальному¹.



Фиг. 14.2.

В практических приложениях можно рассматривать распределение числа скачков $P(r)$ как нормальное при конечном значении N с параметрами $E(r) = (2N - 1)/3$ и $V(r) = (16N - 29)/90$:

$$P(r) \approx N\left(\frac{2N - 1}{3}, \frac{16N - 29}{90}\right). \quad (14.4)$$

Это приближение является хорошим при N порядка 10.

¹ W. A. Wallis, H. V. Roberts, Statistics, a new approach, Free Press, Glencoe, Illinois, 1956.

Рассмотрим вывод значения величины $E(r)$. На фиг. 14.2 приведена последовательность неравных чисел, где точки x_i и x_j представляют собой «точки смены знака». Они непосредственно предшествуют или, наоборот, следуют за точками, которые или выше, или ниже их и которые являются максимумами или минимумами рассматриваемой последовательности.

В случайной последовательности вероятностно того, что любой последовательный набор трех точек содержит точку смены знака, можно определить путем простого перечисления всех возможных вариантов расположения точек. Допустим, что для чисел x_1, x_2, x_3 справедливо неравенство $x_1 < x_2 < x_3$. Все возможные варианты расположения этих трех точек в последовательности приведены на фиг. 14.3. Обратим особое внимание на среднюю точку всех шести полученных последовательностей и подсчитаем число t_i случаев появления точки смены знака, а затем вычислим величину $E(t_i)$:

$$E(t_i) = \sum_{i=1}^6 p(t_i) t_i = \frac{4}{6} = \frac{2}{3},$$

так как на шесть возможных вариантов расположения трех точек приходится четыре точки смены знака. В последовательности из N элементов содержится $N-2$ возможных подпоследовательностей, состоящих из трех элементов. Записав всю последовательность и подсчитав все возможные подпоследовательности из трех элементов, получим

$$E(t) = (N-2)E(t_i) = (N-2) \cdot \frac{2}{3} = \frac{2N-4}{3}. \quad (14.5)$$

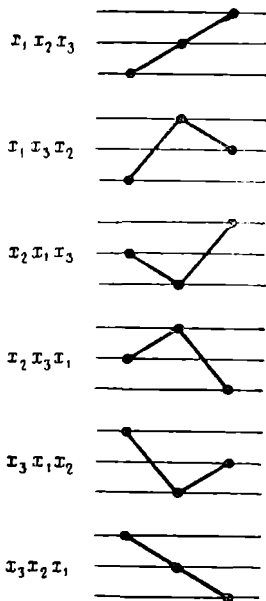
Распределение числа точек смены знака в последовательности из N элементов можно выразить через нормальное распределение (в том же смысле, что и распределение числа скачков вверх или вниз) с параметрами¹

$$E(t) = \frac{2N-4}{3} \quad \text{и} \quad V(t) = \frac{16N-29}{90},$$

т. е.

$$P(t) \approx N\left(\frac{2N-4}{3}, \frac{16N-29}{90}\right).$$

¹ W. A. Wallis, G. H. Moore, A significance test for time-series, Tech. Paper, № 1, Nat. Bureau of Econ. Res., New York, 1941.



Фиг. 14.3.

Число скачков вверх или вниз на единицу меньше числа точек смены знака. Это показано на фиг. 14.1, где $\sum_i(t_i) = 7$, а $r = 8$. Следовательно,

$$\begin{aligned} E(r) &= E(t+1) = E(t) + 1 = \\ &= \frac{2N-4}{3} + 1; \end{aligned} \quad (14.6)$$

$$E(r) = \frac{2N-1}{3}.$$

В результате мы получили математические ожидания числа точек смены знака и числа скачков вверх или вниз в последовательности из N элементов. Дисперсия для обеих случайных величин одна и та же. Таким образом, сосчитав число точек смены знака, можно воспользоваться параметрами $E(t)$ и $V(t)$ или же, сосчитав число скачков вверх или вниз, применить $E(r)$ и $V(r)$. Вероятностный результат будет один и тот же, если учесть, что $P(r) = P(t+1)$.

ПРИЛОЖЕНИЯ

Начиная от основания до верхней части обнажения Скаергаардской интрузии в Гренландии, была взята серия образцов гортнолитового габбро¹. Химические анализы части образцов приведены в табл. 14.1. Результаты анализов на 13 окислов расположены в последовательности от основания интрузии к ее верхней части, причем анализы образцов, взятых в краевых частях интрузии, из рассмотрения исключены.

Проанализируем эти данные с целью выяснения вопроса о наличии или отсутствии тренда. Попутно рассмотрим критерии для проверки гипотезы о тренде определенного вида, когда альтернативность гипотезы о случайном расположении в последовательности кажется очевидной.

Проверка гипотезы о случайном расположении

Результаты анализа на каждый окисел были выражены в виде дихотомической последовательности из n_1 элементов (+) и n_2 (—). Знаком (+) были обозначены содержания, превышающие медианное значение, а знаком (—) — те результаты, которые меньше его. Последовательности знаков (+) и (—), а

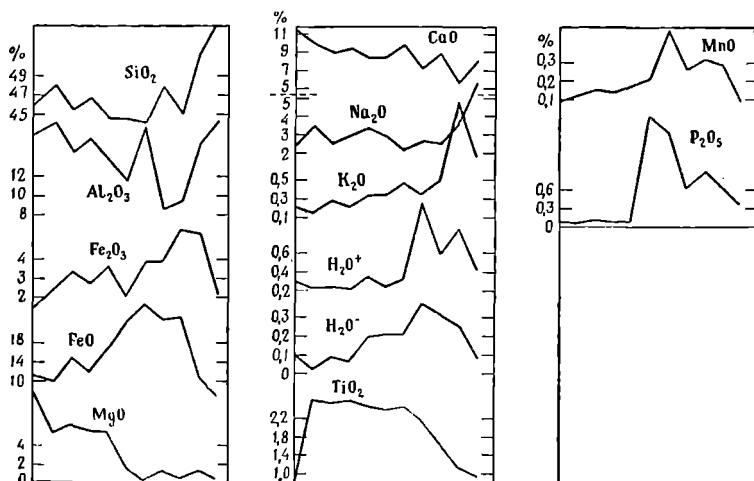
¹ L. R. Wager, W. A. Deer, Geological investigations in East Greenland, Pt. III, the petrology of the Skaergaard intrusion, Meddelelser om Grønland, Bd. 105, № 4, 1939.

Т а б л и ц а 14.1

Окислы	II 4077	III 3662	IV 3661	IVa	V 1907	VI 4145	VII 4142	VIII 1881	IX 4139	X 4137	XI 4136
SiO ₂	46,37	48,15	45,65	46,90	44,81	44,61	44,13	48,27	45,19	52,13	55,30
Al ₂ O ₃	16,82	18,02	15,08	16,55	13,96	11,70	17,88	8,58	9,37	15,87	18,32
Fe ₂ O ₃	1,52	2,52	3,41	2,96	3,75	2,05	4,05	4,06	5,78	5,61	2,18
FeO	10,44	9,50	14,86	12,18	16,66	22,68	26,63	22,89	23,77	11,17	7,47
MgO	9,61	5,25	6,35	5,80	5,54	1,71	0,25	1,21	0,43	1,11	0,21
CaO	11,29	10,17	9,18	9,67	8,53	8,71	10,03	7,42	9,05	5,80	8,20
Na ₂ O	2,45	3,46	2,48	2,97	3,35	2,95	2,15	2,65	2,43	3,63	6,01
K ₂ O	0,20	0,14	0,28	0,21	0,33	0,35	0,47	0,34	0,49	1,38	0,78
H ₂ O ⁺	0,29	0,20	0,22	0,21	0,34	0,22	0,30	1,13	0,57	0,86	0,41
H ₂ O ⁻	0,09	0,02	0,08	0,05	0,19	0,20	0,19	0,37	0,31	0,25	0,06
TiO ₂	0,79	2,64	2,59	2,61	2,55	2,43	2,48	2,20	1,67	1,14	0,94
MnO	0,09	0,12	0,15	0,14	0,17	0,21	0,48	0,26	0,32	0,30	0,09
P ₂ O ₅	0,06	0,05	0,08	0,06	0,08	1,85	1,61	0,65	0,91	0,70	0,42

Оксиды	Медиана		n_1 (+)	n_2 (-)	u	$P(u)$	ρ
SiO ₂	46,37	+ - + - - - + - + +	5	5	7	0,190	0,290
Al ₂ O ₃	15,87	+ + - + - - + - - +	5	5	7	0,190	-0,155
Fe ₂ O ₃	3,41	- - - + - + + + + -	5	5	5	0,190	0,527
FeO	14,86	- - - + + + + + - -	5	5	3	0,032*	0,200
MgO	1,71	+ + + + + - - - - -	5	5	2	0,008*	-0,881
CaO	9,05	+ + + + - - + - - -	5	5	4	0,127	-0,782
Na ₂ O	2,95	- + - + + - - - + +	5	5	6	0,286	0,245
K ₂ O	0,34	- - - - - + + + + +	5	5	2	0,008*	0,946
H ₂ O ⁺	0,30	- - - - - + - + + +	5	5	4	0,127	0,770
H ₂ O ⁻	0,19	- - - - - + + + - -	3	6	3	0,164	0,530
TiO ₂	2,43	- + + + + + - - - -	5	5	3	0,032	-0,482
MnO	0,17	- - - - - + + + + +	5	5	3	0,032*	0,475
P ₂ O ₅	0,42	- - - - - + + + + +	5	5	2	0,008*	0,673

также значения n_1 , n_2 и числа скачков u приведены в табл. 14.2. Так как число элементов последовательности очень невелико,



Фиг. 14.4.

то при вычислении $P(u)$ не применялось нормальное приближение, а все расчеты проводились по формулам (14.1) и (14.2). Результаты этих вычислений приведены в предпоследнем столбце табл. 14.2. На фиг. 14.4 последовательности значений каждого компонента приведены в виде графиков, которые должны

помочь понять табл. 14.2. До того как начнем рассматривать результаты, рассмотрим поведение каждого компонента в отдельности. На фиг. 14.4 видно, что главные компоненты проявляют тенденцию к увеличению или, наоборот, уменьшению значений, близкую к линейной. Например, содержания MgO частично убывают, а для кривой K_2O характерно общее повышение. Проверим, устойчива ли такая тенденция в рассматриваемых последовательностях или же она вызвана случайными изменениями. Так как сумма содержаний всех компонентов в одном образце равна 100%, результаты анализов зависимы внутри строки. Поэтому, хотя приводимые ниже индивидуальные критерии, по-видимому, верны, имеется положительная корреляция. Более подходящим был бы какой-либо совместный критерий.

ПРИМЕНЕНИЕ РАНГОВОГО КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ СПИРМЕНА

Критерий, который будет применяться в дальнейшем, представляет собой оценку коэффициента корреляции между наблюдаемой последовательностью и возрастающим рядом целых чисел. Рассмотрим его подробнее на примере последовательности значений содержания SiO_2 (табл. 14.2 и фиг. 14.4). Эта последовательность содержит 11 значений, так что мы будем сравнивать попарно ранги содержаний SiO_2 с элементами последовательности ряда целых чисел от 1 до 11, расположенных в возрастающем порядке:

Ряд целых чисел	1	2	3	4	5	6	7	8	9	10	11	
Ранг SiO_2	6	8	5	7	3	2	1	9	4	10	11	
d (разность)	-5	-6	-2	-3	2	4	6	-1	5	0	0	$\sum d = 0$

$$\rho = 1 - \frac{6 \sum d^2}{n^3 - n} = 1 - \frac{936}{1320} = 0,2901.$$

Полученное значение ρ свидетельствует о том, что тенденция к увеличению содержания SiO_2 выражена слабо.

Ранговый коэффициент корреляции Спирмена ρ принимает значения в интервале от -1 до 1 , так что сильно выраженному тренду увеличения содержаний соответствуют близкие к 1 значения ρ и, наоборот, тренду, связанному с понижением значений в последовательности, соответствуют величины ρ , близкие к -1 .

Коэффициент корреляции Спирмена был вычислен для всех компонентов этого примера, а результаты вычислений приведены в последнем столбце табл. 14.2.

ВЫВОДЫ

Результаты изучения последовательностей с помощью анализа скачков даны в табл. 14.2 в виде значений вероятности $P(u)$ того, что для наблюдаемого порядка элементов (+) и (—) верна гипотеза о случайном расположении. Звездочкой отмечены те значения вероятностей, которые настолько малы, что не позволяют признать расположение случайным. В соответствии со сформулированным выше определением мы говорим, что в подобных случаях имеет место тренд. Из 12 рассмотренных последовательностей гипотеза об отсутствии тренда в шести случаях подтвердилась, а в шести отвергнута, что обосновывает наличие тренда.

Сравним эти результаты со значениями коэффициента корреляции Спирмена. Разделим все 12 последовательностей на две группы, из которых в одну войдут те, где наличие тренда подтвердилось, а в другую — последовательности со случайным расположением элементов. В группе, для которой характерен тренд, содержание MgO закономерно понижается, тогда как для K_2O и P_2O_5 типично регулярное увеличение их содержания. Для FeO , TiO_2 , MnO регулярное увеличение или уменьшение содержания проявлено не очень отчетливо, однако при рассмотрении графиков видно, что для этих окислов характерно повышение содержания с последующим снижением. Результаты применения обоих критериев представляются состоятельными в случае наличия линейного тренда. Для выявления нелинейного тренда критерий Спирмена непригоден.

Теперь рассмотрим вторую группу последовательностей, ту, в которой расположение элементов последовательностей оказалось случайным. Результаты изучения последовательностей для SiO_2 , Al_2O_3 , Fe_2O_3 и Na_2O не позволяют делать вывода о наличии в них отчетливо выраженных тенденций к возрастанию или снижению, о чем свидетельствуют полученные значения ρ . Однако в случае последовательностей CaO и H_2O^+ критерий, основанный на распределении числа скачков, обязывает сделать вывод о случайном расположении элементов последовательностей, тогда как значения коэффициента корреляции Спирмена свидетельствуют о регулярном понижении содержаний в первом случае и возрастании их во втором. Возникает дилемма: какой критерий заслуживает большего доверия? Авторы считают, что подобные ситуации следует рассматривать как случаи с неявно выраженными результатами. В связи с тем что объемы рас-

сматриваемых выборок малы, имеющийся в действительности тренд может не быть обнаружен, так как результаты изучения выборки дадут большое значение вероятности случайного расположения элементов. Этот вопрос рассмотрен в конце настоящей главы.

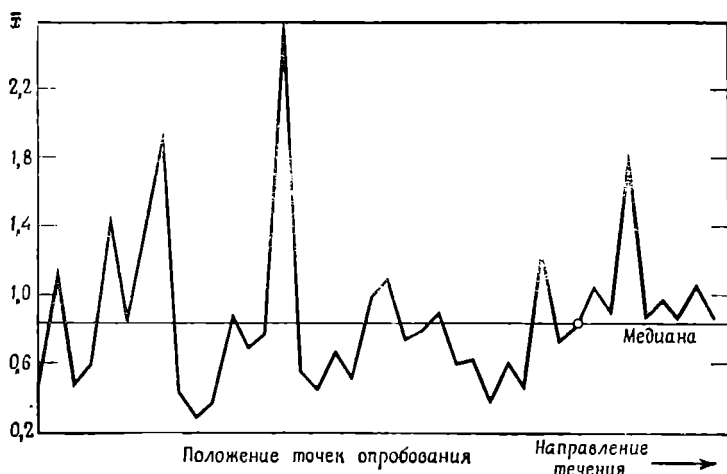
Действительно, в тех задачах, в которых решается вопрос о наличии линейного тренда, критерий, основанный на распределении числа скачков, оказывается слабее критерия Спирмена. Анализ скачков вверх или вниз может дать более надежные результаты, чем критерий Спирмена.

ПРИМЕНЕНИЕ ДИХОТОМИЧЕСКИХ ПРОЦЕССОВ ПРИ ИЗУЧЕНИИ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ИЗ $n_1 + n_2$ ЭЛЕМЕНТОВ ПРИ БОЛЬШИХ ЗНАЧЕНИЯХ $n_1 + n_2$

При больших N последовательность значений случайной величины можно обрабатывать, используя нормальную аппроксимацию для $P(u)$, которая значительно облегчает работу. Приведенный ниже пример иллюстрирует эту процедуру. Лагн¹ для характеристики главного русла Миссисипи получил данные о среднем размере частиц аллювия, собрав образцы через равные интервалы вниз по течению. На фиг. 14.5 приведен график изменения среднего размера частиц аллювия на участке от Дэвенпорта (Айова) до устья Иллинойса. Анализируя график фиг. 14.5, нельзя сделать вывода о явном тренде. Преобразуем эти данные следующим образом. Подсчитаем оценку для медианы, а затем те значения среднего размера аллювия, которые ее превышают, обозначим как (+), а те, которые меньше ее, — как (—). Запишем элементы новой последовательности, сохраняя старый порядок. Она состоит из 40 элементов, из которых 20 (+) и 20 (—). Кроме того, в этой последовательности оказалось 16 скачков. Теперь проверим гипотезу, заключающуюся в том, что наблюдаемый порядок расположения плюсов и минусов случаен. Если число скачков окажется меньше, чем следовало бы ожидать для случайной последовательности, нужно сделать вывод о возможном тренде. Если же число скачков больше допустимого для случайной последовательности, возникает предположение о наличии тренда, связанного с регулярным повышением или понижением значений элементов ряда. Таблицы допустимых значений числа скачков при уровне значимости 0,025 для всех n_1 и $n_2 \leq 20$ и для $n_1 = n_2$ до 100 (здесь n_1 — число плюсов, а n_2 — число минусов) приведены в [79].

¹ A. L. Lugin, Sedimentation in the Mississippi River, Augustana Lib. Publ., № 11, 1927.

В рассматриваемой последовательности для дихотомического преобразования использована медиана; поэтому $n_1 = n_2 = 20$. Так как $u = 16$, то в соответствии с таблицами гипотеза о случайном расположении элементов последовательности должна быть принята при уровне значимости 0,025.



Ф и г. 14.5.

Воспользуемся для проверки этой же гипотезы нормальным приближением. Для этого вычислим

$$E(u) = \frac{2n_1n_2}{n_1 + n_2} + 1 = \frac{2 \cdot 20^2}{40} + 1 = 21;$$

$$V(u) = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)} = \frac{800 \cdot (800 - 40)}{1600 \cdot 39} = 9,74;$$

$$z = \frac{u - E(u)}{\sqrt{V(u)}} = \frac{16 - 21}{3,121} = -1,602.$$

По таблицам нормального распределения находим, что численному значению z соответствует вероятность 0,056 того, что флуктуации значений в последовательности около медианы случайны. Таким образом, значение вероятности мало, чтобы можно было принять отчетливое решение. Поэтому проверку следует продолжить.

Форма графика на фиг. 14.5 позволяет допустить возможность регулярного снижения и повышения значений. Поэтому мы снова проверим гипотезу о случайном расположении элементов последовательности, используя в качестве критерия число точек смены знака. Это число t равно 29. Воспользовавшись

нормальным приближением, для функции $P(t)$ получим:

$$E(t) = \frac{2N-4}{3} = 25,33,$$

$$V(t) = \frac{16N-29}{90} = 6,788,$$

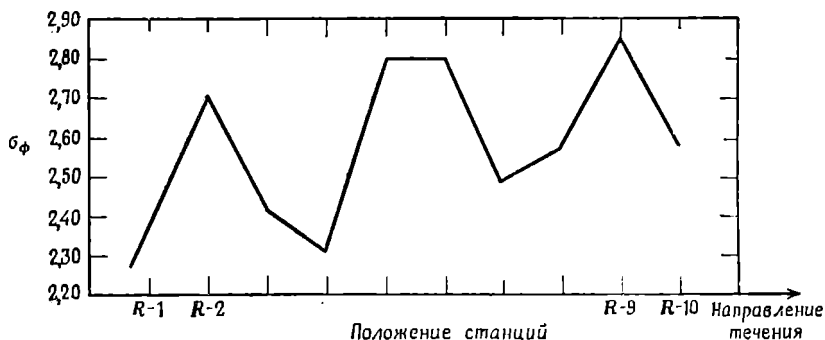
$$z = \frac{t - E(t)}{\sqrt{V(t)}} = \frac{29 - 25,33}{2,605} = 1,332.$$

Из таблицы нормального распределения следует, что вычисленное значение и в этом случае значительно меньше допустимого при уровне значимости 0,025, что обязывает принять гипотезу о случайном расположении элементов последовательности в целом.

В итоге, основываясь на результатах применения анализа числа скачков и точек смены знака, авторы делают вывод об отсутствии тренда для всего рассматриваемого участка реки. Предположения *a priori* о наличии локального тренда, например являющегося следствием значительного привноса терригенного материала притоками Миссисипи, могут привести к изучению подпоследовательностей рассмотренных данных.

СЛУЧАИ, КОГДА $n_1 + n_2$ НЕВЕЛИКО

Рассмотрим пример изучения последовательности с небольшим числом элементов, который послужит дополнением к пре-



Фиг. 14.6. График изменения значений коэффициента отсортированности по результатам последовательности наблюдений вниз по течению.

дыдущему разделу. Исследуем последовательность наблюдаемых значений коэффициента отсортированности σ_ϕ осадка по данным Плумлея [217], собранным вниз по течению реки Бнэр-Бют (Южная Дакота). Графическое изображение результатов измерения этого коэффициента приведено на фиг. 14.6, на

которой видно, что в данной последовательности вполне возможен значительный тренд вверх, несмотря на отчетливую цикличность в последовательной смене увеличения и уменьшения значений коэффициента отсортированности. Сформулируем две гипотезы, подлежащие проверке.

Согласно первой, расположение (+) и (—) элементов последовательности случайно; если эта гипотеза будет отвергнута, следует констатировать наличие тренда. Согласно второй, число точек смены знака не больше или не меньше ожидаемого для случайного расположения. Если эта гипотеза окажется ложной и число точек смены знака чрезмерно большим, из этого последует вывод о возможной цикличности явления, например о периодическом чередовании повышенных и пониженных значений упорядоченных данных [141]. Эти гипотезы будут проверены в том же порядке, в каком они сформулированы.

Критерий числа скачков

С помощью медианы значения наблюдаемого ряда были преобразованы в элементы дихотомической последовательности (+) и (—). Так как число элементов последовательности невелико ($N=10$), то пользоваться нормальным приближением нельзя и значение $P(u)$ нужно вычислять непосредственно из комбинаторных формул. Последовательности — + — — + + — — + + соответствуют величины: $u=6$, $N=10$, $n_1=n_2=5$. Поэтому

$$P(u) = 2 \left[\frac{4!4!5!5!}{2!2!2!2!10!} \right] \approx 0,1428.$$

Таким образом, вероятность того, что наблюдаемое число скачков не противоречит гипотезе о случайном расположении элементов последовательности, достаточно высока, что позволяет принять эту гипотезу как подтвердившуюся. В связи с тем что возможна некоторая периодичность в изменениях значений коэффициента отсортированности, проверим вторую гипотезу о случайном расположении с помощью критерия, основанного на распределении числа точек смены знака. В этом примере $t=5$, а $N=10$; поэтому

$$E(t) = \frac{2N-4}{3} = 5,33,$$

$$V(t) = \frac{16N-29}{90} = 1,455,$$

$$z = \frac{t - E(t)}{\sqrt{V(t)}} = 0,273.$$

В связи с тем что вычисленное значение z намного меньше допустимого (1,96) при уровне значимости 0,05, следует сделать вывод, что при объеме выборки 10 наличие пяти точек

смены знака не дает оснований для признания периодичности колебаний в значениях коэффициента отсортированности.

Выборочное значение r_s рангового коэффициента корреляции Спирмена ρ , применяемого для проверки гипотезы об отсутствии направленного линейного тренда, оказалось равным 0,500, что следует рассматривать как предельное значение. Однако это заставляет предполагать цикличность, наложенную на линейный тренд, характеризующийся повышением значений слева направо, что четко отражает график (фиг. 14.6). Предлагаемая процедура проверки гипотезы о линейном тренде состоит в подборе прямой линии для имеющихся данных по методу наименьших квадратов и проверке ее пригодности в качестве уравнения регрессии.

Если эта линия оказывается удовлетворительной, можно вновь применить критерий, основанный на распределении числа скачков, но на этот раз следует рассматривать отклонения не от медианы, а от линии регрессии. Если число скачков относительно этой линии превысит ожидаемое для случайного расположения (односторонний критерий), следует сделать вывод о наложении цикличности на негоризонтальный линейный тренд.

Скачки в последовательности, содержащей более двух видов независимых элементов

Допустим, что с целью изучения плейстоценовых отложений последовательно взята серия образцов. Отбор производился следующим способом. Через интервалы, равные приблизительно 1,6 км (в зависимости от обнаженности), отбирались образцы ледниковых отложений. Всего было отобрано 20 образцов, лежащих на прямой линии, причем объем каждого из них был приблизительно равен 0,016 м³. Предполагалось, что линия взятия образцов в общем совпадала с осью ледника. В каждом образце изучался состав галек. Доминирующий состав гальки в отдельных образцах был различным: известняки (L), глинистые сланцы (S), кристаллические породы (X).

Результаты определения доминирующего состава

Станция	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Преобладающий состав	L	L	S	L	S	X	X	X	X	S	L	L	S	S	S	S	S	L	L	X

Полученная последовательность состоит из 20 элементов, из которых L наблюдается семь раз (n_L), S — восемь раз (n_S), X — пять раз (n_X). Число скачков для этого ряда равно 10.

Теперь проверим гипотезу, заключающуюся в том, что расположение элементов трех видов случайно. Приведенные ниже уравнения даны без вывода (они взяты из работы Уиллиса и Мура):

$$\begin{aligned} E(u_3) &= \frac{2(n_L n_S + n_L n_X + n_S n_X)}{n_L + n_S + n_X} + 1 = \\ &= \frac{2 \cdot (7 \cdot 8 + 7 \cdot 5 + 8 \cdot 5)}{7 + 8 + 5} + 1 = 14,1; \\ V(u_3) &= \frac{[2(n_L n_S + n_L n_X + n_S n_X)]^2}{(n_L + n_S + n_X)^2 (n_L + n_S + n_X - 1)} - \\ &- \frac{2(n_L n_S + n_L n_X + n_S n_X) + 6n_L n_S n_X}{(n_L + n_S + n_X)(n_L + n_S + n_X - 1)} = 3,922; \\ z &= \frac{u_3 - E(u_3)}{\sqrt{V(u_3)}} = 2,070. \end{aligned}$$

Из таблицы значений нормальной функции распределения следует, что вычисленное значение z превышает допустимое при уровне значимости 0,025, в связи с чем гипотезу о случайном расположении элементов последовательности следует отвергнуть. Расположение L , S и X в последовательности свидетельствует о групповом эффекте. Возможно, что при дальнейших исследованиях можно будет выделить среди изученной толщи пород линейно коррелируемые группы.

В данном случае в свете полученного результата было бы полезно провести более детальное опробование. Кроме того, здесь можно применить так называемый метод «ближайшего соседа», описанный в гл. 16.

При обработке подобных данных следует проявлять некоторую осторожность в выводах. Дело в том, что хотя критерий и отвергает гипотезу о случайном расположении элементов последовательности, он не дает основания для выбора какой-либо конкретной гипотезы из множества альтернатив, которые можно сформулировать при описании наблюдаемого набора данных. В этой ситуации, по мнению авторов, следует рекомендовать продолжать более детальные исследования.

ВЫВОДЫ

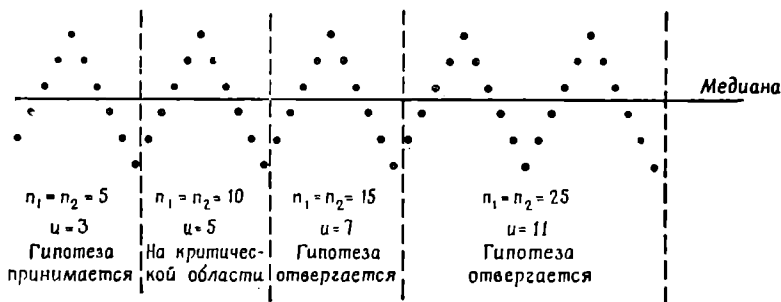
Очень важно иметь в виду, что критерии, основанные на распределении числа скачков и числа точек смены знака, предназначены для проверки гипотезы о случайном порядке. Геолог, применяющий тот или иной из описанных здесь критериев, должен быть уверен в том, что порядок расположения элементов в последовательности имеет геологический смысл. В следующем разделе будет дан обзор последовательностей, упорядоченных

во времени. Упорядочение может быть проведено также на основе данных полевого опробования, например на основе привязки к пересечению или так, как это было сделано при изучении отсортированности осадка по длине русла.

При чтении этой главы нетрудно было заметить, что критерий для проверки гипотезы о случайном порядке выбирался на основании некоторой предполагаемой альтернативы, но эта частная альтернатива не принималась как единственно возможная из всего множества конкурирующих гипотез. Логические построения, связанные с этой процедурой, можно найти в работе Кендалла¹.

Наконец, мы определяем термин *тренд* с помощью *регулярности в расположении*, противопоставленной случайному порядку. Эта регулярность не всегда одинакова и бывает представлена закономерностями, связанными с периодическим увеличением и уменьшением значений, а также линейным возрастанием или снижением. Кроме того, существует еще параболический или экспоненциальный тренд. Для выбора из различных видов тренда, соответствующего изучаемой последовательности, при условии, что гипотеза о случайном порядке отвергнута, требуются дальнейшие исследования, например подбор соответствующей кривой.

Результаты, полученные с помощью критерия, основанного на числе скачков, нужно применять с осторожностью, особенно в условиях малых выборок. На фиг. 14.7 схематически показана последовательность, характеризующаяся чередованием

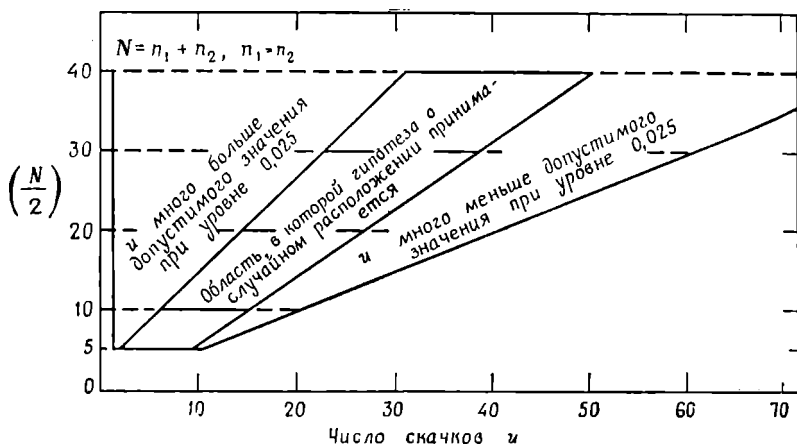


Фиг. 14.7.

возрастания и снижения значений в группах элементов. Предположим, что в этой последовательности нет двух точно совпадающих чисел и что различия между ними превышают аналитические погрешности. Теперь применим критерий, основанный на

¹ M. G. Kendall, The advanced theory of statistics, II, Chas. Griffin, London, 1948.

распределении числа скачков, к возрастающим последовательностям при уровне значимости 0,025. Начав с первого элемента, рассмотрим первые 10 членов последовательности. Для них характерны одно возрастание и одно понижение, отчетливо выражающие осязательную на глаз регулярность. Построим последовательность: — — + + + + — — —, для которой $n_1 = n_2 = 5$ и $u = 3$. При уровне значимости 0,025 мы вынуждены принять гипотезу о случайном расположении! Теперь рассмотрим подпоследовательность из 20 первых элементов. Как и в предыдущем случае, регулярность очевидна, но при $n_1 = n_2 = 10$ и $u = 5$ получается неопределенный результат, при котором

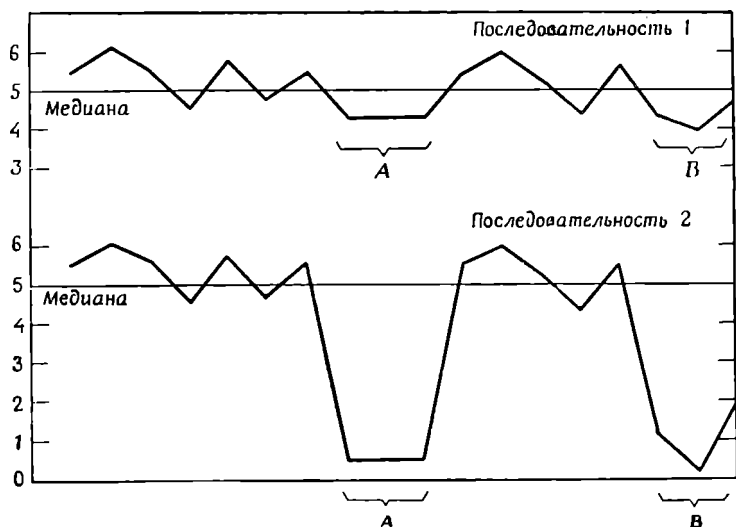


Ф и г. 14.8.

нельзя ни принять, ни отклонить гипотезу о случайном расположении. Однако при анализе подпоследовательности, начиная с 1-го элемента по 30-й, при $n_1 = n_2 = 15$, $u = 7$ и уровне значимости 0,025 оказалось, что проверяемую гипотезу следует отвергнуть, как и ожидали, основываясь на наблюдении. На фиг. 14.8 показана зависимость отчетливости проявления регулярности от длины рассматриваемой последовательности, что иллюстрирует зависимость мощности критерия от объема выборки. На ординате этого графика отложены значения $n_1 = n_2$, а на абсциссе — число скачков. Область, в которой принимается нулевая гипотеза, ограничена линиями, связывающими нижний предел числа скачков и верхний, с различными значениями $n_1 = n_2$ при уровне значимости 0,025. Нетрудно видеть, что при малых значениях $n_1 = n_2$ большая доля возможного числа скачков попадает в область, в которой нулевая гипотеза принимается. Однако с увеличением $n_1 = n_2$ уменьшается доля возмож-

ного числа скачков, соответствующая интервалу принятия гипотезы о случайном расположении.

Во многих случаях геолог может отказаться от окончательного решения по данным анализа первой подпоследовательности фиг. 14.7. С другой стороны, вычислив $P(u)$, он получит точный ответ, заключающийся в утверждении, что число способов расположения пяти плюсов и пяти минусов, дающих три скачка, как это выражено в первой подпоследовательности, велико по сравнению с общим числом всех возможных последовательностей из этих десяти элементов.



Ф и г. 14.9.

Тем не менее использование критерия скачков не вызывает затруднений; его следует рекомендовать, но геолог должен ясно понимать смысл критерия, быть осторожным в своих выводах при использовании критерия в условиях малых выборок, понимать ограничения критерия и располагать некоторыми ожидаемыми результатами, полученными *a priori*.

Интерпретация, характеризующая ограниченность критерия и связанная с величиной расхождения между элементами последовательности, вызывает некоторое затруднение. Допустим, что все различия, изображенные на фиг. 14.9, действительно существуют и не являются следствием ошибок эксперимента или округления.

В результате применения критерия скачков было установлено при $n_1=n_2=9$ и $u=10$, что первую последовательность следует рассматривать как случайную. Вторая последователь-

ность совпадает с первой в том смысле, что ее элементы (+) и (—) занимают те же места, что и в первой. Однако геолог, конечно, рассматривал бы участки *A* и *B* во второй последовательности как неожиданные и важные. В то же время вторая последовательность отличается от первой только изменением в порядке величины различий на участках *A* и *B*; относительные возрастание и понижение кривой остаются неизменными. Поэтому следует иметь в виду, что в подобных случаях при использовании критерия скачков не учитывается величина различий.

При выявлении тренда предлагается следующая процедура анализа:

1. Применить критерий числа скачков или критерий числа точек смены знака.

2. Если представляет интерес выявление линейного тренда, нужно применить ранговый коэффициент корреляции Спирмена или статистику Кендалла τ .

3. В том случае когда гипотеза о случайном расположении отвергается при выполнении пункта 1 или же отмечается высокое значение коэффициента корреляции r_s , можно воспользоваться кривой, подобранной по методу наименьших квадратов. Однако при выборе кривой лучше основываться на предположениях *a priori*, чем на изучении выборочных данных.

4. Пригодность выбранной кривой должна быть проверена, например, с помощью критерия χ^2 .

5. После этого можно воспользоваться критерием скачков для анализа отклонений от подобранной кривой.

Временной ряд — это последовательность наблюдений, упорядоченных во времени. Так как этими рядами давно интересовались ученые, работающие в различных областях, по этому вопросу существует обширная литература и разработаны многочисленные способы анализа.

В предлагаемом обзоре авторы для простоты делят временные ряды на два класса.

класс I

Каждому элементу ряда приписано значение времени, отсчитанного от некоторого начального момента t_0 . Единица измерения времени в этом случае не обязательно должна быть фиксирована. Ее можно выразить как в миллионах лет, так и в микросекундах.

класс II

Точной (или надежной) единицы измерения времени нет. Однако такие серии наблюдений являются также «временными рядами», так как в них соблюдается осмысленное упорядочение во времени. Это значит, что, основываясь на последовательности номеров $1, 2, \dots, i, \dots, j, \dots$, мы говорим, что i -е наблюдение предшествует j -му. Однако измерить время, прошедшее между моментами получения i -го и j -го наблюдений, нельзя.

Очень большое число случаев, с которыми приходится иметь дело геологу, соответствует второму классу. Например, для стратиграфического разреза палеозойских отложений в настоящее время мы располагаем лишь сведениями об «относительных временных рядах». Можно измерить мощность слоев или других стратиграфических единиц и построить упорядоченную последовательность мощностей. Но тем не менее нельзя каждому измерению приписать хронологическое время. Похожая ситуация характерна для большинства палеонтологических задач. Все это

приводит к общему выводу, что большинство временных рядов исторической геологии представлено «относительными временными последовательностями». Существуют надежно проверенные методы определения абсолютного возраста, такие, например, как углеродный, калиево-аргоновый и ураново-свинцовый.

Основная трудность применения методов определения абсолютного возраста в статистическом анализе временных рядов заключается в выборе единиц измерения времени. Дело в том, что чем древнее порода, тем больше бывает ошибка в определении абсолютного возраста. Весьма важная сторона вопроса заключается в том, что выбор единицы измерения времени зависит от характера задачи, а эта единица, в свою очередь, не может быть меньше максимальной ошибки применяемого метода определения возраста.

При изучении временных рядов геолог может ставить перед собой задачу выявить цикличность и тренд, а также решить вопросы, связанные с предсказаниями. Термином «цикл» в геологическом смысле следует пользоваться осторожно. Например, «циклотема», или чередующаяся последовательность слоев морского и континентального происхождения¹, является столь же цикличной, как процесс регулярного увеличения и уменьшения интенсивности солнечных пятен или как колебание средней годовой температуры. Однако в случае циклотемы остаются неизвестными абсолютные хронологические даты, соответствующие отдельным наблюдениям. (Кроме того, здесь не рассматривается еще одна трудность, связанная с правильным расположением отдельных обнажений в хронологической последовательности.)

Рассмотрим другой пример тонкослоистых образований. Если можно предположить, что образование переслаивающихся светлых и темных прослоев обусловлено сезонными изменениями в процессе озерного осадконакопления, то каждую пару таких прослоев можно рассматривать как единицу отложений за год. Следовательно, задачу можно отнести к классу I. Если же тонкослоистая серия пород представляет нерегулярное чередование осадков, связанных то с периодами обильных дождей, то с засушливыми периодами, задачу следует отнести к классу II.

КРАТКИЙ ОБЗОР АНАЛИЗА ВРЕМЕННЫХ РЯДОВ

Временные ряды детально описаны Кендаллом². Он предположил, что наблюдения производятся через равные интервалы

¹ J. M. Weller, *Cyclical sedimentation of the Pennsylvanian period and its significance*, Journ. Geol., 38, 1930.

² M. G. Kendall, *The advanced theory of statistics*, II, New York, 1948.

времени. Материал, излагаемый ниже, взятый в основном из данных Кендалла, знакомит читателя с некоторыми из способов классического анализа временных рядов.

Обычно временные ряды разлагают на три части: тренд, циклы и «случайная» компонента. Отдельный ряд может включать как все три, так и одну или две из этих компонент.

Введем некоторые определения.

Сезонным эффектом называется регулярная изменчивость, выраженная как чередование повышений и понижений значений и возникающая вследствие действия некоторых природных циклических факторов, не зависящих от изучаемой величины. Яркими примерами проявления этого эффекта могут служить циклы годовых колебаний температуры, ежедневные циклы приливов и отливов, а также цикличность в процессе размножения животных. Как Кендалл¹, так и Уоллис и Робертс² предупреждают, что выделить и устранить этот эффект можно лишь при наличии надежных сведений *a priori*. Кендалл предлагает исключать действие этого эффекта путем выбора соответствующего интервала времени. Именно таким образом за счет подбора 12-часового периода наблюдений был устранен эффект действия прилива при изучении уровня воды в Кейп-Коде. К этому вопросу мы вернемся ниже.

