Д. А. РОДИОНОВ

СТАТИСТИЧЕСКИЕ РЕШЕНИЯ В ГЕОЛОГИИ

Д. А. РОДИОНОВ

СТАТИСТИЧЕСКИЕ РЕШЕНИЯ В ГЕОЛОГИИ

МОСКВА «НЕДРА» 1981

УДК 519.2 : 55

Родионов Д. А. Статистические решения в геологии.— М., Недра, 1981.— 231 с.

Содержит методические указания по применению статистических методов в геологии. Описаны статистические методы решения типовых геологических задач с рекомендацией способов, уменьшающих риск, связанный с принятием ошибочных решений. Последовательно рассмотрены конкретные геологические задачи по разграничению геологических объектов, по выделению их информативных признаков и по геологическому прогнозированию. Приведены статистические методы решения задач. Освещено моделирование геологических процессов. Описаны принципы и способы оптимизации принимаемых решений. Даны практические примеры.

Для специалистов геологического профиля, использующих в своей работе математические методы.

Табл. 43, ил. 12, список лит.— 57 назв.

Рецензент д-р геол.-минер. наук *Р. М. Константинов* (ИГЕМ)

Р $\frac{20802\text{-}204}{043(01)\text{-}81}$ 32-81 1904010000

ПРЕДИСЛОВИЕ

Последнее десятилетие характеризуется интенсивным внедрением математических методов и ЭВМ в практику геологоразведочных работ. Достаточно сказать, что в настоящее время число вычислительных центров в Советском Союзе, занятых решением геологических задач, достигает нескольких десятков, причем все они, как правило, оснащены двумя и более современными вычислительными машинами. Это, естественно, создает основу для массовой обработки геологоразведочной информации и позволяет реализовать на ЭВМ весьма сложные алгоритмы, что было бы практически невозможно вручную.

Среди весьма разнообразного арсенала используемых в геологии математических методов существенную роль играют методы математической статистики, среди которых имеются как весьма простые, легко реализуемые на настольных вычислительных машинах, так и достаточно сложные, требующие определенных усилий при их реализации на ЭВМ и серьезных математических знаний при практическом применении.

В связи с большой популярностью статистических методов при геологических исследованиях за последние 15 лет было издано немало различных руководств по их применению при решении геологических задач. Особо следует отметить книги Р. Миллера и Дж. Кана [32], У. Крамбейна и Ф. Грейбилла [27], и др., которые были переведены на русский язык и изданы в СССР, а также ряд работ, ориентированных на конкретные задачи.

Несмотря на то, что все эти руководства были написаны на высоком научном уровне и учитывали множество полезных для геологии задач математической статистики, они в недостаточном объеме охватывали многомерные задачи, а также не всегда делали упор на риск, связанный с принятием статистических решений. Именно это и явилось причиной написания данной книги.

Книга состоит из девяти глав, в которых, начиная с весьма простых понятий, рассматриваются вопросы, связанные с геологическими данными, дается представление о статистических оценках неизвестных параметров распределения, методах их получения и критериях качества. Особое внимание уделено вопросу построения статистических решений в геологии, в связи с чем рассмотрены особенности обоснования выводов по геологическим данным, дано представление о статистических гипотезах и критериях для их проверки, а также принципы построения статистических критериев и выбор критической области. Естественно, что в книге такого рода очень трудно, да и не нужно охватить все возможные варианты статистических гипотез и множество способов их проверки. Поэтому в ней излагаются вначале наиболее простые задачи проверки ги-

потез по одномерным статистическим данным, а затем делается переход к многомерным аналогам этих задач.

В связи с тем, что очень многие геологические исследования при обосновании выводов содержат ссылки на величину различий между сравниваемыми объектами, этому вопросу в книге посвящена особая глава. Кроме того, подробно рассмотрен вопрос о выборе информативных комбинаций признаков, причем изложены новые методы решения этой задачи, более совершенные, чем описанные в более ранней работе автора [38]. Нет надобности особо подчеркивать важность этой задачи для геологии, так как ее постановка встречается в самых разнообразных геологических ситуациях — при прогнозировании месторождений, при диагностике пород, фауны и т. п.

Выбор информативных комбинаций признаков осуществляется для последующего их использования в некоторых классификационных процедурах, примером которых может служить дискриминантный анализ. В связи с этим в книге рассмотрены способы построения линейных и квадратичных решающих правил, позволяющих проводить классификационное отнесение изучаемого объекта к одной из двух заданных групп.

Последняя глава книги содержит подробное рассмотрение статистических методов разграничения геологических объектов по комплексу признаков, причем методы, касающиеся задач разграничения в случае расположения наблюдений на плоскости, т. е. линейно не упорядоченных, существенно усовершенствованы, по сравнению с описанными в более ранней работе автора [38].

Естественно, что в данной книге невозможно было охватить все применяющиеся в геологии статистические методы. Например, в ней совершенно не затрагивается такой серьезный вопрос, как построение уравнений регрессии и построение поверхностей тренда. Это сделано сознательно в связи с тем, что в небольшой главе невозможно с достаточной полнотой охарактеризовать эту большую проблему, которая требует применительно к геологии написания отдельной книги. То же самое можно сказать и о таких разделах математической статистики, как дисперсионный и факторный анализ, а также о случайных процессах.

Данная книга предназначена для геологов самых разнообразных специальностей, использующих математические методы при решении геологических задач, и автор надеется, что она окажется полезной в их практической деятельности.

ГЛАВА I

ГЕОЛОГИЧЕСКИЕ ДАННЫЕ

В основе геологических выводов, которые представляют собой итог всех геологических исследований, лежат результаты наблюдений, которые мы будем называть геологическими данными. Эти данные столь многоаспектны (как и сама геология), что представляется весьма затруднительным рассмотреть все их особенности и свойства, и поэтому мы остановимся только на некоторых наиболее важных вопросах, связанных с ними.

1.1. ИСТОЧНИКИ ГЕОЛОГИЧЕСКИХ ДАННЫХ

Наиболее удачная классификация геологических данных предложена У. Крамбейном и Ф. Грейбиллом [27]. Эти авторы все геологические данные делят на результаты полевых и лабораторных

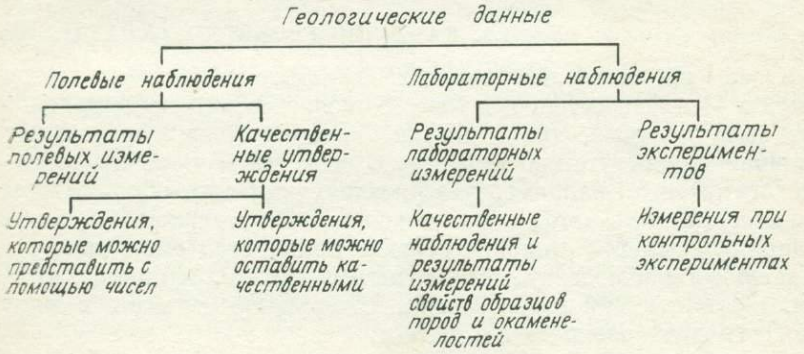

Рис. 1. Классификация геологических данных

наблюдений (рис. 1). Полевые наблюдения подразделяются ими на результаты измерений в поле и качественные утверждения, а последние в свою очередь разделены на утверждения, которые можно представить с помощью чисел, и утверждения, которые остаются чисто качественными.