Термин *тренд* можно понимать двояко. Кендалл полагает, что плавное, постепенное изменение в течение длительного периода времени (по отношению к длине серии) служит показателем тренда. Выражение «по отношению к длине серии» в данном случае можно рассматривать как ключевое. Это определение оказывается нужным в тех случаях, когда ряд задан до начала исследования, как это бывает при изучении стратиграфических разрезов или при изучении изменений ископаемых организмов во времени. Иными словами, Кендалл предлагает, чтобы мы противопоставляли плавное изменение резким скачкам, а не тренд колебаниям.

Тренд можно охарактеризовать путем сглаживания и подбора к имеющимся данным соответствующего полинома. Если ряд очень изменчив, это приводит к полиному высокой степени, подбор которого требует большой затраты труда. Для более объективного выбора полинома нередко используется сглаживание с помощью скользящих средних. Пример применения такого сглаживания будет приведен ниже.

Устранение тренда производится путем вычитания соответствующих значений полинома, описывающего тренд, из наблюдаемых значений временного ряда.

¹ См. сноску 2 на стр. 354.

² W. A. Wallis, H. V. Roberts, Statistics, a new approach, The Free Press, Illinois, 1956.

Стационарные временные ряды характеризуются отсутствием тренда, которое или типично для исходных данных, или же вызвано исключением его из них. Если установлено, что временной ряд стационарен, можно перейти к изучению циклических эффектов. Для описания регулярных повышений и понижений значений элементов последовательности широко используется гармонический анализ Фурье, описывающий эти колебания с помощью суммы циклических компонент¹. Примером применения гармонического анализа в геологических задачах может служить работа Вистелиуса [5].

Очень распространенным свойством большинства временных рядов является так называемая *автокорреляция*. Вместо корреляции в обычном смысле, когда изменения функции зависят от изменений аргумента, здесь рассматривается корреляция между членами ряда. Поясним это на примере скорости движения взвешенной частицы в направленном потоке. Если в данный момент времени скорость превышает среднее значение, то вполне правдоподобно предположить, что в предыдущий и последующий моменты она также будет превышать среднее. В этом весьма практическом смысле последовательно расположенные элементы временного ряда не могут быть независимыми. Естественно, что автокорреляция несколько осложняет анализ временных рядов, и в то же время ею нельзя пренебрегать.

Силу автокорреляционной зависимости во временном ряду можно охарактеризовать путем вычисления коэффициентов автокорреляции согласно описанию Кендаллом², что является весьма трудоемкой процедурой (см. также работу Поупа³).

УТИЛИТАРНАЯ ТОЧКА ЗРЕНИЯ

Авторы считают нужным отметить, что точка зрения, излагаемая Уоллисом и Робертсом⁴, больше отражает их личное мнение относительно анализа временных рядов в геологических задачах, чем объективное мнение специалиста. Временные ряды следует изучать с большой осторожностью. Попытка выделения циклов или тренда при помощи различных аналитических приемов, но без модели, построенной *a priori* на естественных предположениях, очень опасна. При предлагаемых ниже процедурах рекомендуется помнить, что последним судьей исследователя должно быть индуктивное восприятие им предмета. Ре-

¹ N. Wiener, Generalised harmonic analysis, Acta Math., 55, 1930; R. A. Fisher, Tests of significance in harmonic analysis, Proc. Roy. Soc. A, 125, 1929.

² См. сноску 2 на стр. 354.

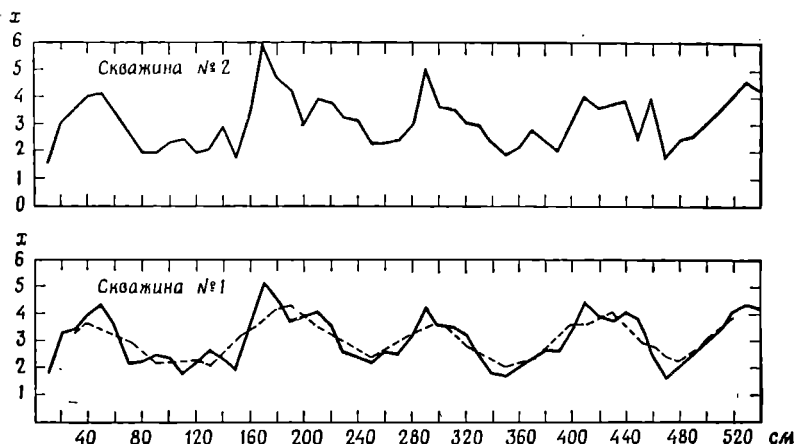
³ J. A. Pore, Errors in the estimation of serial correlations, Nature, 172, 1953.

⁴ См. сноску 2 на стр. 355.

зультаты применения этого набора критериев следует использовать как предпосылки для выбора направления дальнейших исследований природной структуры изучаемого конкретного временного ряда.

ПРИМЕНЕНИЕ

Считая, что любой вид формального анализа временных рядов в геологии следует предпринимать лишь в случае, когда физическая сущность изучаемой последовательности максимально ясна, авторы решили привести гипотетические примеры.



Фиг. 15.1. Изменение переменной x в зависимости от глубины скважин. Глубина рассматривается как эквивалент времени.

Пунктир — сглаживание скользящей средней (интервал 5 единиц).

Предположим, что в центральной части океанского бассейна на расстоянии несколько сотен миль одна от другой были пробурены две глубинные морские скважины № 1 и 2 и что поблизости нет ни островов, ни материков; это предположение позволяет допустить, что осадконакопление происходит в спокойных постоянных условиях в течение того периода, которому соответствует глубина скважин. Таким образом, мы обосновали замену единицы времени единицей глубины скважины. Результаты измерения некоторого свойства осадков (через каждые 10 см) по скважине № 1 приведены в табл. 15.1, в которой глубину в сантиметрах следует понимать как эквивалент времени. Изменения значений признака в обеих скважинах показаны на фиг. 15.1. Заметим, что этот временной ряд рассматривается без учета тренда, и по принятой терминологии его следует считать стационарной временной последовательностью.

Таблица 15.1

Глубина сква- жины, см	x	Знак отклонения от медианы (2,95)	Скользящая средняя по пяти числам		Сезонная циклич- ность; сглажива- ние интервалов в четыре единицы		Уклонение наблюдаемого значения от скользящей средней	Временной ряд, исправ- ленный за счет сезонной цикличности
			сумма	среднее	сумма	среднее		
1	2	3	4	5	6	7	8	9
10	1,9	—						2,31
20	3,4	+						3,41
30	3,5	+	17,2	3,44	12,8	3,20	0,30	3,13
40	4,0	+	18,9	3,78	15,3	3,82	0,18	4,02
50	4,4	+	17,7	3,54	15,5	3,87	0,53	4,81
60	3,6	+	16,5	3,30	17,2	3,55	0,05	3,61
70	2,2	—	15,0	3,00	12,5	3,12	—0,92	1,83
80	2,3	—	13,0	2,60	10,6	2,65	—0,35	2,32
90	2,5	—	11,2	2,24	9,4	2,35	0,15	2,91
100	2,4	—	11,1	2,22	9,0	2,25	0,15	2,41
110	1,8	—	11,5	2,30	8,8	2,20	—0,40	1,43
120	2,1	—	11,4	2,28	9,0	2,25	—0,15	2,12
130	2,7	—	10,9	2,18	9,0	2,25	0,45	3,11
140	2,4	—	12,7	2,54	9,1	2,27	0,13	2,41
150	1,9	—	15,9	3,18	10,6	2,65	—0,75	1,53
160	3,6	+	17,7	3,54	13,2	3,30	0,30	3,62
170	5,3	+	19,1	3,82	15,3	3,82	1,48	5,71
180	4,5	+	21,1	4,22	17,2	4,30	0,20	4,51
190	3,8	+	21,6	4,32	17,5	4,37	—0,57	3,43
200	3,9	+	19,9	3,98	16,3	4,07	—0,17	3,92
210	4,1	+	17,8	3,56	15,4	3,85	0,25	4,51
220	3,6	+	16,4	3,28	14,0	3,50	0,10	3,61
230	2,4	—	14,7	2,94	12,5	3,12	—0,72	2,03
240	2,4	—	13,2	2,64	10,6	2,65	—0,25	2,42
250	2,2	—	12,1	2,42	9,6	2,40	—0,20	2,61
260	2,6	—	12,8	2,56	9,7	2,42	0,18	2,61
270	2,5	—	14,7	2,94	10,4	2,60	—0,10	2,13
280	3,1	+	16,1	3,22	12,5	3,12	—0,02	3,12
290	4,3	+	17,0	3,40	13,5	3,37	0,93	4,71
300	3,6	+	17,8	3,56	14,5	3,62	—0,02	3,61
310	3,5	+	17,1	3,42	14,7	3,67	—0,17	3,13
320	3,3	+	14,6	2,92	12,8	3,20	0,10	3,32
330	2,4	—	12,7	2,54	11,0	2,75	—0,35	2,81
340	1,8	—	11,2	2,24	9,2	2,30	—0,50	1,81
350	1,7	—	10,5	2,10	7,9	1,97	—0,27	1,33
360	2,0	—	10,7	2,14	8,1	2,02	—0,02	2,02
370	2,6	—	11,4	2,28	8,9	2,22	0,38	3,01
380	2,6	—	13,0	2,60	9,7	2,42	0,18	2,61

Глубина скважины, см	x	Знак отклонения от медианы (2,95)	Скользящая средняя по пяти числам		Сезонная цикличность; сглаживание интервалом в четыре единицы		Уклонение наблюдаемого значения от скользящей средней	Временной ряд, исправленный за счет сезонной цикличности
			сумма	среднее	сумма	среднее		
1	2	3	4	5	6	7	8	9
390	2,5	—	15,4	3,08	11,0	2,75	—0,25	2,13
400	3,3	+	17,7	3,54	12,8	3,20	0,10	3,32
410	4,4	+	17,8	3,56	14,1	3,52	0,88	4,81
420	3,9	+	19,3	3,86	15,3	3,82	0,08	3,91
430	3,7	+	19,8	3,96	16,0	4,00	—0,30	3,33
440	4,0	+	17,8	3,56	15,4	3,85	0,15	4,02
450	3,8	+	15,5	3,10	13,9	3,47	0,33	4,21
460	2,4	—	13,8	2,76	11,8	2,95	—0,55	2,41
470	1,6	—	12,2	2,44	9,8	2,45	—0,85	1,23
480	2,0	—	11,2	2,24	8,4	2,10	—0,10	2,02
490	2,4	—	12,1	2,42	8,8	2,20	0,20	2,81
500	2,8	—	14,6	2,92	10,5	2,62	0,18	2,81
510	3,3	+	16,9	3,38	12,6	3,15	0,15	2,93
520	4,1	+	18,7	3,74	14,5	3,62	0,48	4,12
530	4,3	+			15,9	3,97	0,33	4,71
540	4,2	+						4,21

$$r = 10; \quad n_1^+ = n_2^- = 27$$

Проверка гипотезы о случайном расположении элементов ряда

Если основываться на внешнем облике последовательности значений ряда, легко впасть в заблуждение, утверждая, что в ней проявлены тренд или цикличность. Визуальное изучение временных рядов не дает надежной основы для выводов. Первый шаг в анализе ряда представляет проверку нулевой гипотезы, заключающейся в том, что увеличение и уменьшение значений признака имеют случайный характер. Эту гипотезу легко проверить с помощью критерия, основанного на распределении числа скачков, который был рассмотрен в предыдущей главе.

Применяемая здесь при анализе скважины № 1 форма критерия заимствована из работы Уоллиса и Робертса. Спрашивается, можно ли признать наблюдаемое число скачков в последовательности плюсов и минусов возможным для случайного расположения? Для преобразования последовательности значений признака, приведенной в табл. 15.1, в ряд, состоящий из

(+) и (—), использовалась меднапа. В результате при $n_1 = n_2 = 27$ число скачков оказалось равным 10. Далее была вычислена величина

$$K = \frac{(r - 1/2) - M_r}{\sigma_r},$$

которая при условии, что нулевая гипотеза верна, распределена нормально с параметрами 0, 1. В предыдущем выражении

$$M_r = (2n_1 n_2 / n) + 1$$

и

$$\sigma_r = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n)}{n^2 (n - 1)}},$$

где $n = n_1 + n_2$, или

$$K = \frac{n \left(r - \frac{1}{2} \right) - 2n_1 n_2 - n}{\sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n)}{n - 1}}}.$$

Для скважины № 1

$$K = \frac{54 \cdot 9,5 - 1458 - 54}{\sqrt{38 \cdot 623}} = -5,05.$$

Так как вычисленное значение K очень велико по абсолютной величине, гипотезу о случайном расположении элементов последовательности следует отвергнуть, а изучение особенностей ряда продолжить. Необходимо отметить, что сильно выраженная автокорреляция приводит к большой вероятности отрицания нулевой гипотезы, проверяемой с помощью критерия скачков.

Если расположение элементов ряда не «случайно», то или следует продолжить его изучение путем сглаживания на конечном этапе, или же, если сложившаяся ситуация позволяет, перейти к дальнейшему анализу, т. е., исключив эффект, вызванный сезонными флуктуациями, проверить гипотезу о цикличности, а затем о корреляции, как было отмечено выше.

Сглаживание скользящей средней

Сглаживание временных рядов скорее следует рассматривать как описательный прием, а не аналитический. Скользящая средняя не только маскирует пики и минимумы последовательности, которые могут оказаться важными в дальнейших исследованиях, но и обладает свойством сдвигать эти точки от первоначального положения (фиг. 15.1). Скользящая средняя для данной точки ряда образуется путем обобщения результатов

нескольких предшествующих и последующих точек. Сумма всех значений в этих точках называется *скользящей суммой*. Таким образом, скользящая средняя по пяти точкам содержит значение, соответствующее точке, которой приписана средняя, а также величины в двух предшествующих и двух последующих точках. Затем весь интервал из пяти точек сдвигается вниз по столбцу на один шаг и процесс повторяется. Например, в табл. 15.1, в столбце 4, сумма для точки 3 образована значениями в точках 1, 2, 3, 4, 5; для точки 4 — значениями, соответствующими 2, 3, 4, 5, 6, и т. д. Затем результаты, помещенные в столбце скользящих сумм табл. 15.1, преобразуют в скользящие средние, приведенные в столбце 5, путем деления каждой суммы на 5. Эти результаты нанесены на график фиг. 15.1.

Устранение результатов воздействия сезонных флуктуаций

Сезонные флуктуации можно рассматривать как циклические явления, которые отчетливо выражены в виде распространенной в природных ситуациях регулярности. Результаты их воздействия следует ожидать в отдельных рядах наблюдений, подвергающихся исследованию. Таким образом, во многих случаях годовые климатические циклы могут влиять на временные ряды. Однако под термином *сезонный* имеется в виду цикличность, не только связанная с годовым периодом. Например, и 12-часовой цикл океанских приливов или некоторые биологические циклы относятся к этой же категории цикличности.

Допустим, что в рассматриваемом примере имеется достаточно оснований (*a priori*) для включения в цикл четырех элементов ряда. Это значит, что в терминах глубины скважины (эквивалент времени) этот период охватывает четыре точки по длине ряда.

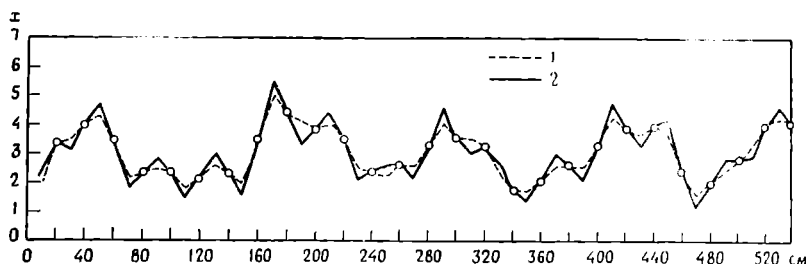
Уоллис и Робертс (1956 г.) разработали два метода, которые можно использовать для «исправления» временного ряда за счет эффекта, вызванного сезонными флуктуациями. В обоих методах предусмотрена поправка, которая вводится в каждый элемент цикла путем усреднения значений этого элемента в повторяющихся циклах. Приведенный ниже метод разностей лучше всего использовать в тех случаях, когда сезонные эффекты рассматриваются как не зависящие от значений элементов изучаемой последовательности, т. е. поправка будет одной и той же для всех элементов независимо от того, какое значение принял элемент.

Так как предполагается, что сезонный цикл состоит из четырех элементов, первый шаг процедуры будет заключаться в вычислении скользящих средних по значениям в четырех точках. В табл. 15.1 скользящая сумма приведена в столбце 6 и первое ее значение расположено против третьего элемента ряда,

**Вычисление средних единичных уклонений для цикла
в четыре единицы**

Глубина скважины	Уклонение	Глубина	Уклонение	Глубина	Уклонение	Глубина	Уклонение
1		2		3		4	
30	0,30	40	0,18	50	0,53	60	0,05
70	-0,92	80	-0,35	90	0,15	100	0,15
110	-0,40	120	-0,15	130	0,45	140	0,13
150	-0,75	160	0,30	170	1,48	180	0,20
190	-0,57	200	-0,17	210	0,25	220	0,10
230	-0,72	240	-0,25	250	-0,20	260	0,18
270	-0,10	280	-0,02	290	0,93	300	-0,02
310	-0,17	320	0,10	330	-0,35	340	-0,50
350	-0,27	360	-0,02	370	0,38	380	0,18
390	-0,25	400	0,10	410	0,88	420	0,08
430	-0,30	440	0,15	450	0,33	460	-0,55
470	-0,85	480	-0,10	490	0,20	500	0,18
510	0,15	520	0,48	530	0,33	540	
Сумма	-4,85		+0,25		+5,36		+0,18
Среднее	-0,373		0,019		0,412		0,014
	-0,37		0,02		0,41		0,01

а в столбце 7 дано соответствующее значение скользящей средней. В столбце 8 приведены разности элементов ряда и вычисленных значений скользящей средней. В табл. 15.2 сгруппированы подобные элементы (в смысле цикла), которым противопоставлены соответствующие отклонения от скользящей сред-



Фиг. 15.2. Исключение сезонной цикличности из исходных данных временного ряда по скважине.

1 — исходный временной ряд; 2 — после устранения влияния сезонной цикличности.

пей. В нижней части табл. 15.2 указаны средние значения сезонных уклонений для каждого из четырех элементов цикла.

Поправки на цикличность для каждого из четырех элементов цикла следующие: $-0,37$ для первого элемента цикла, $+0,02$ — для второго, $+0,41$ — для третьего и $+0,01$ — для последнего. Затем эти поправки были введены в исходную серию данных. Например, в элемент последовательности, соответствующий глубине скважины 30 см , вводится исправление: $3,50 - 0,37 = 3,13$. Полученный «исправленный» ряд значений приведен на фиг. 15.2.

Второй метод сходен с первым, но лишь в пределах первого этапа работы; в дальнейшем, согласно второму методу, надо использовать отношение скользящей средней к исходному значению ряда.

«Метод отношения скользящих средних» особенно рекомендуется, когда сезонный показатель зависит от значений элементов ряда.

Проверка гипотезы о цикличности после введения поправки на сезонные воздействия

Анализируя данные фиг. 15.2 и учитывая информацию *a priori*, можно предположить повторение цикла через интервал 120 см с максимумами, расположенными около точек 50 , 170 , 290 , 410 и 530 см . Проверим справедливость этого предположения для изучаемого ряда с помощью метода, предложенного Мак-Юэном¹, который суммировал подобные элементы последовательных циклов с последующим изучением полученных средних. В нашем иллюстративном примере гипотетический цикл содержит 12 элементов. На фиг. 15.2 первый гипотетический цикл, начавшийся с пика в точке 50 см , заканчивается в точке 160 см . Второй цикл заключен в интервале от 170 до 280 см и т. д. Например, второй элемент первого цикла равен $3,61$ и соответствует точке 60 см , тогда как второй элемент второго цикла равен $4,51$ и находится в точке 180 см . Эти данные приведены в табл. 15.3.

Изучая средние, мы видим, что цикл характеризуется единственным повышением с последующим снижением значений, после которого появляется второй пониженный пик. Заканчивается цикл подъемом. Мак-Юэн предложил проверить реальность этой регулярности с помощью дисперсионного анализа. Для проверки нулевой гипотезы, заключающейся в том, что средние можно рассматривать как равные, воспользуемся однофакторным

¹ G. F. McEwen, The reality of regularities indicated in sequences of observations, Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability, Univ. of Calif. Press, 1949.

Временные ряды

Цикл	Элементы циклов											
	1 ¹	2	3	4	5	6	7	8	9	10	11	12
1	4,81	3,61	1,83	2,32	2,91	2,41	1,43	2,12	3,11	2,41	1,53	3,62
2	5,71	4,51	3,43	3,92	4,51	3,61	2,03	2,42	2,61	2,61	2,13	3,12
3	4,71	3,61	3,13	3,32	2,81	1,81	1,33	2,02	3,01	2,61	2,13	3,32
4	4,81	3,91	3,33	4,02	4,21	2,41	1,23	2,02	2,81	2,81	2,93	4,12
Сумма	20,04	15,64	11,72	13,58	14,44	10,24	6,02	8,58	11,54	10,44	8,72	14,18
Среднее	5,01	3,91	2,93	3,40	3,61	2,56	1,51	2,15	2,89	2,61	2,18	3,55

¹ Начиная с 50 и кончая 530 см, чтобы можно было полностью охватить гипотетические циклы.

дисперсионным анализом, описанным в гл. 7 (табл. 7.12). Элементы последовательности, сгруппированные по циклам,

Таблица 15.4

Дисперсионный анализ

Тип дисперсии	Число степеней свободы	Сумма квадратов	Оценка дисперсии	Проверка гипотезы: все 12 средних равны
Между средними . . .	$12 - 1 = 11$	39,025	3,55	$F''_{36} = \frac{3,55}{0,30} = 11,8^2$
Внутри групп	$12(4 - 1) = 36$	10,708	0,30	
Общая дисперсия . . .	47	49,733	1,57	
Между средними . . .	$\frac{(20,04)^2 + \dots + (14,18)^2}{4} - \frac{(145,14)^2}{48} =$			
	$= 477,892 - 438,867 = 39,025$			
Внутри групп	Разность между нижней и верхней строками = $= 49,733 - 39,025$			
Общая дисперсия . . .	$(4,81)^2 + \dots + (4,12)^2 - \frac{(145,14)^2}{48} = 49,733$			

¹ Вычисления проведены по схеме, приведенной в табл. 7.18.

² При уровне значимости 0,05 допустимое значение $F''_{36, (0,05)} \approx 2,9$.

приведены в табл. 15.3, а результаты дисперсионного анализа — в табл. 15.4; из данных этих таблиц следует, что гипотезу о равенстве средних нужно отвергнуть, т. е. между средними существуют «реальные» различия. Однако это неравенство не означает, что установлена цикличность. Действительно, на глаз видно проявление вторичных пиков. Результаты же проведенного исследования свидетельствуют о том, что цикл из 12 элементов возможен и что необходимо дальнейшее изучение ряда в соответствии с природной ситуацией. Два последующих этапа изучения заключаются в следующем: первый — повторный анализ данных применительно к более короткому циклу; второй — повторное исследование природной ситуации. Можно воспользоваться более детальной выборочной схемой, сгустив сеть опробования на изучаемом участке, или же использовать данные, взятые с соседних участков.

Ситуация, при которой можно применить критерий для проверки гипотезы о цикличности

В южной Калифорнии обнажается формация мощностью 3000 м, представленная слоистыми тонкопесчанистыми сланцами, содержащими обильную фауну планктонных ископаемых фораминифер. Средняя мощность отдельных прослоев 15 см. (фиг. 15.3).

По простиранию отдельного слоя невозможно установить наличия тренда ни в размерах частиц, ни в распространенности планктонных фораминифер. В истории образования формации представляет определенный интерес фактор скорости осаджения осадочного материала. Было установлено, что в качестве единицы осадконакопления можно рассматривать отдельный прослой, так как каждый из них образовался в однотипных условиях.

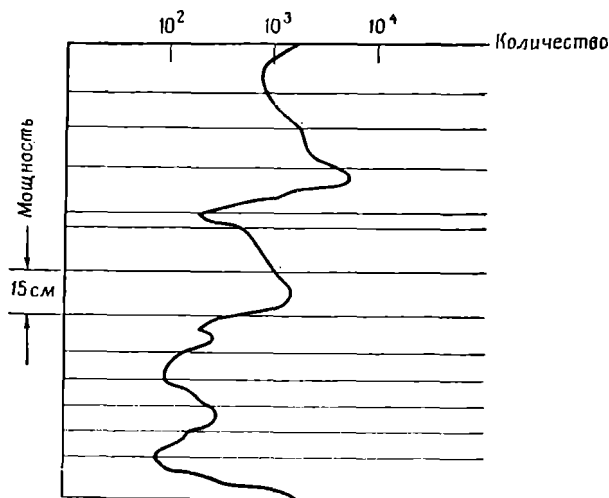
В связи с тем что в интерпретации процесса осадкообразования возникают разногласия, предлагается проверить две гипотезы о скорости осадконакопления.

1. Слои формировались в условиях глубоководного морского бассейна при вертикальном равномерном осаджении частиц на дно. Общепринятая приблизительная оценка (основанная на общих соображениях) скорости осадконакопления равна 20 см за 1000 лет. Таким образом, оценка времени формирования всей толщи оказывается очень большой.

2. Относительная распространенность планктонных фораминифер по слоям свидетельствует о возможном наличии циклических явлений. Согласно экологическим данным, можно допустить существование двух- или трехлетнего цикла, выраженного в увеличении популяции с последующим ее уменьшением. Так как пики распространенности фораминифер приурочены

приблизительно к каждому третьему слою, следовательно, скорость осадконакопления в первом случае значительно занижена.

Очевидно, здесь была бы уместна проверка гипотезы о наличии годового цикла длиной в несколько лет. Кроме того, при



Фиг. 15.3. График количественной распространенности планктонных фораминифер по стратиграфическим единицам.

Вертикальная ось — последовательность стратиграфических единиц (самая поздняя из них расположена сверху); горизонтальная ось — количество фораминифер, приходящееся на стратиграфическую единицу.

отрицательном ответе, проверив гипотезу об отсутствии тренда при альтернативе случайных флуктуаций, можно получить ценные сведения.

КОРРЕЛЯЦИЯ МЕЖДУ ВРЕМЕННЫМИ РЯДАМИ

Проверить гипотезу о том, что два временных ряда независимы, можно с помощью критерия, предложенного Барнеби¹. Обозначим через $E_{r, r+1}$ произведение перекрестных разностей элементов двух временных рядов:

$$E_{r, r+1} = (p_r - q_{r+1})(q_r - p_{r+1}), \quad r < r+1,$$

где $p_1, p_2, \dots, p_r, \dots$ и $q_1, q_2, \dots, q_r, \dots$ — два сравниваемых временных ряда. Если соответствующие элементы p и q отли-

¹ Т. Р. Burnaby, A suggested alternative to the correlation coefficient for testing the significance of agreement between pairs of time series and its application to geological data, Nature, 172, 1953.

чаются друг от друга меньше, чем от соседних членов **своих** рядов, то среднее $E = \frac{1}{n} (E_{1,2} + E_{3,4} + \dots)$ будет положительной величиной. Читатель может проверить этот способ на гипотетической серии. На практике для умеренно больших выборок обычно вводится упрощение, заключающееся в подсчитывании числа положительных, отрицательных и нулевых значений $E_{r,r+1}$. Кроме того, можно также воспользоваться критерием χ^2 . Согласно нулевой гипотезе для пары временных рядов, ряды независимы, и поэтому число плюсов и минусов одинаково [гл. 6, уравнения (6.10) и (6.11)]. Этот ускоренный метод применяется для сравнения двух рядов значений. Ниже приведен пример, иллюстрирующий этот метод, используемый для сравнения проанализированной выше первой скважины со второй, пробуренной в том же морском бассейне на расстоянии нескольких сотен километров от первой.

Результаты использования критерия χ^2 приведены в табл. 15.5. Заметим, что при обработке данных удобно пользоваться столбцами 4 и 5: в первом записываются знаки перекрестных разностей, а во втором — знак произведения.

Авторы обнаружили, что при обработке результатов больших временных рядов этим методом ошибка уменьшается.

В результате применения χ^2 установлено, что сходство между двумя рядами неслучайно. Интерес представляет метод Барнеби при анализе тонкослоистых толщ: он приводит к более общему классу задач. В рассмотренном примере с двумя скважинами пары заранее определены, так как скважины охарактеризованы одинаковым числом наблюдений. Теперь рассмотрим случай, когда число элементов в них неодинаково. Задача заключается в таком их сопоставлении, чтобы полученное расположение отличалось от случайного. Как пример подобной ситуации можно привести две тонкослоистые толщи, не имеющие заранее заданных сопоставимых точек. Признаками, которые рассматриваются при изучении этих рядов, могут служить мощность, средний размер частиц, содержание редкого элемента и т. п. Естественно, что сопоставление можно сначала провести визуально, а затем проверить его результаты с помощью описанного критерия. Однако если заданный при проверке нулевой гипотезы уровень значимости равен P_n , то истинный уровень значимости P_t будет равен произведению P_n и числа N возможных способов сопоставления рядов, т. е.

$$P_t = NP_n.$$

Например, толща A из 200 слоев сравнивается с толщиной B , в которой насчитывается 74 слоя. Существует 127 возможных способов сопоставления толщи B с толщиной A . Вообще если

Глубина скважины, см	Скв. № 1	Скв. № 2	Перекрестные разности ¹	Знак произведения	Глубина скважины, см	Скв. № 1	Скв. № 2	Перекрестные разности ¹	Знак произведения
1	2	3	4	5	1	2	3	4	5
10	1,9	1,7	—	+	290	4,3	5,1	+	+
20	3,4	3,2	—	+	300	3,6	3,7	+	+
30	3,5	3,6	—	+	310	3,5	3,6	+	+
40	4,0	4,1	—	+	320	3,3	3,1	+	+
50	4,4	4,2	+	+	330	2,4	3,0	+	+
60	3,6	3,5	+	+	340	1,8	2,3	+	+
70	2,2	2,0	+	—	350	1,7	1,9	—	+
80	2,3	2,0	+	—	360	2,0	2,1	—	+
90	2,5	2,4	0	0	370	2,6	2,8	+	+
100	2,4	2,5	0	0	380	2,6	2,4	+	+
110	1,8	2,0	—	+	390	2,5	2,0	—	+
120	2,1	2,1	—	+	400	3,3	3,0	—	+
130	2,7	3,0	—	—	410	4,4	4,0	+	+
140	2,4	3,1	—	—	420	3,9	3,6	+	+
150	1,9	1,8	—	+	430	3,7	3,7	—	+
160	3,6	3,5	—	+	440	4,0	3,8	—	+
170	5,3	6,0	+	+	450	3,8	2,5	—	—
180	4,5	4,8	+	+	460	2,4	3,9	—	—
190	3,8	4,4	+	+	470	1,6	1,8	—	+
200	3,9	3,0	+	+	480	2,0	2,4	—	+
210	4,1	4,0	+	+	490	2,4	2,5	—	+
220	3,6	3,9	+	+	500	2,8	2,9	—	+
230	2,4	3,3	—	—	510	3,3	3,5	—	+
240	2,4	3,2	—	—	520	4,1	4,0	—	+
250	2,2	2,3	—	+	530	4,4	4,6	+	+
260	2,6	2,3	—	+	540	4,2	4,3	+	+
270	2,5	2,4	—	+					
280	3,1	3,0	—	+					

¹ Перекрестная разность равна $p_r - q_{r+1}$, или $q_r - p_{r+1}$, где q_r — значение с номером r для скв. № 2, а p_r — соответствующее значение для скв. № 1;

$$\chi^2 = \frac{(22-13)^2}{13} + \frac{(4-13)^2}{13} = 12,46;$$

$\chi^2_{0,995} = 7,88$ при одной степени свободы.

число элементов большей серии обозначить через N_t , а меньшей — N_s , то уровень значимости можно определить как

$$P_t = [(N_t - N_s) + 1] P_n.$$

При использовании упомянутого критерия имеется в виду предположение, согласно которому два сравниваемых временных ряда обладают приблизительно равными средними и дисперсиями. Если этого нет, то Барнеби предлагает предварительно производить некоторое линейное преобразование.

ВЫВОДЫ

Нетрудно видеть, что применение методов анализа временных рядов в геологических задачах — область совершенно неизученная. За исключением работы Барнеби о корреляции двух временных рядов и статей Вистелиуса и Корна, посвященных исследованию тонкослоистых толщ, авторам неизвестны связанные с этим вопросом никакие другие публикации. Это относится также и к другим разделам теории стохастических процессов (например, работы Бартлета¹, Дуба² и Феллера³).

Возможно, что большинство геологических явлений, связанных с временными интервалами, измеряемыми миллионами лет, следует рассматривать скорее не как периодические, а как эпизодические; это значит, что явление повторяется или его интенсивность меняется, но не в смысле регулярности периода между пиками.

С этих позиций утверждение о том, что процессы горообразования и основные оледенения регулярны во времени, т. е. имеют периодический характер, можно рассматривать в совершенно ином аспекте. Нам же нужен только правильный выбор аналитической комбинации, такой, например, как некоторые виды анализа временных рядов, чтобы можно было количественно описать основные, зависящие от времени геологические последовательности.

¹ M. S. Bartlett, *Stochastic Processes*, Cambridge Univ. Press, London, 1955.

² J. L. Doob, *Stochastic Processes*, New York, 1953.

³ W. Feller, *An introduction to probability theory and its applications*, 1, New York, 1950.

Эта глава посвящена в основном обзору некоторых способов проверки гипотезы о случайном расположении точек на плоскости, хотя в ней затрагиваются и некоторые вопросы распределения точек в пространстве и на прямой. При этом рассматриваются две альтернативы: 1) распределение равномерно, когда точки обладают тенденцией к правильному расположению относительно друг друга, и 2) распределения неоднородны со скоплениями, когда точки на плоскости характеризуются тенденцией к группировке. Так как в этой главе используется термин *случайное распределение точек на плоскости*, нужно дать его определение. Особое внимание, уделяемое термину *плоскость*, вызвано тем, что такая размерность широко используется в ряде геологических задач.

В задачах о распределении точек на плоскости слово *случайный* используется в следующем смысле¹:

1) при случайном распределении точек на плоскости все точки имеют одну и ту же вероятность занять любое положение на этой плоскости;

2) некоторый малый участок плоскости характеризуется такой же вероятностью содержать точки, как и любой другой участок такого же размера;

3) положение данной точки на плоскости не влияет на положение всех остальных точек.

Следует обратить особое внимание на границы изучаемой области, так как вполне возможно, что данная область, характеризующаяся случайным расположением точек, заключена в пределах большей области, в которой распределение точек неслучайно.

В этой главе описаны модели наиболее распространенных ситуаций, а также способы построения этих моделей, которые помогают их понять. Прежде чем перейти к описанию моделей,

¹ P. J. Clark, F. C. Evans, Distance to nearest neighbor, as a measure of spatial relationships in populations, *Ecology*, 35, 1954.

дадим определения некоторых вспомогательных терминов, которые будут применяться в дальнейшем изложении.

Картис и Макинтош¹ детально рассмотрели меры, применяемые при анализе распределения точек на плоскости в экологических исследованиях. Некоторые из этих мер можно использовать при решении геологических задач. Из всего списка определений, приведенных в указанной работе Картиса и Макинтоша, нас интересуют следующие.

1. Плотность $D = \frac{\text{Общее число точек}}{\text{Общая площадь}}.$

Таким образом, D — число точек, приходящееся на единицу площади.

2. Средняя площадь, равная $1/D$.

3. Квадрат — выборочная единица. Картис и Макинтош¹ рассмотрели также задачу о размере квадрата при изучении данного района. Они пришли к выводу, согласно которому оптимальный квадрат должен иметь площадь, равную удвоенной средней площади для данного района. Таким образом, размеры этого квадрата зависят как от площади района, так и от числа точек. Эту зависимость можно выразить в следующем виде:

$$\text{Площадь оптимального квадрата} = \frac{2 \cdot (\text{общая площадь})}{\text{Общее число точек}}.$$

Это выражение будет использовано в конце главы в конкретной задаче.

ПУАССОНОВСКАЯ МОДЕЛЬ

Распределение Пуассона было описано в гл. 2. Теперь мы рассмотрим модель, которая приводит к этому распределению и широко используется при изучении природных явлений². Несмотря на то что авторы, описывая эту модель на примере интервалов прямой линии, имеют в виду одномерный случай, принцип подхода к задаче можно распространить и на большее число измерений.

Предположим, что отрезок прямой разделен на n равных интервалов длиной $1/n$ каждый. Пусть на этот отрезок наудачу брошено некоторое число точек. Тогда каждый из интервалов, на которые разбит отрезок, или содержит точку, или не содержит. Если интервалы сделать достаточно малыми, вероятность

¹ J. T. Curtis, R. P. McIntosh, The interrelations of certain analytic and synthetic phytosociological characters, *Ecology*. 31, 1950.

² W. Feller, An Introduction to probability theory and its applications, 1950.

нахождения нескольких точек в одном интервале станет ничтожной. Обозначим вероятность того, что в одном интервале содержится точка, через p_n и допустим, что эта вероятность не влияет на вероятности нахождения или отсутствия точки в других интервалах. Заметим также, что, так как все интервалы равны, равны и соответствующие им вероятности p_n . Описанные условия допускают применение биномиальной модели, и, следовательно, вероятность того, что в отрезок, состоящий из n интервалов, попадет точно k точек, будет равна значению биномиальной функции $b(k; n, p_n)$. Если $np_n \rightarrow \lambda$, то $b(k; n, p_n)$ можно записать как выражение $b(k; n, \lambda/n)$, которое приближается к распределению Пуассона $p(k; \lambda)$. Теперь можно показать, что вероятность события, заключающегося в том, что на отрезке, состоящем из n интервалов, находится точно k точек, будет определяться выражением $p(k; \lambda) = e^{-\lambda} (\lambda^k / k!)$; $k=0, 1, \dots, n$. Положим, что длина нашего отрезка равна 1, и рассмотрим t таких отрезков. Если каждый из них разделен на n интервалов, общее число интервалов будет nt , а длина нового большого

отрезка определится как $\sum_1^{nt} \frac{1}{n} = t$. Параметр λ , который характеризует среднюю плотность точек для нового отрезка, нужно умножить на t . Следовательно, вероятность того, что на отрезке длиной t есть точно k точек, равна

$$p(k; \lambda t) = \frac{e^{-\lambda t} (\lambda t)^k}{k!}. \quad (16.1)$$

В этой модели (16.1) предусмотрены два основных предположения:

1. Если интервал мал, вероятность нахождения в нем более одной точки мала по сравнению с длиной интервала.

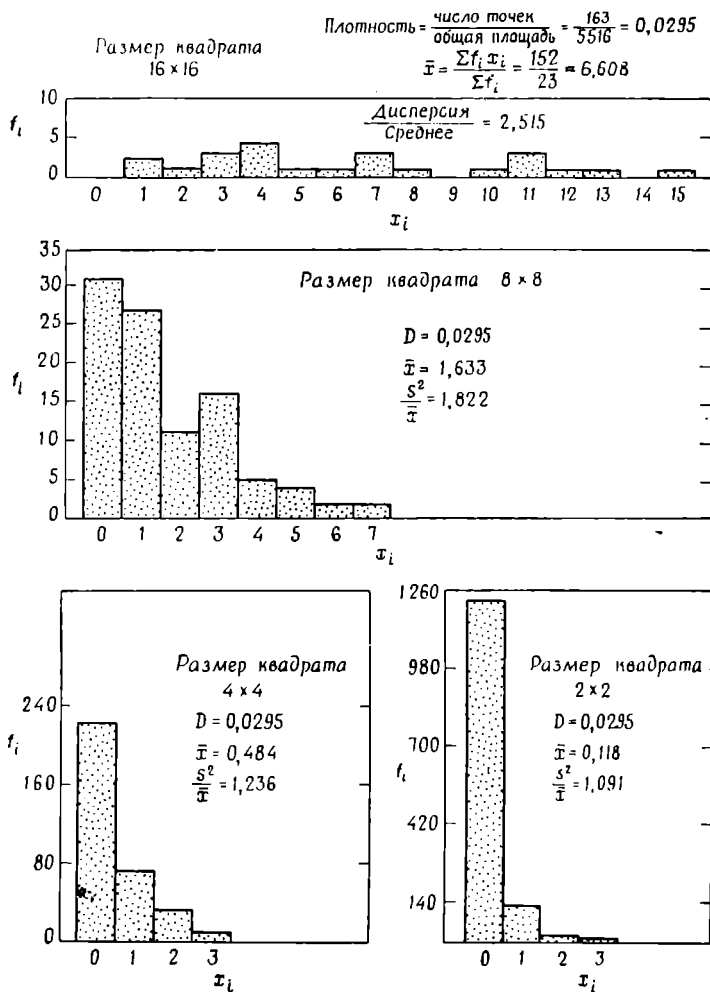
2. Интервалы взаимно независимы: число точек в одном интервале нельзя определить по информации о любом другом интервале.