Лабораторные наблюдения, как и полевые, делятся на две группы, из которых первая также представляет результаты измерений, а вторая — результаты экспериментов.

Комплекс полевых наблюдений чрезвычайно разнообразен. К нему относятся самые разнообразные утверждения о внешнем виде объектов исследования, которыми могут быть породы, минералы или другие образования, утверждения об их составе, степени выветривания или других вторичных изменений и т. п. Кроме того,

в процессе полевых наблюдений нередко проводятся и различные измерения, например замеры азимутов простирания и углов падения слоев пород или трещин, или же секущих их магматических образований. Все результаты наблюдений подобного рода несут огромную геологическую информацию, сосредоточенную в полевых дневниках, которая обычно только в небольшой своей части подвергается достаточно полной обработке.

Современные аналитические и другие средства лабораторных исследований дают возможность получать очень обширную информацию о химическом и минеральном составе и физических свойствах пород, а также о наличии в них остатков органического мира как животного, так и растительного.

За последние годы значительно увеличилась роль данных, получаемых непосредственно с карт, аэрофотоснимков, а также со снимков, сделанных с космических спутников.

Таким образом, современный поток геологических данных столь интенсивен, а эти данные так разнообразны, что все это требует разработки общих подходов к их представлению и методам их обработки при обосновании геологических выводов.

1.2. КАЧЕСТВО ГЕОЛОГИЧЕСКИХ ДАННЫХ

Одна из наиболее важных характеристик геологических данных — их качество. В процессе геологических наблюдений могут быть получены такие данные, которые только констатируют наличие или отсутствие некоторого свойства изучаемого геологического объекта, например утверждение типа «этот образец известняка серый» или «этот образец диорита не содержит биотитовых шлиров». С другой стороны, нередко приходится иметь дело с результатами наблюдений, выраженными числом, например азимут простирания дайки равен 84° или содержание монацита в данной пробе гранита равно 10 г/т, и т. п.

Данные первого типа характеризуют только наличие или отсутствие некоторого свойства или качества изучаемого объекта и называются качественными данными. Данные же второго типа представлены числом и их принято называть количественными. Кроме того, существуют еще и промежуточные данные, называемые полуколичественными. Примером их может служить набор образцов породы (допустим, гранита), упорядоченный по изменению цвета от бледно-серого до темно-серого, причем каждому образцу приписывается номер, характеризующий его место, который может подвергаться обработке как числовая характеристика. Примером полуколичественных данных являются также использованные автором [38] показатели численности фораминифер в образцах. Если данный вид фораминифер в образце отсутствовал, то это обозначалось нулем, если число особей данного вида не превышало первого десятка — ставилась единица, если численность приблизительно соответствовала интервалу от 10 до 30, то образцу приписывалось

значение 2, если же число изучаемых особей попадало в интервал от 30 до 100, то это кодировалось тройкой, интервал от 100 до 300 обозначался цифрой 4, и, наконец, если данный вид изобиловал в образце, это обозначалось цифрой 5. Такое представление позволяло с помощью всего шести значений (0, 1, 2, 3, 4, 5) проводить статистическую обработку очень большого биостратиграфического материала без трудоемких затрат на точные подсчеты.

Таким образом, все геологические данные можно разделить на качественные, полуколичественные и количественные. У. Крамбейн и Ф. Грейбилл [27] выделяют четыре шкалы значений геологических данных.

Первая шкала, номинальная, соответствует выделенным нами качественным данным. Она применяется для классификации объектов по признаку равенства их свойств, причем безразлично, какой номер будет приписан тому или иному классу. Например, первый класс может объединять образцы гранита светло-серого цвета, второй — розового и третий — серого. Смысл классификации в данном случае не изменяется, если нумерацию изменить.

Вторая шкала, порядковая, соответствует выделенным нами полуколичественным данным. Она применяется в тех случаях, когда изучаемые объекты можно расположить по увеличению (усилению) или убыванию (ослаблению) какого-либо свойства, приписав такой последовательности наблюдений номера от 1 до N.

Третья шкала, интервальная, как и порядковая, охватывает некоторые разновидности полуколичественных данных. Примером может служить описанное нами представление численности фауны фораминифер. Для интервальной шкалы характерно, что нулевую точку можно выбрать произвольно.

Четвертая, наивысшая, шкала, или шкала отношений, соответствует количественным данным. Для этой шкалы характерно, что любое число x, выбранное на этой шкале, можно заменить новым числом y, причем эти числа будут связаны следующим соотношением:

$$x = ay \text{ при } a > 0. \tag{1.1}$$

Так, например, при переходе от содержаний, выраженных в процентах, к содержаниям, выраженным в граммах на тонну, константа $a = 10\,000$.

Естественно, что данные, представляющие собой числа, выраженные на шкале отношений, т. е. количественные данные, обычно воспринимаются геологом с большим доверием, чем качественные и полуколичественные данные. Однако это совершенно не означает, что выводы, полученные по количественным данным, будут более надежными, чем выводы, в основе которых лежит качественная и полуколичественная информация. В дальнейшем мы подробно остановимся на этом вопросе и покажем, что обоснованные геологические выводы можно получить и по качественной информации, если соблюдать некоторые несложные статистические правила.

1.3. ОБЪЕКТЫ НАБЛЮДЕНИЯ

Объект наблюдения в геологических исследованиях можно определить самыми разнообразными способами. Так, например, при изучении размеров и форм гранитных массивов некоторого интрузивного комплекса единичным объектом наблюдения будет гранитный массив. При изучении процесса эволюции состава математического очага, породившего многофазный гранитный интрузив, объектом наблюдения будут пробы, отобранные из пород фаз, образующих изучаемый интрузив. Если же задача заключается в исследовании распределения некоторого комплекса минералов или элементов в пределах одной изучаемой интрузивной фазы, то объектами наблюдения будут пробы, отобранные из пород только этой фазы. Естественно, что в подобной ситуации методика отбора проб должна быть одна и та же как при исследованиях в пределах одной фазы, так и для многофазного интрузива в целом, чтобы обеспечить возможность сравнения полученных результатов.