Для практических приложений этой модели необходимо еще показать, что параметр λt можно приближенно выразить через величину T/N , где T — общее число точек в N отдельных интервалах, каждый из которых имеет фиксированную длину t . Подсчитаем число точек, приходящихся на каждый из интервалов. Процедура этого подсчета показана на фиг. 16.1. Обозначим через N_i число интервалов длиной t , содержащих по i точек, и пусть T — число всех точек. Тогда

$$0N_0 + 1N_1 + 2N_2 + \dots + iN_i + \dots + kN_k = T, \quad (16.1a)$$

Проверка пригодности распределения Пуассона

Для проверки гипотезы о согласованности распределения Пуассона и реальных данных Клаппем¹ воспользовался отноше-



Фиг. 16.2. Влияние размера квадрата на плотность, среднее и отношение дисперсии к среднему по данным распределения шипов *Lystrocanthus* (фиг. 16.6).

нием μ_2/μ_1' , где $\mu_2 = \sigma^2$ (дисперсия) и $\mu_1' = \mu$ (среднее). Известно (гл. 3), что в распределении Пуассона отношение μ_2/μ_1'

¹ A. R. Clapham, Over dispersion in grassland communities and the use of statistical methods in plant ecology, Journ. Ecology, 24, 1936.

равно единице. Если это отношение превышает единицу, точки (например, на плоскости) «рассеяны сильнее», чем при случайном расположении. Если отношение μ_2/μ_1' меньше единицы, это свидетельствует о тенденции к групповому расположению точек. Так как практически обычно исследуют не всю совокупность, а выборку, то вместо параметров μ_2 и μ_1' используют их оценки. В этом случае оценкой для отношения μ_2/μ_1' будет служить величина

$$\frac{\sum_{i=0}^k (x_i' - \bar{x})^2}{\bar{x} (n - 1)},$$

где n — общее число точек.

К сожалению, это отношение, очень удобное для вычислений, зависит от размера квадрата (выборочной единицы). Эванс¹ исследовал зависимость формы кривой распределения от площади квадрата. При увеличении размера последнего соответственно возрастает вероятность нахождения в нем большого числа точек и, наоборот, уменьшается вероятность того, что квадрат окажется пустым.

Таким образом, увеличение размера квадратов ведет к уменьшению числа тех, в которых совсем нет точек или есть лишь одна точка. Однако в связи с тем, что общее число квадратов также уменьшается, относительная частота таких квадратов возрастает при уменьшении относительной частоты тех, в которых содержится большое число точек (фиг. 16.2).

Показатель рассеяния

Для проверки гипотезы о распределении Пуассона Фишер [95] предложил метод, основанный на так называемом показателе рассеяния. Эта статистика при условии, что проверяемая гипотеза верна, распределена как χ^2 с $N - 1$ степенями свободы. Она дается выражением

$$\sum_{i=0}^k \frac{[f_i (x_i - \bar{x})]^2}{\bar{x}},$$

где $x_i = 0, 1, 2, \dots, k$; f_i — число квадратов, содержащих x_i точек; $N = \sum f_i$, а $\bar{x} = \frac{1}{N} \sum f_i x_i$. Этот показатель можно запи-

¹ F. C. Evans. The influence of size of quadrat on the distributional patterns of plant populations, Contr. Lab. Verb. Biol., № 54, Michigan, 1952.

сать в следующей, более удобной для вычислений форме:

$$\frac{\sum f_i x_i^2 - \bar{x} \sum f_i x_i}{x}.$$

В более поздней работе Фишер¹ говорит о трудностях, возникающих в тех случаях, когда значения x малы, а число классов невелико. Тогда отклонения от распределения Пуассона могут быть вызваны значительной долей тех классов, которые имеют теоретические значения частот меньше 5. Вследствие этого обстоятельства результаты применения показателя рассеяния вызывают сомнение. Хотя Фишер и не считает эту трудность серьезной, он все же предлагает необходимые изменения для малой численности классов.

Проверка гипотезы о распределении Пуассона с помощью обычного критерия

Применяя метод χ^2 для проверки гипотезы о непротиворечивости распределения Пуассона выборочным данным, когда теоретические значения частот в краевых классах распределения значительно меньше 5, геолог встречает те же трудности. Как отмечает Фишер, объединяя в таких случаях несколько интервалов, мы понижаем чувствительность критерия; вывод о пригодности теоретического распределения может быть сделан вследствие пренебрежения индивидуальными значениями частот в классах.

Техника проверки гипотезы заключается в следующем. Так как распределение Пуассона дается выражением

$$p(x) = e^{-\lambda} \frac{\lambda^x}{x!},$$

параметр λ оценивается с помощью среднего арифметического $\bar{x}$. Подставив в это выражение вместо λ значение $\bar{x}$, получим оценку для $p(x)$, которая равна

$$e^{-\bar{x}} \frac{\bar{x}^x}{x!};$$

умножив эту величину на объем выборки n , получим теоретическое значение частоты $n(\bar{x})^x / e^{\bar{x}x}$ для заданного x . Процедура вычисления упростится, если это отношение прологарифмировать:

$$\log \left\{ \frac{[n(\bar{x})^x]}{e^{\bar{x}x}} \right\} = (\log n + x \log \bar{x}) - [\log(x!) + \bar{x} \log e]$$

(табл. 16.3).

¹ R. A. Fisher, Significance of deviations from expectation in a Poisson series, Biometrics, 6, 1950.

В ряде природных ситуаций модель Пуассона нельзя применить непосредственно. В тех случаях, когда требование независимости точек или выборочных единиц (таких, как единичный квадрат) не выполняется, распределение Пуассона не подходит в качестве модели *a priori*. Например, групповое расположение точек, возникающее благодаря взаимодействию, привело к возникновению термина *эффект заражения*. Для вероятностного описания подобных случаев предложено много моделей, из которых две, нередко используемые в экологии растений, представляются пригодными для решения геологических задач.

Распределение заражения типа А

Эту модель предложил в 1939 г. Нейман¹. Термин *заражение* был использован для описания класса распределений, для которых характерна тенденция образовывать естественные скопления или группы точек². Согласно этой модели, вероятность того, что в выборочном квадрате окажется ровно x точек, например $x=2$, дается выражением

$$P(x=2) = e^{-m_1(1-e^{-m_2})} \frac{m_2^2}{2!} (m_1^2 e^{-2m_2} + m_1 e^{-m_2}),$$

где m_1 и m_2 — параметры распределения, из которых первый характеризует среднее число групп в выборочной единице, или квадрате, а второй — среднее число точек в группе. В 1952 г. Томсон³ описал метод применения этой модели в эмпирических задачах. Для вычисления различных характеристик в уравнении Неймана он пользуется полиномами. В 1949 г. Шентон⁴ предложил другое решение, основанное на методе максимального правдоподобия.

«Двойная пуассоновская» модель Томаса

Рассмотрим более сложный случай применения модели Пуассона. Допустим, что в некоторой области случайно распределено множество точек. Предположим также, что каждая из них связана с другими точками, число которых представляет собой случайную величину, распределенную по закону Пуассона.

¹ J. Neyman, On a new class of contagious distributions applicable in entomology and bacteriology, Ann. math. Stats, X, 1939.

² G. Polya, Sur quelques points de la théorie des probabilités, Ann. de l'Institut Henri Poincaré, 1, 1931.

³ G. W. Thomson, Measures of plant aggregation based on contagious distribution, Contr. Lab. Vert. Bil., Michigan, № 53, 1952.

⁴ L. R. Shenton, On the efficiency of the method of moments and Neymans Type A distribution, Biometrika, 36, 1949.

Таким образом, мы должны рассмотреть два взаимодействующих распределения точек. Разделим эту площадь на небольшие квадраты. Для вычисления вероятности того, что данный квадрат содержит $k=0, 1, 2, \dots$ точек, Томас¹ предложил выраже-

ние P (в квадрате k точек) $= \sum_{r=1}^k \frac{m^r e^{-m} (r\lambda)^{k-r} e^{-r\lambda}}{r! (k-r)!}$, где m — пара-

метр, характеризующий среднее число групп в квадрате, а λ — среднее число растений в группе, уменьшенное на единицу. Позднее Томсон (1952) предложил метод применения модели «двойных пуассоновских» рядов к реальным задачам. Как и для распределения типа А по Нейману, для упрощения вычислений он использовал полиномы.

Согласно результатам, полученным Скиллемом², модели этого типа не рекомендуются в тех случаях, когда группы некомпактны или когда процесс настолько растянут во времени, что появляются последующие дополнительные точки. Это обстоятельство несколько ограничивает возможность применения указанных моделей в геологических задачах, связанных с изучением изменений во времени.

В этой же работе Скиллем предложил модель, в которой учтены случаи некомпактности групп. Квадраты в этой модели предполагаются сравнительно малыми. Кроме того, Скиллем рассмотрел модель для описания процессов во времени.

Ивенс³ привел пример, когда, несмотря на равенство среднего и дисперсии, распределение Пуассона непригодно в качестве модели эмпирического распределения. Томсон, поясняя этот пример, отмечает, что, хотя сильное отклонение отношения дисперсии к среднему от единицы и указывает на отличие расположения точек от случайного, обратное заключение необоснованно. Это значит, что из близости этого отношения к единице не следует с необходимостью пуассоновский характер распределения.

Плотность точек не зависит от размера выборочного квадрата, если она выражена числом точек, приходящихся на единицу площади. Однако, если эта величина представлена как число точек на квадрат, она пропорциональна площади этого квадрата.

Вероятностное распределение числа точек в квадратах также зависит от площади квадрата (ср. текст и фиг. 16.2). В связи с этим Ивенс для изучения степени отклонения распределения

¹ M. Thomas, A generalization of Poisson's binomial limit for use in ecology, *Biometrika*, 36, 1949.

² J. G. Skellam, Studies in statistical ecology, I, Spatial pattern, *Biometrika*, 39, Pts. 3, 4, 1952.

³ P. C. Evans, The influence of size of quadrat on the distributional patterns of plant populations, *Contr. Lab. Vert. Biol., Michigan*, № 54, 1952.

точек от случайного предложил проводить повторный анализ одной и той же области, меняя размер квадратов. В разработанном им методе на график наносятся значения плотности точек, соответствующих различным размерам квадратов в зависимости от частоты встречаемости. Затем кривая, описывающая полученные точки, визуальнo сравнивается с теоретической случайной кривой, значения которой F вычисляются с помощью выражения

$$F = 100(1 - e^{-D}),$$

где D — плотность¹. Хотя авторы и находят этот способ приемлемым, они считают необходимым отметить, что он весьма трудоемок и что в конечном итоге он дает скорее качественные, чем количественные ответы.

ПРИМЕРЫ ПРИМЕНЕНИЯ НЕКОТОРЫХ КРИТЕРИЕВ ПРОВЕРКИ ГИПОТЕЗЫ О РАСПРЕДЕЛЕНИИ ПУАССОНА

Приведенные ниже примеры содержат данные изучения глинистых сланцев Мекки пенсильванского возраста².

По выборке, содержащей 163 плавниковых шипа *Lystrocanthus*, с площади 5×6 м была проверена гипотеза о согласованности распределения этих шипов с законом Пуассона. Данные, приведенные в табл. 16.1, собраны с квадратов оптимальной площади, длина сторон которых вычислена методом Картиса и Макинтоша. Эти же данные были использованы и для иллюстрации других критериев.

В табл. 16.2 приведены результаты проверки гипотезы о согласованности выборочного распределения числа шипов на квадрат с законом Пуассона. Порядок вычисления показан в табл. 16.3. При вычислении значения критерия последние три класса объединены в один (табл. 16.2). Так как после объеди-

Таблица 16.1
Распределение числа шипов
на квадрат

Число шипов в квадрате x_i	Число квадратов f_i
0	31
1	27
2	11
3	16
4	5
5	4
6	2
7	2
$\Sigma f_i = 98$	

¹ A. R. Clapham, Over-dispersion in grassland communities and the use of statistical methods in plant ecology, Journ. Ecology, 24, 1936.

² R. L. Miller, Speculation on water currents in a black shale environment, by use of orientation and dispersion of fossil fragments, Trans. Am. Geophys. Union, 37, 1956.

нения трех последних классов их стало шесть, число степеней свободы критерия стало равным четырем ($6 - 2 = 4$). Допустимое значение χ^2 при уровне значимости 0,005 и 4 степенях свободы равно 14,86, а вычисленное значение $\chi^2 = 38,10$, следовательно, гипотезу о согласованности распределения шипов с законом Пуассона следует отвергнуть.

Т а б л и ц а 16.2

Результаты проверки гипотезы о распределении Пуассона с помощью критерия χ^2

Классы — число точек, наблюдаемых в квадрате, x_i	Наблюдаемые частоты — число квадратов, содержащих x_i точек, f_i	Теоретические частоты — вычисления, основанные на распределении Пуассона, n_i	$f_i - n_i$	$\frac{(f_i - n_i)^2}{n_i}$
0	31	18,58	12,42	8,30
1	27	30,83	—3,83	0,48
2	11	25,70	—14,70	8,41
3	16	14,24	1,76	0,22
4	5	5,92	—0,92	0,14
5	4	1,97	7,38	20,55
6	2	0,55		
7	2	0,13		
Число степеней свободы равно 4	$\sum_{i=0}^7 f_i = 98$	97,92		$\chi^2 = 38,10$

Примечание. Вычисления в этой таблице проведены по форме, данной в табл. 16.3.
 $\bar{x} = \frac{1}{\sum f_i} \sum x_i f_i = \frac{163}{98} = 1,663$, $n = \sum f_i = 98$, $k = 6$; после объединения классов число степеней свободы равно $k - 2 = 4$.

В данном случае распределение Пуассона соответствует случайному расположению точек на плоскости, поэтому в связи с отрицательным результатом проверки можно сделать вывод о неслучайном расположении шипов *Lystrocanthus*. Раз критерий не дает указаний на возможную альтернативную интерпретацию, то, прежде чем перейти к такой интерпретации, нужно вновь изучить данные табл. 16.1 с помощью двух других описанных выше критериев.

Применение «показателя рассеяния» Фишера

По формуле

$$I_D = \frac{1}{x} \left[\sum_{i=0}^k f_i x_i^2 - \bar{x} \sum_{i=0}^k f_i x_i \right].$$

Таблица 16.3

	Наблюдаемая частота	Логарифм ожидаемой частоты	$\begin{pmatrix} \text{Наблюдаемая} \\ \text{частота} \end{pmatrix} - \begin{pmatrix} \text{Ожидаемая} \\ \text{частота} \end{pmatrix}$	$\frac{[\begin{pmatrix} \text{Наблюдаемая} \\ \text{частота} \end{pmatrix} - \begin{pmatrix} \text{Ожидаемая} \\ \text{частота} \end{pmatrix}]^2}{\text{Ожидаемая частота}}$
0	n_0	$\log n - x \log e$		
1	n_1	$\log n_1 + 1 (\log x) - \log (1!) + x \log e$		
2	n_2	$\log n_2 + 2 (\log x) - \log (2!) + x \log e$		
3	n_3	$\log n_3 + 3 (\log x) - \log (3!) + x \log e$		
.	.			
.	.			
.	.			
i	n_i			
.	.			
.	.			
.	.			
k	n_k			
k — число классов величин	$\sum_{i=1}^N n_i = n = \sum f_i$			$\chi^2 = \sum \frac{[\begin{pmatrix} \text{Наблюдаемая} \\ \text{частота} \end{pmatrix} - \begin{pmatrix} \text{Ожидаемая} \\ \text{частота} \end{pmatrix}]^2}{\text{Ожидаемая частота}}$

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

Примечание. Вообще число степеней свободы равно числу классов минус единица и минус число наложенных линейных связей. В данном случае единственное линейное ограничение связано с оценкой $\bar{x}$. Таким образом, число степеней свободы равно $k - 1 - 1 = k - 2$ (гл. 6).

где $l=0, 1, \dots, k$, используя данные табл. 16.1, вычислим значение I_D :

$$\sum f_l x_l = 163, \quad \sum f_l x_l^2 = 565, \quad \bar{x} = \frac{\sum_{l=0}^k f_l x_l}{\sum_{l=0}^k f_l} = \frac{163}{98} = 1,663;$$

следовательно,

$$I_D = \frac{565 - 1,663 \cdot 163}{1,663} = 176,75.$$

Если распределение близко к закону Пуассона, то величина I_D распределена как χ^2 с $\left(\sum_{l=0}^k f_l\right) - 1$ степенями свободы. Так как вычисленная величина $I_D = 176,75$ сильно превышает допустимую при уровне значимости 0,005 и 97 степенях свободы, равную 125,0, гипотезу о распределении Пуассона нужно отвергнуть и, следовательно, также отвергнуть гипотезу о случайном расположении шипов. Заметим, что при рассмотрении показателя рассеяния Фишера альтернативная гипотеза не затрагивалась.

Применение отношения Клапема

Характерная особенность распределения Пуассона заключается в том, что величина отношения среднего к дисперсии равна единице.

Следовательно, критерий для проверки гипотезы о законе Пуассона, основанный на величине отношения выборочного среднего к выборочной дисперсии, представляется весьма желательным. Результат применения этого критерия допускает высказывания относительно альтернативной гипотезы. Если отношение среднего к дисперсии превышает единицу, можно сделать грубый вывод о влиянии группового эффекта на расположение точек; если же это отношение меньше единицы, в расположении точек существует тенденция к «взаимному отталкиванию».

1. Допустим, что дисперсия мала по сравнению со средним. Значит, число точек в исследуемых квадратах близко к среднему, и, следовательно, точки равномерно распределены на плоскости. В этом случае отношение меньше единицы.

2. Допустим, что дисперсия больше среднего. Значит, число точек в отдельных квадратах сильно колеблется, т. е. точки «стремятся» к группировке. В этом случае отношение превышает единицу.

Вычислим выборочное значение этого отношения по данным табл. 16.1. Подставив их в формулу

$$\frac{\sum_{i=1}^k (x_i - \bar{x})^2}{\bar{x}(n-1)} = \frac{\sum f_t x_t^2 - \bar{x} \sum f_t x_t}{\bar{x} (\sum f_t - 1)},$$

где $\bar{x} = 1,663$, найдем, что это отношение равно 1,822. Согласно полученному результату, гипотеза о случайном расположении точек отвергается, а альтернатива о наличии группового эффекта (а не взаимного отталкивания) принимается.

Однако, как уже было отмечено, на результаты влияет размер квадрата, принятого в качестве выборочной единицы. На фиг. 16.2 показаны формы распределений и приведены соответствующие им значения плотности, среднего и отношения среднего к дисперсии для четырех различных размеров квадрата. Эти размеры, выраженные в единицах площади, были следующими: 8×8 (оптимальный размер по Картису и Макинтошу¹), 16×16 , 4×4 и 2×2 . Ясно, что, если на размер квадрата не наложено никаких ограничений, значение отношения может быть как близким к единице (что соответствует случайному расположению точек), так и значительно ее превышать, что служит указанием на групповой эффект для одного и того же объекта исследования.

Вполне естествен вопрос: на много ли превышает единицу выборочное значение отношения и можно ли на основе этого отклонения принять альтернативную гипотезу? Ответ на этот вопрос можем получить, если используем свойство показателя Фишера; в условиях нулевой гипотезы он распределен по закону χ^2 .

МЕТОД «БЛИЖАЙШЕГО СОСЕДА»; ВЗАИМНАЯ СВЯЗЬ ТОЧЕК В ПРОСТРАНСТВЕННЫХ РАСПРЕДЕЛЕНИЯХ

Другой аналитический прием изучения пространственных распределений, позволяющий избежать трудности, связанные с размерами выборочного квадрата, основан на непосредственном рассмотрении пространственных зависимостей между точками. Метод, который будет описан в этом разделе, можно легко распространить на многомерные распределения.

¹ J. T. Curtis, R. P. McIntosh, The interrelations of certain analytic and synthetic phytosociological characters, *Ecology*, 31, 1950.

Вывод функции распределения расстояния до «ближайшего соседа»

Приведенный ниже вывод основан на результатах, полученных различными авторами, в частности Скиллемом¹, Муром² и Кларком и Ивенсом³.

Рассмотрим множество точек, фиксированных в некоторой принятой системе координат. В исследуемой области наудачу выбирается точка и измеряется расстояние между нею и ближайшей соседней точкой. Этот процесс повторяется много раз. Полученное в результате распределение измеренных расстояний характеризует пространственную зависимость между точками данной области. Пусть ρ — средняя плотность точек изучаемого распределения. Об определении этой величины говорилось выше, т. е. она представляет собой число точек на единицу площади, или

$$\frac{\text{Общее число точек}}{\text{Общая площадь}}.$$

Пусть r — расстояние от данной точки до ближайшей к ней точки в любых линейных единицах.

Если точки размещены на плоскости случайно, вероятность того, что выбранная наудачу единица площади содержит точно x точек в соответствии с распределением Пуассона, будет равна

$$P(x) = e^{-\lambda} \frac{(\lambda)^x}{x!}, \quad (16.2)$$

где λ — средняя плотность.

Допустим, что в качестве единичной площади рассматривается сектор круга радиусом r . Сектор выбран так, что он умещается в круге k раз.

Если ρ — число точек, приходящихся на единицу площади, или средняя плотность, то величина $\rho k^{-1} \pi r^2$ — средняя плотность сектора (единичная площадь). Эта величина совпадает с параметром λ . Подставив значение λ в выражение (16.2), получим

$$P(x) = e^{-\rho \pi k^{-1} r^2} \frac{(\rho \pi k^{-1} r^2)^x}{x!}.$$

Положив $x=0$, сможем подсчитать вероятность того, что сектор не будет содержать точек:

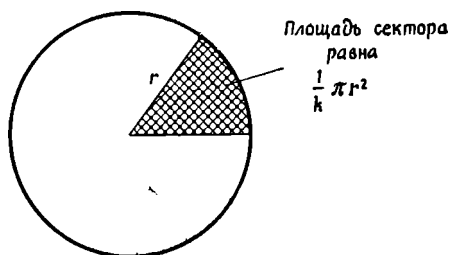
$$P(0) = e^{-\rho k^{-1} \pi r^2}.$$

¹ J. G. Skellam, Studies in statistical ecology, I, Spatial pattern, Biometrika, 39, Pts. 3, 4, 1952.

² P. J. Moore, Spacing in plant populations, Ecology, 35, 1954.

³ P. J. Clark, F. C. Evans, Distance to nearest neighbor as a measure of spatial relationship in populations, Ecology, 35, 1954; On some aspects of spatial pattern in biological populations, Sci., 121, 1955.

На фиг. 16.3 показано: вероятность того, что сектор не будет содержать точек на протяжении расстояния r , эквивалентна вероятности того, что сектор вообще не содержит точек. Таким образом, мы получили вероятность того, что расстояние от точки, помещенной в центр круга, до ближайшей соседней точки внутри сектора окажется большим или равным r . Кроме того, вероятность события, заключающегося в том, что сектор содержит одну или более точек, равна разности между единицей и вероятностью того, что сектор не содержит ни одной точки. В свою очередь, эта разность эквивалентна вероятности



Ф и г. 16.3.

того, что в секторе будет находиться точка, расположенная на расстоянии от центра круга, меньшем или равном радиусу r . Следовательно, мы получили долю расстояний до ближайшей соседней точки, меньших или равных r , во всем множестве этих расстояний. Это можно записать как

$$1 - e^{-\rho k^{-1} \pi r^2} = P(r);$$

плотность вероятности этого распределения примет вид

$$\frac{dP(r)}{dr} = e^{-\rho k^{-1} \pi r^2} \cdot 2\rho k^{-1} \pi r,$$

или

$$dP(r) = 2\rho k^{-1} \pi r e^{-\rho k^{-1} \pi r^2} dr.$$

Математическое ожидание изучаемой случайной величины определяется путем интегрирования

$$E(r) = \int_0^{\infty} 2\rho k^{-1} r^2 e^{-\rho k^{-1} \pi r^2} dr = \frac{\sqrt{k}}{2\sqrt{\rho}}. \quad (16.3)$$

Дисперсию можно найти из соотношения $E(r^2) - [E(r)]^2$, т. е.

$$E(r^2) = \int_0^{\infty} r^2 g(r) dr = \int_0^{\infty} 2\rho k^{-1} r^3 e^{-\rho k^{-1} \pi r^2} dr = \frac{k}{\rho \pi}.$$

Следовательно,

$$\text{var}(r) = \frac{k}{\rho\pi} - \frac{k}{4\rho} = \frac{(4-\pi)k}{4\rho\pi}. \quad (16.4)$$

Распространение понятия «ближайшая точка» на « n -ю ближайшую точку»; свойство отражения

Вполне естественно распространить понятие ближайшей точки к точке, которая выбрана наудачу, на последовательность ближайших точек, которым приписаны соответствующие номера. Например, можно говорить о второй ближайшей точке, о третьей и т. д. Данные, приведенные ниже, основаны на работе Кларка¹.

Рассмотрим плоскость, на которой распределены точки. Выберем одну из них и обозначим ее X_0 . Если X_0 — ближайшая точка к X_1 , а X_1 — ближайшая точка к X_0 , то они симметричны друг другу. Допустим, что точки, окружающие X_0 , рассматриваются в порядке их близости. Если X_1 — ближайшая точка к X_0 , то следующая ближайшая точка будет X_2 , третья — X_3 и т. д. Понятие симметрии в расположении точек теперь можно распространить на ближайшую точку с номером n . Это означает, что если X_n — ближайшая точка к X_0 , то, согласно определению, X_n будет ближе к X_0 , чем ко всем другим точкам, за исключением точек $X_1, X_2, \dots, X_{n-1}$, которые образуют последовательность ближайших к X_0 точек, упорядоченную по величине близости.

Для случая бесконечной совокупности точек, случайно расположенных на плоскости, Кларк составил таблицы теоретических значений вероятностей появления симметрично связанных точек.

Ниже показан вывод этих вероятностей для пар типа X_0, X_1 . Он поможет читателю понять смысл математической модели, положенной в основу критерия. Вывод заимствован из работы Кларка и Ивенса².

Допустим, что в совокупности точек со средней плотностью ρ рассматриваются две точки — X_0 и X_1 . Точки, образующие совокупность, расположены на плоскости, т. е. ситуация остается такой же, как и в рассмотренном выводе распределения расстояния до первой ближайшей точки.

Две выбранные точки можно рассматривать как центры, вокруг которых описаны окружности радиусом r (фиг. 16.4). Вероятность того, что X_1 — ближайшая точка к X_0 , равна вероятности предположения, что в пределах круга, описанного вокруг

¹ P. J. Clark, Grouping in spatial distributions, Sci., 123, 1956.

² P. J. Clark, F. C. Evans, On some aspects of spatial pattern in biological populations, Sci., 121, 1955.

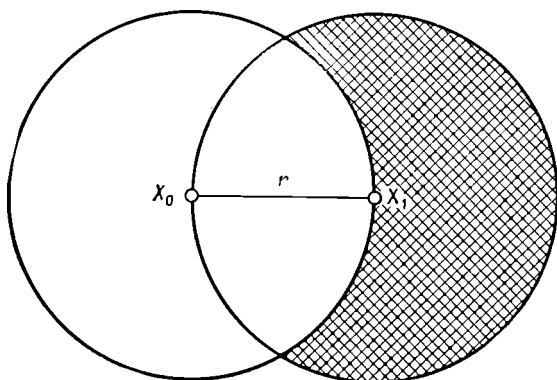
X_0 , нет других точек. Как отмечалось выше, вероятность того, что в секторе, размер которого $1/k$ от всего круга радиусом r , нет других точек, дается выражением

$$dP(r) = 2\pi k^{-1} r e^{-\rho k^{-1} \pi r^2} dr.$$

Для всего круга получим

$$dP(r) = 2\rho\pi r e^{-\rho\pi r^2} dr.$$

Если дано, что X_1 — ближайшая точка к X_0 , то вероятность того, что X_0 — также ближайшая точка к X_1 , равна вероятности события, заключающегося в том, что в той части круга



Фиг. 16.4.

(описанного вокруг X_1), не принадлежащей кругу, описанному вокруг X_0 , нет других точек. На фиг. 16.5 эта часть круга показана перекрестной штриховкой. Ее площадь, равную $r^2(\pi/3 + \sqrt{3}/2)$, можно определить следующим образом: судя по симметричной фиг. 16.5, площадь области с кривой штриховкой равна $1/4$ площади области пересечения двух кругов, поэтому искомая площадь будет равна

$$\pi r^2 - 4 \int_0^{r/2} \sqrt{r^2 - (x-r)^2} dx = r^2 \left(\frac{\pi}{3} + \frac{\sqrt{3}}{2} \right).$$

Так как величина ρ — средняя плотность района, параметр λ распределения Пуассона будет равен

$$\rho r^2 \left(\frac{\pi}{3} + \frac{\sqrt{3}}{2} \right).$$

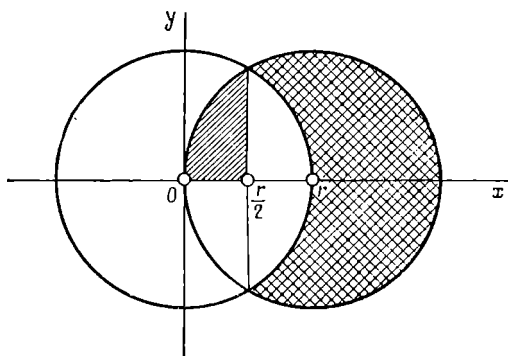
Подставив его в выражение

$$P(x) = e^{-\lambda} \frac{\lambda^x}{x!},$$

получим

$$P(0) = \exp \left[-\rho r^2 \left(\frac{\pi}{3} + \frac{\sqrt{3}}{2} \right) \right]. \quad (16.5)$$

Эта величина представляет вероятность того, что X_0 — ближайшая точка к X_1 при условии, что X_1 — ближайшая точка к X_0 и что она находится от X_0 на расстоянии r . Умножая на $dP(r)$



Ф и г. 16.5.

и интегрируя по r от 0 до ∞ , находим, что та доля P совокупности, для которой связь ближайших точек является симметричной, определяется следующим выражением:

$$\begin{aligned} P &= \int_0^{\infty} 2\rho\pi r e^{-\rho\pi r^2} \exp \left[-\rho r^2 \left(\frac{\pi}{3} + \frac{\sqrt{3}}{2} \right) \right] dr = \\ &= \int_0^{\infty} 2\rho\pi r \exp \left[-\rho r^2 \left(\frac{4\pi}{3} + \frac{\sqrt{3}}{2} \right) \right] dr = \\ &= \exp \left[-\rho r^2 \left(\frac{4\pi}{3} + \frac{\sqrt{3}}{2} \right) \right] \frac{\pi}{\frac{4\pi}{3} + \frac{\sqrt{3}}{2}} \bigg|_0^{\infty} = \frac{6\pi}{8\pi + 3\sqrt{3}}. \end{aligned}$$

Обобщенное выражение для распределения расстояний r до n -й ближайшей точки в k -мерном случае дал Кларк¹:

$${}_k P_n = \left[\frac{\frac{1}{2} \Gamma \left(\frac{k+1}{2} \right)}{\Gamma \left(\frac{k+1}{2} \right) - \pi^{-1/2} \Gamma \left(\frac{k}{2} + 1 \right) \int_{1/2}^1 (1-x^2)^{\frac{(k-1)}{2}} dx} \right]^n, \quad (16.6)$$

где k — размерность пространства, в котором расположены точки. Величина ${}_kP_n$ характеризует теоретическую долю точек в бесконечной совокупности со случайным распределением, для которых связь между X_0 и n -й ближайшей точкой симметрична. Подставив в выражение (16.6) $k=2$, получим значение этой доли для плоскости

$${}_2P_n = \left[\frac{\frac{1}{2} \Gamma\left(\frac{3}{2}\right)}{\Gamma\left(\frac{3}{2}\right) - \frac{\Gamma(2)}{\sqrt{\pi}} \int_{1/2}^1 (1-x^2)^{1/2} dx} \right]^n. \quad (16.7)$$

Так как

$$\Gamma\left(\frac{3}{2}\right) = \frac{1}{2} \sqrt{\pi}, \quad \Gamma(2) = 1 \quad \text{и} \quad \int_{1/2}^1 (1-x^2)^{1/2} dx = \frac{\pi}{6} - \frac{\sqrt{3}}{8}.$$

Т а б л и ц а 16.4

Вероятности появления точек, связанных с n -й ближайшей точкой обратной зависимостью в бесконечно большой совокупности

n -я ближайшая точка	На линии	На плоскости	В трехмерном объеме
1	0,6666	0,6215	0,5926
2	0,4444	0,3863	0,3512
3	0,2963	0,2401	0,2081
4	0,1975	0,1492	0,1233
5	0,1317	0,0927	0,0731
6	0,0878	0,0576	0,0433
7	0,0585	0,0358	0,0257
8	0,0390	0,0223	0,0152
9	0,0260	0,0138	0,0090
10	0,0173	0,0086	0,0053
11	0,0116	0,0053	0,0032
12	0,0077	0,0033	0,0019
13	0,0051	0,0021	0,0011
14	0,0034	0,0013	0,0007
15	0,0023	0,0008	0,0004
16	0,0015	0,0005	0,0002
17	0,0010	0,0003	0,0002
18	0,0007	0,0002	0,0001
19	0,0005	0,0001	0,0000
20	0,0003	0,0001	0,0000
21	0,0002	0,0000	0,0000

Примечание: В этой таблице значения для плоскости взяты из работы Кларка, а для линии и объема вычислены из уравнения (16.6), которое в одномерном случае сокращается до ${}_1P_n = (2/3)^n$, а в трехмерном — до ${}_3P_n = (16/27)^n$.

$${}_2P_n = \left[\frac{\frac{\sqrt{\pi}}{4}}{\frac{\sqrt{\pi}}{2} - \frac{1}{\sqrt{\pi}} \left(\frac{\pi}{6} - \frac{\sqrt{3}}{8} \right)} \right]^n = \left(\frac{6\pi}{8\pi + 3^{3/2}} \right)^n. \quad (16.8)$$

При $n=1$, т. е. для первой ближайшей точки, это выражение примет следующий вид:

$$\frac{6\pi}{8\pi + 3^{3/2}} = 0,6215;$$

при $n=2$, что соответствует второй ближайшей точке,

$$\frac{(6\pi)^2}{(8\pi + 3^{3/2})^2} = 0,3863 \text{ и т. д.}$$

Значения этих величин для одно-, двух- и трехмерного случаев бесконечной совокупности приведены в табл. 16.4. Этой таблицей можно пользоваться для изучения результатов малых выборок при установлении тенденций к группированию или, наоборот, рассеянию точек по сравнению со случайным расположением. В табл. 16.5 показана последовательность вычислительных операций определения этих пропорций по выборочным данным.

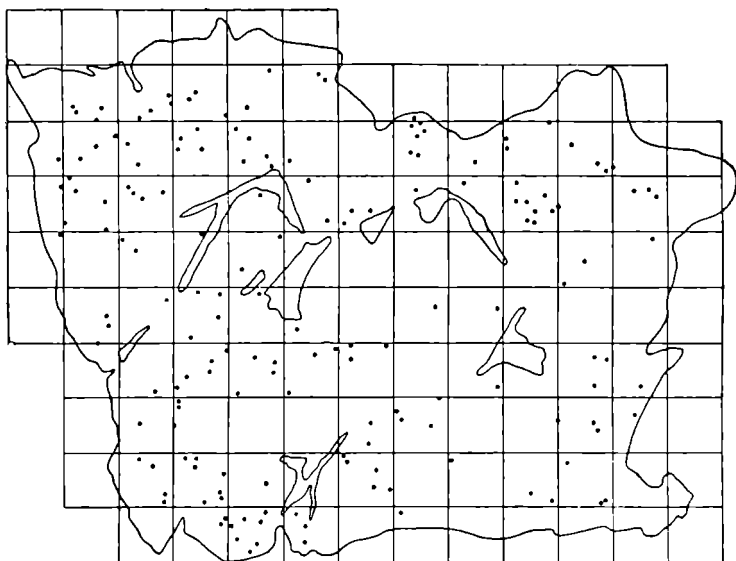
Если выборочное значение вероятности намного превышает ожидаемую для случайного расположения величину, это свиде-

Таблица 16.5

Вычислительная карта

Выбранная наудачу точка X_i	Ближайшая точка				
	X_1	X_2	X_3	...	X_n
Первая точка	Если X_0 и X_i связаны симметрично, то этот результат обозначается через r и в противном случае через нуль				
Вторая точка					
.					
.					
n -я точка					
Наблюдаемая доля в выборке объемом N	$\frac{\sum r_{X_1 X_0}}{N}$	$\frac{\sum r_{X_2 X_0}}{N}$	$\frac{\sum r_{X_3 X_0}}{N}$	...	$\frac{\sum r_{X_n X_0}}{N}$
Доля, ожидаемая при случайном расположении	0,6215	0,3863	0,2401	...	0,0000

тествует о наличии тенденции к группировке точек на плоскости. Если же выборочное значение вероятности сильно отличается от теоретического в меньшую сторону, из этого следует, что расположение точек на плоскости имеет тенденцию к взаимному отталкиванию. Форма для вычислений, приведенная



Фиг. 16.6. Распределение шипов *Lystrocanthus* на площади определенного стратиграфического уровня в сланцах Мекки пенсильванского возраста (показано точками).
Общая площадь, выраженная в произвольных единицах, составляет 5516. Картис и Макинтош определяют оптимальный размер квадрата как

$$\frac{2 \cdot (\text{общая площадь})}{\text{Общее число точек}} = \frac{2 \cdot 5516}{163} = 67,67$$

квадратных единиц, или приблизительно как квадрат 8×8 единиц. При увеличении размера квадрата возникают трудности. При большем размере квадраты, расположенные по краям контура, лишь частично покрывают область развития сланцев, тогда как установленное правило требует не учитывать в подсчете квадраты, для которых более 50% площади лежит за пределами контура. В результате теряется информация.

в табл. 16.5, остается такой же при любой размерности пространства, в котором расположены точки, за исключением того, что в основании таблицы следует привести теоретические значения вероятностей.

Свойство обратной зависимости между точками на плоскости было использовано для изучения распределения шипов *Listrocanthus* (фиг. 16.6) по данным, которые уже изучались

Таблица 16.6

Наудачу выбранная точка X_0	Ближайшая точка												
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}
1 точка	r	0	0	0	0	r	0	0	r	0	0	0	0
2 точка	r	r	0	0	r	0	0	0	0	r	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
89-я точка	0	0	0	r	0	0	0	0	0	0	0	0	0
Число r для i -й бли- жайшей точки	59	27	33	22	7	11	1	5	3	7	5	1	1
Наблюдаемая пропорция	0,663	0,303	0,370	0,247	0,078	0,121	0,011	0,056	0,034	0,078	0,056	0,011	0,011
Ожидаемая пропорция	0,621 +	0,386 —	0,240 +	0,149 +	0,093 —	0,058 +	0,036 —	0,022 +	0,014 +	0,009 +	0,005 +	0,003 +	0,002 +

(фиг. 16.2). Однако в этом случае учитывались различия в стратиграфическом положении. Результаты исследования приведены в табл. 16.6. Сравнение двух нижних строк таблицы осуществлялось следующим образом. Если выборочное значение вероятности превышало теоретическое, оно обозначалось знаком плюс, а если выборочное значение было меньше ожидаемого при случайном расположении точек — знаком минус. В результате было установлено, что в этом примере выборочные значения вероятностей сильно превышают теоретические, что свидетельствует о тенденции к группированию в распределении шипов *Listrocanthus*.

ВЫВОДЫ

Авторы считают, что модель, основанная на рассмотрении числа симметрично расположенных ближайших точек, и применяемый при этом критерий, предложенный Кларком, обладают важными преимуществами по сравнению с большинством других методов изучения распределения точек в пространстве. Эти преимущества сводятся к следующему.

1. Модель и критерий не зависят от размера единичного квадрата, так как вместо него используются расстояния между точками.

2. Предлагаются альтернативы к гипотезе случайности так же, как и в случае метода, в котором используется отношение среднего к дисперсии.

3. Возможность изучения тенденций к образованию групп точек заданного размера, таких, как пары, тройки и т. д.

Следует отметить и некоторые недостатки этой модели, например отсутствие единого критерия для сравнения выборочных значений вероятностей с теоретическими при проверке гипотезы о случайном расположении точек в пространстве и при альтернативных гипотезах о наличии эффекта группирования или, наоборот, взаимного рассеяния; кроме того, процедура измерения расстояний между точками более трудоемка, чем подсчет точек в квадратах.

Название и содержание этой главы несколько не соответствуют одно другому. Однако под этим названием можно объединить те работы, в которых в той или иной степени отражено применение статистических методов при картировании. В последнее время появилось много работ, в которых их авторы предлагают различные методы статистического анализа, используемые при решении задач, связанных с геологическим картированием. Многие из этих методов, снабженные соответствующими дополнениями, рассмотрены в настоящей главе. При этом основное внимание уделяется следующим двум направлениям, которые не исключают друг друга:

1. Различные количественные методы, с помощью которых составляются карты, что и является окончательным результатом.

2. Статистические методы, используемые как вспомогательное средство анализа и интерпретации карт. Следует отметить, что термин *карта* не ограничивается географическими координатами некоторой части земной поверхности, хотя очень многие приложения относятся именно к этой категории: его можно использовать более широко.

В нескольких разделах мы будем иметь дело с изометричными картами. Термин *изометрия* в буквальном смысле означает равенство мер соответствующих объектов. Этот термин используется в настоящей главе для обозначения карт, построенных по истинным значениям величины z , являющейся функцией координат x и y . Обычно предполагается, что z измерена без значительных погрешностей и что путем объединения в виде контуров равных значений z можно представить трехмерную поверхность. На практике для проведения контурных линий применяется интерполяция между наблюдаемыми значениями z . Основой для такой интерполяции служат значения z , измеренные в рассеянных точках. Это делает метод субъек-

тивным, так как окончательный результат картирования, представляемый в виде непрерывной поверхности, не является итогом интерполяции между точками, нередко распределенными нерегулярно.

Заметим, что координаты x и y необязательно должны быть координатами карты в географическом смысле. Например, Крамбейн [159] при изучении связи между мокрыми и сухими зернами сыпучих тел рассматривает поверхность тренда как функцию двух координат — влажности и среднего геометрического диаметров частиц.

МЕТОДЫ ГРАФИЧЕСКОГО И АРИФМЕТИЧЕСКОГО СГЛАЖИВАНИЯ

Метод профилей

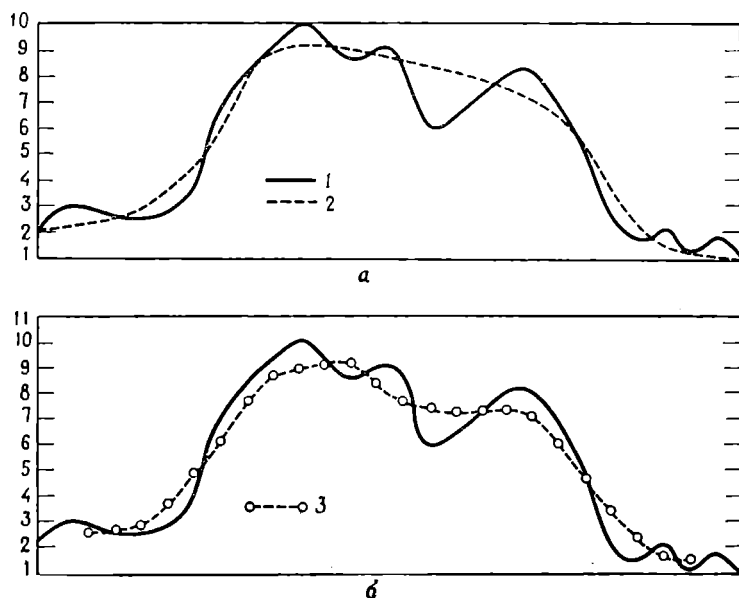
Этот графический способ сглаживания локальных колебаний в контурах на карте оставляет неустранимым только значительно выраженный тренд. Первый шаг этого метода заключается в нанесении на карту сетки линий, по которым будут строиться профили. После этого для каждой линии сетки, про-

Таблица 17.1

Построение сглаженного профиля с помощью скользящей средней по пяти точкам

Исходные данные	Σ	Скользя- щая средняя	Остаток	Исходные данные	Σ	Скользя- щая средняя	Остаток
1	2	3	4	1	2	3	4
2,0				8,6	38,6	7,70	0,90
3,0				6,0	37,3	7,45	—1,45
2,8	12,9	2,58	0,22	6,5	36,5	7,30	—0,80
2,5	13,7	2,74	—0,24	7,2	35,9	7,17	0,03
2,6	14,7	2,94	—0,34	8,2	36,4	7,28	0,92
2,8	18,9	3,78	—0,98	8,0	34,9	6,98	1,02
4,0	24,4	4,87	—0,87	6,5	29,7	5,98	0,52
7,0	30,8	6,15	0,85	5,0	23,0	4,60	0,40
8,0	38,0	7,60	0,40	2,0	17,2	3,42	—1,42
9,0	43,2	8,64	0,36	1,5	11,9	2,38	—0,88
10,0	44,7	8,93	1,07	2,2	8,7	1,74	0,46
9,2	45,7	9,12	0,08	1,2	7,7	1,54	—0,34
8,5	45,3	9,06	—0,56	1,8			
9,0	41,3	8,26	0,74	1,0			

веденной с востока на запад, строится профиль по значениям пересекаемых ею контуров. Точно так же строятся профили, соответствующие линиям, проведенным с севера на юг. Естественно, что число профилей определяется густотой сетки, которая выбирается согласно здравому смыслу. Затем каждый профиль сглаживается на глаз так, как это описывает Крамбейн¹ (см. также фиг. 17.1), который отмечает, что, несмотря на субъективный характер такого сглаживания, оно помогает сократить до минимума число точек перегиба.



Фиг. 17.1. 1 — исходный профиль; 2 — сглаженный на глаз; 3 — сглаженный при помощи скользящей средней по интервалу 5 единиц.

Выравнив все профили, надо совместить значения в точках пересечения линий, идущих в разных направлениях. Полученные после сглаживания значения следует нанести на карту в узлах сетки и по этим данным провести изолинии. Кроме полученной таким образом карты «регионального» тренда, можно построить карту остаточного, или локального, эффекта. Такая карта строится путем нанесения в узлах сетки разностей между значениями исходных и выравненных профилей с последующим проведением изолиний (см. ниже).

¹ W. C. Krumbein, Regional and local components in facies maps, Bull. Am. Assoc. Petrol. Geol., 40, 1956.

Авторы полагают, что субъективный эффект выравнивания на глаз можно было бы уменьшить, сгладив отдельные профили с помощью скользящей средней. Этот метод, кроме карты тренда, помогает построить также карту остаточного эффекта; способ ее построения аналогичен способу, рассмотренному выше. На фиг. 17.1, а приведен выборочный профиль, для которого сглаживающая линия проведена на глаз. Этот же выборочный профиль выравнивался скользящей средней, вычисленной по пяти точкам (фиг. 17.1, б). Нетрудно видеть, что разница между этими результатами весьма мала. В табл. 17.1 приведены данные, иллюстрирующие сглаживание с помощью скользящей средней. В столбце 1 помещены значения исходного профиля; в столбце 2 — скользящие суммы по пяти точкам; в столбце 3 — значения средней, которые нанесены на график в виде выравнивающего профиля; в столбце 4 — разности между значениями столбцов 1 и 3 с соответствующим знаком.

РЕГРЕССИОННЫЕ ПОВЕРХНОСТИ

Мы отличаем построения регрессионных поверхностей графическим и арифметическим методами сглаживания потому, что в первом методе ответ получается в форме уравнения, а во втором такого уравнения нет. Ясно, что если при исследовании требуется лишь приблизительно выровнять данные карты для облегчения дальнейшей интерпретации, нет особой необходимости в весьма трудоемкой процедуре вычисления регрессионной поверхности. Действительно, субъективный фактор играет в подобном исследовании весьма малую роль, так как поверхности, построенные с помощью методов регрессии, мало отличаются от поверхностей, полученных путем графических построений. В этом случае такими методами, как, например, сокращенный регрессионный или метод «ожидаемых значений», следует пользоваться с большей осторожностью, чем любым графическим или полным регрессионным методом, так как в них не учтен тренд, непараллельный линиям выборочной сетки.

Таким образом, регрессионные методы являются «аналитическими», дающими в результате их применения уравнение, описывающее карту. Если такое уравнение необходимо для дальнейших исследований, применение регрессионных методов целесообразно. Кроме того, эти методы позволяют построить доверительный интервал для оцененной по выборке регрессионной поверхности.

В настоящее время в геологической литературе почти нет работ, в которых затрагивались бы вопросы применения регрессионных поверхностей.

Поверхность отклика

Бокс¹ рассматривал такие поверхности в связи с задачами статистического анализа химических данных. Несколько позднее Крамбейн² использовал их при изучении фациальных карт. Таким образом, под термином *поверхность отклика* имеется в виду поверхность, построенная по геологическим факторам, контролирующим осадкообразование, тектонику, эрозионные процессы, меняющие состав и геометрию геологических тел.

В связи с этим рассмотрим некоторые термины, которые встречаются в геологической литературе и которые были использованы Крамбейном. Поверхность отклика можно рассматривать как композицию локального и регионального эффектов. Термин *региональный эффект* эквивалентен понятию *карта тренда*, введенному Миллером³. Эти термины относятся к сглаживаемой поверхности, из которой исключены локальные флуктуации, называемые *шумом*. Термину *локальный*, или «*остаточный*», эффект, предложенному Крамбейном, эквивалентно понятие *ошибка определения*, введенное Миллером. В обоих случаях этот термин означает разность между значениями исходной поверхности и величиной, вычисленной для данной точки.

Крамбейн рассматривает также различные интерпретации терминов *карта остаточного эффекта* и *карта тренда*. Основная трудность интерпретации заключается в том, что результаты построения карты остаточного эффекта во многом зависят от метода составления карты тренда. Крамбейн приводит пример построения по одним и тем же исходным данным карт тренда и остаточного эффекта с помощью трех различных способов. Получаемые при этом карты тренда оказываются очень близкими между собой, тогда как карты остаточного эффекта значительно отличаются одна от другой.

Картирование тренда

Локальная нерегулярность часто затемняет, а не проясняет любую попытку выделения диагностической части конкретных районов. Таким образом, ищут «тренд», чтобы можно было описать адекватно распределенные значения изучаемого параметра по географическому региону и устранить эффект чисто местных иррегулярностей, т. е. «шума». Так как при картировании приходится иметь дело с системой трех координат (по двум координатам определяют точку на карте, по одной — значения

¹ G. E. P. Box, The exploration and exploration of response surfaces, some general considerations and examples, *Biometrics*, **10**, 1954.

² См. сноску на стр. 396.

³ R. L. Miller, Trend surfaces: their application to analysis and description of environments of sedimentation, *Journ. Geol.*, **64**, 1956.

признака в этой точке), то самый простой тип тренда в подобной ситуации будет выражаться плоскостью. В ряде случаев представление тренда в виде плоскости является значительным упрощением, как, например, при изучении распределения размеров частиц в осадках шельфа. Тем не менее наклон плоскости выражает в качестве первого приближения скорость изменения значений. Эту плоскость можно подобрать с помощью метода наименьших квадратов и затем в дальнейшем использовать для более детальных исследований тренда. Если же нерегулярность проявлена очень сильно, необходимо применить полиномы высокой степени. Такую поверхность можно получить аналитическим путем в виде уравнения, по которому затем вычислить соответствующие значения признака в точках карты для проведения изолиний. Вычисленные таким образом значения признака в точках карты образуют базис для графического изображения вычисленной поверхности. Это позволяет избежать неопределенности в проведении изолиний, так как из уравнения можно получить сколь угодно густую сеть значений изучаемого признака.

Грант¹ предложил формальное определение тренда, рассматривая последний как функцию координат карты x и y , описывающую поведение значений признака, независимо от характера эксперимента, например метода полевых наблюдений. Он использовал также метод наименьших квадратов в такой же форме, как это сделано в работе Дурбина и Кенделла², в которой функцию $t(x, y)$ эти авторы оценивают с помощью наблюдения значений T_i . Значения функции t отличаются от T_i , и математическое ожидание их разности равно нулю (гл. 8). Наилучшим методом построения эмпирического уравнения в данном случае следует считать метод полиномов.

РЕГРЕССИОННЫЕ МЕТОДЫ

Методы анализа карт, в которых используются поверхности регрессии, для простоты делят на две группы. К первой из них относят методы, требующие сбора данных через равные интервалы, ко второй — все те, которые не требуют выполнения этого условия.

Методы, использующие сетку

Большое число работ, посвященных этому вопросу, принадлежат геофизикам, так как такой обработке легко поддаются

¹ F. Grant, A problem in the analysis of geophysical data, *Geophysics*, 22, 1957.

² J. Durbin, M. G. Kendall, On the method of least squares. *Biometrika*, 38, 1951.

основные типы геофизических данных. Так, Симпсон¹ описал метод, приспособленный к быстродействующим вычислительным машинам, Олдхэм и Сатерленд² предложили способ, основанный на построении ортогональных полиномов. Кроме того, Крамбейн³ описал два регрессионных метода, применяемых при анализе фациальных карт, а также сделал достаточно детальный общий обзор применяемых в этой области статистических методов. Первый из них называется *сокращенным регрессионным методом*. Особенность этого метода заключается в том, что смешанные произведения в полиномах не принимаются во внимание. Это настолько сокращает объем вычислений, что их можно провести на клавишной электрической вычислительной машине. Однако сокращенный регрессионный метод не рекомендуется для тех регионов, где направление тренда не параллельно направлениям сетки, которая обычно задается произвольно.

Во втором методе смешанные произведения полиномов учитываются. Однако в этом случае Крамбейну удалось показать на примере предварительного изучения исходной карты с помощью дисперсионного анализа, что в этом методе имеют решающее значение только некоторые смешанные произведения, поэтому объем вычислений сильно сокращается.

Детальный обзор и анализ приложения методов, основанных на ортогональных полиномах, к задачам картирования принадлежит Гранту⁴. В той же работе этот автор рассматривает методологию альтернативного случая, когда точки, в которых проведены наблюдения, распределены неравномерно.

Вистелиус [277, 278] также пользуется ортогональными полиномами при изучении спектральной яркости песков, которую измеряли с помощью фотометра. В этой работе ортогональные полиномы Вистелиус использовал для выявления зависимости между спектральной яркостью и условиями осадко-накопления.

В подавляющем большинстве различных видов геологических карт нет данных, расположенных по сетке с равными интервалами. Так, в упомянутой работе Крамбейна в рассматриваемых примерах используются как отношения размеров кластического материала, так и мощности песков. В обоих случаях вместо неравномерно расположенных фактических точек наблюдения необходимо проводить интерполяцию, начиная с построения изолиний с равными интервалами. Крамбейн расцени-

¹ S. M. Simpson, Least squares polynomial fitting to gravitational data, and density plotting by digital computers, *Geophysics*, 19, № 2, 1954.

² C. H. G. Oldham, D. B. Sutherland, Orthogonal polynomials: their use in estimating the regional effect, *Geophysics*, 20, № 2, 1955.

³ См. сноску на стр. 396.

⁴ См. сноску 1 на стр. 399.

вает эту интерполяцию как недостаток метода и для ее устранения предлагает пользоваться статистическим приемом, описанным Гулденом¹, позволяющим вычислить те значения, которых не хватает, чтобы можно было получить равномерную сетку. Однако Крамбейн отмечает, что данные карты, составленной путем интерполяции, сходны с данными графических эквивалентов, полученными по более формальным статистическим методам.

Следует отметить, что метод наименьших квадратов теряет свои оптимальные свойства в том случае, когда вместо исходных данных, полученных в точках наблюдения, используются какие-либо другие, вычисленные данные. Однако если известную поверхность можно хорошо описать с помощью полинома, то результаты ее аппроксимации по методу наименьших квадратов в условиях равномерной сетки могут оказаться вернее аналогичных результатов, полученных при неравномерном расположении точек измерения.

Вполне очевиден другой подход к этой задаче: или полный отказ от сетки профилей, или вычисление уравнения регрессионной поверхности непосредственно по данным замеров в неравномерно расположенных на карте точках. Переходим к обзору этих методов.

Методы, при которых не надо разбивать пространство на квадраты

В случае нахождения коэффициентов полинома

$$X_M = a + bx_2 + cx_3 + dx_2x_3 + ex_2^2 + fx_3^2 + gx_2x_3^2 + hx_2^2x_3 + \dots,$$

где M — измеряемый признак; X_M — значения этого признака в точке с координатами (x_2, x_3) ; за основу принимают метод наименьших квадратов. При этом методе не требуется разбивать изучаемое пространство сеткой: сразу же можно обрабатывать данные, полученные эмпирическим путем. Преимущества такого подхода по сравнению с методами, в которых используют скользящую среднюю и разбивают пространство сеткой, рассмотрел Грант². Однако поиск коэффициента полинома становится весьма трудоемким, если этот полином имеет степень выше третьей и при этом становятся необходимыми быстродействующие вычислительные машины. Этот метод состоит в построении так называемых нормальных уравнений, которые для вычисления коэффициентов решаются совместно (гл. 10). Однако число уравнений и, следовательно, коэффи-

¹ С. Н. Goulden, *Methods of statistical analysis*, New York, 1952.

² См. сноску 1 на стр. 399.

циентов быстро растет при увеличении степени полинома, что в свою очередь приводит к необходимости вычисления детерминантов высокого порядка. Сделать это без применения электронной техники практически невозможно. Например, если дана система кубических уравнений и некоторые из них имеют 10 коэффициентов, нужно вычислить несколько детерминантов порядка 10×10 ; если же рассматриваются уравнения четвертой степени, порядок детерминантов будет 15×15 .

Применение ортогональных полиномов не столь трудоемко, но требуется разбить изучаемое пространство сеткой (в связи с тем что хотя ортогонализация возможна также и для неравномерно расположенных точек, ее преимущества в этом случае исчезают).

В литературе, касающейся применения математики в геологии, описаны три метода решения этой задачи.

1. Первый метод описали Джилкрест и Крессман¹ в применении к обработке метеорологических данных. Указанные авторы разработали объективный переход от неравномерно расположенных исходных полевых данных к фиксированной равномерной сетке.

Аналогичный переход применил в своей работе и Грант. Располагая объективными данными, расположенными по сетке, можно использовать удобный в работе метод ортогональных полиномов. На последнем этапе работы целесообразно применить быстродействующую электронную машину.

2. На электронных вычислительных машинах подсчитывают детерминанты. Так поступили Миллер и Зейглер [185] при изучении зависимости между подвижной и неподвижной частями осадка в зоне приобья. При этом они подсчитывали (на машинах AVIDAK и IBM 650) детерминанты для системы уравнений третьей и четвертой степеней с учетом всех смешанных произведений. В результате была построена поверхность, нанесенная на график в виде карты тренда.

Уиттен² проанализировал свойства гранитов Донегала в Ирландии. Кроме того, он изучил с последующей интерпретацией поверхности, построенные по полиномам различных степеней. При этом Уиттен уделял особое внимание остаточным данным, по которым также составлялись карты. Он интерпретировал эти карты как «следы стратиграфии», предшествовавшей процессам гранитизации. При обработке данных этот автор использовал электронные вычислительные машины.

¹ B. Gilchrist, G. P. Cressman, An experiment in objective analysis, *Tellus*, 6, № 2, 1954.

² E. H. T. Whitten, Composition trends in a granite; modal variation and ghost stratigraphy in part of the Donegal granite, *Eire, Journ. Geophys. Res.*, 64, 1959.

В работе Крамбейна¹ дан обзор литературы, в которой освещаются вопросы картирования тренда в связи с геологическими задачами, а также рассматривается подход как к анализу поверхностей, построенных по полиномам различных степеней, так и к остаточным поверхностям. В этой же работе Крамбейн предлагает программу вычисления поверхностей регрессии на быстродействующей машине IBM 650.

Электронная вычислительная техника, без сомнения, будет интенсивно проникать в геологию по мере того, как геологи будут узнавать возможности этой техники при решении геологических задач. Кристаллографы, например, неоднократно использовали электронную технику при решении задач, связанных с методом Фурье². О некоторых других приложениях электронных вычислительных средств упоминается в гл. 10.

3. Третий метод рассматривается более детально, так как он основан на приближенном построении регрессионных поверхностей с помощью метода наименьших квадратов, и все связанные с ним вычисления можно провести на клавишных машинах, не прибегая к электронной технике. Этот метод применялся Миллером³ при изучении им зависимости между движением волн, системой течений, размерами частиц осадка и характеристикой степени отсортированности осадочного материала в прибрежном районе Ла-Холья (Калифорния).

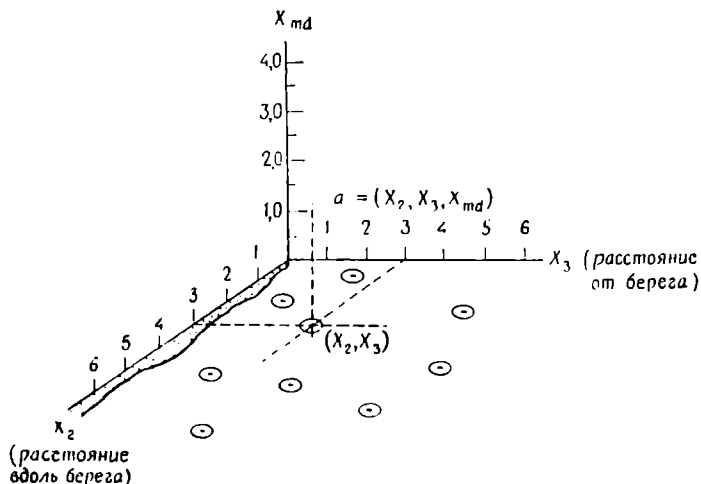
В качестве примера применения третьего метода рассмотрим работу Инмена⁴, изучавшего современные морские прибрежные отложения по выборкам, взятым в их основании. В каждом образце определялось медианное значение размеров частиц осадка. На фиг. 17.2 показано расположение выборочных точек, охарактеризованное координатами, принятыми в данном примере. Если значение третьей координаты (a) поставить в соответствие каждой из этих точек, получим распределение точек в пространстве, для которого нужно провести изолинии. Первый этап работы заключается в выборе плоскости, которая ближе всего подходит к распределению этих точек. Эту плоскость можно рассматривать как первое приближение для поверхности тренда. Нужно найти такую плоскость, для которой сумма квадратов разностей между наблюдаемыми и теоретическими значениями в точках будет минимальной. Это показано на

¹ W. C. Krumbein, Trend surface analysis of contour-type maps with irregular control point spacing, Journ. Geophys. Res., 64, 1959.

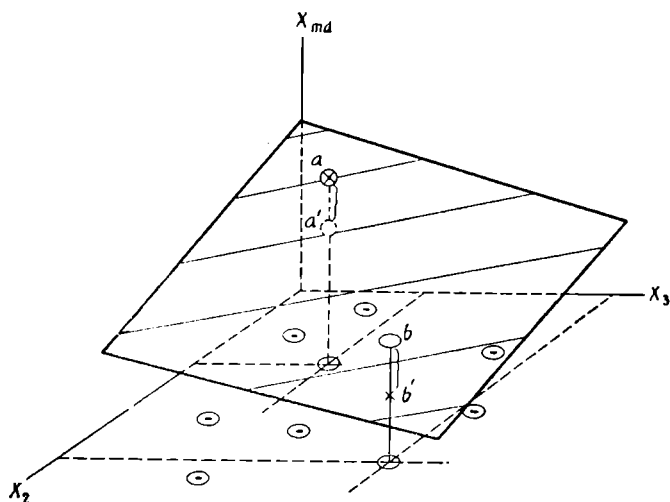
² V. V. V. and, R. P. P. s k y, The statistical approach to X-ray structure analysis, X-ray and Crystal Anal. Lab. Dept. Physics, Pennsylvania State College, Pennsylvania, 1953.

³ См. сноску 3 на стр. 398.

⁴ D. L. Inman, A real and seasonal variations in beach and nearshore sediments at La Jolla, California, U. S. Beach Erosion Board, Tech. memo 34, 1953.



Фиг. 17.2. X_2 — расстояние вдоль берега, X_3 — расстояние от берега, X_{md} — медианное значение размеров частиц, выраженное в фи-единицах. Допустим, что $X_2 = 3$, $X_3 = 3$, а $X_{md} = 2,5$. Тогда точка a расположена в трехмерном пространстве на расстоянии 2,5 по перпендикуляру от плоскости X_2X_3 .



Фиг. 17.3. Плоскость наилучшего приближения для трехмерных данных. Наблюдаемое значение a расположено ниже вычисленного a' , находящегося в этой плоскости, а наблюдаемое значение b расположено выше вычисленного b' .

фиг. 17.3. Заметим, что точка a лежит ниже соответствующего теоретического значения a' на плоскости, а точка b — выше теоретического b' . Естественно, что все точки, соответствующие значениям медианного размера частиц в образцах, не могут лежать в одной плоскости. Поэтому нужно выбрать такую плоскость в заданном пространстве, для которой сумма квадратов разностей $(a - a')^2 + (b - b')^2 + \dots$ обращалась бы в минимум. Очень хороший метод вычисления уравнения такой плоскости описан Крокстоном и Коуденом [72].

Вычислим теперь для каждого наблюдаемого значения величину $Z = X_{md} - X_{md}^{pl}$, где X_{md} — значение медианного размера частиц осадка, измеренное в точке (X_1, X_2) плоскости или карты (X_{md} соответствует X_M приведенного выше полинома); $X_{md}^{pl} = a + bx_2 + cx_3$, т. е. X_{md}^{pl} — вычисленное значение. Выберем на оси X_3 произвольные интервалы $\Delta_i X_3$. Каждому из них будет соответствовать объем, ограниченный плоскостями, параллельными $X_2 X_{md}$, как показано на фиг. 17.4. Вычислим для каждого i -го интервала величину

$$\bar{Z}_i = \sum_{s=1}^{n_i} \frac{Z_{\Delta_i X_3}}{n_i},$$

где n_i — число величин $Z_{\Delta_i X_3}$, соответствующих i -му интервалу на оси X_3 . Полученные значения $\bar{Z}_i$ нанесем на график (фиг. 17.5), где на оси X_{md} откладываются Z_i , соответствующие серединам интервалов $\Delta_i X_3$. Полученный график характеризует отклонения от точной линии регрессии $X_{md}^{r.p.}$, которая определяется как

$$\bar{X}_{md}^{r.p.} = a + bX_2 + c\bar{X}_3.$$

Аналогичный график строится и в плоскости $X_2 X_{md}$ (фиг. 17.5). Если кривую на плоскости $X_{md} X_3$ обозначить через $F'(X_3)$, а на плоскости $X_{md} X_2$ — через $F'(X_2)$, то первую приближенную поверхность можно выразить уравнением

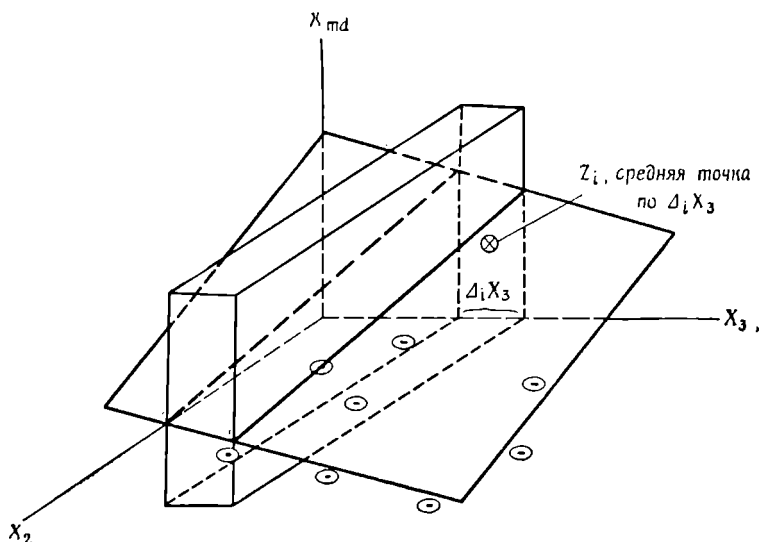
$$X'_{md} = a + F'(X_2) + F'(X_3),$$

где

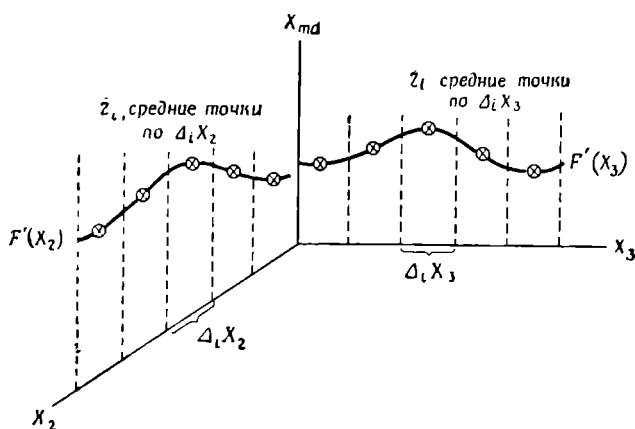
$$a = \bar{X}_{md} - \frac{\sum_{j=1}^n [F'(X_{2j}) + F'(X_{3j})]}{n}, \quad \bar{X}_{md} = \frac{1}{n} \sum_{j=1}^n (X_{md})_j,$$

n — объем всей выборки.

Таким образом, получено первое приближение поверхности. В данном случае аналитическое выражение полинома не дано, так как к построению второго приближения мы перейдем непосредственно от графиков $F'(X_2)$ и $F'(X_3)$.

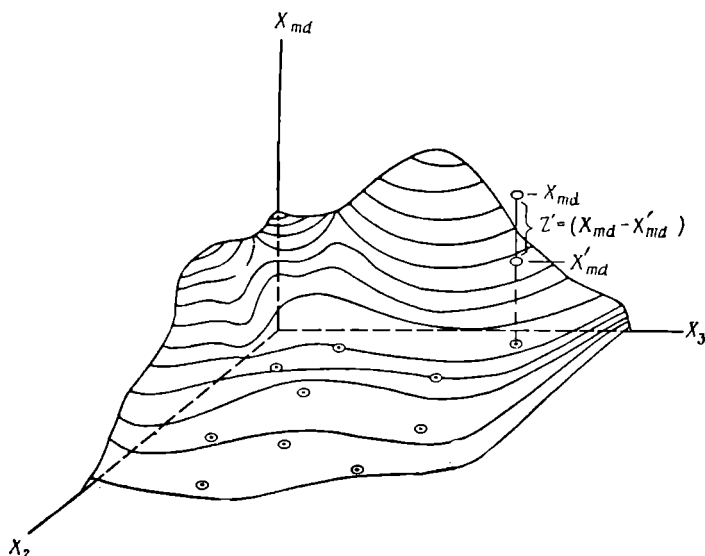


Фиг. 17.4. Геометрическое изображение точки $\bar{Z}_i$ на плоскости $X_{md}X_3$. Ширина блока определяется шириной интервала $\Delta_i X_3$. По значениям $Z = X_{md} - X_{md}^{\text{pl}}$, попавшим в блок, определяется среднее значение для i -го блока $\bar{Z}_i$.

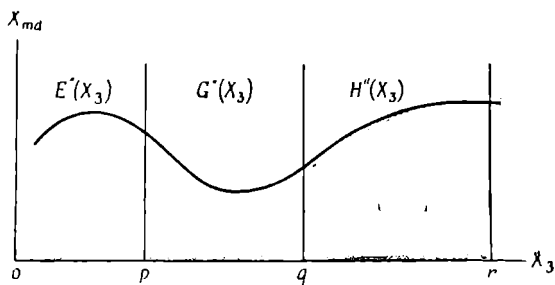


Фиг. 17.5. Каждая точка на плоскости $X_{md}X_3$, полученная, как показано на фиг. 17.4, представляет среднее значение по блоку. Аналогично получены точки на плоскости $X_{md}X_2$. Кривые, проведенные через точки, как показано, соединяют точки $F'(X_2)$ и $F'(X_3)$.

Второе приближение для поверхности тренда получается следующим образом. Каждому наблюдаемому значению X_{md} соответствует вычисленное значение X'_{md} . Следует ожидать, что



Ф и г. 17.6. Поверхность второго приближения тренда. Величина $Z' = X_{md} - X'_{md}$ характеризует расхождение между наблюдаемым и вычисленным значениями точки.



Ф и г. 17.7. Произвольное разбиение функции на участки, описываемые уравнением второй степени или первой.

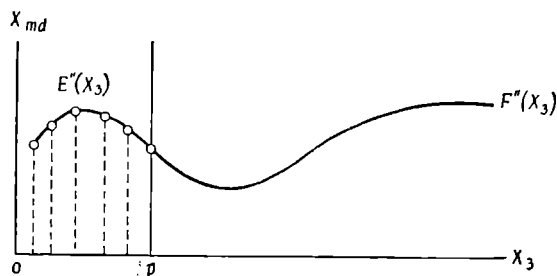
сумма абсолютных значений разностей $X_{md} - X'_{md}$, которые мы обозначим Z' , будет меньше суммы величин $|X_{md} - X^{pl}_{md}| = Z$, так как в данном случае вместо плоскости используется графическая

чески улучшенное приближение искривленной поверхности (фиг. 17.6). Заметим, что характер искривления приближенной поверхности зависит от положения наблюдаемых точек в рассматриваемом трехмерном пространстве. Как и в случае первого приближения, ось X_3 разбивается на интервалы, для которых подсчитываются суммы $\bar{Z}_{\Delta_i X_3}$, соответствующие срединам интервалов на графике. Для оси X_2 проводится та же процедура.

Однако, чтобы нанести на график поверхность второго приближения, нужно аналитически выразить новые кривые $F''(X_2)$ и $F''(X_3)$. Связанные с этим построением вычисления описаны ниже.

ПОСТРОЕНИЕ ЭМПИРИЧЕСКИХ ФУНКЦИЙ

Для функций, выраженных графически, нужно подобрать эмпирические функции. Чтобы по возможности сократить объем вычислений и избежать в них ошибок, рекомендуется применять следующую методику.



Фиг. 17.8. Выбор точек для эмпирического подбора кривой на участке $F''(X_3)$ в интервале $0p$ оси X_3 .

1. Воспользуемся методом средних в уравнениях не выше второй степени. Разобьем изучаемую кривую на участки, содержащие не более одной пограничной точки (фиг. 17.7 и 17.8). Таким образом, функция $E''(X_3)$ будет определена в области от 0 до p , функция $G''(X_3)$ — от p до q , а $H''(X_3)$ — от q до r .

2. В шести¹ выбранных точках эмпирической кривой значения точек для функций последовательно подставляются в квадратное уравнение $X_{md} = a + bX_3 + cX_3^2$, что приводит к шести эмпирическим уравнениям. Затем последовательные пары уравнений складываются, что дает в результате систему из трех уравнений с тремя неизвестными.

¹ Для удобства число шесть выбрано произвольно.

3. Значения неизвестных коэффициентов a , b , c определяются путем решения системы уравнений методом детерминантов. При такой обработке функция $F''(X_3)$ бывает представлена в виде последовательности функций, являющихся частями $F''(X_3)$:

$$F''(X_3) = E''(X_3)|_o^p, \quad G''(X_3)|_p^q, \quad H''(X_3)|_q^r$$

и

$$I''(X_2) = D''(X_2)|_o^l, \quad K''(X_2)|_l^m, \quad L''(X_2)|_m^n.$$

Для построения поверхности тренда окончательные значения вычисляются из уравнения

$$X''_{md} = a'' + F''(X_3) + I''(X_2),$$

где

$$a'' = \bar{X}_{md} - \frac{\sum_{j=1}^n [F''(X_{3j}) + I''(X_{2j})]}{n},$$

n — общее число точек, а $F''(X_3)$ и $I''(X_2)$ подсчитываются из приведенных выше уравнений.

МОЗАИЧНЫЙ МЕТОД

При построении поверхностей тренда изучаемую область можно разбить на несколько более мелких регионов и для каждого из них получить с помощью полиномов первого или второго порядка выражение для описания поверхности. Такой способ исследования, когда каждая подобласть анализируется отдельно от других, но при этом может быть построено единое уравнение для всей области, называется *мозаичным*.

Нарушение непрерывности при использовании мозаичного метода может играть очень существенную роль в теоретических построениях, сказываясь на стыках отдельных подобластей. Однако на практике можно получить хорошие поверхности, если их рассматривать только как приближения. Степень нарушения непрерывности частично зависит от степени равномерности и близости расположения выборочных точек, которое в свою очередь отражается на точности графических методов.

Следует отметить, что трудно перейти от интерполяции к экстраполяции. Возникает вопрос: насколько близко должны быть расположены точки в пространстве, чтобы полученные результаты оказались надежными? Допустим, например, что большой (в географическом смысле) бассейн активного осадконакопления был опробован, но так, что выборочные точки были значительно дальше расположены друг от друга, чем в упоминавшемся выше исследовании Инмена (см., например, работу

Стетсона¹⁾. Будут ли в таком случае изолинии, отражающие поверхность тренда, обоснованными? Точность интерполяции между выборочными точками при любом методе проведения изолиний, очевидно, является функцией как пространственного размещения, так и числа выборочных точек. Ситуации, когда в одних частях региона выборочных точек много, а в других мало, представляют особые случаи этой задачи.

Та же трудность возникает при экстраполяции за пределы опробованной площади. Спрашивается, сколь далеко за пределы опробованной области можно распространить полученную поверхность тренда? Как отмечает Хальд [19]: «Экстраполяция содержит элемент недоверности, который не может быть статистически изучен». Если предположения о форме функции за пределами опробованного района не основаны на надежных эмпирических данных, экстраполяция становится весьма сомнительной процедурой.

СТРУКТУРНЫЕ ДИАГРАММЫ

Обзор статистического анализа структурных диаграмм дается по работе Чейза²⁾ (эти диаграммы можно рассматривать как карты двух- или трехмерной ориентации в выбранной системе координат).

В двухмерном случае для сравнения результатов наблюдения с теоретической моделью (например, изотропной, т. е. такой, в которой все множество равных азимутальных подразделений характеризуется одинаковыми частотами) используется критерий χ^2 .

Роза-диаграмма, составленная Шмидтом³⁾, в данном случае служит аналогом обычной гистограммы. При построении таких диаграмм круг разбивается произвольно на равные секторы. Длины радиусов, разделяющих каждый сектор пополам, выбираются пропорционально соответствующим частотам и их концы соединяются. С помощью нормальной кривой Гаусса Шмидт построил теоретические розы-диаграммы для различных степеней отклонения от изотропного случая. Это позволило производить графические сравнения эмпирических круговых диаграмм с теоретическими.

Чейз рассмотрел трехмерную изотропную модель.

¹ H. C. Stetson, The sediments of the western Gulf of Mexico, Pt. I, Papers in Phys. Ocean. and Meteorol., 12, № 4, Mass. Inst. Tech. and Woods Hole Oceanographic Inst., 1953.

² F. Chayes, Structural Petrology of Deformed Rocks, Cambridge, Mass., ch. 22, 23, 1949.

³ W. Schmidt, Statistische Methoden beim Gefugestudium kristalliner Schiefer, Sitz, Kaiserl. Akad. Wiss. Wien, Math-Nat. Kl. Abt., 1, 1918.

Винчелл¹ для сравнения результатов наблюдения с гипотетическим распределением использовал сетку Шмидта с последующим применением критерия χ^2 .

«Обобщенный критерий», по Винчеллу, основан на другом подходе к задаче. Винчелл рассматривал большое число малых выборок, каждая из которых представлена на точечной диаграмме. Затем эту диаграмму он делил на несколько малых квадратов, таких, что среднее число точек, попавших в квадрат, стало очень малым. В этих условиях можно воспользоваться моделью Пуассона и подсчитать вероятность того, что в квадрате будет находиться 0, 1, 2, ... точек. Вычисленные для этой модели теоретические частоты сравниваются с наблюдаемыми с помощью критерия χ^2 (гл. 16).

Для определения преобладающего направления Чейз² предложил воспользоваться корреляционными методами. Он предположил, что эти методы окажутся более чувствительными при выявлении преобладающих направлений, тогда как, применяя обобщенный критерий Винчелла, следует совокупность классифицировать как изотропную.

Чейз обратил также внимание на работу Стьюдента³, в которой тот предложил метод проверки гипотезы о наличии локальной анизотропии в распределении кровяных телец на площади, разделенной на квадраты. В данном квадрате подсчитывалось число кровяных телец (точек). Эта же величина определялась в четырех соседних квадратах, окружающих первый. В результате получились четыре пары наблюдений. Такая процедура повторяется для каждого из квадратов. В случае локальной неоднородности следует ожидать сильную положительную корреляцию между числами телец каждого квадрата и этими же величинами соседних четырех квадратов. Чейз использовал эту процедуру для изучения локальной неоднородности структур и считал сильно выраженную положительную корреляцию достаточной для вывода о преобладающем направлении ориентировки.

Флинн⁴ сделал обзор критериев, применяемых для выявления преобладающего направления ориентировки в трехмерных структурных диаграммах. Он пришел к выводу о неудовлетворительности этих критериев и вместо них предложил непосредственно сравнивать фактические структурные диаграммы пород

¹ H. Winchell, New method of interpretation of petrofabric diagrams, Am. Min., 22, 1937.

² F. Chayes, Application of the coefficient of correlation to fabric diagrams, Trans. Am. Geophys. Union, 27, 1946.

³ Student, On the error of counting with a haemocytometer, Biometrika, 5, 1907.