Таким образом, нетрудно видеть, что определение объекта наблюдения в каждом конкретном случае зависит от специфики поставленной геологической задачи и четкой формулировки некоторого комплекса условий (обозначим его G), при которых производится единичное наблюдение. Кроме того, необходимо отметить, что если объектом наблюдения является проба фиксированного веса, которая по заданной методике отбирается из изучаемой породы, то предполагается, что таких проб можно отобрать бесконечно много и в случае надобности повторить эксперимент. Естественно, что последнее не всегда выполняется в геологической практике, особенно для образований, имеющих небольшие размеры. В таких случаях нужно отдавать себе отчет, что не все те закономерности, которые действительны для совокупностей бесконечного объема, приложимы к таким ограниченным наборам наблюдений. В связи с этим будет полезно рассмотреть понятие «геологическая совокупность», что будет сделано в следующем разделе.

1.4. ГЕОЛОГИЧЕСКИЕ СОВОКУПНОСТИ

Под термином «совокупность» обычно понимается набор объектов, объединенных каким-либо единым признаком. Например, совокупность образцов гранита, совокупность галек руслового аллювия данной реки, совокупность ксенолитов изучаемого интрузива и т. п. В первом случае отличительный признак — это то, что все объекты являются образцами гранита, во втором — гальками русла данной реки, а в третьем — ксенолитами интрузива.

Таким образом, под геологической совокупностью мы будем понимать множество геологических объектов, объединенных каким-либо признаком.

Формально геологическую совокупность можно представить как некоторое множество T точек t, т. е. $t \in T$, причем T может быть как непрерывным, так и дискретным. Примером последнего

может служить конечный набор N образцов, моделью которого будет просто натуральный ряд чисел 1, 2, . . . , t, . . . , N, соответствующий номерам, образующим совокупность объектов. В качестве примера, когда множество T непрерывно, можно привести совокупность точек оси скважины, по которой построена каротажная диаграмма. Однако в дальнейшем в подавляющем большинстве случаев мы будем использовать в качестве модели геологической совокупности конечное дискретное множество T точек t. Дело в том, что геолог, как правило, при решении различных геологических задач располагает конечным числом наблюдений, которое хотя и может быть весьма большим, но не является бесконечным. Более того, все геологические наблюдения обычно делаются в дискретном множестве точек, что и определяет дискретность модели геологической совокупности.

Весьма важной особенностью геологических исследований является то, что не весь изучаемый объект обычно доступен наблюдению. Так, например, отбирая образцы для фаунистических исследований из некоторого стратиграфического подразделения, геолог зачастую вынужден довольствоваться только обнажениями, расположенными в оврагах и по берегам рек, которые весьма неполно представляют всю площадь развития исследуемых отложений. Более того, обнажения, как правило, бывают так расположены по площади, что не дают возможности провести равномерное территориальное опробование.

Таким образом, выводы, касающиеся всей области развития изучаемых отложений, геолог вынужден делать только по результатам опробования обнажений, характеризующих лишь незначительную часть исследуемых пород и расположенных весьма неравномерно. Из этого примера следует, что необходимо четко различать изучаемую и опробуемую совокупности и всегда отдавать себе отчет в том, насколько опробуемая совокупность представительна по отношению к изучаемой.

Следует отметить, что каких-либо стандартных рецептов получения ответа на этот вопрос не существует. В каждом конкретном случае геолог должен на основе своего опыта и имеющихся дополнительных геологических сведений определить степень представительности опробуемой совокупности по отношению к изучаемой. Во всяком случае, об этом никогда не следует забывать, так как недоучет этого фактора может привести к необоснованным геологическим выводам, несмотря на самые хитроумные расчеты и сложные математические методы обработки данных. К сожалению, источник возможных ошибок не ограничивается только несовпадением опробуемой и изучаемой совокупностей и поэтому необходимо ввести понятие выборочной совокупности, под которой мы будем понимать множество наблюдений, произведенных над опробуемой совокупностью. Естественно, что выборочная совокупность, или как мы будем ее нередко называть, выборка, обычно во много раз меньше опробуемой совокупности, и выводы, полученные по выборочным

данным, распространяются не только на опробуемый набор объектов, но и на изучаемую совокупность. На практике обычно забывают о промежуточном звене опробуемой совокупности и сразу же делают заключения непосредственно об изучаемой совокупности. Невнимание к этому обстоятельству может также привести к ошибочным заключениям.

В геологической практике очень часто ставится вопрос — сколько нужно иметь проб или наблюдений в выборке, чтобы получаемые на основе статистических методов выводы были обоснованы? Следует особо подчеркнуть, что было бы серьезной ошибкой пытаться ответить на этот впопрос. Дело в том, что статистические методы дают возможность проверять по эмпирическим данным различные предположения или гипотезы, т. е. отвечать на вопрос — противоречит или не противоречит проверяемая гипотеза имеющимся результатам наблюдений. Причем применяемые статистические критерии учитывают риск, связанный с принятием ошибочных решений, что позволяет даже при малом числе данных делать достаточно обоснованные выводы. Подробно эти вопросы будут рассмотрены в дальнейшем.

1.5. ОБОБЩЕННАЯ МОДЕЛЬ ГЕОЛОГИЧЕСКОГО НАБЛЮДЕНИЯ

В разделе 1.4 в качестве математической модели геологической совокупности мы рассматривали множество T точек t. Обозначим через J множество свойств j, которые можно измерить на каждом геологическом объекте. Естественно, что множество J дискретно и $j = 1, 2, \ldots, m$. Тогда любому геологическому объекту с номером $t \in T$ можно поставить в соответствие некоторое множество $J_k \subset J$, состоящее из $k \leqslant m$ элементов и являющееся подмножеством J.

Обозначим через x_{tj} результат наблюдения признака с номером j на объекте с номером t, а через $X_t(J_k)$ — набор наблюдений k признаков. Тогда

$$X_t(J_k) = \{x_{t1}, x_{t2}, \ldots, x_{tj}, \ldots, x_{tk} | J_k\} \qquad (1.2)$$

будет представлять собой k-мерный вектор — строку или вектор-столбец, являющийся моделью геологического наблюдения при условии, что наблюдения проводятся над заданным множеством признаков J_k. В дальнейшем, если набор J_k задан, мы не будем использовать его в записи, и формула (1.2) примет вид

$$X_t = \{x_{t1}, x_{t2}, \ldots, x_{tj}, \ldots, x_{tk}\}. \qquad (1.3)$$

Необходимо отметить, что в данной модели учтены как качественные и полуколичественные признаки, так и количественные.

Если признак с номером j качественный, то он принимает только два значения 0 и 1

$$x_{ij} = \begin{cases} \text{данное качество проявлено} \\ \text{данное качество отсутствует.} \end{cases}$$

Так, например, при изучении образцов гранита может учитываться степень альбитизации, причем выделены только две градации — сильная и слабая альбитизация. Тогда наличие сильной альбитизации будет обозначаться как 1, а слабой — как 0. Если таких качественных признаков несколько или же весь набор J_k является качественным, то каждое наблюдение X_i будет частично или полностью представлять собой последовательность из нулей и единиц. Полуколичественные и количественные данные также представлены с помощью модели (1.3).