⁴ D. Flinn, On tests of significance of preferred orientation in three-dimensional fabric diagrams, Journ. Geol., 66, № 5, 1958.

с искусственно построенными диаграммами со случайным расположением элементов. В работе описан также графический метод, облегчающий такое сравнение. См., кроме того, [11, 15].

ПРИМЕНЕНИЕ ДИСПЕРСИОННОГО АНАЛИЗА В КАРТИРОВАНИИ

В специальной литературе описано много моделей, связанных с дисперсионным анализом. Некоторые из них были рассмотрены в гл. 7. Другие планы эксперимента, а также примеры из различных областей можно найти в литературе, приведенной в конце книги. Естественно, что примеры моделей, связанные с задачами картирования, рассматриваются в этой главе, а не в главе о дисперсионном анализе. Мы используем здесь процедуры, описанные в гл. 7, хотя в этих работах может не быть ни формальной постановки задачи, ни вычисленных соответствующим способом статистик.

СЕЛЕКТИВНАЯ СОРТИРОВКА ГАЛЕКИ НА ОТМЕЛИ ОЗЕРА МИЧИГАН

Крамбейн и Миллер [153] при изучении галечников озера Мичиган разделили отмель длиной 50 футов на пять сегментов. Из каждого сегмента они отобрали по пять выборок из 20 галек, причем каждая выборка была взята наудачу, но в узком гранулометрическом классе от 16 до 32 мм. Затем в каждой выборке, состоящей из 20 галек, подсчитывалось число кремневых галек. Задача заключалась в выяснении наличия вдоль отмели селективной транспортировки галечного материала, выражающейся в существенной разнице количества кремневых галек в сегментах. Эмпирические данные, представляющие значения числа кремневых галек в пяти выборках, отобранных в каждом сегменте отмели, приведены в табл. 17.2.

Таблица 17.2
Число кремневых галек в выборках объемом 20
из галечной отмели северного побережья
Уилметта (Иллинойс)

6	7	2	6	5
2	9	8	8	7
5	8	8	10	9
6	6	4	8	3
6	5	10	9	6
25	35	32	41	30

Нетрудно видеть, что выборочная схема этого примера соответствует модели I дисперсионного анализа, которая описана в табл. 7.12 и 7.18.

Таким образом, проверяемая гипотеза заключается в том, что неизвестные средние значения числа кремневых галек в пяти сегментах отдели равны. Вычисления были проведены по способу, описанному в табл. 7.18, а их результаты приведены в табл. 17.3.

Таблица 17.3

Дисперсионный анализ кремневых галек

Тип изменчивости	Число степеней свободы	Сумма квадратов	Оценка дисперсии	<i>t</i>
Среди групп	4	28,2	7,05	1,50
Внутри групп	20	94,0	4,70	—
Общая	24	122,2	—	—

$$F \left[\begin{matrix} 4 \\ 20 \end{matrix} \right]_{0,05} = 2,87.$$

Так как допустимое значение $F \left[\begin{matrix} 4 \\ 20 \end{matrix} \right]_{0,05} = 2,87$ превышает вычисленное $F=1,50$, нулевую гипотезу следует принять как подтвердившуюся. С геологической точки зрения этот результат можно интерпретировать как отсутствие селективной сортировки по отношению к кремневым галькам.

Более полный обзор приложения такого анализа к задачам фациального картирования можно найти в работе Крамбейна [155].

ВЫБОРОЧНЫЕ СЕТКИ И ДВУХФАКТОРНЫЙ ДИСПЕРСИОННЫЙ АНАЛИЗ

Наиболее простой формой двухфакторного дисперсионного анализа является сравнение степени изменчивости по строчкам и столбцам. При изучении осадочных образований отбор образцов можно производить по сетке, одна ось которой совпадает с направлением транспортировки осадка, а другая — перпендикулярна к ней.

Анализ без повторения

Опробование по сетке можно использовать при изучении локальных или региональных изменений в характеристиках осадочных образований. Рассмотрим комплекс руслового аллювия, **413**

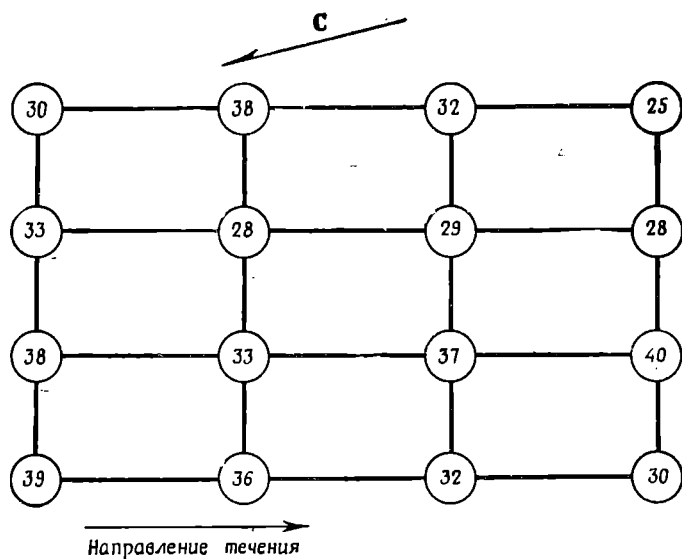
гранулометрический состав которого меняется от чисто галечникового до песчаного, и проследим его на несколько десятков километров вниз по течению. «Региональная» изменчивость на этом большом участке осложнена локальными изменениями, такими, как песчаные и галечные карманы, а также другими особенностями, которые могут быть столь сильно проявлены, что будут маскировать тренд на отдельных участках. Цель данного исследования заключается в отделении регионального тренда от локальных флуктуаций. Чтобы освободиться от локальной изменчивости, было решено отобрать серии близко расположенных проб по сетке через каждые 1,5 км, затем усреднить полученные данные, а результаты усреднения использовать для выделения регионального тренда. В каждой точке сетки отбиралась одна проба, в которой определялись различные характеристики (размеры частиц, их форма, литологический состав, пористость, содержание воды и т. п.). Затем числовое значение соответствующего свойства пробы записывалось в той точке сетки, откуда была взята проба. Статистическую задачу можно сформулировать следующим образом: можно ли эту группу чисел рассматривать как однородную, т. е. представляет ли она единую локальную совокупность, и ее характеристики можно распространить на всю изучаемую совокупность в целом.

Число, входящее в узел сетки, вообще представляет результат единичного наблюдения в пробе, как, например, характеристика пористости. Однако при изучении размеров частиц может быть использован средний размер, который вычисляется при усреднении замеров многих зерен. В последнем случае вступает в действие закон больших чисел, что дает нам основание применять как средние, так и единичные значения в узлах сетки. Но все же и здесь могут быть некоторые ограничения.

На фиг. 17.9 показана сеть опробования участка Арройо-Секо в Калифорнии, где величины в точках взятия 16 проб характеризуют число галек гранодиоритового состава на 100 наудачу отобранных галек. Густота сети опробования была выбрана произвольно. Пробы вдоль участка отбирались через 7,6 м, а поперек — через 5 м. Эти пробы были отобраны для других целей, но результаты их анализа могут служить яркой иллюстрацией поставленной задачи.

В данном случае действовали два фактора изменчивости: первый — основное течение, второй — поперечные флуктуации течения. В результате совместного действия этих факторов состав руслового аллювия мог сильно меняться. В применяемой модели дисперсионного анализа сравнивались эффекты по строкам и по столбцам с «остаточной» изменчивостью, являющейся следствием случайной изменчивости, независимо от различий строк и столбцов.

Если окажется, что как вдоль долины, так и поперек ее изменения несущественны при заданном уровне значимости, можно будет сделать следующий вывод: изменения значений в отдельных пробах являются следствием воздействия случайных факторов. С другой стороны, если между данными строк и столбцов разница окажется существенной, значит сеть опробования столь обширна, что улавливается региональный тренд в локальных выборках. Однако следует помнить, что, применяя



Ф и г. 17.9.

статистический критерий, мы проверяем нулевую гипотезу против множества всех альтернатив, но не можем принять какую-либо конкретную альтернативу.

В данном случае для любого из экспериментов принимается гипотеза, согласно которой значения средних одинаковы как вдоль, так и поперек долины (план эксперимента дан в табл. 7.14). Вычисления в этом примере проведены согласно модели, описанной в табл. 7.19, а результаты даны в табл. 17.4. Так как вычисленные значения для строк и столбцов оказались меньше допустимого при 5-процентном уровне значимости, то средние являются однородными. Короче, можно уверенно усреднить все 16 проб и в дальнейших региональных исследованиях рассматривать их как одну точку.

В связи с этим осмысливание остаточной изменчивости требует некоторых пояснений. Вообще говоря, остаточная дисперсия представляет сумму случайных отклонений в совокупности

и дисперсии, вызванной действием некоторых других факторов, которые не учтены при постановке эксперимента. В данном примере предполагается, что лабораторные аналитические погрешности невелики, т. е. остаточная дисперсия не играет существенной роли.

Таблица 174

Результаты дисперсионного анализа¹

Тип изменчивости	Число степеней свободы	Сумма квадратов	Оценка дисперсии	F
Между столбцами	3	39,5	13,17	< 1
Между строками	3	131,0	43,88	2,8
Остаточная	9	139,0	15,44	—
Общая изменчивость . . .	15	310,0	—	—

¹ По данным фиг. 17.9; $F \left[\begin{smallmatrix} 3 \\ 9 \end{smallmatrix} \right]_{0,05} = 4,26$.

Если отношение дисперсий для строк или для столбцов окажется значимым, то в данном случае будет более приемлема меньшая по площади сеть. Следует отметить, что рассмотренный эксперимент относится только к совокупности гранодиоритовых галек. Может оказаться, что для других характеристик аллювия изменчивость будет большей или, наоборот, меньшей. Общая задача опробования осадочных образований может потребовать проведения ряда подобных исследований с целью получения четко сформулированных выводов об изменчивости в осадках. Обзор тематики, связанной с изучением изменчивости в различных отложениях, принадлежит Крамбейну¹; позднее Гриффитс [104], применяя дисперсионный анализ при исследовании погрешностей выборочного метода, также сделал подобный обзор.

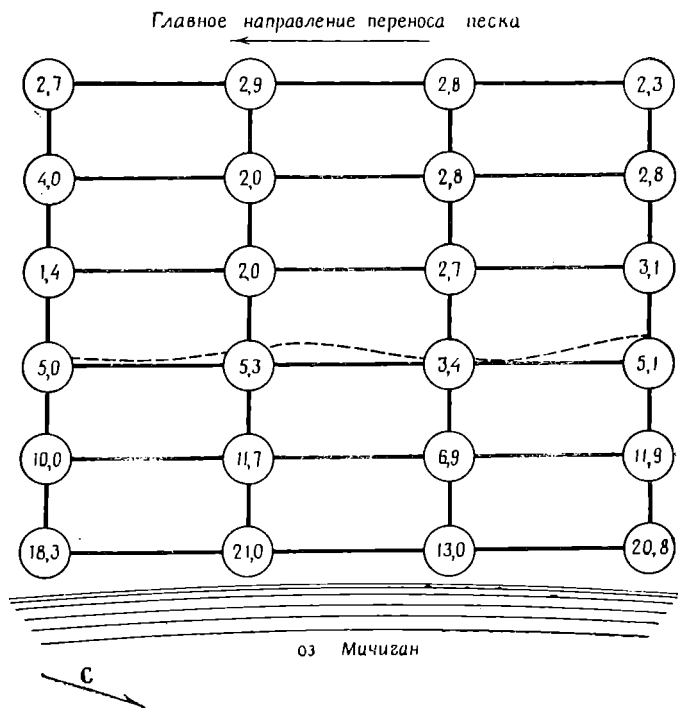
В некоторых примерах результаты опробования по сетке явно свидетельствуют о наличии существенных изменений в изучаемой совокупности. Так, на фиг. 17.10 (показано расположение сетки относительно уреза воды) приведены результаты определения влажности в пробах, собранных в Уилметте (Иллинойс) по сетке. Этот пример выбран потому, что изменчивость по строкам (с удалением от уреза воды) выражена весьма четко, тогда как по столбцам эта изменчивость выражена не очень

¹ W. C. Krumbein, The probable error of sampling sediments for mechanical analysis, Am. Journ. Sci., 227, 1934.

сильно. В табл. 17.5 приведены результаты дисперсионного анализа этих данных.

Из таблицы видно, что дисперсия между столбцами несущественна, тогда как между строками она значительна.

Затруднения, которые встают перед геологом при интерпретации им результатов опробования по такой сетке, заключаются в двойственном характере изучаемого участка. Волны озера омывают этот участок в направлении, перпендикулярном урезу



Фиг. 17.10. (По Крамбейну и Миллеру [15].)

воды, но только во время штормов волны переходят границу обычных сезонных колебаний. Разделим сеть опробования на две части по линии, параллельной урезу воды, и проведем дисперсионный анализ верхней половины, содержащей четыре столбца и три строки.

Результаты этого анализа см. в табл. 17.6; нетрудно видеть, что никаких существенных изменений в выборках по любому направлению нет. Аналогичный анализ нижней части сетки, т. е. тех выборок, которые были подвержены действию волн, показал бы, что между строками еще имеются существенные различия.

Результаты дисперсионного анализа влажности [153]¹

Тип изменчивости	Число степеней свободы	Сумма квадратов	Оценка дисперсии	F
Между столбцами	3	21,46	7,15	2,53
Между строками	5	798,45	159,69	56,43
Остаточная	15	42,46	2,83	—
Общая изменчивость	23	862,37	—	—

$$F \left[\begin{matrix} 3 \\ 15 \end{matrix} \right]_{0,05} = 3,29, \quad F \left[\begin{matrix} 5 \\ 15 \end{matrix} \right]_{0,01} = 4,56$$

¹ По данным фиг. 17.10.

Таблица 17.6

Дисперсионный анализ верхней части фиг. 17.10

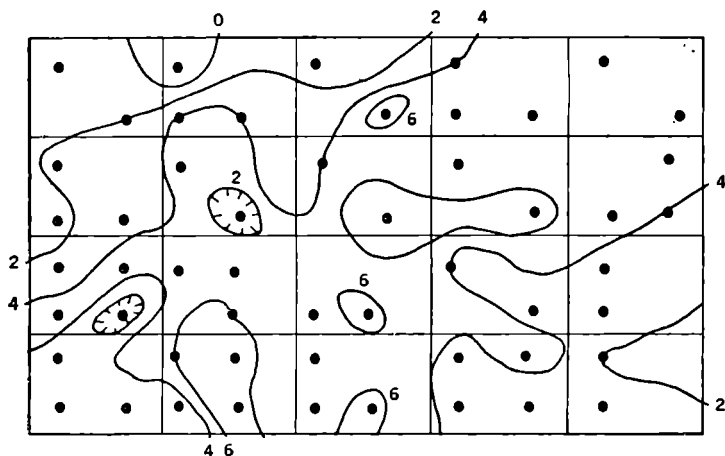
Тип изменчивости	Число степеней свободы	Сумма квадратов	Оценка дисперсии	F
Между столбцами	3	0,43	0,143	< 1
Между строками	2	0,73	0,65	< 1
Остаточная	6	3,52	0,586	—
Общая изменчивость	11	4,68	—	—

Таким образом, установлено, что принятой в данном эксперименте выборочной сетью можно пользоваться для изучения влажности на достаточно большом расстоянии от береговой линии, в то время как вблизи последней такая сеть оказывается непригодной. Однако при изучении других свойств осадка, например степени окатанности частиц, дисперсии могут оказаться как сходными, так и различными по отношению к береговой линии. Поэтому при планировании схемы опробования необходимо рассмотреть изучаемые характеристики в тех частях берега, на которых будет проводиться опробование. Кохрен¹ рассмотрел как рандомизированные выборки внутри квадратов сетки, так и систематизированный выбор точек.

¹ W. G. Cochran, Sampling techniques, New York, 1953.

Основная форма двухфакторного анализа с повторением была использована Крамбейном¹ при изучении площадной изменчивости числа песчаных слоев зоны Арнольда (Оклахома).

На фиг. 17.11 показано расположение скважин внутри квадратов, на которые разбит изучаемый район. Число песчаных



Фиг. 17.11. Карта числа песчаных слоев. Оконтурированы участки зоны Арнольда с одинаковым числом песчаных слоев.

слоев подсчитывалось в каждой скважине с помощью электрокаротажа. Результаты приведены в табл. 17.7 в порядке расположения квадратов. В связи с тем что эти данные представлены

Таблица 17.7

Числа песчаных слоев в зоне Арнольда (по Крамбейну)

Число слоев					Квадратные корни				
1	4	1	3	4	1,00	2,00	1,00	1,73	2,00
2	4	6	4	4	1,41	2,00	2,45	2,00	2,00
2	5	4	4	4	1,41	2,24	2,00	2,00	2,00
3	1	6	6	5	1,73	1,00	2,45	2,45	2,24
и т. д.					и т. д.				

малыми значениями, для стабилизации дисперсии в квадрате данные были преобразованы, что заключалось в извлечении из них квадратного корня табл. 7.21. В табл. 17.8 приведены

¹ См. сноску на стр. 416,

Результаты дисперсионного анализа числа песчаных слоев в зоне Арнольда (по Крамбейну)

Тип изменчивости	Сумма квадратов	Число степеней свободы	Оценка дисперсии	F
Между столбцами	1,0138	4	0,2534	1,41
Между строками	0,5337	3	0,1779	< 1
Смешанная	1,8982	12	0,1582	< 1
Суммарная	3,4457	19	—	—
Остаточная	3,5953	20	0,1798	—
Общая изменчивость . . .	7,0410	39	—	—

результаты дисперсионного анализа, на основании которых можно сделать вывод об отсутствии тренда в изученной области. Крамбейн, интерпретируя этот факт, выдвигает предположение об однообразии геологических условий при формировании зоны Арнольда.

Случай разного числа наблюдений в ячейке

В рассмотренных примерах была использована техника дисперсионного анализа, базирующаяся на равном числе наблюдений в ячейках. Во многих случаях полевых исследований, в частности при изучении больших площадей, размер выборочной сетки невозможно подобрать так, чтобы в одну ячейку попало равное число наблюдений. Такая ситуация может возникнуть в случаях, когда одна часть исследуемого региона хорошо обнажена, а в другой встречаются лишь отдельные обнажения. При использовании данных бурения исследователь сталкивается чаще с групповым расположением скважин, чем с равномерным.

Известны только два способа учета неравномерного расположения точек наблюдения при дисперсионном анализе.

1. На месте пробелов в распределении данных на площади проводить изолинии и недостающие значения получать путем интерполяции [153]. Ясно, что такой метод субъективен и основан на предположении, что в тех участках, где отсутствуют данные, в действительности нет ничего необычного. Однако такое предположение должно быть основано на надежной информации, полученной *a priori*.

2. Второй метод основан на вычислении «эффективного» объема выборки. Детальное описание этого метода с числовыми примерами его применения при изучении косої слоистости приведено в работе Олсона и Поттера [199]. Другой способ вычисления «эффективного» объема описал Снедекор [260], а обобщенные уравнения и доказательства можно найти в работе Кемпторна¹. Критический обзор статистического анализа косо-слоистых толщ сделал Нихофф [196].

ДИСПЕРСИОННЫЙ АНАЛИЗ ПРИ «ГНЕЗДОВОМ ОПРОБОВАНИИ»; ПРИМЕНЕНИЕ КЛАССИФИКАЦИИ ПО ПОДЧИНЕННОСТИ В КАРТИРОВАНИИ

За последние несколько лет статистический анализ геологических свойств региона был упрощен за счет применения выборочного метода, который называют *гнездовым* или *многостадийным*. Цель этих исследований — насколько возможно широко охватить опробованием обширную территорию; при этом обращается большое внимание на локальную изменчивость. Обычно изучаемый район делят на несколько равных по площади участков, что рассматривается как первая стадия. Каждый выделенный таким образом участок можно разделить на множество более мелких участков второй стадии. Некоторое количество этих малых участков, соответствующих одному подразделению первой стадии, представляет выборку из множества всех возможных разбиений этой территории. Процесс такого последовательного дробления участков можно продолжать до тех пор, пока не будет достигнут нижний предел, имеющий геологический смысл.

Например, Поттер и Сивер² при изучении стратиграфической изменчивости использовали следующую выборочную схему:

1. Участок высшей стадии равен 9 супертоуншипам (1 супертоуншип равен 3×3 тоуншипам)³.

2. Из каждого супертоуншипа выбираются два участка по 1 тоуншипу.

3. Из каждого выбранного тоуншипа выбираются два участка, равных 1 кв. миле каждый.

4. На выбранных участках, равных 1 кв. миле, задаются по две скважины.

¹ O. Kempthorne, The design and analysis of experiments, New York, 1952.

² P. E. Potter, R. Siever, A comparative study of upper Chester and lower Pennsylvanian stratigraphic variability, Journ. Geol., 63, 1955.

³ Тоуншип (township) — участок площадью 36 кв. миль, или 93.236 кв. км. — Прим. ред.

Такой подход позволил охарактеризовать региональные тенденции, не пренебрегая локальной изменчивостью. Кроме того, для интерпретации в распоряжении геолога имеются оценки изменчивости на различных стадиях выборки. Из наиболее ранних работ, посвященных применению гнездового опробования с последовательным дисперсионным анализом в геологических задачах, можно назвать статью Олсона и Поттера [199]. Особое внимание названные авторы уделяли изучению двух толщ косо-слоистых песчаников пенсильванского возраста. При этом они разработали следующую схему гнездового опробования:

1. Серия тоуншипов, а для обнаженного района серия горных хребтов (высшая стадия).

2. Несколько участков в 1 кв. милю в одном и том же тоуншипе или горном хребте.

3. Одно или более обнажений в каждом участке, равном 1 кв. миле.

4. Каждое обнажение было представлено разным числом стратиграфических единиц.

5. В одной и той же стратиграфической единице для разных обнажений подсчитывалось число слоев.

В результате применения такой выборочной схемы и дисперсионного анализа была составлена карта направлений косой слоистости, в которой наряду со средними направлениями и их доверительными пределами для отдельных участков представлены средние направления и величины, характеризующие изменчивость на более обширных площадях (более высокие уровни выборки). Примеры применения гнездового выборочного метода можно найти в работах [155—157, 220, 221].

ПРИМЕНЕНИЕ СТАТИСТИЧЕСКИХ МЕТОДОВ В ЗАДАЧАХ ГЕОМОРФОЛОГИИ

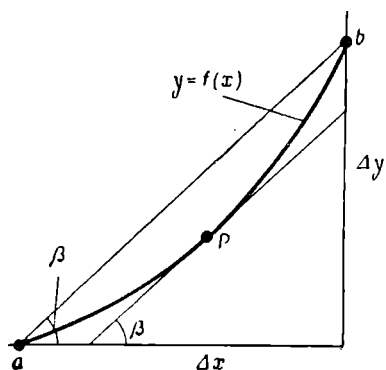
Стралер [267] описал два параметра, которые с успехом можно использовать при решении задач геоморфологического картирования, и провел их статистический анализ. Существуют различные параметры для описания наклона земной поверхности. Следует отметить, что наклон считается важным геоморфологическим свойством, оказывающим влияние на солифлюкцию, движение почвенных растворов и сток.

Первый параметр — это наклон касательной. Если профиль участка земной поверхности рассматривать как кривую $y=f(x)$, изображенную на фиг. 17.12 [267], то тангенс угла наклона прямой ab , соответствующей малому участку $f(x)$, будет равен $\Delta y/\Delta x$. Это отношение при $\Delta x \rightarrow 0$ превращается в производную от $f(x)$ в точке P . Такое заключение основано на теореме, которая утверждает, что если функция $y=f(x)$ непрерывна, а

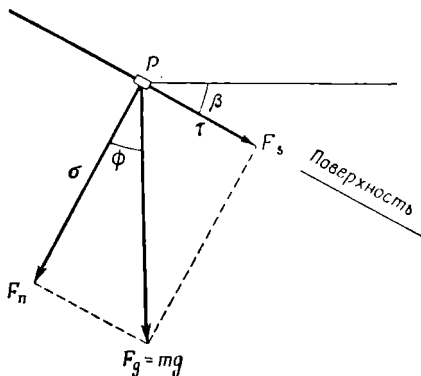
$\Delta y/\Delta x$ — тангенс угла наклона прямой ab , то на $f(x)$ существует по крайней мере одна точка P ($a \leq P \leq b$), в которой производная dy/dx равна $\Delta y/\Delta x$.

С помощью описанной процедуры можно построить карту равных тангенсов, применяя ее во множестве точек контуров карты или же в ее отдельных участках. Соединив одинаковые значения (вместо контуров равных высот), получим контуры равных тангенсов. Стралер предложил называть карту равных тангенсов *картой первой производной* или *картой скорости изменений*.

Второй параметр представлен тригонометрическим синусом. Точка P (фиг. 17.13) [267]) находится на наклонной поверхности,



Фиг. 17.12.



Фиг. 17.13.

которая образует с горизонтальной прямой угол β . Вектор силы тяжести F_g можно разложить на составляющие — перпендикуляр F_n к поверхности и вектор F_s , параллельный склону. Согласно Стралеру, обозначим длину F_n через σ , которая является величиной, прямо пропорциональной силе сцепления частицы с поверхностью. Длину вектора F_s обозначим через τ , которая выражает силу скольжения частицы по поверхности. На фиг. 17.13 видно, что угол ϕ равен углу β , из чего следует зависимость

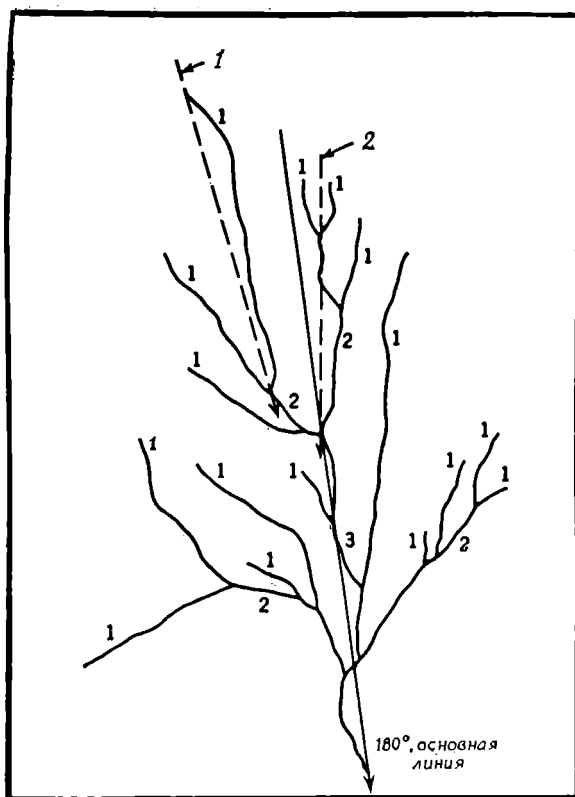
$$\sin \beta = \sin \phi = F_s/F_g.$$

Оконтуривая площади, характеризующиеся равными значениями $\sin \beta$, можно построить карту равных синусов. Эту карту можно интерпретировать как графический аналог распределения той составляющей F_g , которая двигает частицу по плоскости. Естественно, что карта равных синусов идентична карте равных тангенсов. Авторы считают, что следует рассмотреть

физический смысл карты равных синусов. Хаккер¹ также рассматривал силы, действующие на частицу, находящуюся на склоне. Однако он пришел к выводу, что существуют четыре фактора, определяющих стабильность массы породы верхнего слоя. Эти факторы следующие: угол наклона поверхности наименьшего сцепления, сила сцепления, коэффициент трения и угол склона. Таким образом, Стралер и Хаккер измеряли один и тот же угол, но последний пришел к выводу, что вес массы породы не является определяющим фактором, «так как сила, толкающая частицу вниз по наклонной поверхности, и сила трения увеличиваются с одной скоростью при увеличении веса массы пород». Стралер, продолжая статистический анализ склонов, измерил на карте равных тангенсов или синусов площади между последовательными контурами и на основе полученных данных построил гистограммы. Последние представлены в виде процентов от общей площади изучаемого района аналогично весовым гистограммам, используемым при анализе размеров частиц осадочных образований. Стралер считает, что значения параметров, полученные по таким гистограммам, можно рассматривать почти как точные μ , σ^2 , σ , а не как оценки $\bar{x}$, s^2 , s . Он обосновывает свое утверждение тем, что подсчетом он охватил всю площадь и, следовательно, погрешности зависят только от точности измерения площадей в пределах контуров. Следует отметить, что такое заключение справедливо лишь для некоторых региональных исследований. Стралер описал также выборочную схему, основанную на случайном выборе координат точек при анализе углов наклона [152, 153]. Как и в предыдущем случае, Стралер пользовался гистограммами, построенными по результатам измерения площадей на картах равных тангенсов. Используя выборочные методы, основанные на рандомизации точек, он сравнивал выборочные статистики $\bar{x}$ и s^2 , полученные по выборкам разного объема, с приближенными генеральными значениями параметров этого же региона, которые были определены при помощи замеров площадей. Этот способ позволяет однозначно оценить правильный объем выборки при построении карт равных тангенсов и равных синусов. В результате было установлено, что при точечном способе сильно экономится время, так как обработка даже столь больших выборок, как состоящих из 1000 точек, менее трудоемка, чем измерение площади с помощью планиметра. Читателю предлагается исследовать мощность критерия. Сравнение генеральных значений параметров распределения наклонов поверхности для различных регионов можно провести с помощью критерия χ^2 . Кроме того, были проведены интересные исследо-

¹ W. Hacker, Overloading as a motor of mass movement, Ann. Assoc. Am. Geogr., 20, № 4.

вания влияния на точечный метод персонального фактора при выборке объемом 100. В результате установили, что расхождения между выборочными средними несущественны, по крайней мере для выборок, сделанных двумя исполнителями. Интересное статистическое сравнение подводных каньонов с субаэральными дренажными бассейнами провел Шайкинд¹. Он считает,



Фиг. 17.14. Измерения направлений течения в молодом дендритовом бассейне (по Шайкинду, 1956).

1 — направление течения притоков первого порядка; 2 — направление течения притоков второго порядка.

что если изучаются достаточно малые геоморфологические единицы бассейна, то на схеме бассейна можно выделить структурные группы. В указанной работе Шайкинд касался также геометрических свойств схем дренажных бассейнов, включая

¹ E. B. Shykind, Quantitative studies in geomorphology: subaerial and submarine erosional environments, Unpublished doctoral dissertation, Univ. of Chicago, Bull. GSA, 56, 1956.

и свойство симметрии. Он изучал характер распределения направлений течения, в частности, на всех участках русла бассейна.

На данном участке русла азимут направления течения измерялся как угол между прямой, соединяющей верхний и нижний концы рассматриваемого участка, и основной линией. Согласно Стралеру [266], ось симметрии бассейна, по крайней мере в принципе, совпадает с основной линией. На практике в качестве основной линии принимается прямая, соединяющая устье главного русла с его верхним концом. Азимут этой линии обычно считают равным 180° (фиг. 17.14).

Шайкинд предложил ввести следующую терминологию: если в системе бассейна участки русла «ориентированы случайно», такой бассейн называется *дендритовым* или *древовидным*, а в случае отклонения от случайного расположения — *недендритовым*. Это определение можно усилить с помощью двух гипотез.

1. Если бассейн дендритовый, ориентация отрезков русел, упорядоченных по длине, меняется случайно относительно направления главного русла.

2. Если бассейн дендритовый, то длины русел, упорядоченных в соответствии с их ориентацией, будут случайно меняться относительно медианной длины.

Для статистической проверки этих гипотез можно воспользоваться критерием скачков (гл. 14). При сравнении схем симметричного и асимметричного бассейнов (штат Юта) были получены результаты, показанные в табл. 17.9.

Таблица 17.9

	Симметричная схема	Асимметричная схема
Ориентировка относительно оси симметрии с азимутом 180° при упорядочении по длине	Случайная ¹	Неслучайная
Расположение длин относительно медианы при упорядочении по ориентировке	Случайное	Случайное

¹ Результаты получены методом упорядоченных русел Стралера.

Дальнейшее изучение результатов измерения направлений течения производилось следующим образом. Распределение ориентировки участков русла в системе с произвольным порядком сравнивалось с распределением ориентировки участков фиксированной длины также без учета порядка в системе. Пример для этого исследования был заимствован из работы Шамма [243].

Затем распределение ориентировки 483 участков русла произвольной длины сравнивалось с распределением ориентировки 62 участков упорядоченной системы. При этом был использован критерий Колмогорова — Смирнова, который, будучи непараметрическим, позволяет сравнивать общие выборочные распределения, а не описательные статистики, такие, как оценки средних или дисперсий. В результате было установлено, что оба выборочных распределения отличаются одно от другого существенно. Это позволило Шайкинду сделать вывод о том, что, изучая направления течений, можно измерять длину участков упорядоченной системы, что значительно сокращает труд.

МНОГОКОМПОНЕНТНЫЕ СИСТЕМЫ

В геологическом картировании есть класс задач, содержащих три и более компонент, выраженных процентными величинами, сумма которых равна 100%. Ситуация бывает аналогична, если величины выражены в долях единицы, а их сумма равна 1. Например, тройные диаграммы, используемые петрологами в случае трехкомпонентных систем, относятся к этому классу. Процедура построения таких диаграмм была распространена Мерти¹ сначала на четыре переменные, затем на семь компонент. Крамбейн² изучал свойства тетраэдра с точки зрения задач фациального картирования. В указанной работе он достаточно подробно рассматривал вопросы применения неортогональных координатных систем как наглядного способа графического изображения четырех и менее переменных. Вполне возможно, что геометрические построения, основанные на ортогональных системах координат, могут оказаться не только более удобными, но также и открывающими зависимости, совершенно не ожидаемые в различных геологических задачах. Например, можно упростить последующий статистический анализ.

Крамбейн и Тьюки [156] описали метод анализа, предназначенный для решения задач, связанных с изучением площадной изменчивости состава. Однако этот метод можно использовать и в других задачах. Он позволяет рассматривать любое число переменных, выраженных процентными величинами и составляющих в сумме 100%. Основное внимание эти авторы уделили вопросам изучения гомогенности и гетерогенности различных площадей — от отдельных участков до региона.

¹ J. B. Mertie, Charting five, six and seven variables on hypertetrahedral faces, *Am. Min.*, 33, 1948; Charting five and six variables on the bounding tetrahedra of hyper-tetrahedra, *Am. Min.*, 34, 1949.

² W. C. Krumbein, The tetrahedron as a facies mapping device, *Journ. Sed. Petrol.*, 24, 1954.

Данные, предназначенные для исследования, отбираются по многостадийной выборочной схеме (см. раздел, посвященный применению дисперсионного анализа в картировании). При этом выделяются интересные в геологическом отношении категории и определяются проценты членов, образующих категорию в каждой выборочной точке. Например, комплекс тяжелых минералов можно рассматривать как категорию и каждую выборочную точку некоторого региона анализировать на содержание каждого из тяжелых минералов, являющихся компонентами категории.

Статистическая модель, описанная в приложениях, позволяет использовать при решении многомерных задач одномерные методы. При этом используется однофакторный гнездовой выборочный метод в сочетании с перечнем компонент, что приводит к таблице с двумя входами. Основное внимание уделяется взаимодействиям. Предварительный шаг в обработке исходных данных, выраженных процентными величинами, заключается в их преобразовании в арксинус квадратного корня. Например, при изучении химического состава пород особый интерес могут представлять данные о содержании редких элементов. Данные, полученные путем логарифмического преобразования, могут быть использованы впоследствии для «взвешивания» процентов. Однако, по мнению Крамбейна¹, в случае большого размаха значений процентных величин, что наблюдалось при изучении главных элементов наряду с редкими, наилучшим преобразованием исходных данных можно считать арксинус (гл. 7).

Следует отметить, что особое внимание должно быть уделено выборочной схеме и исходному выбору переменных, а также последующей интерпретации различных шагов анализа. Ясно, что оптимальные результаты при применении этого метода скорее всего будут получены у исследователей, имеющих опыт применения дисперсионного анализа или же пользующихся советами статистиков.

Применение теории информации

Рассмотрим работу Пелто², в которой при картировании многокомпонентных систем используется теория информации. Пелто предложил два способа картирования систем, содержащих много компонент, но мы сосредоточим внимание на втором из них, в котором используется функция, подобная энтропии и

¹ W. C. Krumbein, Comparison of percentage and ratio data in facies mapping, Journ. Sed. Petrol., 27, 1957.

² C. R. Peltó, Mapping of multicomponent systems, Journ. Geol., 62, № 5, 1954.

определенная по Шеннону:

$$H = - \sum_i p_i \log p_i,$$

где p_i — доля i -го компонента в системе, а $\sum_i p_i = 1$. Если это выражение умножить на константу, оно будет эквивалентно обычной энтропии

$$S = - k \sum_i f_i \log f_i,$$

где f_i — вероятность того, что система находится в i -й фазе пространства, а k — постоянная Больцмана, известная из термодинамики. Однако в данном случае термин *энтропия* выражает неопределенность. Читателю, который хочет подробнее ознакомиться с теориями связи и информации в применении к геологии, рекомендуем ознакомиться с работами Шеннона [246], Винера [284] и Мак-Дональда [176].

Нижеследующий обзор дается по работе Пелто. Он отмечал, что энтропия, будучи величиной, характеризующей степень неопределенности, может быть с успехом использована при картировании многокомпонентных систем. Таким образом, зоны, характеризующиеся высокой энтропией (пониженной информационной способностью), являются природными барьерами, отделяющими районы развития тех фаций, которые обладают значительной информацией. Например, когда в каждой из множества выборочных точек карты наблюдение ведется над пятью компонентами, правильным будет считать минимум «информации» соответствующим точкам, обладающим равными пропорциями всех пяти признаков. В данном случае в связи с отсутствием преобладающей компоненты невозможно классифицировать эту точку на карте в терминах какой-либо одной компоненты или их группы.

С другой стороны, максимальная «информация» достигается в том случае, когда доля одной компоненты равна единице (100%), а остальные четыре равны нулю.

Пелто ввел понятие *относительная энтропия*, представляющее отношение наблюдаемого значения энтропии к максимальному, которое можно получить для данного числа компонент:

$$100 H_r = \frac{- 100 \sum_{i=1}^N p_i \ln p_i}{H_m},$$

где p_i — доля i -й компоненты в N -компонентной системе, а H_m — максимально возможная энтропия, которая дается выражением

$$H_m = \sum_{i=1}^N \frac{1}{N} \ln \frac{1}{N} = \ln N.$$

Таблица для вычисления энтропии
 H_m (максимальная энтропия)

Число компонент	H_m
3	1,0986
4	1,3863
5	1,608
6	1,795
7	1,943
8	2,076
9	2,187
10	2,303

p	$-p \ln p$	p	$-p \ln p$	p	$-p \ln p$	p	$-p \ln p$
0,001	0,007	0,19	0,315	0,46	0,356	0,73	0,230
0,002	0,012	0,20	0,322	0,47	0,355	0,74	0,223
0,003	0,017	0,21	0,328	0,48	0,354	0,75	0,216
0,004	0,022	0,22	0,333	0,49	0,350	0,76	0,208
0,005	0,027	0,23	0,340	0,50	0,343	0,77	0,200
0,006	0,031	0,24	0,342	0,51	0,344	0,78	0,194
0,007	0,035	0,25	0,346	0,52	0,340	0,79	0,186
0,008	0,038	0,26	0,350	0,53	0,336	0,80	0,178
0,009	0,042	0,27	0,353	0,54	0,334	0,81	0,170
0,01	0,046	0,28	0,357	0,55	0,329	0,82	0,162
0,02	0,084	0,29	0,359	0,56	0,325	0,83	0,152
0,03	0,106	0,30	0,361	0,57	0,320	0,84	0,147
0,04	0,129	0,31	0,364	0,58	0,316	0,85	0,139
0,05	0,150	0,32	0,365	0,59	0,312	0,86	0,126
0,06	0,168	0,33	0,366	0,60	0,307	0,87	0,121
0,07	0,186	0,34	0,366	0,61	0,302	0,88	0,113
0,08	0,202	0,35	0,3675	0,62	0,297	0,89	0,102
0,09	0,216	0,36	0,368	0,63	0,291	0,90	0,094
0,10	0,230	0,37	0,370	0,64	0,286	0,91	0,086
0,11	0,243	0,38	0,368	0,65	0,280	0,92	0,077
0,12	0,254	0,39	0,3675	0,66	0,274	0,93	0,067
0,13	0,265	0,40	0,366	0,67	0,268	0,94	0,058
0,14	0,275	0,41	0,366	0,68	0,261	0,95	0,050
0,15	0,285	0,42	0,364	0,69	0,256	0,96	0,039
0,16	0,293	0,43	0,364	0,70	0,250	0,97	0,029
0,17	0,301	0,44	0,361	0,71	0,243	0,98	0,019
0,18	0,309	0,45	0,360	0,72	0,236	0,99	0,010

Таким образом, для случая $N = 5$

$$H_m = -5 \cdot 0,20 \ln 0,20 \approx 1,608.$$

Тогда величина $100H_r$, являющаяся относительной энтропией, выраженной в процентах, будет максимальна, если

$$100H_r = \frac{-100 \sum_{i=1}^5 0,20 \ln p_i}{-\sum_{i=1}^5 0,20 \ln p_i} = 100,$$

и минимальна, если

$$100H_r = \frac{-100 \cdot (1,00 \ln 1,00 + 0 \ln 0 + \dots + 0 \ln 0)}{1,608}.$$

Для быстрого вычисления значений энтропии авторы составили табл. 17.10.

ПРИМЕНЕНИЕ КАРТЫ ОТНОСИТЕЛЬНОЙ ЭНТРОПИИ ПРИ ФАЦИАЛЬНЫХ ИССЛЕДОВАНИЯХ

Процедура обработки результатов наблюдения показана авторами на примере построения карты относительной энтропии по данным (заимствованным из работы Паркера и др.¹) экологического изучения фораминифер из бухты Сан-Антонио. В данном примере использована лишь часть рассмотренного ими региона. Паркер и др. на основе изучения большого числа образцов разделили наблюдаемые виды фораминифер на семь биофаций. В табл. 17.11 приведена характеристика биофаций бухты Сан-Антонио и части острова Матагорда. Те виды, которые были обнаружены только в других участках изученного общего региона, в этой таблице не приведены; те виды, которые обнаружены в бухте Сан-Антонио и на острове Матагорда, но не вошли ни в одну из выделенных биофаций, объединены в восьмую категорию, которая названа *недиагностируемой*. Таким образом, нам нужно построить карту относительной энтропии для восьмикомпонентной системы. В табл. 17.12 приведены примеры выборочных станций и показаны некоторые детали вычислений. Все значения $-p \ln p$ даны в табл. 17.10. Однако применение таблиц натуральных логарифмов обеспечивает почти такую же скорость работы. В данном примере использованы результаты наблюдений на 122 станциях. На фиг. 17.15 показана часть района, изученного Паркером и др.

¹ F. L. Parker, F. B. Phleger, J. F. Peirson, Ecology of foraminifera from San Antonio Bay and environs, Southwest Texas, Spec. Publ. № 2, Cushman Foundation for Foraminiferal Res., 1953.

Таблица 17.11

Болото	Бухта I	Открытый залив	Бухта II	Береговая линия	Глубинная часть бухты	Недиагностируемые
<i>Ammonoastura inerta</i>	<i>Ammobaculites dilatatus</i>	<i>Bolivina lowmani</i>	<i>Discorbis floridana</i>	<i>Quinqueloculina seminulum</i>	<i>Bolivina striatula</i>	<i>Ammonoastura pseudospiralis</i>
<i>Arenoparrella mexicana</i>	<i>A. salsus</i> and variants	<i>Cibicides strattoni</i>	<i>Elphidium discaudale</i>	<i>Q. wiesneri</i>	<i>Elphidium</i> sp. cf.	<i>Elphidium gunteri</i>
<i>Miliammina fusca</i>	<i>A. sp. A.</i>	<i>Elphidium advenum</i>	<i>E. incertum mexicanum</i>	<i>Q. sp. cf. Q. compta</i>	<i>E. Koeboeense</i>	<i>E. matagordani</i>
<i>Trochammina macroscelus</i>	<i>A. sp. A.</i>	<i>Gaudryina exilis</i>	<i>Quinqueloculina cultrata</i>		<i>Elphidium poeyanum</i>	<i>Massilina protea</i>
Болото острова Матагорда	<i>Elphidium delicatulum</i>	<i>Massilina peruviana</i>	<i>Q. funajutiensis</i>		« <i>Rotalia</i> » <i>beccarii</i> A	<i>Quinqueloculina poeyana</i>
<i>Discorinopsis aguayoi</i>	<i>E. galvestonensis</i>	<i>Nonionella atlantica</i>	<i>Triloculinella obliquinoda</i>		« <i>R</i> » <i>beccarii</i> B	<i>Q. rhodiensis</i>
<i>Triloculina fiterrei meningori</i>	<i>E. sp. cf. E. tumidum</i>	<i>Reussella atlantica</i>			<i>Triloculina si-debottomi</i>	« <i>Rotulia</i> » <i>beccarii</i> C.
Река	<i>Palmerinella palmerae</i>	<i>Textularia</i> sp. cf. <i>T. majori</i>				<i>Trochammina inflata</i>
<i>Protonina lagernaria</i>		<i>Bigenerina irregularis</i>				
<i>Urnulina compressa</i>		<i>Quinqueloculina lamarckiana</i>				

Биофации

№ станции	Болото	Река	Бухта I	Открытый залив	Бухта II	Прибрежная зона	Глубинная часть бухты	Не диагностируемая часть	$100 \sum p_i \ln p_i$	$100 H_r$
60	2,0	0	85,0	0	0	2,0	9,0	2,0	60,7	29,2
58	5,0	0	35,6	2,0	5,0	0,6	16,0	34,0	144,2	69,3

Результаты, помещенные в таблице, получены авторами по данным Паркера и других, представляющим общую таблицу совокупности всех видов, наблюдаемых на данной станции. Доля каждой биофации выражена в процентах. Некоторые особи, меньшие 0,149 мм, не точно определенные, здесь не использованы.

Последовательность вычислений.

1. $0,02 \ln 0,02 = -0,084$. Аналогичный подсчет производится для всех данных станции № 60. Умножая сумму полученных значений на 100, найдем число 60,7.

2. Максимальная энтропия для восьми компонент $H_m = -8 \cdot 0,125 \ln 0,125 = 2,076$.

$$3. 100 H_r = \frac{-100 \cdot (0,02 \ln 0,02 + 0 \ln 0 + 0,85 \ln 0,85 + \dots)}{2,076} = \frac{60,7}{2,08} = 29,2.$$

Таким образом, для станции № 60 характерно небольшое значение относительной энтропии. Преобладающая биофация будет фацией бухты, занимающая 85% всей популяции. Станция № 58 отличается высокой относительной энтропией и не имеет преобладающей фации.

Кроме положения станций, на этой фигуре оконтурены участки с относительной энтропией больше 60% (заштрихованы в двух направлениях) и участки с относительной энтропией больше 50% (заштрихованы в одном направлении). В связи с тем что участки, характеризующиеся высокими значениями относительной энтропии, обладают значительной степенью неопределенности, их следует интерпретировать как природные зоны, отделяющие области с низкой энтропией, в пределах которых доминируют одна или две биофации. Таким образом, зоны, характеризующиеся наиболее низкими значениями энтропии, следует рассматривать как «фации». Например, в правой верхней части карты изображено несколько зон с очень низкой энтропией, в которых доминирует речная биофация. В некоторых других участках преобладает глубоководная фация, тогда как внизу карты в середине показано развитие болотной биофации. Это исследование позволяет сделать ряд интересных выводов.

1. Карта относительной энтропии в общем согласуется с интерпретацией результатов исследования этой области Паркером и др.

2. Отчетливых в географическом смысле областей развития одной фации, т. е. имеющих четкую границу, в данном примере не существует. Поэтому понятие *области развития фации* как определенной зоны не является в данном случае универсальным. Тем не менее в бухте существуют также участки, где переход от высокой энтропии к низкой выражен отчетливо, как в случае болотной и речной фаций. Вообще полученная схема выгля-

дит пятнистой с извилистыми контурами, и ее следует рассматривать как типичную карту скорее с постепенными переходами, чем с резкой сменой фаций.

3. Вытянутая узкая зона высокой энтропии внизу карты хорошо согласуется с географической границей острова Матагорда и бухты Сан-Антонио, что служит подтверждением обоснованности предположения, согласно которому зоны высоких значений энтропии можно рассматривать как природный барьер между фациями.

4. Паркер и др. предполагали, что между типами донных осадков и биофациями существует корреляционная зависимость. Детальные исследования такого типа можно сильно упростить, совместив карту распределения осадков с картой относительной энтропии биофаций. Аналогично можно сопоставить карты распределений солености и температур. Возможны также и другие интересные приложения карт относительной энтропии. Теперь мы перейдем к изучению карт распределения осадка, выраженных в терминах функций энтропии.

Относительная энтропия как мера «площадной отсортированности» осадка

Относительную энтропию H , применяют, кроме того, при измерении степени «площадной отсортированности» осадка по размерам осадочных частиц. В этом случае функция H , играет роль описательного параметра; по его значениям строятся контуры карты. В нижеследующем примере мы охарактеризуем площадную отсортированность как распределение осадочных частиц по их размерам в терминах относительной энтропии. Область значений, которая разделяется на классы, фиксируется с помощью наименьшего и наибольшего наблюдаемых значений размеров частиц¹ осадка в изучаемом районе, во всех последующих вычислениях принимающихся за константы. Те классы группировки, которые оказываются пустыми, рассматриваются в качестве нулевых элементов². Такое включение пустых классов, как внутренних, так и внешних, в максимально и минимально заполненные классы возможно для любой точки.

Такая мера площадной отсортированности удобна, например, когда нужно определить характер распределения частиц по

¹ При выделении интервалов на практике используется механический гранулометрический анализ.

² В данной задаче необходимо рассмотреть число гранулометрических классов. Если задача заключается в изучении распределения осадка по максимальному фиксированному числу классов, сравниваются результаты наблюдения в отдельных точках площади. Если же основное внимание сосредоточено на изменениях в отдельной точке, число гранулометрических классов

размеру при каком-либо физическом процессе, протекавшем в исследуемом регионе.

Следующие свойства относительной энтропии весьма облегчают ее приложение:

1. Область значений H_r находится в интервале 0—1. Нуль соответствует случаю наилучшей отсортированности, а единица — наихудшей. Максимальное значение $H_r=1$ наблюдается, когда вероятности для всех классов равны. Минимальное значение $H_r=0$ соответствует случаю, когда вся масса распределения сосредоточена в одном классе.

2. Величина H_r безразмерна и не зависит от среднего размера частиц осадка.

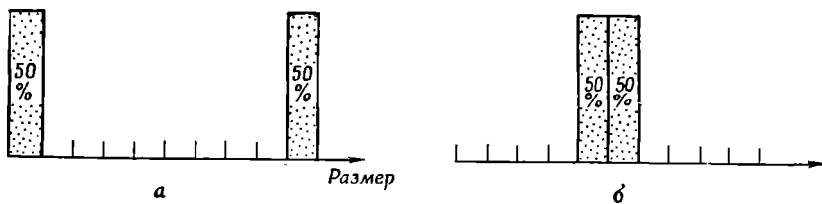
3. Вычисление этой величины крайне простое.

Детальное аналитическое сравнение H_r с дисперсией данного распределения размеров частиц было бы весьма интересным. Мы приведем две контурные карты одного и того же района. Одна из них построена по значениям стандартного отклонения размеров частиц в миллиметрах, а вторая — по значениям H_r . Пример вычислений приведен в табл. 17.13.

Изученный регион расположен в южной части озера Мичиган (Дюнас-Стейт-Парк). При его изучении прямоугольная сеть опробования была выбрана так, что в него вошли дюна, удаленный от берега участок, плоский участок, прибрежная зона, уступ и зона размыва, т.е. прямоугольник размером 60×120 футов охватил самые разнообразные формы рельефа. На фиг. 17.18 приведена топографическая карта этого участка, на которой показано положение точек опробования и динамических зон. Ситовой анализ, топографическую съемку и главную часть полевого опробования выполнил Н. Колимэн (Чикагский университет). На фиг. 17.19 приведена контурная карта стандарт-

будет меняться в зависимости от числа заполненных и пустых классов. В обоих случаях возникают некоторые трудности.

Допустим, что из 10 возможных классов заполнены только два. Тогда, если число классов фиксировано, значение H_r будет одинаковым для двух крайних случаев (фиг. 17.16).

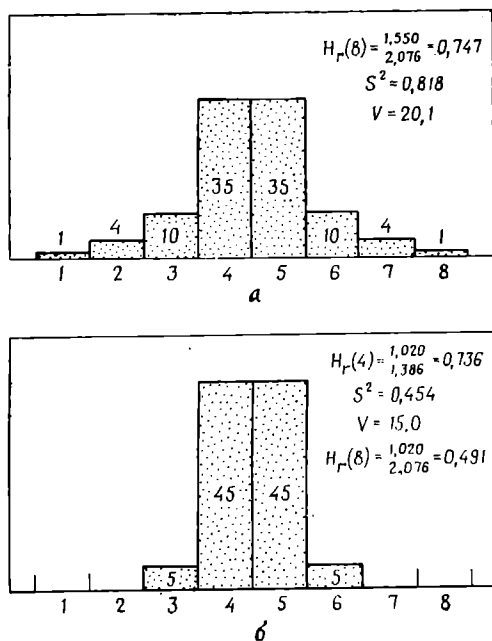


Фиг. 17.16.

Если же при вычислении H_r использовать все 10 классов для случая *a*, а для случая *б* только два класса, полученные результаты будут более реальными. Однако ситуация *a* практически неправдоподобна.

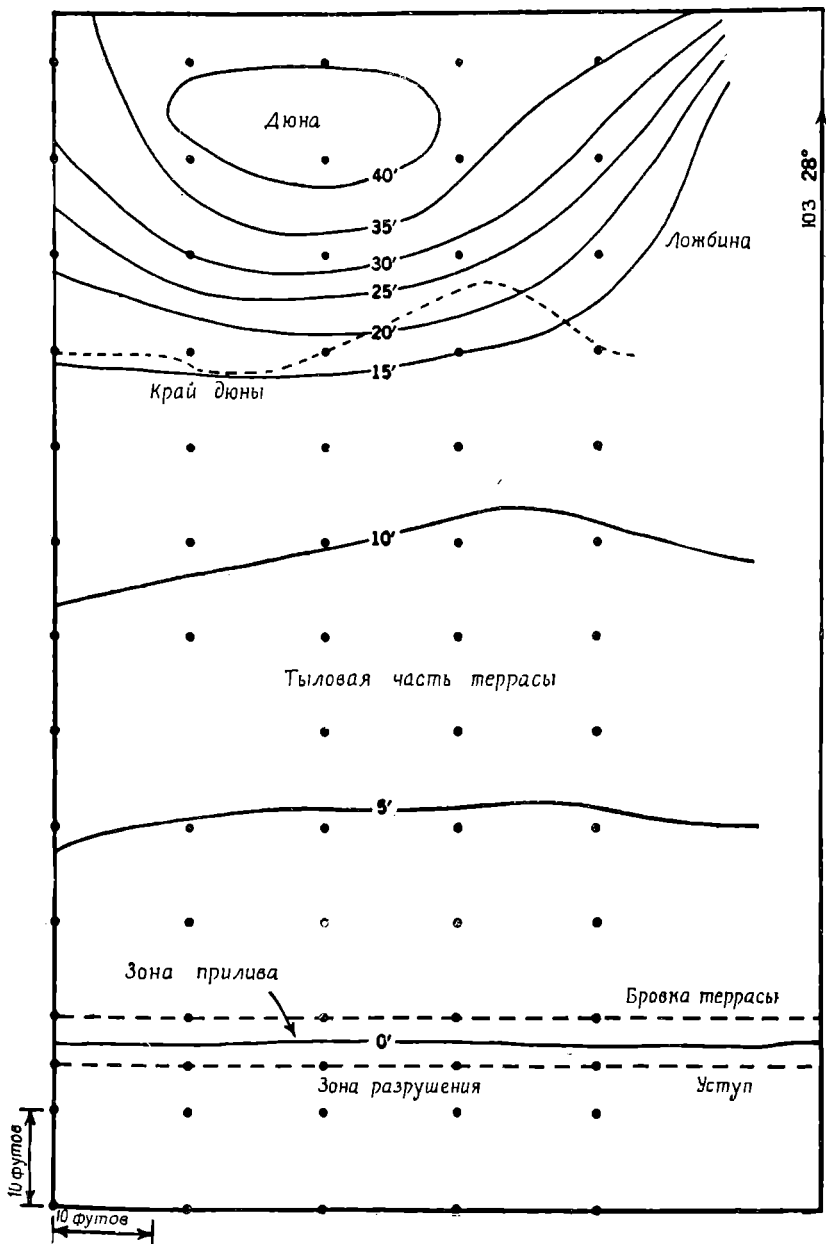
ных отклонений, выраженных в миллиметрах, иллюстрирующая некоторые общие трудности, возникшие при картировании стандартных отклонений. Так как источником материала служили прибрежная зона и зона разрушения, контуры различных значений стандартных отклонений также тяготеют к этим участкам независимо от динамического эффекта. Наоборот,

С другой стороны, если H_r мы вычислим на основе числа заполненных классов (но включив пустые интервалы), то на практике может возникнуть следующая ситуация. На фиг. 17.17 приведены две гистограммы (а и б). На гистограмме а представлено распределение частиц осадка (в процентах) по восьми гранулометрическим классам. Середины классов для простоты обозначены числами натурального ряда. На гистограмме б показано распределение, в котором участвуют только четыре класса.

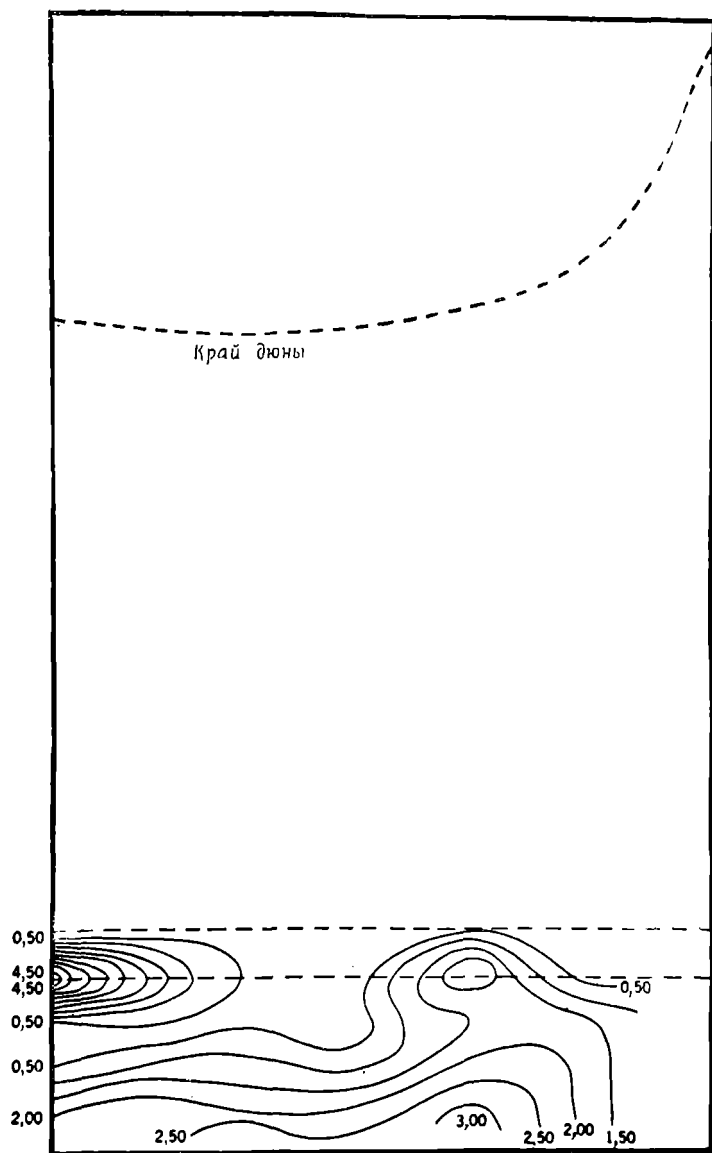


Фиг. 17.17.

Ситуация б скорее всего будет рассматриваться как случай лучшей отсортированности, чем а. Дисперсия для гистограммы а равна 0,818, а для б — 0,454. Значения коэффициентов вариации близки. Однако, если при вычислении H_m в случае а будет использовано восемь классов, а в случае б — четыре класса, результат окажется явно неудовлетворительным, так как H_r примет значения 0,747 для а и 0,736 для б, что практически одно и то же. С другой стороны, если в обоих случаях используется одно число классов, например восемь, то значение H_r для б снизится до 0,491, что более правдоподобно по сравнению с а, где $H_r = 0,747$.



Ф и г. 17.18.



Ф и г. 17.19. Контурная карта стандартного отклонения размеров частиц (в миллиметрах) через интервал 0,50 мм.

Примеры вычисления значений H_r

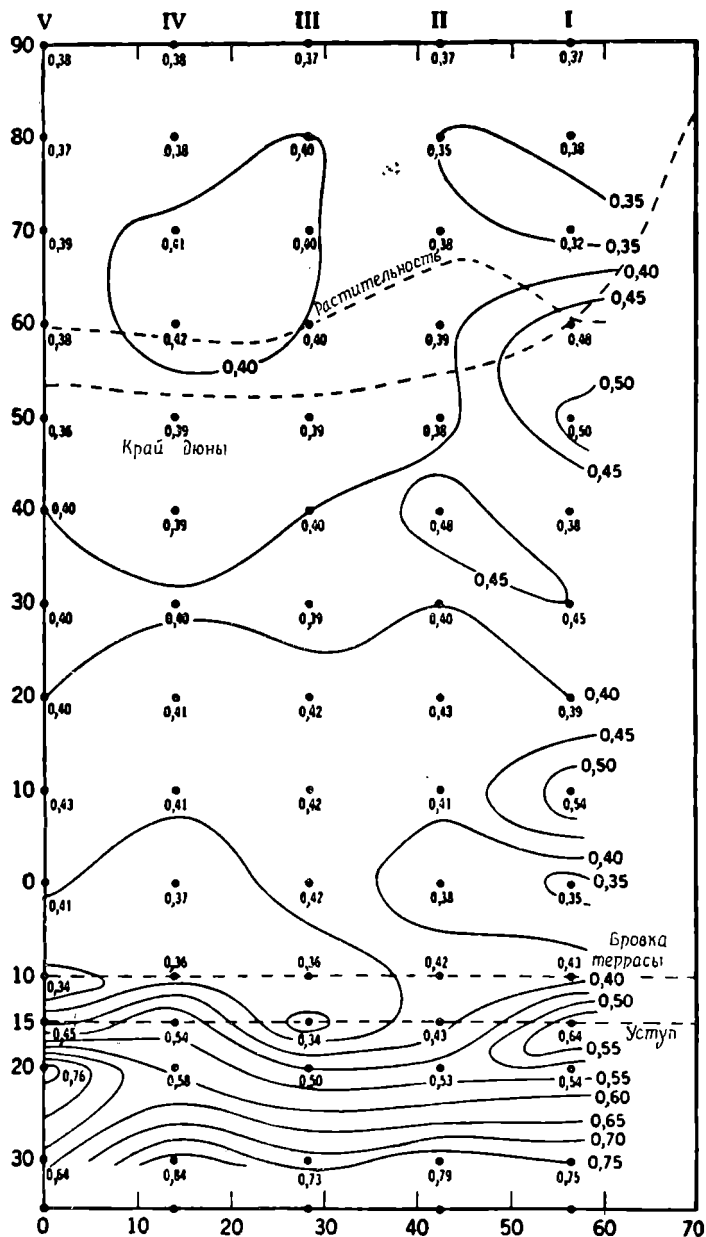
Размеры зерен, мм	Выборка А		Выборка В	
	вес. %	$-p_i \ln p_i$	вес. %	$-p_i \ln p_i$
> 15,850				
15,850—11,100			2,265	0,084
11,100—7,925			1,909	0,084
7,925—5,613			2,201	0,084
5,613—3,962			4,338	0,129
3,962—2,794			4,350	0,129
2,794—1,981			3,591	0,129
1,981—1,397			4,660	0,150
1,397—1,000			5,004	0,150
1,000—0,701	0,024	—	7,106	0,186
0,701—0,500	0,341	0,017	13,297	0,265
0,500—0,350	7,320	0,186	22,284	0,333
0,350—0,250	37,040	0,370	18,611	0,315
0,250—0,175	49,951	0,350	9,197	0,216
0,175—0,125	4,616	0,150	0,589	0,031
0,125—0,088	0,523	0,027	0,233	0,012
0,088—0,061	0,073	0,007	0,029	—
< 0,061	0,048	—	0,005	—
$-\sum p_i \ln p_i = 1,107$			$-\sum p_i \ln p_i = 2,297$	
Выборка А $S = \sqrt{\frac{9,875-9,140}{99}} = 0,086$ $\bar{x} = 0,30 \text{ мм}$ $H_r = \frac{1,107}{2,904} = 0,381$			Выборка В $S = \sqrt{\frac{1228,890-320,440}{99}} = 3,120$ $\bar{x} = 1,79 \text{ мм}$ $H_r = \frac{2,297}{2,904} = 0,791$	

Значения $-p_i \ln p_i$ взяты из табл. 17.10. Для каждой выборки приведено вычисление оценки стандартного отклонения. Выборка А взята из верхней части дюны, а выборка В — из зоны разрушения. Заметим, что $H_m = 2,904$ соответствует 18 гранулометрическим классам. Это число классов — наибольшее из всех наблюдаемых в точках опробования. В данном случае

$$H_m = -18 \cdot 0,055 \ln 0,055 \approx 2,904.$$

в пределах дюны и удаленного от берега участка, где средние размеры частиц меньше, изменчивость не проявлена.

На фиг. 17.20 приведена карта отсортированности по площади. Эта карта построена по значениям относительной



Фиг. 17.20. Карта относительной энтропии для 18 гранулометрических классов. Контуры проведены через интервал 0,05.

энтропии. Соответствующие данные приведены в табл. 17.13. Так как значения относительной энтропии не зависят от размеров частиц, площадная изменчивость на этот раз проявилась на участках, удаленных от берега, и в районе дюны. Кроме того, эта изменчивость несколько сглажена в береговой зоне и зоне разрушения. В данном случае интерпретация результатов более легкая. Большая отсортированность по мере увеличения расстояния от берега отмечается в зоне разрушения, охватывающей ряд других участков [185]. Нерегулярность отсортированности по площади заметна вдоль правой стороны карты в части, удаленной от берега, которую можно попытаться интерпретировать как локальные концентрации, обусловленные действием дождевых потоков по линии сочленения впадины и дюны (фиг. 17.18).

Хотя такая интерпретация и субъективна, авторы считают, что значительный контраст между картой, построенной по значениям H_r , и картой стандартных отклонений наглядно характеризует преимущества H_r как меры сортировки материала по площади, особенно при решении задач, подобных описанной. Следует отметить, что в тех случаях, когда подразделения (в этом примере — классы размеров) осмысленно упорядочены, величина H_r (и дисперсия) нечувствительна к таким важным эффектам, как полимодальность. Возможно, что наилучшее решение при этом будет заключаться в отыскании дополнительных параметров, так как невозможно достаточно полно описать ситуацию с единственной характеристикой.

В основном статистическим анализом пользуются в тех случаях, когда необходимо получить числовые характеристики. Поэтому мы начнем с предположения о наличии числовых данных. Согласно экспериментам (возможно, что это утверждение рискованно), тщательные измерения морфологических характеристик бывают особенно необходимы при изучении низших таксономических уровней. За основу различий между особями на более высоких таксономических уровнях принято наличие или отсутствие одного морфологического признака или комбинации нескольких признаков.

Например, наша задача заключается в следующем.

Допустим, что мы обнаружили разнообразные ископаемые остатки. Заметим, что вопросы биостратиграфии здесь не рассматриваются. Вполне возможно, что будут установлены таксономически близкие по крайней мере три новых рода и значительно большее число новых видов. Для простоты предположим, что все индивиды можно отнести к одному семейству на основе четко выраженных главных морфологических характеристик. В связи с тем что было условлено касаться лишь количественной стороны вопроса, мы будем рассматривать его именно с этих позиций.

ЧТО ИЗМЕРЯТЬ?

Чаще всего приходится иметь дело с остеометрическими исследованиями или, как в случае беспозвоночных, с сохранившимися частями скелета. Однако при соответствующем выборе измеряемых признаков косвенно может приниматься во внимание и анатомия мягких частей организма, например расстояние от начала до места прикрепления мышцы или расстояние между различными отверстиями, через которые проходили нервы.

К выбору измеряемых признаков при решении таксономической задачи можно подойти по-разному.

1. Если полученные *a priori* сведения, определяющие выбор измеряемых признаков, отсутствуют, измеряются все признаки, которые только можно измерить.

Эти измерения необходимы, потому что, если измеряемые признаки невозможно выбрать заранее, при таком подходе можно будет изучить даже те зависимости, которые не подзревались.

Непрактичность такого подхода, дающего в результате огромное число характеристик, очевидна. Даже если допустить, что можно преодолеть огромную трудоемкость анализа, все равно возникнут большие затруднения при интерпретации.

2. Для измерений выбирают только те части ископаемого организма, которые возможно хотя бы приблизительно морфологически интерпретировать. Интерпретировать можно на основе функционального взаимодействия между твердыми и мягкими частями организма или на основе тех изменений частей скелета, которые не зависят от мягких тканей.

а) При использовании первой группы упомянутых признаков необходимо до их выбора располагать возможно более детальными сведениями об анатомии мягких тканей (нервов, мышц, циркуляторной и пищеварительной систем). Такой подход применим к тем животным, у которых мягкие ткани в значительной степени не зависят от частей скелета, как, например, у семейства фораминифер.

б) Выделить подходящие характеристики частей скелета, которые не отражают функциональных связей с мягкими тканями, — задача весьма трудная. В случае давно вымерших форм, для которых отсутствуют близкие современные аналоги, наилучшая возможная интерпретация обычно основана на «грамотных догадках». В подобной ситуации для выбора измеряемых признаков должны служить основой сведения *a priori*, полученные палеонтологом.

Предположим, например, что рассматриваются два вида палеозойских брахиопод рода *Chonotes*. В обоих случаях замочная линия осложнена разнообразными скульптурными деталями. Кроме того, у обоих видов на внешней поверхности створок расположены выступы, количество которых сильно меняется. Предположим, устанавливать зависимость между скелетными формами и мягкими тканями, что неизбежно приводит к потере сведений об анатомическом строении, нет оснований. Если изменчивость скульптуры замочной линии внутри каждой группы незначительна, между этими группами возможны сильные расхождения. В этом случае целесообразно проводить измерение и статистическое сравнение. Если выступы на поверхности ство-

рок заметно меняются внутри групп, интерпретация результатов количественного анализа осложняется.

3. Выбираются те характеристики, ценность которых можно признать при таксономическом разделении наиболее близких форм¹. Отбор и отбраковка измеряемых признаков в данном случае делаются на основании пунктов 1 и 2. Главное затруднение, связанное с этим подходом, заключается в возможности пропустить информативные признаки, которые или уже были пропущены исследователем при изучении наиболее близких форм, или у данных форм отсутствовали.

Подводя итог, можно отметить, что в настоящее время общего подхода к выбору признаков, который бы гарантировал охват исследованием частей скелета и одновременно являлся бы практичным, не существует. Авторы предпочитают второй подход по двум причинам. В большинстве известных авторам случаев описываются результаты измерения признаков, выбранных по третьему методу в сочетании с первым методом, в котором предусмотрены основы для выбора наиболее полезных признаков из их большой совокупности. При этом подходе будет ошибкой исключить из рассмотрения очень важные признаки, которые до проведения обработки данных остались неинтерпретируемыми, что иногда делает этот подход несостоятельным.

ПОВТОРЯЕМОСТЬ

Линейные измерения следует производить только между такими точками, положение которых точно определено и в том случае, если повторные измерения дали то же расстояние. Таким образом, измеренное расстояние между швами на черепе позвоночного животного нельзя считать достаточно точным. Дело в том, что расстояния должны измеряться не между линиями, а между точками. Степень точности частично зависит и от типа измеряемых признаков. Однако при этом большую роль играет округление результатов измерения. Округление производится по общему правилу: если результат измерения выражен нецелым числом наименьших единиц, так что дробная часть меньше половины такой единицы, округление производится в меньшую сторону; если же дробная часть больше половины — в большую сторону. В тех случаях, когда дробная часть равна половине наименьшей единицы, округление производится в четную сторону. При этом не менее важно соблюдать четко определенные условия эксперимента в их строгой последовательности.

¹ C. D. Michener, R. R. Sokal, A quantitative approach to a problem in classification, *Evolution*, XI, 1957.

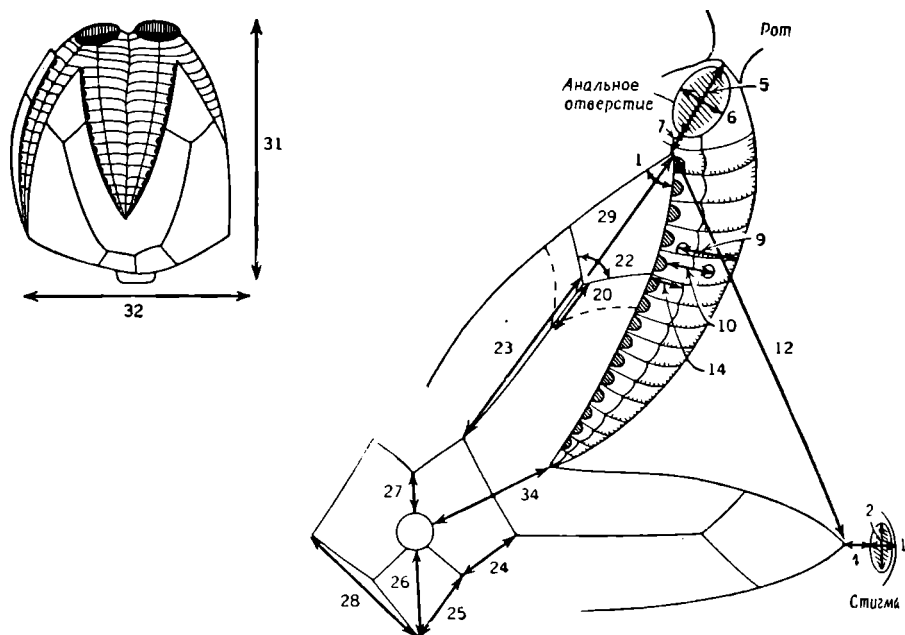
Каких-либо работ, посвященных вопросам изменчивости, в которых бы говорилось о значении персонального фактора в палеобиометрических задачах, насколько известно авторам, не существует. Однако с уверенностью можно утверждать, что некоторые виды измерений частей скелета сильнее каких-либо других измерений зависят от персонального фактора.

ТИПЫ ИЗМЕРЕНИЙ ЧАСТЕЙ СКЕЛЕТА

Примеры

Приведенный ниже список типов измерений нельзя считать полным.

1. *Линейные измерения.* Например, на черепе белки [202].
2. *Подсчеты.* Например, число ребер спинной створки *Spirifer pennatus*.



Фиг. А.1. Плоская диаграмма *Pentremites*. Числами обозначены те признаки, которые измерялись при изучении этой формы (см. табл. А.1). Показаны не все признаки, приведенные в табл. А.1).

3. *Углы.* Например, у *Spirifer pennatus* измеряется угол между линией замкового шва и краем вентрального синуса.

4. *Площадь*. Например, наружная площадь передней кости, которая измеряется планиметром.

5. *Объем*. Например, объем черепа белки, который измеряется путем заполнения зернами.

6. *Сферичность*. Например, у брахиопод измеряется степень сходства со сферой (детальную серию измерений *Pentremites calyx* см. табл. А.1, фиг. А.1).

Вернувшись к поставленной задаче, сформулируем первую основную предпосылку.

Будем считать наименьшей различимой таксономической единицей такую совокупность, в которой индивидуумы отличаются статистически несущественно один от другого по всем рассмотренным морфологическим характеристикам (за исключением стадий линьки, а также стадий морфологического роста, как, например, у фораминифер).

Эти наименьшие единицы приняты за исходные элементы для построения таксономических единиц более высокого порядка. Следует отметить, что они не имеют генетического значения.

Введем вторую основную предпосылку. Каждая особь представляет собой выборку, состоящую из одного элемента и взятую из неизвестной статистической совокупности. Рассмотрим термин *биологическая совокупность*, или *популяция*.

Термин *биологическая совокупность*, или *популяция*, применяется при описании группы родственных особей. Например, вид можно рассматривать как биологическую совокупность, члены которой могут производить жизнеспособное потомство. Подразделения вида описываются как популяции, характеризующиеся физическими различиями, на основании которых производится географическое разграничение. При этом могут быть использованы даже весьма тонкие различия. Действительно, термин *популяция* означает самые различные совокупности — от непосредственно генетических подразделений, связанных с наличием или отсутствием отдельного признака, до географических районов, например совокупности горных хребтов вулканического острова.

Биометрия основана на тех же понятиях. Известно, что популяция — это совокупность объектов с одним или более общими признаками. Однако в биометрии это положение еще более абстрагировано. Например, набор чисел, полученный в результате измерения некоторого общего для всех индивидов признака, очень удобно рассматривать как популяцию. При такой абстракции легко перейти к таким понятиям, как функция распределения популяции, параметры популяции и т. п. Все это без особого труда можно использовать при рассмотрении одновременно нескольких признаков. Естественно, что если мы примем вторую предпосылку, она приведет нас к необходимости

Измерения *Pentremites*
(по Олсону и Миллеру [202])

№ измерения	Описание
1	Внешний край стигмы по краю орального отверстия
2	Ширина стигмы
3	Расстояние между двумя стигмами
4	Обнаженный конец дельтоидной пластинки в оральном отверстии по линии, разделяющей стигму на две части
5	Внешний край анального отверстия по оральному краю
6	Ширина анального отверстия
7	Обнаженный конец дельтоидной пластинки с внешней стороны анального отверстия
8	Длина амбулякральной борозды
9	Длина пищевой борозды
10	Длина дыхательной борозды
11	Расстояние между соседними амбулякральными гребнями
12	Ширина амбулякра, измеренная как расстояние между кошками соседних дельтоидных пластинок
14	Ширина внутренних и внешних пластинок, примыкающих к ланцетовидной пластинке
15	Ширина ланцетовидных пластин
20	Ширина дельтоидной пластинки, прикрытой радиальными пластинками
21	Угол вершины дельтоидной пластинки
22	Угол между сторонами радиальных пластинок, обнажающих дельтоидную
23	Длина межлучевого шва
24	Длина основного лучевого шва, ближайшего к непарной пластинке
25	Длина шва лучевой непарной пластинки
26	Длина непарной пластинки, измеряемой от середины основания до вершины
27	Длина базального шва
28	Расстояние между ближайшими углами лучевых пластин с непарной базальной стороны
29	Длина открытой части дельтоидной пластинки
30	Ширина дельтоидной пластинки, измеряемая от концов латерального края пары лучевых пластин
31	Общая длина чашечки
32	Максимальная ширина чашечки
34	Основание амбулякра в основании чашечки
35	Число рядов внутренних и внешних пластин, примыкающих к ланцетовидной пластинке

формулировать статистические гипотезы с последующим выбором подходящих критериев для их проверки (при этом весьма важно сделать обзор принятых определений и объяснить, что понимается под термином *популяция*). Таким образом, задача распадается на две стадии: формулировка палеонтологических гипотез с последующим переводом всех формулировок или некоторой их части на точный статистический язык и выбор соответствующего статистического критерия.

Согласно палеонтологической гипотезе, пробное разделение рассматриваемой совокупности индивидуумов на группы является правильным.

1. Первый подход к задаче зависит от наличия общих сведений и навыка палеонтолога, располагающего особи в группы на основе субъективно определяемого сходства.

2. Другой подход основан на выборе таких характеристик, которые позволили бы построить группы особей. Эти группы должны обладать следующим свойством: изменчивость характеристик внутри рассматриваемых групп должна быть небольшой, а между группами — значительной. Пример практического применения такого подхода демонстрировал Вестолл¹, который рассматривал гистограммы отношения толщины к длине раковин пермских брахиопод. Полученную форму гистограмм этот автор взял за основу дальнейшего анализа. Двухвершинные гистограммы Вестолл был склонен интерпретировать скорее как следствие полового диморфизма, чем смешение совокупностей особей двух видов. Если бы он рассматривал каждую характеристику порознь, это его привело бы к жестким ограничениям в поиске формальной процедуры при изучении групп. Таким образом, при первом подходе к задаче в качестве основы решения будет часто приниматься комбинация характеристик формы и т. п. Даже если все последующие статистические гипотезы можно будет легко сформулировать и проверить, изучение лишь одной характеристики приведет к искусственным палеонтологическим гипотезам.

¹ T. S. Westoll, Some aspects of growth studies in fossils, Proc. Roy. Soc., B 137, 1950.