Таким образом, в математической модели геологического наблюдения, охарактеризованного k признаками мы будем рассматривать k-мерный вектор-строку или вектор-столбец, независимо от того, с количественными, полуколичественными или качественными данными мы имеем дело.

1.6. ОСНОВНЫЕ ПОНЯТИЯ ТЕОРИИ ВЕРОЯТНОСТЕЙ

Объем геологических данных, используемых в практической деятельности геологов, обычно столь велик, что возникает естественное желание вместо обширных таблиц использовать некоторые обобщающие характеристики, позволяющие судить об особенностях всего массива информации. Наиболее удобными в данном случае являются некоторые статистические характеристики, которые в настоящее время широко применяются в геологической практике для свертывания информации. Однако в основе математической статистики лежит теория вероятностей и поэтому мы рассмотрим некоторые понятия этой науки, с которыми нам придется сталкиваться в дальнейшем. Естественно, что в данной книге возможно только весьма краткое рассмотрение упоминаемых ниже понятий. Более подробное их рассмотрение можно найти в работах [6, 11, 17, 26, 34].

1.6.1. События

Выбор из опробуемой совокупности некоторого о ъекта последующим наблюдением изучаемых свойств мы будем называть экспериментом. Будем также предполагать, что эксперимент в принципе можно провести достаточно большое число раз при условии соблюдения заданного комплекса условий G, который мы определили в разделе 1.3. Примером эксперимента может служить взятие шлиха из руслового аллювия с последующим определением в нем наличия знаков золота, можно так же как эксперимент рассматривать взятие образца, допустим, из палеогеновых отложений Турк-

мении с последующим определением в нем показателей распространенности заданных видов микрофауны. Такие примеры весьма разнообразны, и читатель, естественно, сам может привести их из самых различных геологических ситуаций. Результат (исход) эксперимента мы будем называть событием. Так, появление (или отсутствие) знаков золота в изучаемом шлихе будет событием. Практика показывает, что многие события, как, например, упомянутое в результате повторения эксперимента, могут происходить, а могут и не происходить, причем нельзя заранее точно предсказать, произойдет оно или нет. Такие события мы будем называть случайными.

В подобной ситуации вполне естественно желание иметь некоторую количественную меру возможности появления данного события при повторении экспериментов. В качестве такой меры можно использовать частоту появления события, т. е. отношение числа экспериментов, в которых появилось событие, к их общему числу. Такая мера носит название вероятности. Вообще же, более строго, вероятность появления события в серии экспериментов — это число; обозначим его P, которое поставлено в соответствие данному событию. При неограниченном повторении экспериментов относительная частота события будет сколь угодно близка к P. Так, например, если в среднем на 100 шлихов, взятых из аллювия, приходится 10 золотоносных, то можно записать

$$P(A) \simeq 0{,}10,$$

где A — событие, заключающееся в том, что один взятый наудачу из аллювия данного участка реки шлих окажется содержащим хотя бы один знак золота.

Не следует особо доказывать, что вероятность всегда положительна и принимает значения в интервале от 0 до 1, причем может быть равной или 0 или 1, т. е.

$$0 \leqslant P(A) \leqslant 1. \tag{1.4}$$

В тех случаях, когда событие обязательно наступает, сколько бы ни повторялся эксперимент, $P(A) = 1$, и такое событие называют достоверным. Если же событие в серии экспериментов никогда не происходит, его называют невозможным и $P(A) = 0$.

При рассмотрении многих вероятностных задач в теории вероятностей используются так называемые урновые схемы, в которых в качестве модели изучаемой совокупности рассматривают урну с различно окрашенными шарами. В геологии такой схеме соответствует ряд реальных задач с изучением петрографического состава галечников. Допустим, что изучаемые нами галечники образованы гальками четырех петрографических разновидностей — гранита, диорита, аплита и диабаза. Обозначим событие, заключающееся в том, что выбранная наудачу галька окажется гранитного состава, через A_1, диоритового — A_2, аплитового — A_3 и диа-

базового A_4, а соответствующие вероятности появления этих событий через p_1, p_2, p_3, p_4, т. е.

$$P(A_1) = p_1,\ P(A_2) = p_2,\ P(A_3) = p_3,\ P(A_4) = p_4. \quad (1.5)$$

Введем еще три градации для наших галек по зернистости — крупнозернистые, среднезернистые и мелкозернистые. Обозначим событие, заключающееся в том, что взятая из совокупности наудачу галька окажется крупнозернистой, через B_1, среднезернистой — B_2 и мелкозернистой B_3, а соответствующие этим событиям вероятности через q_1, q_2, q_3, т. е.

$$P(B_1) = q_1,\ P(B_2) = q_2,\ P(B_3) = q_3. \quad (1.6)$$

Заметим, что $p_1 + p_2 + p_3 + p_4 = 1$ и $q_1 + q_2 + q_3 = 1$. Например, рассмотрим событие, заключающееся в том, что выбранная наудачу галька окажется диоритом с мелкозернистой структурой; таким образом, нас интересует совместное наступление событий A_2 и B_3, вероятность которого мы обозначим через α_{23}, т. е.

$$P(A_2B_3) = \alpha_{23}. \quad (1.7)$$

В общем случае для любых A_i и B_j можно записать

$$P(A_iB_j) = \alpha_{ij}, \quad (1.8)$$

где $i = 1, 2, 3, 4$, $j = 1, 2, 3$.

Эта вероятность называется вероятностью совместного появления событий A_i и B_j.

Введем условие, заключающееся в том, что выбранная наудачу галька окажется диабазом (A_4). Спрашивается, какова вероятность того, что при выполнении этого условия A_4, галька окажется среднезернистой (B_2), т. е. требуется определить $P(B_2|A_4)$. Эта вероятность называется условной вероятностью появления события B_2 при условии, что A_4 наступило. В общем виде можно записать $P(A_i/B_j)$ — условная вероятность появления события A_i при условии, что B_j наступило. Вероятности $P(A_i|B_j)$ и $P(A_i|B_j)$ связаны следующим соотношением:

$$P(A_i \mid B_j) = \frac{P(A_iB_j)}{P(B_j)}. \quad (1.9)$$

Необходимо отметить, что если события A_i и B_j независимы в вероятностном смысле, то

$$P(A_i) = P(A_i \mid B_j) \quad (1.10)$$

и тогда

$$P(A_iB_j) = P(A_i)\,P(B_j). \quad (1.11)$$

Полученное нами равенство (1.11) очень важно в практических приложениях, в частности при проверке предположения о независимости двух или более факторов по качественным данным.

Вернемся к нашим данным, определенным формулами (1.5) и (1.6). Их можно представить в виде табл. 1.

Таблица 1

Вероятности совместного появления событий A_i и B_j

Зернистость	Типы пород				
	Гранит A_1	Диорит A_2	Аплит A_3	Диабаз A_4	Сумма
Крупнозернистые B_1	α_{11}	α_{12}	α_{13}	α_{14}	q_1
Среднезернистые B_2	α_{21}	α_{22}	α_{23}	α_{24}	q_2
Мелкозернистые B_3	α_{31}	α_{32}	α_{33}	α_{34}	q_3
Сумма	p_1	p_2	p_3	p_4	

Из этой таблицы нетрудно видеть, что

$$\sum_{j=1}^{3} \alpha_{ij} = p_i, \tag{1.12}$$

$$\sum_{i=1}^{4} \alpha_{ij} = q_j. \tag{1.13}$$

Кроме того, если зернистость не зависит от типов пород, слагающих гальки, будет верно равенство

$$\alpha_{ij} = p_i q_j. \tag{1.14}$$

Ниже, в табл. 2, приведен числовой модельный пример распределения вероятностей, когда два фактора A и B независимы.

Таблица 2

Вероятности в условиях независимости факторов A и B

	A_1	A_2	A_3	A_4	Сумма
B_1	0,15	0,06	0,06	0,03	0,30
B_2	0,30	0,12	0,12	0,06	0,60
B_3	0,05	0,02	0,02	0,01	0,10
Сумма	0,50	0,20	0,20	0,10	1,00

Из этой таблицы видно, что соотношения (1.12), (1.13) и (1.14) выполнены для всех ее элементов.

В заключение этого раздела заметим, что для любых двух событий A и $\overline{A}$, где $\overline{A}$ — событие, противоположное A, имеет место равенство

$$P(A) + P(\overline{A}) = 1. \tag{1.15}$$

Кроме того, отметим, что вероятность появления хотя бы одного из набора независимых случайных событий A_1, A_2, . . . , A_i, . . . , A_k равна сумме вероятностей этих событий, т. е.

$$P(A_1 \text{ или } A_2 \text{ или } \ldots \text{ или } A_k) = \sum_{i=1}^{k} P(A_i). \qquad (1.16)$$

Это соотношение наглядно иллюстрируется табл. 2.

Заметим, также что равенство (1.14) можно распространить на любое число независимых событий A_1, A_2, . . . , A_k, т. е.

$$P(A_1, A_2, \ldots, A_k) = \prod_{i=1}^{k} P(A_i), \qquad (1.17)$$

что иногда называют теоремой умножения вероятностей.

1.6.2. Случайные величины и функции распределения

В разделе 1.6.1 в качестве примера случайного события мы рассматривали наличие или отсутствие знаков золота в шлихе. Однако число золотинок может быть равно 0, 1, 2, 3 и т. д. Таким образом, несколько детализируя нашу задачу, мы можем рассматривать некоторую дискретную величину (обозначим ее ξ), которая принимает значения 0, 1, 2, . . . , N, но значения эти нельзя предсказать точно.

Еще один пример. Содержания оливина в пробах перидотита, отобранных даже из весьма однородного массива, будут, как и в примере с золотом, меняться случайным образом, только их значения будут сплошь заполнять некоторый интервал, а предсказать их точно заранее, как и в первом примере, невозможно.

Таким образом, необходимо ввести новое понятие «случайная величина», т. е. величина, меняющаяся от эксперимента к эксперименту, значения которой нельзя предсказать точно, а можно их получить только приписав им соответствующую вероятность. Такое определение случайной величины ближе к интуитивному, чем к строго формальному, но для наших целей оно достаточно. Заинтересованный читатель может найти более строгое рассмотрение данного понятия в других работах [17, 26, 34].

Вернемся к нашему первому примеру. В качестве модели числа золотинок в одном наудачу взятом шлихе будем рассматривать случайную величину ξ. Тогда значениям ее $x_i = 0, 1, \ldots, N$ можно поставить в соответствие события $\xi = 0$, $\xi = 1$, $\xi = 2$, . . . , приписав им вероятности

$$P(\xi = x_i) = P_i. \qquad (1.18)$$

Зная значения p_i для всех x_i, можно определить вероятность события $\xi \leqslant x_k$, для всех $k = 1, 2, \ldots$, т. е.

$$F_\xi(x_k) = \sum_{i=1}^{k} P(\xi = x_i) = P(\xi \leqslant x_k). \qquad (1.19)$$

Функцию $F_\xi(x_k)$, определенную формулой (1.19), называют функцией распределения случайной величины ξ. В том случае, когда ξ непрерывна и $F_\xi(x) = P(\xi \leqslant x)$ имеет производную, $f_\xi(x) = dF_\xi(x)/dx$,

$$E_\xi(x) = \int_{-\infty}^{x} f(v)\,dv. \tag{1.20}$$

Таким образом, функция распределения случайной величины ξ как непрерывной, так и дискретной есть вероятность события, заключающегося в том, что в результате единичного эксперимента ξ примет значение, меньшее или равное заданному x, причем эта функция определена для всех x из области значений ξ. Следует отметить, что если заданы два значения x_1 и x_2, то вероятность события, заключающегося в том, что значение случайной величины ξ будет принадлежать интервалу от x_1 до x_2, определится выражением

$$P(x_1 \leqslant \xi \leqslant x_2) = F_\xi(x_2) - F_\xi(x_1). \tag{1.21}$$

В том случае, когда ξ дискретна

$$F_\xi(x_2) - F_\xi(x_1) = \sum_{x_1}^{x_2} P(\xi \leqslant x), \tag{1.22}$$

а когда непрерывна

$$F_\xi(x_2) - F_\xi(x_1) = \int_{x_1}^{x_2} f_\xi(v)\,dv. \tag{1.23}$$

Заметим также, что для дискретной случайной величины функция $F\,\xi(x)$ будет кусочно-постоянной. Кроме того, для непрерывной случайной величины характерно, что $P(\xi = x) = 0$, так как

$$\int_{x}^{x} f_\xi(v)\,dv = 0. \tag{1.24}$$

На рис. 2 приведены примеры дискретного и непрерывного распределения.

Одна из характеристик случайных величин, которая является наиболее важной,— это ее математическое ожидание или ее среднее значение. В дискретном случае эта характеристика определена выражением

$$M\xi = \sum_{k=1}^{\infty} x_k p_k. \tag{1.25}$$

Таким образом, математическое ожидание ($M\,\xi$) дискретной случайной величины (ξ) — это сумма произведений всех значений данной величины (x_k) на соответствующие им вероятности (p_k).

В случае непрерывной случайной величины

$$M\xi = \int_{-\infty}^{\infty} x f_\xi(x)\,dx. \tag{1.26}$$

Простейшие свойства математического ожидания сводятся к следующему.