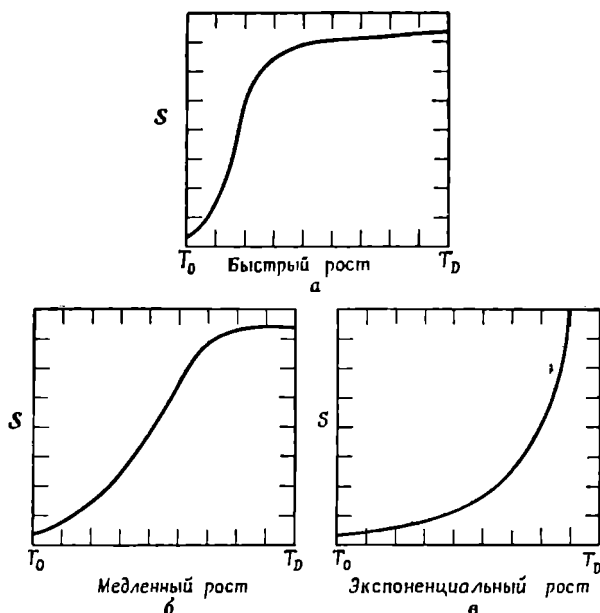
Первая тема этого раздела посвящена природе популяций. Под термином *популяция* в данном случае имеется в виду наименьшая гомогенная совокупность ископаемых организмов, которую можно установить. В нее входят все сохранившиеся члены биологической популяции, представленные в настоящее время ископаемыми остатками. При полевых исследованиях из этой популяции отбирается выборка. По выборке делаются выводы не о биологической популяции, объединяющей совокупность животных в период их жизни, а о популяции всех сохранившихся до настоящего времени ископаемых индивидов.

В связи с этим Олсон¹ отмечал, что переход от выводов, сделанных по статистикам, полученным на основании выборок из ископаемой популяции, к породившей ее биологической популяции нередко весьма труден. Последнее связано с тем, что наиболее мелкие особи (наиболее молодые) сохраняются хуже крупных, что приводит к нарушению первичного соотношения размеров. Олсон показал также, что в некоторых случаях переход от ископаемой выборки к биологической популяции совершенно необоснован.

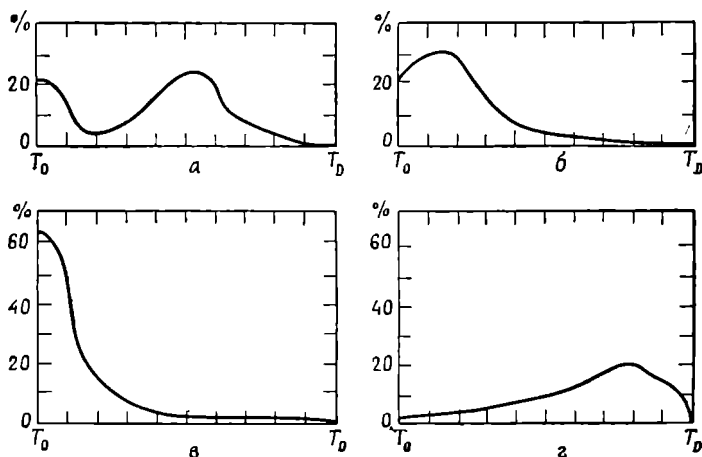
Выразив графически комбинацию размера и частоты, рассматриваемых как функции времени, Олсон построил модель исходной биологической популяции (фиг. В.1 и В.2; см. также работу Куртена²). В обоих случаях выбор функции распределения основан на биологических предположениях. Таким образом, существуют различные теоретические формы функций распределения размеров особей в зависимости от времени, например соответствующих медленному, быстрому и экспоненциальному росту (фиг. В.1). Аналогичные теоретические функции распределения частот в зависимости от времени (возраста) включают

¹ См. общий обзор в работе J. B. S. Haldane, *Biometry, Sankhya*, 16, № 3, 1956.

² B. Kurten, *Population dynamics*, part II, 1953.

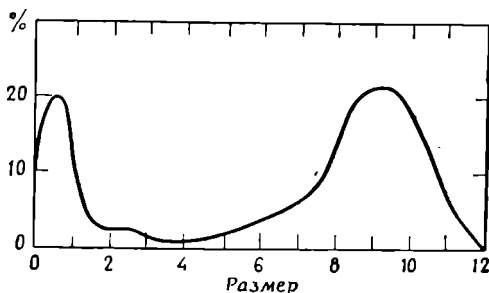


Ф и г. В.1. Три типа кривых роста, применяемых при анализе распределений в различных однородных совокупностях U_B .
 Время рождения (или время появления первых сохранившихся частей) обозначено через T_0 . Величина T_D — максимальное время жизни представителей вида. S — размер, полученный на основе большого числа измерений.



Ф и г. В.2. a — низкая ранняя смертность (при большой скорости роста); $б$ — умеренная ранняя смертность; $в$ — высокая ранняя смертность; $г$ — медленно возрастающая смертность.

также и случаи низкой ранней смертности при высокой рождаемости, медленно возрастающей смертности и т. п. Эти построения сделаны в предположении, что для единой генерации достигнута стабильность и в последующих генерациях форма функций распределения не будет изменяться. В итоге при такой комбинации получают теоретическую форму распределения размеров особей в биологической популяции (фиг. В.3)¹, которую Олсон обозначил как U_B . На фиг. В.3 дана окончательная



Фиг. В.3.

форма кривой распределения размеров особей в биологической популяции U_B , характеризующейся медленным ростом и низкой ранней смертностью. Естественно, что в большинстве случаев такое распределение ископаемых остатков не сохраняется. С другой стороны, его нельзя получить по выборке ископаемых остатков, собранных при полевых исследованиях, за исключением очень маловероятных ситуаций, при которых частоты, соответствующие даже самым маленьким особям, остались неизменными. Авторы считают, что Олсон затронул проблему, которая представляет настоящее бедствие для палеонтологов, желающих применять в своей работе статистические методы. Общая трудность заключается в том, что популяция подвергается сначала естественному селективному опробованию и уже во вторую очередь искусственному, т. е. человеком. Однако, установив, что подобные трудности существуют, перейдем к вопросу применения статистических методов при исследовании распределений размеров особей. При этом особое внимание мы уделим популяции ископаемых форм и физическим условиям транспортировки, осаднения и последующего диагенеза, которые в различной степени воздействуют на все ископаемые остатки.

Случай I: биоценоз

Этот термин широко используется для обозначения ископаемых совокупностей, которые захоронены в основном на месте своего обитания. Такое определение нельзя применять в точном смысле, так как все степени различий меняются в течение времени, прошедшего с момента гибели особи до того, как ее стал наблюдать палеонтолог, когда она уже стала ископаемым остатком.

Случай II: тонатоценоз

Этот термин охватывает все те ископаемые особи, совокупности которых были перенесены (например, водными потоками) с места их прижизненного распространения в другие места.

При попытке определить границы области распространения данной совокупности при жизни особей сразу же наталкиваемся на трудность. Таким образом, свободно плавающий морской организм может быть отложен в любом биоценозе или тонатоценозе донных форм. Для области распространения биоценозов такого типа будет характерно широкое развитие морских осадочных образований. Точно так же широко распространенные биоценозы сухопутных плотоядных в областях своего развития могут характеризоваться наличием субаэральных отложений.

Рассмотрим фацию осадочных образований, одновременно используя результаты изучения осадка и животных. Несмотря на детальный анализ и разделение осадочных образований на фации, никоим образом нельзя считать, что осадочные образования хорошо изучены, тем более что эта область исследования является относительно новой.

Чтобы можно было сделать обоснованное сравнение выборок из различных областей согласно современному учению о фациях, необходимы следующие условия:

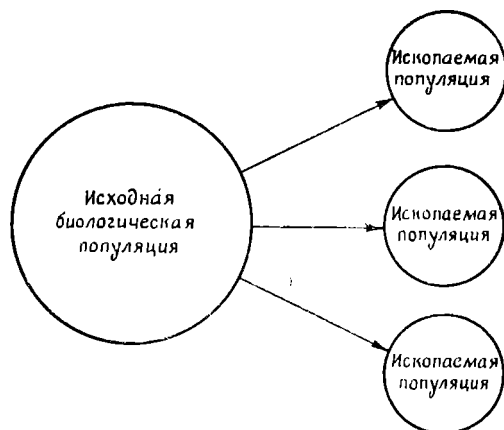
1. Фации должны быть весьма сходными между собой по тем физическим свойствам, которые влияют на транспортировку и захоронение.

2. Литологический состав сопоставляемых отложений должен соответствовать последующей геологической истории.

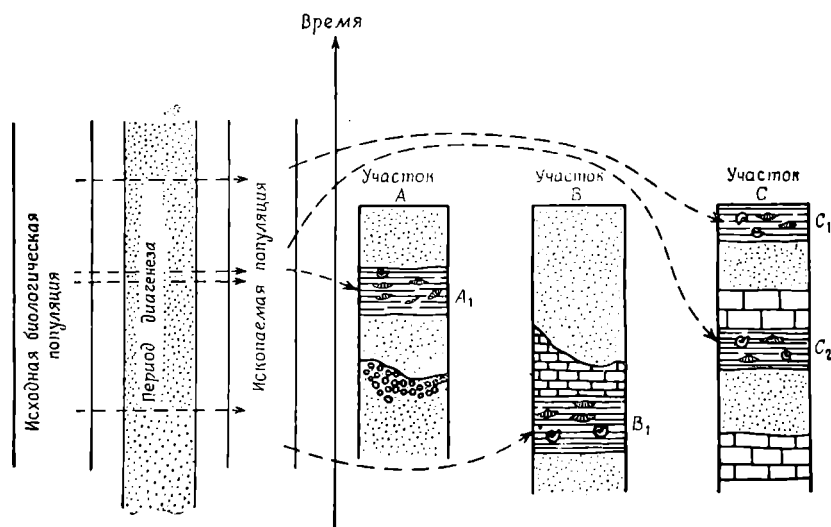
3. Результаты диагенетического воздействия должны быть незначительными (например, процессы растворения или вторичного замещения, такого, как окремнение).

4. Геологические временные различия не должны быть большими.

Если выполняются эти условия, изменения групп, принадлежащих единой биологической популяции и характеризующихся после захоронения сходной геологической историей, окажутся сходными; это сходство будет касаться частотных распределе-



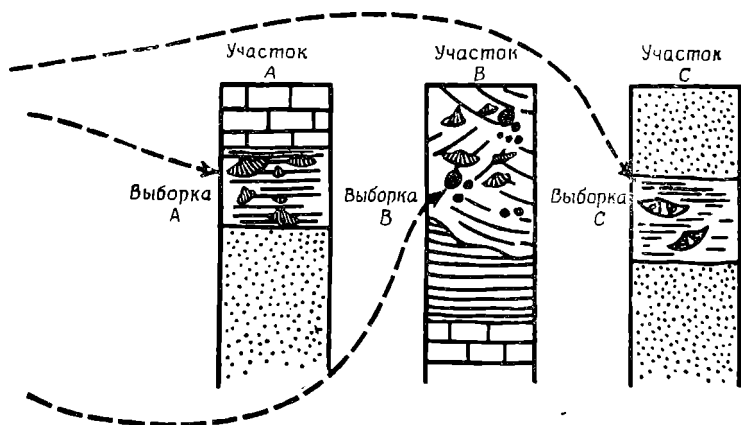
Ф и г. В.4.



Ф и г. В.5. Зависимость между исходной биологической популяцией, ископаемой популяцией и местом захоронения. A_1 , B_1 , C_1 и C_2 — выборки из ископаемой популяции, которая в свою очередь является выборкой из биологической популяции. Как исходная биологическая, так и ископаемая популяция непрерывны во времени, но не в пространстве. Палеонтологические гипотезы: выборки A_1 , B_1 , C_1 и C_2 (все они взяты в сходных литологических условиях; они явно не были переложены) отобраны из одной популяции (ископаемой) и представлены полным набором онтогенетических стадий. Статистические гипотезы: например, все выборочные средние и оценки дисперсий различаются несущественно.

ний как исходной биологической популяции, так и групп, образующих ископаемую совокупность.

Рассмотрим следующую схему. Исходная биологическая популяция существует некоторый период времени. Ее можно рассматривать как аналогию открытой системы, характеризующейся постоянным появлением новых членов (рождаемость) и постоянным выбыванием некоторых из них (смертность). Хотя есть основание рассматривать этот процесс как непрерывный,



Фиг. В.6. Зависимости те же, что и на фиг. В.5, но при различных местах захоронения.

Палеонтологические гипотезы: выборка А взята в глинистых сланцах и представлена всеми онтогенетическими стадиями. Экземпляры хорошей сохранности и явно не переотложены. Выборка В содержит плохо сохранившиеся экземпляры и наряду с ними крупные формы хорошей сохранности. Она взята из песчаных конгломератов, в которых происходило активное переотложение и сортировка материала. Выборка С взята из выветрелых сланцев. Процессы растворения исказили исходную картину, и поэтому были найдены только крупные представители.

Статистические гипотезы: оценить средние и дисперсии; проверить те же нулевые гипотезы, которые указаны в подписи к фиг. В.5.

для простоты остаточные члены можно принять за «макроколичество», т. е. совокупность, образованную множеством дискретных событий. Такое дискретное представление совокупности основано на том, что выборки, собранные в поле палеонтологами, являются дискретными во времени и пространстве (фиг. В.4). Таким образом, в течение конкретного периода времени исходной биологической популяцией обусловлено образование подгрупп погибших индивидуумов, подверженных транспортировке, захоронению и диагенезу. В связи с локальным изменением физических условий особи могут быть перенесены на различные расстояния, затем переотложены и подвергнуты диагенетическим воздействиям в новой последовательности. В дан-

ном случае задача заключается в том, чтобы установить связь ископаемых групп и всей совокупности этих групп с ископаемой популяцией, что должно быть отражено в палеонтологически правильной формулировке гипотез.

Множество ситуаций можно представить наглядно. На фиг. В.5 и В.6 показаны две из них. Более подробно эти вопросы рассмотрены в работах Миллера и Олсона [180] и Джонсона [134].

Теперь рассмотрим некоторые общие модели ископаемой популяции, которые связаны со статистической или какой-либо другой количественной обработкой данных. В этих моделях рассматривается изменение только одной характеристики во времени.

К Л А С С I

Этот класс объединяет все те ископаемые популяции, для которых палеонтологи считают возможным предположить в качестве теоретического распределения нормальный закон (например, для единых онтогенетических стадий или для окаменелостей биоценоза). К этому классу относятся также те распределения, которые не очень сильно отличаются от нормального, например имеющие форму, близкую к симметричной. При изучении ископаемых популяций класса I нужно учитывать следующее:

1. Данная одномерная характеристика распределена нормально; несколько изучаемых характеристик описываются по многомерному нормальному закону.

2. Одной из основных характеристик одномерного нормального распределения является среднее μ или в случае многомерного распределения набор средних $\mu_1, \mu_2, \dots$.

3. Вторая основная характеристика одномерного нормального распределения — мера рассеяния, или дисперсия σ^2 . Для многомерного случая (если признаки независимы) используется набор дисперсий $\sigma_1^2, \sigma_2^2, \dots$.

4. Толерантные пределы.

5. Коэффициенты вариации $V_1, V_2, \dots$.

Так как ископаемая популяция представлена только выборкой, каждый параметр следует охарактеризовать соответствующей выборочной статистикой. В тех случаях, когда из популяции взято несколько выборок, оценки можно улучшить.

Следует отметить, что, помимо оценок параметров, желательно приводить соответствующие им оценки стандартных отклонений, что позволило бы судить об относительной точности.

Для различных видов наилучших оценок особого внимания заслуживает построение параметров популяции как обладающих большей информационной способностью. Помимо трудностей, связанных с отсчетом времени и районированием, реше-

ние таких фундаментальных задач биометрии, как изучение относительных скоростей эволюции и типов эволюционных процессов, осложняется недостаточностью сравниваемых серий отдельных выборок. Кроме простых статистических приемов, при решении этих задач можно применять более сложные методы, такие, как дискриминантный анализ, построение обобщенного расстояния и факторный анализ (например, ρF -метод).

В работе Бармы [40] можно найти интересные примеры приложения одномерного последовательного анализа при определении видов.

К Л А С С 11

К этому классу относятся те популяции ископаемых видов, относительно которых палеонтолог имеет достаточно серьезные основания *a priori* не принимать нормальный закон в качестве модели распределения.

1. Палеонтолог обнаружил, что некоторое преобразование (например, логарифм или арксинус) исходных данных приводит к распределению, близкому к нормальному, что позволило отнести его к классу I.

2. Палеонтолог не хочет делать никаких предположений о характере распределения.

В случаях, подобных второму, можно воспользоваться непараметрическими критериями, для которых форма распределения не играет большой роли.

В данной книге приведены непараметрические методы проверки гипотезы о равенстве средних, способы изучения корреляционных зависимостей, критерии знаков, а также приемы сравнения функций распределения. Большинство непараметрических критериев, за исключением критерия χ^2 , является новым для биометрии. Работы, в которых освещены эти вопросы, приведены в конце книги.

Авторы считают, что значение непараметрических методов в биометрии должно возрастать, так как они не требуют каких-либо предположений о виде функций распределения.

Рассмотрим распределение отношения

$$z = \frac{x_1/n_1}{x_2/n_2},$$

где x_1 и x_2 независимы и распределены по закону χ^2 с числом степеней свободы, равным n_1 и n_2 соответственно.

Распределение случайной величины

$$y_1 = \frac{x_1}{n_1}$$

может быть получено следующим образом:

$$y_1 = \frac{x_1}{n_1}, \quad dy_1 = \frac{dx_1}{n_1}.$$

Из равенства

$$\hat{f}(y_1) = n f(ny_1)$$

выводим, что плотность $\hat{f}(y_1)$ величины y_1 равна

$$\frac{1}{\left(\frac{n_1-2}{2}\right)! \left(\frac{2}{n_1}\right)^{n_1/2}} y_1^{(n_1/2)-1} e^{-n_1 y_1/2}, \quad y_1 > 0,$$

т. е. она представляет собой γ -распределение с параметрами $\beta = 2/n_1$, $\alpha = (n_1 - 2)/2$. Аналогично для $y_2 = x_2/n_2$

$$f(y_2) = \frac{y_2^{(n_2/2)-1} e^{-n_2 y_2/2}}{\left(\frac{n_2-2}{2}\right)! \left(\frac{2}{n_2}\right)^{n_2/2}}, \quad y_2 > 0.$$

По предположению о независимости y_1 и y_2 их совместная плотность равна

$$g(y_1, y_2) = f(y_1) f(y_2) = \frac{y_1^{(n_1/2)-1} y_2^{(n_2/2)-1} \exp \left[-\frac{n_2 y_2 + n_1 y_1}{2} \right]}{\left(\frac{n_1-2}{2}\right)! \left(\frac{n_2-2}{2}\right)! \left(\frac{2}{n_1}\right)^{n_1/2} \left(\frac{2}{n_2}\right)^{n_2/2}}.$$

Мы можем теперь вернуться к задаче определения плотности случайной величины $z = \frac{x_1/n_1}{x_2/n_2} = y_1/y_2$. Определим новые переменные равенствами $y_1 = v$, $y_2 = v/z$. Якобиан преобразования равен $J = y_2^2/y_1$. Из соотношения

$$g(y_1, y_2) dy_1 dy_2 = h(z, v) dz dv,$$

которое можно переписать в виде

$$g(y_1, y_2) J = h(z, v),$$

получаем

$$h(z, v) = C z^{-2 - \frac{n_2}{2} + 1} v^{\frac{n_1 + n_2}{2} - 1} \exp \left[-\frac{\frac{n_2}{2} + n_1}{z} v \right].$$

Интегрируя обе части последнего равенства по v в пределах от 0 до ∞ и замечая, что

$$\int_0^{\infty} v^{p-1} \exp[-av] dv = \frac{\Gamma(p)}{a^p},$$

находим для плотности вероятности $f(z)$ случайной величины z выражение

$$f(z) = C_1 \frac{z^{\frac{n_1}{2} - 1}}{(n_2 + n_1 z)^{\frac{n_1 + n_2}{2}}}, \quad z > 0,$$

где

$$C_1 = \frac{n_1^{n_1/2} n_2^{n_2/2} \left(\frac{n_1 + n_2 - 2}{2} \right)!}{\left(\frac{n_1 - 2}{2} \right)! \left(\frac{n_2 - 2}{2} \right)!}.$$

Распределение с плотностью $f(z)$ называют *F-распределением с числом степеней свободы* (n_1, n_2) .

Вывод χ^2 -распределения

Пусть случайная величина x распределена по закону $N(0, 1)$, т. е. имеет плотность вероятности

$$f(x) = \frac{1}{(2\pi)^{1/2}} e^{-x^2/2}, \quad -\infty < x < +\infty. \quad (\text{D. 1})$$

Каково распределение случайной величины $z = x^2$? Неравенство

$$z < a, \quad a > 0, \quad (\text{D. 2})$$

равносильно неравенству

$$-\sqrt{a} < x < \sqrt{a} \quad (\text{D. 3})$$

и, следовательно, имеет вероятность

$$\frac{1}{(2\pi)^{1/2}} \int_{-\sqrt{a}}^{\sqrt{a}} e^{-u^2/2} du = \frac{2}{(2\pi)^{1/2}} \int_0^{\sqrt{a}} e^{-u^2/2} du.$$

Произведя в последнем интеграле замену

$$v = u^2, \quad dv = 2u du = 2\sqrt{v} du,$$

приведем его к виду

$$\frac{2}{(2\pi)^{1/2}} \int_0^a \frac{e^{-v/2} dv}{2\sqrt{v}}. \quad (\text{D. 4})$$

Таким образом, вероятность неравенства (D. 2), с одной стороны, дает нам значение функции распределения величины z в точке a , с другой — эта вероятность равна (D. 4). Дифференцируя (D. 4) по a , получим значение плотности вероятности случайной величины z в точке $a > 0$:

$$g(a) = \frac{1}{(2\pi)^{1/2}} \frac{e^{-a/2}}{\sqrt{a}}, \quad 0 < a < \infty. \quad (\text{D. 5})$$

Ясно, что в других случаях

$$g(a) = 0.$$

Рассмотрим производящую функцию моментов, соответствующую плотности $g(z)$, т. е.

$$Ee^{tz} = \int_0^{\infty} \frac{1}{(2\pi)^{1/2}} \frac{e^{-(z/2-tz)}}{\sqrt{z}} dz = \int_0^{\infty} \frac{1}{(2\pi)^{1/2}} \frac{e^{-(z/2)(1-2t)}}{\sqrt{z}} dz.$$

По Ходгману¹, находим

$$m(t) = (1 - 2t)^{-1/2}. \quad (\text{D. 6})$$

Заметим, что эта функция является частным случаем производящей функции моментов γ -распределения

$$h(x) = \frac{1}{\alpha! \beta^{\alpha+1}} x^{\alpha} e^{-x/\beta}, \quad x > 0.$$

Действительно, последняя равна

$$m(t) = (1 - \beta t)^{-(\alpha+1)},$$

что при $\alpha = -1/2$, $\beta = 2$ сводится к (D. 6).

Рассмотрим теперь производящую функцию моментов для случайной величины:

$$\chi^2 = \sum_{i=1}^n x_i^2 = \sum_{i=1}^n z_i.$$

Напомним, что плотность каждой из величин z_i определяется формулой

$$g(z_i) = \begin{cases} \frac{1}{(2\pi)^{1/2}} \frac{e^{-z_i/2}}{\sqrt{z_i}} dz_i, & i = 1, 2, \dots, n, \quad 0 \leq z_i < \infty, \\ 0 & \text{во всех других случаях.} \end{cases} \quad (\text{D. 7})$$

Допустим, что величины z_i независимы в статистическом смысле (гл. 1, стр. 25). Тогда

$$E \left[\exp \left(t \sum_{i=1}^n z_i \right) \right] = E(e^{tz_1}) E(e^{tz_2}) \dots E(e^{tz_n}). \quad (\text{D. 8})$$

Производящая функция моментов каждой из величин z_i равна $(1 - 2t)^{-1/2}$, и потому производящая функция моментов для χ^2 есть

$$E[e^{t\chi^2}] = (1 - 2t)^{-n/2}. \quad (\text{D. 9})$$

¹ C. D. Hodgman, Handbook of chemistry and physics, eq. 429, Chem Rubb. Publ. Co., Cleveland, Ohio, 1956.

Это, очевидно, производящая функция моментов γ -распределения с параметрами

$$\alpha = \frac{n-2}{2}, \quad \beta = 2.$$

Следовательно, случайная величина χ^2 , равная сумме квадратов n независимых случайных величин с распределением $N(0, 1)$, имеет плотность вероятности

$$f(\chi_n^2) = \begin{cases} \frac{1}{\left(\frac{n-2}{2}\right)!} \frac{1}{2^{n/2}} (\chi_n^2)^{(n-2)/2} e^{-\chi_n^2/2}, & \chi_n^2 > 0, \\ 0 & \text{во всех других случаях.} \end{cases}$$

Математическое ожидание и дисперсия χ_n^2

Из (D.9) получаем

$$m'(t) = \frac{2n}{2(1-2t)^{n/2+1}},$$

$$m'(0) = n = \mu'_1,$$

$$m''(t) = 2\left(\frac{n}{2} + 1\right)n(1-2t)^{-(n/2)-2},$$

$$m''(0) = 2\left(\frac{n}{2} + 1\right)n = n(n+2) = \mu'_2,$$

$$\mu_2 = \mu'_2 - (\mu'_1)^2 = n(n+2) - n^2 = 2n.$$

Следовательно,

$$E[\chi_n^2] = n,$$

$$\text{var}(\chi_n^2) = \mu_2 = 2n.$$

Таким образом, величина, распределенная по закону χ^2 с n степенями свободы, имеет математическое ожидание n и в два раза большую дисперсию $2n$.

Эта функция имеет вид

$$F(\chi_n^2) = \int_0^{\chi_n^2} \frac{1}{\left(\frac{n-2}{2}\right)! 2^{n/2}} (\chi_n^2)^{(n-2)/2} e^{-\chi_n^2/2} d\chi_n^2.$$

Здесь $F(\chi_n^2)$ зависит еще и от n числа степеней свободы. Для данного n

$$P\{b \leq \chi_n^2 \leq a\} = a - F(b) = P\{\chi_n^2 \leq a\} - P\{\chi_n^2 \leq b\}.$$

Следовательно, для вычисления вероятности подобных неравенств мы должны знать a , b и n . Существуют достаточно удобные для использования таблицы этой функции распределения. Поэтому вероятности событий, связанных с χ_n^2 , можно определить по таблицам, если известно число степеней свободы. По этим же таблицам можно определять значения χ_n^2 , соответствующие заданным вероятностям. Читатель должен обращать внимание на то, приведены ли в таблице значения

$$\int_0^{\chi_n^2} f(\chi_n^2) d\chi_n^2 = F(\chi_n^2)$$

или значения

$$\int_{\chi_n^2}^{\infty} f(\chi_n^2) d\chi_n^2,$$

т. е.

$$1 - F(\chi_n^2).$$

„Аддитивное свойство“ случайных величин, распределенных по закону χ^2 , весьма важно, например, в дисперсионном анализе. Рассмотрим закон распределения случайной величины

$$v = \chi_{n_1}^2 + \chi_{n_2}^2,$$

где

$$\chi_{n_1}^2 = \sum_{i=1}^{n_1} x_i^2$$

и

$$\chi_{n_2}^2 = \sum_{j=1}^{n_2} x_j^2$$

и где величины x взаимно независимы и распределены по закону $N(0, 1)$. Производящая функция моментов для v равна

$$m(t) = E \exp(tv) = E \exp(t\chi_{n_1}^2 + t\chi_{n_2}^2).$$

В силу независимости $\chi_{n_1}^2$ и $\chi_{n_2}^2$

$$E \exp(t\chi_{n_1}^2) \exp(t\chi_{n_2}^2) = E \exp(t\chi_{n_1}^2) E \exp(t\chi_{n_2}^2);$$

поэтому

$$E \exp(tv) = (1 - 2t)^{-(n_1 + n_2)/2}.$$

Полагая $n_1 + n_2 = n$, мы видим, что $m(t) = (1 - 2t)^{-n/2}$ и является производящей функцией моментов для распределения величины χ_n^2 . Следовательно, если случайные величины $\chi_{n_1}^2$ и $\chi_{n_2}^2$ независимы и имеют распределение χ^2 с n_1 и n_2 степенями свободы соответственно, их сумма имеет распределение $\chi_{n_1 + n_2}^2$.

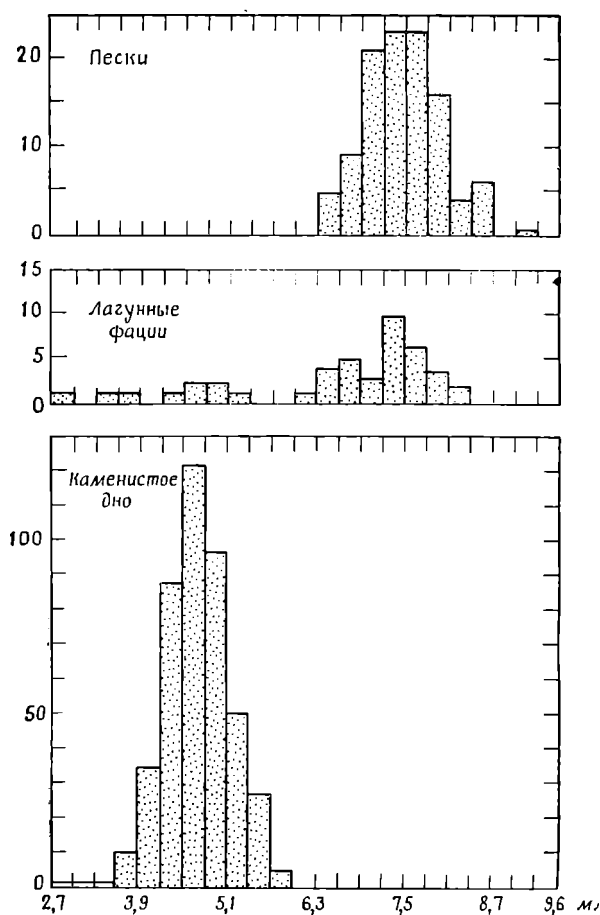
Статистика Колмогорова — Смирнова может быть использована для проверки гипотезы, согласно которой две эмпирические функции распределения соответствуют выборкам из одной и той же генеральной совокупности. Используя графики эмпирических функций распределения, определяем d_n , т. е. максимальное расхождение по вертикали между этими графиками. Если значение выше определенной границы, которая устанавливается по таблицам² по заданному уровню значимости, соответствующие выборки считаются взятыми из разных генеральных совокупностей.

Этот критерий имеет ряд положительных качеств. Его легко использовать, так как вся процедура, почти не требующая вычислений, сводится к графикам; большое количество выборок можно сравнивать попарно на одном и том же чертеже; критерий непараметрический — не нужно делать каких-либо специальных предположений о теоретическом распределении (что, по мнению авторов, особенно важно для геолога); предположение о нормальности часто не только оказывается неоправданным (за недостатком информации), но теоретически бывает неприемлемым: ряд теоретических выводов приводит к распределениям, существенно отличающимся от нормального; наконец, критерий не подвержен ограничениям, которые связаны, например, с использованием критерия χ^2 (математическое ожидание числа наблюдений в каждом классе должно быть не меньше 5).

¹ Некоторые применения статистики Колмогорова — Смирнова в геологических задачах можно найти в работах [160] и E. T. Degen s, E. G. Williams, M. I. Keith, Environmental studies of Carboniferous sediments, I. Geochemical criteria for differentiating marine from fresh water shales, Bull. A. A. P. G., 41, 1957.

² Критерий и таблицы критических значений d_n кратко рассматривают Диксон и Мессе [79, стр. 256—257].

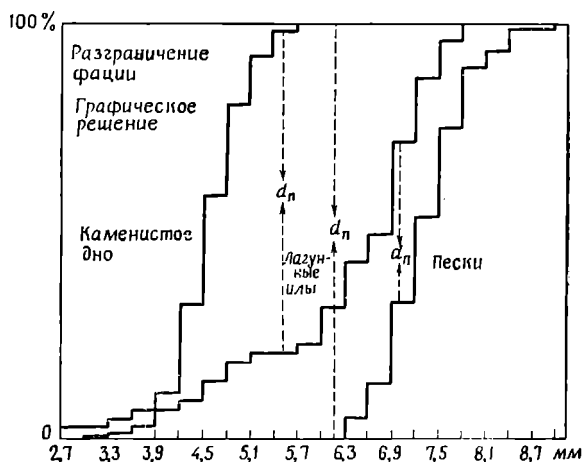
Следующий пример иллюстрирует процедуру применения критерия. Гистограммы, приведенные на фиг. G.1, дают представление о распределении раковин *Mytilus* по их размерам в трех подфациях. Мы хотим проверить гипотезу о том, что все



Ф и г. G.1.

три выборки взяты из совокупностей с одним и тем же теоретическим распределением. На фиг. G.2 нанесены все три соответствующие эмпирические функции распределения. Деления на оси ординат каждый раз определяются объемом выборок. Так, для случаев каменистого дна отрезок 0—100% делится на 431 часть, для лагунных илов — на 45 частей и т. д.

Максимальные расхождения по вертикали d_n указаны на фигуре. Отметим, что d_n для песчаного и каменистого дна равно



Ф и г. G.2.

100%. Результаты анализа сведены в табл. G.1. Для сравнения в таблице указано минимальное расхождение (при уровне значимости 0,05), которое должно восприниматься как значимое.

Таблица G.1

Результат применения критерия Колмогорова — Смирнова при сравнении распределений размеров современных *Mytilus* из трех различных фаций

Песчаное дно — илистое дно	Песчаное дно — каменистое дно	Илистое дно — каменистое дно
$d_n = 0,3933$	$d_n = 1,00$	$d_n = 0,777$

Значимое отклонение при уровне 0,05 равно 0,2407

Установлено различие

Объемы выборок: илы — 45, пески — 105, каменистое дно — 431.

Использование номограмм (фиг. G.3 и G.4)

Ниже объясняется, как следует использовать номограммы для сравнения эмпирической функции распределения с полностью определенной теоретической функцией $F(x)$. Принимаем доверительный уровень равным 0,99.

1. Согласно нулевой гипотезе, неизвестное теоретическое распределение в точности совпадает с $F(x)$.

2. Взяв число N в качестве абсциссы, находим соответствующую ординату точки, лежащей на кривой, помеченной числом 99. Пусть эта ордината будет $\epsilon_{N, 0,99}$.

3. Найденное значение $\epsilon_{N, 0,99}$ удовлетворяет соотношению $P\{D_N < \epsilon_{N, 0,99}\} = 0,99$. Если наблюдаемое максимальное расхождение D_N превосходит $\epsilon_{N, 0,99}$, мы отвергаем нулевую гипотезу на том основании, что в 99% случаев мы ожидаем при ее справедливости расхождений, меньших фактически наблюдаемых.

Номограммы также могут быть использованы при сравнении двух эмпирических функций распределения, построенных по независимым выборкам объемов n_1 и n_2 соответственно.

1. Нулевая гипотеза состоит в том, что две теоретические функции распределения, предполагаемые непрерывными, совпадают. Таким образом, если наблюдаемое максимальное расхождение слишком велико, нулевая гипотеза отвергается.

2. Вычисляется отношение $n_1 n_2 / (n_1 + n_2)$, и его значение N принимается в качестве абсциссы на соответствующем графике. Использование указанного отношения к задаче сравнения двух выборок асимптотически правильно¹.

Результаты сравнения этого метода с другими методами приведены в табл. G.2. Сопоставляются три метода: точный метод, предложенный Мессе² и рассмотренный также Гудменом³; приближенный метод, основанный на предельной формуле Колмогорова (Смирнов, 1948), и описанный выше приближенный метод, разработанный Бирнбаумом⁴.

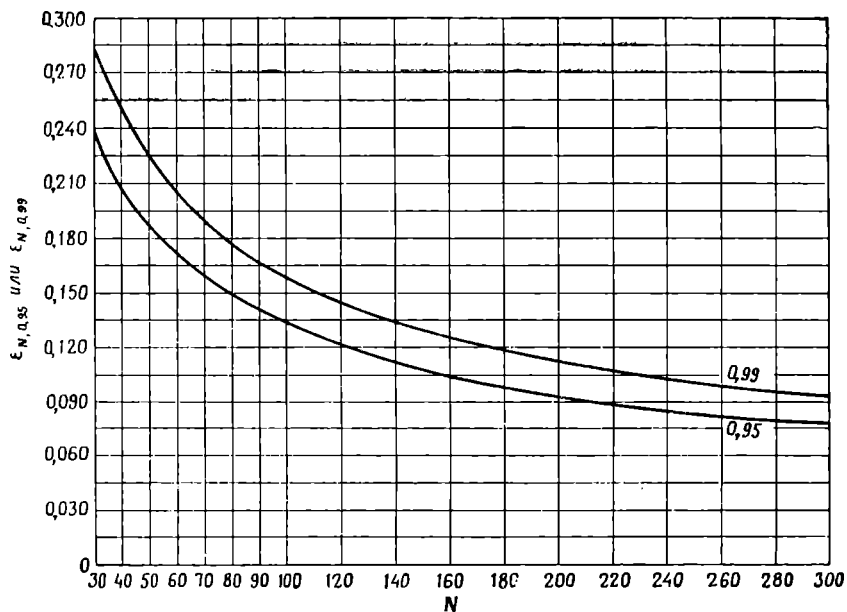
В таблице приведены найденные тремя методами значения величин $\epsilon_{n_1 n_2, 0,95}$ и $\epsilon_{n_1 n_2, 0,99}$, определяемых из уравнений $P_r\{D_{n_1 n_2} < \epsilon_{n_1 n_2, 0,95}\} = 0,95$ и $P_r\{D_{n_1 n_2} < \epsilon_{n_1 n_2, 0,99}\} = 0,99$, где $D_{n_1 n_2}$ —

¹ N. Smirnov, Table for estimating the goodness of fit of empirical distributions, Ann. Math. Stats., 19, 1948.

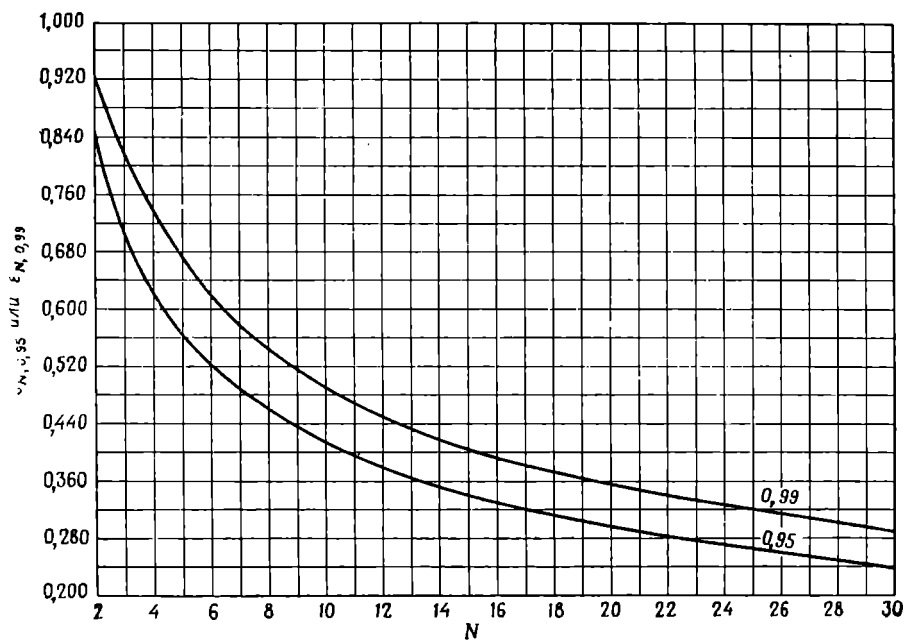
² F. J. Massey, Distribution of maximum deviation between two sample cumulative step functions, Ann. Math. Stats., 22, 1951.

³ L. A. Goodman, Kolmogorov — Smirnov tests for psychological research, Psych. Bull., 51, 1954.

⁴ Z. W. Birnbaum, Numerical tabulation of the distribution of Kolmogorov's statistic for finite samples sizes, Journ. Am. Stats. Assoc., 47, 1952.



Ф и г. G.3.



Ф и г. G.4.

Т а б л и ц а Г.2

**Сравнение результатов использования
графической аппроксимации с точными
значениями и аппроксимацией,
предложенной Смирновым**

(объемы выборок равны $n_1 = n_2$)

$n_1 = n_2$	$\varepsilon_{n_1 n_2, 0,95}$	$\hat{\varepsilon}_{n_1 n_2, 0,95}$	$K_{n_1 n_2, 0,95}$
7	0,714	0,668	0,720
10	0,600	0,566	0,607
22	0,409	0,394	0,410
33	0,333	0,326	0,333
39	0,308	0,300	0,300

$n_1 = n_2$	$\varepsilon_{n_1 n_2, 0,99}$	$\hat{\varepsilon}_{n_1 n_2, 0,99}$	$K_{n_1 n_2, 0,99}$
7	0,857	0,784	0,870
10	0,700	0,676	0,728
13	0,615	0,556	0,638
16	0,562	0,542	0,576
20	0,500	0,490	0,516
27	0,444	0,426	0,443

максимальное расхождение между эмпирическими функциями распределения, построенными по выборкам с объемами n_1 и n_2 соответственно. При этом $\varepsilon_{n_1 n_2, 0,95}$ и $\varepsilon_{n_1 n_2, 0,99}$ — точные значения, взятые из таблиц Месси; $\hat{\varepsilon}_{n_1 n_2, 0,95}$ и $\hat{\varepsilon}_{n_1 n_2, 0,99}$ — значения, найденные графически; $K_{n_1 n_2, 0,95}$ и $K_{n_1 n_2, 0,99}$ — приближенные значения, заимствованные из таблиц Смирнова.