1. Математическое ожидание постоянной a равно этой постоянной, т. е.

$$M\,a = a. \tag{1.27}$$

2. Математическое ожидание суммы двух случайных величин ξ_1 и ξ_2, имеющих математические ожидания $M\,\xi_1$, $M\,\xi_2$, равно сумме этих математических ожиданий, т. е.

$$M\,(\xi_1 + \xi_2) = M\,\xi_1 + M\,\xi_2. \tag{1.28}$$

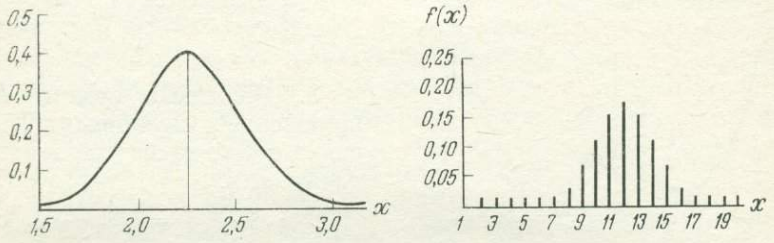

Рис. 2. Примеры непрерывного и дискретного распределений

3. Математическое ожидание произведения независимых случайных величин ξ_1 и ξ_2 равно произведению их математических ожиданий, т. е.

$$M\,(\xi_1\xi_2) = M\,\xi_1 \cdot M\,\xi_2. \tag{1.29}$$

Весьма важными характеристиками распределений случайных величин являются моменты. Моментом порядка k (или k-м моментом) случайной величины ξ относительно константы a, называется число, определенное выражением

$$M_a^k = M\,(\xi - a)^k. \tag{1.30}$$

Если $a = 0$, то момент называется начальным (обозначим его (μ_k) и определяется выражением

$$\mu_k = M\xi^k, \tag{1.31}$$

а если $a = M\,\xi$, то момент называется центральным. Его мы будем обозначать ν_k

$$\nu_k = M\,(\xi - M\xi)^k. \tag{1.32}$$

Таким образом, $\mu_1 = M\,\xi$, т. е. первый начальный момент случайной величины совпадает с ее математическим ожиданием.

Весьма важной характеристикой распределения случайной величины является ее второй центральный момент (ν_2), который представляет собой меру рассеяния значений случайной величины от-

носительно среднего значения. Эту характеристику называют дисперсией и обозначают $D\xi$ или σ^2, т. е.

$$D\xi = M(\xi - M\xi)^2 = M\xi^2 - (M\xi)^2 = \mu_2 - \mu_1^2. \quad (1.33)$$

Заметим, что $\sqrt{D\xi} = \sigma$ называется стандартным отклонением. Простейшие свойства дисперсии сводятся к следующим.

1. Дисперсия постоянной величины равна нулю, т. е.

$$Da = 0. \quad (1.34)$$

2. Дисперсия произведения постоянной a и случайной величины (ξ) определена выражением

$$D(a\xi) = a^2 D\xi = a^2\sigma^2. \quad (1.35)$$

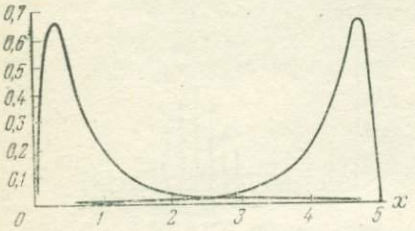

Рис. 3. Положительно и отрицательно асимметричные распределения

3. Дисперсия суммы двух независимых случайных величин равна сумме дисперсий этих величин, т. е.

$$D(\xi_1 + \xi_2) = D\xi_1 + D\xi_2. \quad (1.36)$$

Кроме среднего значения и дисперсии, следует остановиться еще на двух характеристиках распределений случайных величин — асимметрии и эксцессе. Величина

$$\gamma_3 = \frac{\nu_3}{\delta^3}, \quad (1.37)$$

т. е. равная отношению третьего центрального момента к кубу стандартного отклонения, называется асимметрией. Эта величина является мерой асимметричности распределения. Если $\gamma_3 = 0$, распределение симметрично относительно среднего значения, если $\gamma_3 > 0$, то максимум кривой плотности вероятности смещен влево относительно среднего, а если $\gamma_3 < 0$, то он смещен вправо. Примеры приведены на рис. 3.

Эксцессом называют величину

$$\gamma_4 = \frac{\nu_4}{\delta^4} - 3, \quad (1.38)$$

где γ_4 — четвертый центральный момент; σ^4 — квадрат дисперсии. Эта характеристика является мерой крутизны кривой плотности распределения, причем за эталон взята нормальная кривая, которая подробно будет рассмотрена в дальнейшем. Если $\gamma_4 = 0$, кривая по крутизне не отличается от нормальной, если $\gamma_4 > 0$, распределение островершинное, а если $\gamma_4 < 0$, оно имеет уплощенную вершину по сравнению с плотностью нормального распределения.

В связи с тем что в дальнейшем нам неоднократно придется сталкиваться с различными функциями распределения случайных величин в самых разнообразных практических задачах применения статистических методов при обосновании геологических выводов, необходимо кратко рассмотреть наиболее распространенные виды этих функций и условия их возникновения.

Биномиальное распределение

Рассмотрим следующую ситуацию, взятую из геологической практики. Допустим, что мы занимаемся изучением акцессорных минералов в гранитах, для чего производим количественно-минералогический анализ фракций концентрата, выделенного из пробы изучаемой породы. Обычно в процессе такого анализа из изучаемой фракции под бинокуляром путем квартования отделяется проба, в которой отсчитывается n зерен (обычно 500). Среди них определяется число зерен, принадлежащих к тому или иному минералу, после чего определяются содержания последних во фракции.

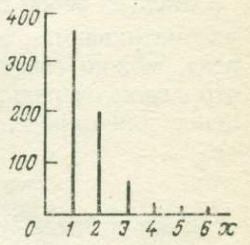

Рис. 4. Биномиальное распределение

Допустим, что в пробе, состоящей из n зерен мы подсчитываем число зерен ильменита. Обозначим событие, заключающееся в том, что взятое наудачу из фракции зерно окажется ильменитом, через A, а противоположное событие, заключающееся в том, что оно не будет ильменитом, через $\overline{A}$. Число зерен во фракции мы можем считать бесконечно большим. Вероятность появления события A в результате выбора из фракции одного зерна мы обозначим через p, а $P(\overline{A}) = 1-p = q$.

Допустим, что мы произвели подсчет n зерен, среди которых k раз наблюдали ильменит. Если фракцию перемещать и снова повторить подсчет n зерен, то мы получим некоторое новое значение числа зерен ильменита, т. е. числа наступления события A. Таким образом, число зерен ильменита в подсчете из n зерен можно рассматривать как случайную величину, которую мы обозначим ξ. Если для каждого $k = 1, 2, \ldots, n$ определить вероятность события $\xi = k$, то мы получим распределение случайной величины ξ. Эти вероятности будут определены следующим выражением:

$$P(\xi = k) = P(k) = C_n^k p^k (1-p)^{n-k}, \qquad (1.39)$$

где C_n^k — число сочетаний из n по k,

$$P = P(A),\ 1-p = q = P(A).$$

Соответственно функция распределения $F(m)$ определится следующим выражением:

$$F(m) = P(\xi \leqslant m) = \sum_{k=0}^{m} P(k) = \sum_{k=0}^{m} C_n^k (1-p)^{n-k} p^k. \qquad (1.40)$$

На рис. 4 приведен пример биномиального распределения. Математическое ожидание $M\xi$ случайной величины ξ, распределенной по биномиальному закону, будет равно np, а дисперсия — npq, т. е.

$$D\xi = npq. \qquad (1.41)$$

Заметим, что биномиальное распределение иногда называют распределением Бернулли.

Следует особо подчеркнуть, что к описанной выше схеме, когда рассматриваются только два события A и $\overline{A}$, т. е. «успех» и « неуспех», можно свести многие реализуемые на практике процедуры, что в свою очередь, позволяет использовать биномиальное распределение для определения соответствующих вероятностей.

Распределение Пуассона

Если в предыдущем примере принять, что вероятность p мала, а число n достаточно велико, то формула (1.39) будет давать плохое приближение. В данном случае для определения вероятности того, что число зерен ильменита примет значение k, лучше воспользоваться формулой

$$P(\xi = k) = P(k) = \frac{a^k}{k!} e^{-a}, \quad (k = 0, 1, 2, \ldots) \qquad (1.42)$$

где a — единственный параметр этого распределения и математическое ожидание случайной величины ξ, т. е.

$$a = M\xi = \sum_{k=0}^{\infty} kP(k). \qquad (1.43)$$

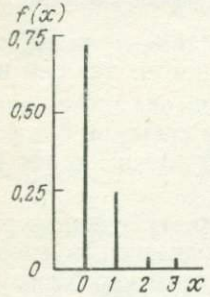

Рис. 5. Распределение Пуассона

Заметим также, что дисперсия этого распределения также равна a. т, е.

$$D\xi = M\xi = a. \qquad (1.44)$$

Функцию распределения Пуассона можно выразить формулой

$$F(m) = P(\xi \leqslant m) = \sum_{k=0}^{m} \frac{a^k}{k!} e^{-a}. \qquad (1.45)$$

На рис. 5 приведено графическое изображение распределения Пуассона.

В геологических задачах распределение Пуассона можно использовать в качестве модели распределения числа наступления редких событий, при описании некоторых геологических процессов, процессов поисков и процессов анализа минерального вещества.

Равномерное распределение

Распределение случайной величины ξ называют равномерным в интервале (a, b), если соответствующая ей плотность вероятности определена выражением

$$f_\xi(x) = \begin{cases} \dfrac{1}{b-a} & \text{при } a \leqslant x \leqslant b \\ 0 \text{ при } x < a,\ x > b. \end{cases} \tag{1.46}$$

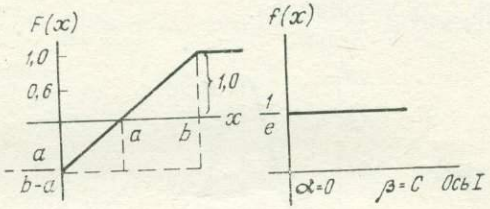

Рис. 6. Равномерное распределение

Тогда функция распределения $F_\xi(x)$ будет

$$F_\xi(x) = P(\xi \leqslant x) = \frac{1}{b-a}\int_a^x dv = \frac{x-a}{b-a}. \tag{1.47}$$

На рис. 6 приведены графики плотности вероятностей и функции распределения равномерного распределения. Следует отметить, что иногда это распределение называют прямоугольным.

Используя формулы (1.26) и (1.33), нетрудно подсчитать, что для равномерно распределенной случайной величины ξ

$$M\xi = \frac{a+b}{2}, \tag{1.48}$$

$$D\xi = \frac{(a-b)^2}{12}. \tag{1.49}$$

Нормальное распределение

Если для случайной величины ξ функция распределения $F_\xi(x)$ определена выражением

$$F_\xi(x) = \frac{1}{\delta\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{(v-\mu)^2}{2\delta^2}}\, dv, \tag{1.50}$$

то такое распределение называется нормальным.

Соответствующая функция плотности вероятности будет иметь вид

$$f_\xi(x) = \frac{1}{\delta\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\delta^2}}, \qquad (1.51)$$

где μ и σ^2 — параметры распределения. На рис. 7 приведены графики функций $F_\xi(x)$ и $f_\xi(x)$, определенных формулами (1.50) и (1.51). Нетрудно видеть, что кривая $f_\xi(x)$ является симметричной, причем достигающей максимума в точке $x = \mu$. Следует отметить, что в случае нормального распределения $M\xi = \mu$, а $D\xi = \sigma^2$. Кроме того, асимметрия $\gamma_3 = 0$, и эксцесс $\gamma_4 = 0$, причем отношение $\nu_4/\sigma^4 = 3$. Величина σ, будучи мерой рассеяния отдельных значений ξ относительно среднего μ, в значительной мере влияет на форму кривой $f_\xi(x)$. На рис. 7 хорошо видно, что кривая с меньшим значением σ более островершинная.

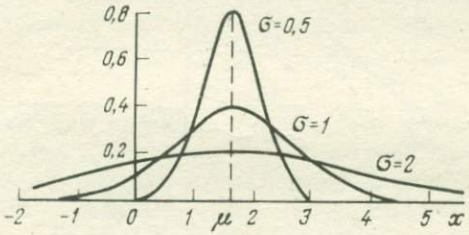

Рис. 7. Нормальные плотности вероятности

Рассмотрим вместо ξ новую случайную величину

$$\tau = \frac{\xi - \mu}{\delta},$$

которая будет, как и ξ, распределена нормально, но с математическим ожиданием, равным нулю, и дисперсией, равной 1, т. е.

$$P(\tau \leqslant t) = \Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{v^2}{2}} dv. \qquad (1.51')$$

Функция плотности примет вид

$$f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}. \qquad (1.52)$$

Такое распределение называется стандартным нормальным распределением, для которого составлены таблицы [6, 11, 13], сокращенный вариант которых приведен в конце книги (см. прил. 1).

Характерно, что сумма нормально распределенных случайных величин также будет распределена нормально. Более того, если $\xi_1, \xi_2, \ldots, \xi_i, \ldots, \xi_n$ — набор случайных величин, которые независимы, одинаково распределены и равномерно малы, то распределение их суммы будет приближаться к нормальному с ростом n. Условия сходимости распределений для сумм случайных величин к нормальному закону определяются группой теорем, объединенных под общим названием центральной предельной теоремы. В данном случае мы ограничимся только рассмотрением условий, являющихся достаточными для возникновения нормального распределения.