Графики (фиг. Г.3 и Г.4) рекомендуется использовать в случаях, не требующих особой точности.

ЛИТЕРАТУРА

1. Вальд А., 1960. Последовательный анализ, М., Физматгиз.
2. Вистелиус А. Б., 1947. О корреляционной связи между апатитом и нефелином в Кукисвимчорр-Юкспорском сфеновом месторождении (Хибинские тундры), ДАН СССР, **56**, № 2.
3. Вистелиус А. Б., 1948. О распределении магния в палеозое востока Русской платформы, Второй симпозиум ВНИГРИ.
4. Вистелиус А. Б., 1949. К вопросу о механизме слоеобразования, ДАН СССР, **65**, № 2.
5. Вистелиус А. Б., 1949. К вопросу о механизме связи при слоеобразовании, ДАН СССР, **65**, № 4.
6. Вистелиус А. Б., 1950. К вопросу о связи между содержанием меди в буровых водах Азербайджана и степенью их минерализации, ДАН АзССР, **6**, № 1.
7. Вистелиус А. Б., 1951. «Еще раз о связи» — ответ С. В. Константинову, Зап. Всес. минер. об-ва, **80**, вып. 1.
8. Вистелиус А. Б., 1951. О палеозойских гальках продуктивных слоев Апшеронского полуострова, ДАН СССР, **79**, № 3.
9. Вистелиус А. Б., 1952. Природа парагенезиса некоторых компонентов нефтей Азербайджана, ДАН АзССР, **VIII**, № 1.
10. Вистелиус А. Б., 1952. Кирмакинская свита Восточного Азербайджана. Основные черты литологии, ДАН АзССР, **VIII**, № 1.
11. Вистелиус А. Б., 1953. Об обработке микроструктурных диаграмм, Зап. Всес. минер. об-ва, **82**, вып. 4.
12. Вистелиус А. Б., 1954. Пески Средней и Нижней Волги, ЛАЭМ.
13. Вистелиус А. Б., 1957. Расчленение пемых толщ по количественным минералогическим, петрографическим и химическим признакам, Зап. Всес. минер. об-ва, **86**, вып. 1.
14. Вистелиус А. Б., 1957. Региональная литостратиграфия и условия продуктивной толщи юго-восточного Кавказа, Тр. Ленингр. об-ва естествоисп., **69**, вып. 2.
15. Вистелиус А. Б., 1957. К статистике микроструктурных диаграмм, Зап. Всес. минер. об-ва, **86**, вып. 6.
16. Крамер Г., 1948. Математические методы статистики, М., ИЛ.
17. Колмогоров А. Н., 1951. Об одном механизме слоеобразования, ДАН СССР.
18. Феллер В., 1964. Введение в теорию вероятностей и ее приложения, М., изд. «Мир».
19. Хальд А., 1956. Математическая статистика с техническими приложениями. М, ИЛ.
20. Ahrens L. H., 1954. The lognormal distribution of the elements (A fundamental law of geochemistry and its subsidiary), *Geochim. et Cosmochim. Acta*, **5**.
21. Ahrens L. H., 1954. The lognormal distribution of elements, *Geochim. et Cosmochim. Acta*, **6**.

22. Ahrens L. H., 1957. Lognormal distributions, *Geochim. et Cosmochim. Acta*, **11**.
23. Allen P., 1945. Sedimentary variation: some new facts and theories, *Journ. Sed. Petrol.*, **15**.
24. Allen P., 1947. Correlation between allogenic grade size and allogenic frequency in sediments, *Journ. Sed. Petrol.*, **17**.
25. Allen P., 1949. Wealdon petrology: the top Ashdown pebble bed and the top Ashdown sandstone, *Quart. Journ. Geol. Soc., London*, **54**.
26. Allen W., 1951. Rough and ready statistical tests, Paper pres. at the Fifth An. Conv. of the Am. Soc. for Quality Control.
27. Alling H. L., 1944. Grain analysis of minerals of sand size in Ball Mills, *Journ. Sed. Petrol.*, **14**.
28. Andel van T. J. H., 1954. Roundness and shape of marine gravels from Urk (Netherlands) a comparison of several methods of investigation, *Journ. Sed. Petrol.*, **24**.
29. Antews E., 1938. Rainfall and tree growth in the Great Basin, *Am. Geogr. Soc. Spec. Publ.*, **21**.
30. Bader R., 1955. Variability and evolutionary rate in the Oreodonts, *Evolution* **IX**, **2**.
31. Bankier J. D., 1955. The theory of thin section analysis: a reply, *Journ. Geol.*, **63**.
32. Baquslavsky G. W., 1956. Statistical estimation of the size of a small population, *Sci*, **124**.
33. Brigt R. C., 1959. A paleoecologic and biometric study of the middle Cambrian Trilobite *Elrathia kingi* (Week), *Journ. Paleont.*, **33**.
34. Brinkmann R., 1929. Statistisch-Biostratigraphische Untersuchungen an mitteljurassischen Ammoniten... *Abh. Ges. Wiss. Göttingen, Math.-phys. Kl. N. F.*, **13**, Heft 3.
35. Bronowski J., 1951. Statistical methods in anthropology, *Nature*, **168**.
36. Bronowski J., Long W. M., 1952. Statistics of discrimination in anthropology, *Am. Journ. Phys. Anthropol.*, **10**.
37. Bronowski J., Long W. M., 1953. Australopithecine milk canines, *Nature*, **172**.
38. Brown H., Patterson C., 1947. The composition of meteoritic matter. I. The composition of silicate of stony meteorites, *Journ. Geol.*, **55**.
39. Bruckner W. D., 1953. *Journ. Sed. Petrol.*, **23**.
40. Burma B. H., 1948. Studies in quantitative paleontology; Some aspects of the theory and practice of quantitative invertebrate paleontology, *Journ. Paleont.*, **22**.
41. Burma B. H., 1949. Multivariate analysis — a new analytical tool for paleontology and geology, *Journ. Paleont.*, **23**.
42. Burma B. H., 1953. Studies in quantitative paleontology; An Application of sequential analysis to the comparison of growth stages and growth series, *Journ. Geol.*, **61**.
43. Cadigan R. A., 1954. Testing raphical methods of grain-size analysis of sandstones and siltstones, *Journ. Sed. Petrol.*, **24**.
44. Carozzi A., 1951. Rytmes de Sedimentation dans le Cretace Helvetique, *Geol. Rund.*, **39**.
45. Carpenter R. L., 1939. Bacterial counts in the of Crystal Lake, etc., *Journ. Sed. Petrol.*, **9**.
46. Carroll D., 1957. A statistical study of heavy minerals in sands of the South river, Augusta Country, Virginia, *Journ. Sed. Petrol.*, **27**.
47. Carter D. J., 1953. A statistical study of Operculina, *Journ. Paleont.*, **27**.
48. Cattell R. B., 1944. A note on correlation clusters and cluster search methods, *Psychometrika*, **9**, № 3.
49. Cattell R. B., 1952. Factor analysis, Harper Bros, New York.

50. Chapman S., Bartels J. Geomagnetism, 11, analysis of the data and physical theories, Oxford Univ. Press, London.
51. Chave K. E., 1954. Aspects of the bigeochemistry of magnesium; Calcareous marine organics, Jour. Geol., 62.
52. Chaves F., 1946. Linear analysis of a medium grained granite, Am. Min., 31.
53. Chaves F., 1949. Ratio correlations in petrography, Journ. Geol., 57.
54. Chaves F., 1950. Comparison of the granites of Westerly and Bradford, Rhode Island, Am. Journ. Sci., 248.
55. Chaves F., 1950. On a distribution between late magnetic and post-magnetic replacement reactions, Am. Journ. Sci., 248.
56. Chaves F., Fairbairn H. W., 1951. A test of the precision of thin section analysis by point counter, Am. Min., 36.
57. Chaves F., 1952. The finer grained calcaline granites of New England, Journ. Geol., 60.
58. Chaves F., 1954. Effect of change origin on mean and variance of two-dimensional fabrics, Am. Journ. Sci., 252.
59. Chaves F., 1954. The lognormal distribution of elements; a discussion, Geochim et Cosmochim. Acta, 6.
60. Chaves F., 1954. The theory of thin section analysis, Journ. Geol., 62.
61. Chaves F., 1956. Petrographic Modal Analysis, New York.
62. Chebotarev I. F., 1955. Metamorphism of natural water in the crust of weathering, Geochim. et Cosmochim. Acta, 8.
63. Chenometh P. A., 1952. Statistical methods applied to Trentonian stratigraphy in New York, Bull. G. S. A., 63.
64. Chorley R. J., Malm D. E. G., Pogorzelski H. A., 1957. A new standart for estimating drainage basin shapes, Am. Journ. Sci., 255.
65. Chorley R. J., 1957. Climate and morphometry, Journ. Geol., 65.
66. Chronic H., 1952. Molluscan fauna from the Permian Kaibab formation Walnut canyon Arizona, Bull. G. S. A., 63.
67. Chung J. H., Frase D. A. S., 1958. Randomization tests for a multivariate two-sample problem, Journ. Am. Stat. Assoc., 53.
68. Cohee G. V., 1938. Sediments of the submarine canyons off the california coast, Journ. Sed. Petrol., 8.
69. Collins, Russel, Farguecher, 1953. Maximum age of elements and age the earth's crust, Canad. Journ. Phys., 31.
70. Conway E. J., 1945. Mean Losses of Na, Ca, etc., in one weathering cycle and potassium removal from the ocean, Am. Journ. Sci., 243.
71. Court A., 1952. Some new statistical techniques in geophysics, Advances in Geophysics, 1.
72. Croxton F. E., Cowden D. J., 1941. Applied general statistics, Prentice-Hall, Englewood Cliffs, New Jersey.
73. Curray J. R., Griffiths J. C., 1955. Sphericity and roudness of quarin grains in sediments, Bull. G. S. A., 66.
74. Curray J. R., 1956. Dimensional grain prientation studies of recent coastal sands, Bull. AAPG, 40.
75. Curray J. R., 1956. An analysis of two dimensional orientation data, Journ. Geol., 64.
76. Davis D. H., 1954. Estimating porosity of sedimentary rocks from bulk density, Journ. Geol., 62.
77. Deevey E. S., 1940. Limnological studies in Counecticut, a contribution of regional limnology, Am. Journ. Sci., 238.
78. Dixon W. J., 1950. Analysis of extreme values, Annal. Math. Stats., 21.
79. Dixon W. J., Massey F. J., 1951. Introduction to statistical analysis, New York.
80. Donner J. J., Kurten R., 1958. The floral and faunal succession of «Cueva de Toll» Spain, Eiszeitalter u. gegenwart, 9.
81. Dryden L., 1935. A statistical method for the comparison of heavy mineral suites, Am. Journ. Sci., 29, № 173.

82. Durand D., Greenwood J. A., 1958. Modifications of the Rayleigh test for uniformity in analysis of two-dimensional orientation data, *Journ. Geol.*, **66**.
83. Đurovič S., 1959. Contribution to the lognormal distribution of elements, *Geochim. et Cosmochim. Acta*, **15**.
84. Eissenhart C., 1935. A test for significance of lithological variation, *Journ. Sed. Petrol.*, **5**.
85. Elderton W. P., 1938. Frequency curves and correlation, third edition, Cambridge Univ. Press.
86. Emery J. R., Griffiths J. C., 1954. Differentiation of oil-bearing from barren sediments by quantitative petrographic analysis, *Producers Monthly*, **19**.
87. Emery K. O., 1954. Some characteristics of southern California Sediments, *Journ. Sed. Petrol.*, **24**.
88. Emery K. O., 1955. Grain size of marine beach gravels, *Journ. Geol.*, **63**.
89. Emilian C., 1955. Pleistocene temperatures, *Journ. Geol.*, **63**.
90. Ewing C. J. C., 1931. A comparison of the methods of heavy mineral separation, *Geol. Mag.*, **68**.
91. Fairbairn H. F., 1953. Precision and accuracy of chemical analysis in silicate rocks, *Geochim. et Cosmochim. Acta*, **4**.
92. Feller W., 1940. On the logistic law of growth and its empirical verification, *Acta Biometrica*, **5**, № 2.
93. Fisher R. A., Tippet L. C. H., 1928. Limiting forms of the frequency distribution of the largest or smallest member of a sample, *Proc. Camb. Phill. Soc.*, **24**.
94. Fisher R. A., 1936. Use of multiple measurements in taxonomic problems, *Annals of Eugenics*, **7**.
95. Fisher R. A., 1949. Statistical methods for research workers, London.
96. Flanagan F. J., Kellagher R. C., Smith U. L., 1959. The slotted cone splitter, *Journ. Sed. Petrol.*, **29**.
97. Flinn D., 1959. An application of statistical analysis to petrochemical data, *Geochim. et Cosmochim. Acta*, **17**.
98. Folk R., Ward W. C., 1957. Brazes River bar, A study in the significance of grain size parameters, *Journ. Sed. Petrol.*, **27**.
99. Friedman G. M., 1958. Determination of sieve-size distribution from the thin-section data for sedimentary petrological studies, *Journ. Geol.*, **66**.
100. Giles E., 1960. Multivariate analysis of Pleistocene and recent coales. (*Canis latrans*) from California, *Univ. Calif. Publ. in Geol. Sci.*, **36**.
101. Gray S. W., 1946. Relative growth in a phylogenetic series, *Am. Journ. Sci.*, **244**.
102. Greenman N. N., 1951. The mechanical analysis of sediments from the thin section data, *Journ. Geol.*, **59**.
103. Griffiths J. C., 1951. Size versus sorting in some Caribbean sediments, *Journ. Geol.*, **59**.
104. Griffiths J. C., 1958. Estimation of error in grain size analysis, *Journ. Sed. Petrol.*, **23**.
105. Griffiths J. C., Rosenfeld M. A., 1953. A further test of dimensional orientation of quartz grains in Bradford Sand, *Am. Journ. Sci.*, **251**.
106. Griffiths J. C., Rosenfeld M. A., 1954. Operator variations in experimental research, *Journ. Geol.*, **62**.
107. Griffiths J. C., 1958. Petrography and Porosity of the Cow Run sand. St. Marys, West Virginia, *Journ. Sed. Petrol.*, **28**.
108. Griffiths J. C., 1959. Size and shape of rock fragments in the Tuscarora Scree, *Journ. Sed. Petrol.*, **29**.
109. Griffiths J. C., 1960. Frequency distributions on accessory mineral analysis, *Journ. Geol.*, **68**.
110. Groba E., 1958. Methodische Schwierigkeiten bei der Untersuchungen Litoraler Sedimentie, *Geol. Rund.*, **47**, Heft 1.

111. Gumbel E. J., 1941. Probability interpretation of the observed return periods of floods, *Trans. Am. Geophys. Union*, Pt. 3.
112. Gumbel E. J., 1954. Statistical theory of extreme values and some practical applications, *Nat. Bur. Standards, Applied Math.*, Ser. № 33.
113. Gumbel E. J., Lieblein J., 1954. Some application of extreme-value methods, *Am. Statistician*, 8.
114. Guthrie R. K., Greenberger M. H., 1955. The use of multiple correlation analysis for interpreting petroleum engineering data, Division of production, A. P. I., Dallas, Texas.
115. Hald A., 1952. *Statistical Theory with Engineering Applications*, New York.
116. Haldane I. B. S., 1955. The measurement of variation, *Evolution*, 1X.
117. Harris S. A., 1958. Probability curves and the recognition of adjustment of depositional environment, *Journ. Sed. Petrol.*, 28.
118. Harris S. A., 1958. Differentiation of various Equiptian Aolian micro-environments by mechanical composition, *Journ. Sed. Petrol.*, 28.
119. Harrison P. W., 1957. New techniques for three-dimensional fabric analysis of till and englecial debris containing particles from 3 to 40 mm in size, *Journ. Geol.*, 65.
120. Herdan, 1953. *Small particle statistics*, Elsevier Press, New York.
121. Hilliard H. E., Cahn J. W., 1959. An evaluation of procedures in quantitative metallography I volume-fraction analysis, General Electric Research Laboratory, Report № 59-RL-2294M, Sgenectady, New York.
122. Hoel P. G., 1947. *Introduction to mathematical statistics*, New York.
123. Holzinger K. J., Harman H. H., 1941. *Factor analysis*, Univ. of Chicago Press.
124. Horber B., Potter P. E., 1955. Stratigraphic and sedimentologic aspects of Lemont Drift of Northeastern Illinois, *Illinois State Geol. Rept. of Invest.*, 185.
125. Horst P., 1932. A short method for solving for a coefficient of multiple correlation, *Ann. of Math., Stats.* 3.
126. Hospers J., 1955. Rock magnetism and polar wandering, *Journ. Geol.*, 63.
127. Hutta J. J., Griffiths J. C., 1955. Directional permeability of sandstones; a test of technique, 19, № 11, 12.
128. Huxley I. S., 1932. *Problems of relative growth*, London.
129. Imbrie J., 1955. Quantitative lithofacies and biofacies study of florena shale (Permian) of Kansas, *Bull. A. A. P. G.*, 39.
130. Imbrie J., 1956. The place of biometrics in taxonomy, *Bull. Am. Mus. Nat. Hist.*, 108, art. 2.
131. Inman D. L., 1952. Measures for describing the size distribution of sediments, *Journ. Sed. Petrol.*, 22.
132. McIntyre D. D., 1959. The hydraulic equivalence and size distributions of some mineral grains from a beach, *Journ. Geol.*, 67.
133. Jizba Z. V., 1953. Mean and standart deviation of certain geologic data; a discussion, *Am. Journ. Sci.*, 251.
134. Johnson R. G., 1960. Models and methods for the analysis of mode of formation of fossil assemblages, *Bull. G. S. A.*, 71.
135. Jolicoeur P., 1959. Multivariate geographical variation in the painted turtle: a principal component analysis.
136. Jonson W., 1940. Application of the non-central *t*-distribution, *Biometrika*, 31.
137. Jizba Z. V., 1959. Frequency distribution of elements in rocks, *Geochim. et Cosmochim. Acta*, 16.
138. Kahma A., Mokka T., 1946. A statistical method for quantitative refractive index of minerals in rocks, *Compt. Rend. de la Soc. geologique de Finlande*, XIX.
139. Kahn J. S., 1959. Anisotropic sedimentary parameters, *Trans. New York Acad. Sci.*, Ser. 11, 21, № 5.

140. Kapteyn J. C., 1933. Skew frequency curves in biology and statistics, Noordhoff, Groningen.
141. Kendall M. G., 1955. Rank correlation methods, 2nd ed., New York.
142. Kendall M. G., 1957. Course in multivariate analysis, London.
143. Kermack K. A., Haldane J. B. S., 1950. Organic correlation and allometry, *Biometrika*, **37**.
144. Kermack K. A., 1954. A biometrical study of *Micraster corangrinum* and *M. Isomicraster senonenis*, *Phil. Trans. Roy. Soc., London, Ser. B*, **237**.
145. Korn H., 1938. Schichtung und absolute zeit U. S. V. Neues Jahrb. f. Min. Geol. Pal. Abt. A, **74**, Heft 1.
146. Kotaka T., 1953. Variation of Japanese *Anadara granosa*, *Trans. Proc. Paleont. Soc., Japan*, **10**.
147. Krinsley D., 1960. Magnesium, strontium and aragonite in the shells of certain littoral gastropods, *Journ. Paleont.*, **34**.
148. Krumbein W. C., 1934. Size frequency distributions of sediments, *Journ. Sed. Petrol.*, **4**.
149. Krumbein W. C., 1936. Application of logarithmic moments to size frequency distributions of sediments, *Journ. Sed. Petrol.*, **6**.
150. Krumbein W. C., 1932. Size frequency distributions of sediments and the normal Phi curve, *Journ. Sed. Petrol.*, **8**.
151. Krumbein W. C., 1953. Latin square experiments in sedimentary petrology, *Journ. Sed. Petrol.*, **23**.
152. Krumbein W. C., 1953. Statistical designs for sampling beach sand, *Trans. Am. Geophys. Union*, **34**.
153. Krumbein W. C., Miller B. L., 1953. Design of experiments for statistical analysis of geological data, *Journ. Geol.*, **62**.
154. Krumbein W. C., 1955. Experimental design in the earth sciences, *Trans. A. G. U.*, **36**.
155. Krumbein W. C., 1956. Statistical analysis of facies maps, *Journ. Geol.*, **63**.
156. Krumbein W. C., Tukey J. W., 1956. Multivariate analysis of mineralogic, lithologic and chemical composition of rock bodies, *Journ. Sed. Petrol.*, **36**.
157. Krumbein W. C., Slack H. A., 1956. Statistical analysis of low level radioactivity of Pennsylvanian black fissile shale in Illinois, *Geol. Soc. Am. Bull.*, **67**.
158. Krumbein W. C., Lieblein J., 1956. Geological application of extreme-value methods to interpretation of obbles and boulders in gravel deposits, *Trans. Am. Geoph. Union*, **37**.
159. Krumbein W. C., 1959. The sorting out of geological variables illustrated by regression analysis of factors controlling beach firmness, *Journ. Sed. Petrol.*, **29**.
160. Kruskal W., 1953. On the uniqueness of the line of organic correlation, *Biometrics*, **9**.
161. Kruskal W., Usable but not widely known statistical techniques, Work sponsored by Joint Services Advisory Comm. for Res. Groups in Applied Math. and Statistics Contract № 60ri-02035.
162. Kurten B., 1954. Observation on allometry in mammalian dentitions. its interpretation and evolutionary significance, *Acta Zool.*
163. Kurten B., 1955. Sex dimorphism and size trends in the cave bear, *Ursus spelaeus*, *Acta Zool Fennica*, **90**.
164. Kurten B., 1957. On the bears of the Holsteinian interglacial, *Stockholm Contrib. in Geol.*, **11**, № 5.
165. Kurten R., 1960. Chronology and faunal evolution of the earlier European glaciations, *Soc. Sci. Fennica Comm. Biol.*, **XXI**, № 5.
166. Leitch D., 1940. A statistical investigation of the *Antracomyas* of the basal *Simulis Pulchra* zone of Scotland, *Quart. Journ. Geol.*, **96**.
167. Lohse E. A., 1955. A theoretical curve for statistical analysis of sediments, *Journ. Sed. Petrol.*, **25**.

168. Lonquet-Higgins J. S., 1952. On the statistical distribution of the heights of sea waves, *Journ. Mar. Res.*, **11**.
169. Lorenz P., 1935. Annual survey of statistical technique: trends and seasonal variations, *Econometrica*, **3**.
170. Lumer H. et al., 1942. On the significance of the constant b in law of allometry y equals b times x to the sigma, *Am. Nat.*, **76**.
171. Lundelius E. L., 1955. Skeletal adaptations in two species of *Sceloporus* as determined by regression lines, Ph. D. Thesis, Univ. of Chicago library.
172. Manning J. C., 1953. Application of statistical estimation and hypothesis testing to geologic data, *Journ. Geol.*, **62**.
173. Marsal D., 1949. Zur Metodik der Paleontologie; Die statistische Sicherung Mittelwerten und Korrelationsziffern, *Neuen Jahrbuch F. Mineralogie etc. Abt. B, Heft 8*.
174. Marschner A. W., 1953. A Method for the size analysis of sand on a number frequency basis, *Journ. Sed. Petrol.*, **23**.
175. Matson C. C., Folk R. L., 1958. Differentiation of beach, dune and aeolian flat environments by size analysis, Mustang Island, Texas, *Journ. Sed. Petrol.*, **28**.
176. McDonald D. K. C., 1952. Information theory and its application to taxonomy, *Journ. App. Phys.*, **23**.
177. McKay A. T., 1932. Distribution of the coefficient of variation and the extended « t » distribution, *Journ. Roy. Stat. Soc.*, **95**.
178. McKelvey V. E., 1940. Beach sediments of Trouf Lake, Wisconsin, *Journ. Sed. Petrol.*, **10**.
179. Mengershausen H., 1937. Annual survey of statistical technique; Methods of computing and eliminating changing seasonal fluctuations, *Econometrica*, **5**.
180. Miller R. L., Olson E. C., 1955. The statistical stability of quantitative properties as a fundamental criterion for the study of environments, *Journ. Geol.*, **63**.
181. Miller R. L., 1949. An application of the analysis of variance to paleontology, *Journ. Paleont.*, **23**.
182. Miller R. L., 1954. A model for the analysis of environments of sedimentation, *Journ. Geol.*, **42**.
183. Miller R. L., Goldber E. D., 1955. The normal distribution in geochemistry, *Geochim. et Cosmochim. Acta*, **8**.
184. Miller R. L., Olson E. C., 1955. The statistical stability of quantitative properties as a fundamental criterion for the study of environments, *Journ. Geol.*, **4**.
185. Miller R. L., Zeigler J. M., 1958. A model relating dynamics and sediment pattern in equilibrium, in the region of shoaling waves, breaker zone and foreshore, *Journ. Geol.*, **66**.
186. Morisawa M., 1958. Measurement of drainage-basin outline form, *Journ. Geol.*, **66**, № 5.
187. Motohashi H., 1930. Craniometrical studies on skulls of wild asses of West Mongolia, *Mem. Tottori Ag. College of Nippon*, № 1.
188. Nunroe M. E., 1951. Theory of probability, New York.
189. U. S. National Bureau of Standards, 1953. Probability tables for the analysis of extreme-value data, *Nat. Bur. Stand., App. Math. Ser.* № 22.
190. Neyman, Pearson E., 1929. On the use and interpretation of certain test criteria for purposes of statistical inference, Parts I, II, *Bio-metrica*, **20A**.
191. Neyman, Pearson E., 1933. On the problem of the most efficient tests of statistical hypotheses, *Phil. Trans. Roy. Soc. A*, **231**.
192. Neyman, Pearson E., 1936. Contributions to the theory of testing statistical hypothesis, *Stat. Res. Mem.* **1**.
193. Neyman, Pearson E., 1936. Sufficient statistics and uniformly most powerful tests of statistical hypotheses, *Stat. Res. Mem.*, **1**.

194. Neyeman, Pearson E., 1938. Contributions to the theory of testing statistical hypotheses, Stat. Res. Mem., 2.
195. Nichol D., 1944. New West American species of the foraminiferal genus *Eplphidium*, Journ. Paleont., 18.
196. Niehoff W., 1958. Die Primar Gerichten Sedimentstrukturen, etc. Geol. Rund., 47, Heft 1.
197. Olson E. C., Miller R. L., 1951. Relative growth in paleontological studies, Journ. Paleont., 25.
198. Olson E. C., Miller R. L., 1951. A mathematical model applied to a study of the evolution of species, Evolution, 5.
199. Olson J. S., Potter P. W., 1954. Variance components of crossbedding direction in some basal Pennsylvanian sandstone the eastern interior basin; statistical methods, Journ. Geol., 62.
200. Olson E. C., 1957. Size frequency distribution in samples of extinct organisms, Journ. Geol., 65.
201. Olson J. C., 1958. Lake Michigan dune development I wind velocity profiles, Journ. Geol., 66.
202. Olson E. C., Miller R. L., 1958. Morphological integration, University of Chicago Press, Ch. IV.
203. Orcutt G. H., James S. F., 1948. Testing the significance of correlation between time series, Biometrika, 35.
204. Packham G. H., 1955. Volume, weight, and member frequency analysis of sediments from thin section data, Journ. Geol., 63.
205. Parkinson D., 1958. Quantitative studies of brachiopods from the lower Carboniferous reef limestones of England *Schizophoria resupinata* (Martin), Journ. Paleont., 28.
206. Pastiels A., 1953. Etude biometrique des Anthracosiidae du Westphalien, A de la Belgique, Assoc. Etude Paleont., Stratigraphie Houilleres Publ., № 16.
207. Pearson H. S., 1928. Chinese fossil Suidae, Paleontologia, Sinica, Ser. C, V, Fasc. 5.
208. Pearson K., 1894. Contribution to the mathematical theory of evolution, Phill. Trans. Roy. Soc., London, 185.
209. Pelto C. R., 1952. The mechanical analysis of sediments from thinsection data; a discussion, Journ. Geol., 60.
210. Pettijohn F. J., Lundahl A. C., 1943. Shape and roundness of Lake Eric beach sands, Journ. Sed. Petrol., 13.
211. Phleger F. B., Jr., 1940. Relative growth and vertebrata phylogeny, Am. Journ. Sci., 238.
212. Phleger F. B., Putnam W. S., 1942. Analysis of Merychiododon skulls, Am. Journ. Sci., 240.
213. Piggot C. S., 1938. Radium in rocks; The radium content of the four groups of PreCambrian granites of Finland, Am. Journ. Sci., 350.
214. Pincus H. J., 1951. Statistical methods applied to the study of rock fractures, Bull. G. S. A., 62.
215. Pincus H. J., 1953. The analysis of aggregates of orientation data in the earth sciences, Journ. Geol., 61.
216. Pincus H. J., 1956. Some vector and arithmetic operations on twodimensional orientation variates with applications to geological data, Journ. Geol., 64.
217. Plumley W. J., 1948. Black Hills terrace gravels, a study in sediment transport, Journ. Geol., 56.
218. Plumley W. J., Davis D. H., 1956. Estimation of recent-sediment size parameters from a triangular diagram, Journ. Sed. Petrol., 26.
219. Pollard H. S., 1934. On the relative stability of the median and arithmetic mean, etc., Ann. Math. Stat., 5.
220. Potter P. E., Olson J. S., 1954. Variance components, etc., Geol. appl. Journ. Geol., 62.

221. Potter P. E., Siever R., 1956. Sources of basal Penn. sediments in the eastern interior basin I. Cross bedding, Journ. Geol., **64**.
222. Preston F. W., 1948. The commonness, and rarity, of species, Ecology, **29**.
223. Pye W. D., 1943. Rapid methods sedimentational analysis of arenaceous sediments, Journ. Sed. Petrol., **13**.
224. Quenouille M. H., 1952. Associated measurements, Acad. Press, New York, Ch. II, Time series Analysis.
225. Ramberg H., De Vore G., 1951. The distribution of Fe^{++} and Mg^{++} in coexisting olivines and pyroxenes, Journ. Geol., **59**.
226. Rao C. R., 1952. Advanced Statistical Methods in Bimetric Research, New York.
227. Raup D. M., 1958. The relation between water temperature and morphology in Dendroaster, Journ. Geol., **66**.
228. Richards O. W., 1936. Analysis of the constant differential growth ratio, Carnegie Inst., Washington, Publ. 29.
229. Reeve E. C. R., 1941. A statistical analysis of taxonomic differences within the genus *Tamandua*, Royal Zool. Soc. London, Proc., Ser. A, **111**.
230. Reeve E. C. R., 1940. Relative growth in the shout of the anteaters, Proc. Roy. Zool. Soc., London, S. A., **3**.
231. Riley N. A., 1941. Projection sphericity, Journ. Geol., **11**.
232. Rittenhouse G., Connaughton M. P., 1944. Errors of sampling sands for mechanical analysis, Journ. Sed. Petrol., **14**.
233. Robb R. C., 1929. On the nature of hereditary size-limitations; The growth of parts in relationship to the whole, Brit. J. Exper. Biol., **6**.
234. Rominger J. F., 1954. Relationships of the plasticity and grain size of Lake Agassiz sediments, Journ. Geol., **62**.
235. Rosenfeld M. A., Jacobsen L., Ferm J. C., 1953. A comparison of sieve and thin section technique for size analysis, Journ. Geol., **61**.
236. Rosenfeld M. A., Griffiths J. C., 1953. An experimental test of visual comparison technique in estimating two dimensional sphericity and roundness of quartz grains, Am. Journ. Sci., **251**.
237. Roy S. N., Mitra S. K., 1956. An introduction to some non-parametric generalizations of analysis of variance and multivariate analysis, Biometrika, **43**.
238. Rusnak G. A., 1957. The orientation of sand grains under conditions of unidirectional fluid flow; Theory and experiment, Journ. Geol., **65**.
239. Rusnak G. A., 1957. A fabric and petrologic study of the Pleasantview Sandstone, Journ. Sed. Petrol., **27**.
240. Sauramo M., 1954. Das Rätsel des Ancyclussees, Geol. Rund., **42**.
241. Schlect W. G., 1949. Probable error of a geochemical analysis, Bull., **992**. Contribution to geochemistry, U.S.G.S.
242. Schmid K., 1934. Biometrische Untersuchungen und Foraminiferen aus dem Pliacsen von Ceram, Ecol. Geol. Helv., **27**, № 1.
243. Schumm S. A., 1956. Evolution of drainage systems and slopes in badlands at Perth Amboy N. J., Bull. G.S.A., **67**.
244. Schwarzscher W., 1953. Gross-bedding and grain size in the lower K. sands of East Anglia, Geol. Mag. XC.
245. Sewell, Wright, 1921. Correlation and causation, Journ. Ag. Res., **20**.
246. Shannon C. E., Wlaver W., 1943. The mathematical theory of communication, Univ. of Illinois Press, Urbana.
247. Shaw D. M., Bankier J. D., 1954. Statistical methods applied to geochemistry, Geochim. et Cosmochim. Acta, **5**.
248. Shaw D. M., Harrison W. D., 1955. Determination of the mode of a metamorphic rock, Am. Min., **40**.
249. Shaw A. B., 1956. Quantitative trilobite studies; The statistical description of trilobites, Journ. Paleont., **30**.
250. Shaw A. B., 1957. Measurement of the dorsal Shell of non Agnostidean Trilobites, Journ. Paleont., **31**.

251. Silver R., Potter P. E., 1956. Sources of basal Pennsylvanian sediments in the eastern interior basin; *Sedimentary petrology*, Journ. Geol., **64**.
252. Simpson G. G., 1935. The Tiffany fauna, Upper Paleocene, *Am. Mus. Novit.*, **817**.
253. Simpson G. G., 1937. The Fort Union of the Crazy Mountain field Montana, and its mammalian faunas, *U.S. Nat. Mus. Bull.*, **169**.
254. Simpson G. G., 1947. Note on the measurement of variability and on relative variability of fossil mammals, *Am. Journ. Sci.*, **245**.
255. Simpson G. G., Roe A., Lewontin R. C., 1960. *Quantitative Zoology*, rev. ed., New York.
256. Sloan R. E., 1955. Paleocology of the Pennsylvanian marine shales of Palo Pinto Co., Texas, *Journ. Geol.*, **63**.
257. Small J., 1948. Details of the history of diatoms, *Evolution IX—XIII*, *Proc. Roy. Irish. Acad.*, **51**.
258. Small J., 1949. The shape and pattern of evolution, *Quantitative evolution XVII*, *Phyton*, **1**.
259. Small J., 1950. Increase of species in diatoms, *Quantitative evolution XVI*, *Ann. Bot.*, **XIV**.
260. Snedecor G. W., 1946. *Statistical methods*, 4th. ed. Collegiate Press, Ames, Iowa.
261. Sneed E. D., Folk F. L., 1958. Pebbles in the lower Colorado River, Texas, a study in particle morphogenesis, *Journ. Geol.*, **66**.
262. Strahler A. N., 1950. Equilibrium theory of erosional slopes approached by frequency distribution analysis, *Am. Journ. Sci.*, **248**, pp. 673—696.
263. Strahler A. N., 1950. Equilibrium theory of erosional slopes approached by frequency distribution analysis, *Am. Journ. Sci.*, **248**, pp. 800—814.
264. Strahler A. N., 1952. Hypsometric (area-altitude) analysis of erosional topography, *Bull. G.S.A.*, **63**.
265. Strahler A. N., 1952. Quantitative geomorphology of erosional landscapes, *Comp. Rend.*, 19th Intern. Geol. Congres., 1952, sec. 13, pt. 3.
266. Strahler A. N., 1954. Statistical analysis in geomorphic research, *Journ. Geol.*, **62**.
267. Strahler A. N., 1956. Quantitative slope analysis, *Bull. G.S.A.*, **67**.
268. Swineford A., Swineford F., 1946. A comparison of three sieve shakers, *Journ. Sed. Petrol.*, **16**.
269. Swineford A., Frye J. C., Leonard A. B., 1955. Petrography of the late Tertiary volcanic ash fall Central Great Plains, *J. Sed. Petrol.*, **25**.
270. Tanner W. F., 1958. Comparison of phi percentile deviations, *Journ. Sed. Petrol.*, **28**.
271. Tesser G., 1948. La relation d'allometric, sa signification statistique et biologique, *Biometrics*, **4**.
272. Thurstone L. L., 1947. *Multiple factor analysis*, Univ. of Chicago Press.
273. Tippet L. H. C., 1945. *Methods of Statistics*, Williams and Norgate, Ltd., London.
274. Tippet L. C. H., 1950. *Technological applications of statistics*, New York.
275. Trask P. D., 1932. Origin and environment of source sediments of petroleum, *Gulf. Publ. Co.*, Houston.
276. Tryon R. C., 1939. *Cluster analysis*, Ann. Arbor, Michigan.
277. Vistelius A. B., 1958. Volume frequency analysis of sediments from thin section data; a discussion, *Journ. Geol.*, **66**.
278. Vistelius A. B., 1958. Paragenesis of sodium, potassium and uranium in volcanic rocks of Lassen Volcanic National Park, California, *Geochim. et Cosmochim. Acta*, **14**.
279. Vistelius A. B., 1961. Sedimentation time trend functions and their application for correlation of sedimentary deposits, *Journ. Geol.*, **69**, № 6.
280. Weiner N., 1948. *Cybernetics*, New York.
281. Wimberley C. S., 1955. Marine sediments north of Seripps submarine Canyon, La Jolla, *Journ. Sed. Petrol.*, **25**.

- 282. Whitworth W. A., 1942. Choice and chance, New York, and DCC exercises, New York.
- 283. Wickman F. E., 1954. The total amount of sediments and the composition of the average igneous rock, *Geochim. et Cosmochim. Acta*, 5.
- 284. Wiener N., 1949. Extrapolation interpolation and smoothing of stationary time series, New York.
- 285. Wilks S. S., 1932. Movements and distributions of estimates of population parameters from fragmentary samples, *Ann. of Math. Stats.*, 3.
- 286. Wright H. E., Jr., 1956. Origin of the Chuska sandstone Arizona, New Mexico, etc., *Bull. G.S.A.*
- 287. Yalovsky R., 1957. The relationship between paleotemperature and carbonate content in a deep-sea core, *Journ. Geol.*, 65.

ОГЛАВЛЕНИЕ

Предисловие	5
Предисловие авторов	9
Глава 1. Теория вероятностей	13
Глава 2. Функции плотности вероятности	28
Глава 3. Моменты и математические ожидания	52
Глава 4. Статистические выводы; оценки параметров и их распределения	60
Глава 5. Статистические выводы; проверка гипотез о параметрах распределения	79
Глава 6. Статистические выводы; проверка некоторых гипотез	92
Глава 7. Статистические выводы; оценка параметров и проверка гипотез	142
Глава 8. Линейный регрессионный анализ	193
Глава 9. Регрессионный анализ при палеобиометрических исследованиях	208
Глава 10. Дисперсионный и регрессионный анализы	219
Глава 11. Полное статистическое исследование	232
Глава 12. Многомерный анализ	250
Глава 13. Применение многомерного и факторного анализов	287
Глава 14. Одномерный тренд (включая временные ряды)	330
Глава 15. Временные ряды	353
Глава 16. Распределение точек на плоскости и в пространстве; обзор моделей, критериев и областей применения	370
Глава 17. Применение некоторых статистических методов в задачах картирования	394
Приложение А. Разделение коллекции ископаемых организмов	443
Приложение В. Модели ископаемых популяций	450
Приложение С. F-распределение	458
Приложение D. χ^2 -распределение	460
Приложение Е. Функция распределения для распределения χ^2	463
Приложение F. Теорема сложения для распределения χ^2	464
Приложение G. Статистика Колмогорова — Смирнова; номограммы для 95- и 99-процентных доверительных областей	465
Литература	471

Р. Миллер и Дж. Кан
СТАТИСТИЧЕСКИЙ АНАЛИЗ
В ГЕОЛОГИЧЕСКИХ НАУКАХ

Редактор *В. К. Знаменская*
Художник *Н. К. Саложников*
Художественный редактор *Е. И. Вескова*
Технический редактор *Ю. И. Экке*

Сдано в производство 23/VI 1965 г.
Подписано к печати 13/XI 1965 г.
Бумага 60×90¹/₄, 15.13 бум. л.
30,25 печ. л.
Уч.-изд. л. 27,0 Изд. № 5/2907
Цена 2 р. 08 к. Зак. 1672
(Темплан 1965 г. изд-ва „Мир“ пор. № 141)

ИЗДАТЕЛЬСТВО „МИР“
Москва, 1-й Рижский пер., 2
Ленинградская типография № 2
имени Евгении Соколовой
Главполиграфпрома
Государственного комитета
Совета Министров СССР по печати,
Измайловский проспект, 29