Пусть $\xi_1, \xi_2, \ldots, \xi_i, \ldots, \xi_n$ — последовательность независимых случайных величин, имеющих математические ожидания $\mu_i = M\xi_i$ и дисперсии $\sigma_i^2 = D\xi_i$. Кроме того, $|\xi_i| \leqslant K$, где K — константа. Если дисперсию суммы $\sum_{i=1}^{n} \xi_i$ обозначить через B_n^2, то можно записать

$$B_n^2 = \sum_{i=1}^{n} \delta_i^2. \tag{1.53}$$

Соответственно тому, как это указано в работе А. Бэрри [50], для случайной величины

$$\frac{1}{B_n} \sum_{i=1}^{n} (\xi_i - \mu_i) \tag{1.54}$$

будет справедливо неравенство

$$\left| P\left\{ \frac{1}{B_n} \sum_{i=1}^{n} (\xi_i - \mu_i) < x \right\} - \Phi(x) \right| \leqslant C \frac{K}{B_n}, \tag{1.55}$$

где $\Phi(x)$ — нормальная функция распределения, C — константа (положим $C = 7{,}5$), P — вероятность события, указанная в фигурных скобках. Это неравенство означает, что если константа K мала по сравнению с B_n, т. е. каждая случайная величина $\xi_i - \mu_i$ мала по сравнению с B_n, то распределение случайной величины

$$\frac{1}{B_n} \sum_{i=1}^{n} (\xi_i - \mu_i) \tag{1.56}$$

будет близко к нормальному. Это условие заключается в требовании равномерной малости всех слагаемых в сумме.

Логарифмически-нормальное распределение

Пусть ξ — непрерывная положительная случайная величина. Если ее логарифм распределен нормально, то мы говорим, что величина ξ распределена логарифмически-нормально, т. е. если

$$P(\ln \xi \leqslant \ln x) = \frac{1}{\delta\sqrt{2\pi}} \int_{-\infty}^{\ln x} e^{-\frac{(\ln v - \mu)^2}{2\delta^2}} d \ln v = N(\ln x;\ \mu,\ \sigma^2), \tag{1.57}$$

$$P(\xi \leqslant x) = \frac{1}{\delta\sqrt{2\pi}} \int_0^x \frac{1}{v} e^{-\frac{(\ln v - \mu)^2}{2\delta^2}} dv = \lambda(x;\ \mu,\ \delta^2). \qquad (1.58)$$

Соответствующая $\Lambda(x;\ \mu,\ \sigma^2)$ функция плотности вероятности будет определена выражением

$$f_\xi(x) = \frac{1}{x\delta\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\delta^2}}. \qquad (1.59)$$

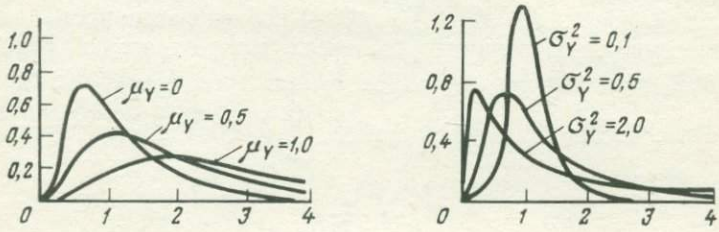

Рис. 8. Логарифмически нормальные плотности вероятности. Слева при постоянном значении σ и разных μ, справа — при постоянном μ и разных σ

Эта функция изображена на рис. 8. Нетрудно видеть, что она положительно асимметрична и достигает максимума в точке $x = e^{\mu-\sigma^2}$, где $\mu = M \ln \xi$, $\sigma^2 = D \ln \xi$. Необходимо отметить, что нередко в геологической практике приходится сталкиваться с ошибочным мнением, что мода логарифмически-нормального распределения равна величине e^{μ}. В действительности, мода, равная $e^{\mu-\sigma^2}$, меньше, чем e^{μ}, причем e^{μ} является медианой этого распределения.

Математическое ожидание логнормально распределенной случайной величины ξ будет определено следующим выражением:

$$M\xi = e^{\mu+\frac{1}{2}\delta^2}, \qquad (1.60)$$

а дисперсия

$$D\xi = e^{2\mu+\delta^2}(e^{\delta^2}-1), \qquad (1.61)$$

где $(e^{\delta^2}-1)$ — квадрат коэффициента вариации. Подробно вопрос о моментах логнормального распределения рассмотрен в книге Дж. Ачисона и Дж. Брауна [48] и в одной из работ автора [37].

Логарифмически-нормальное распределение нередко используется в качестве модели наблюдаемых геологических характеристик, особенно содержаний элементов и минералов в породах. Бо-

лее того, в геологической практике встречаются также распределения, моделями которых служат некоторые распределения, связанные с логнормальными. Рассмотрим некоторые из них.

Пусть ξ — положительная случайная величина, распределенная логнормально с параметрами μ и σ^2, т. е.

$$P(\xi \leqslant x) = \Lambda(x; \mu, \sigma^2). \qquad (1.62)$$

Рассмотрим новую случайную величину $\eta = 1/\xi = \xi^{-1}$. Без особых затруднений можно записать следующую цепочку равенств:

$$P(\eta \leqslant z) = \lambda(z; -\mu, \delta^2) = \lambda(x^{-1}; \mu, \delta^2) = \frac{1}{\delta\sqrt{2\pi}} \int_0^z e^{-\frac{(\ln v - \mu)^2}{2\delta^2}} \frac{dv}{v}. \qquad (1.63)$$

Соответствующая этой функции распределения плотность вероятности будет определена выражением

$$f_\eta(z) = \frac{1}{z\delta\sqrt{2\pi}} e^{-\frac{(\ln z + \mu)^2}{2\delta^2}}. \qquad (1.64)$$

Используя полученное выражение плотности вероятности (1.64), нетрудно подсчитать значения математического ожидания и дисперсии случайной величины η:

$$M_\eta = e^{\frac{1}{2}\delta^2 - \mu}, \qquad (1.65)$$

$$D\eta = e^{\delta^2 - 2\mu}\left(e^{\delta^2} - 1\right). \qquad (1.66)$$

Рассмотрим еще одну случайную величину $\frac{1}{\xi}$. Можно показать, что

$$\lambda\left(\frac{1}{x}; -\mu, \delta^2\right) = 1 - \lambda(x; \mu, \delta^2) \qquad (1.67)$$

или

$$P(\xi \leqslant x) = \lambda(x; \mu, \delta^2) = P\left(\frac{1}{\xi} \geqslant \frac{1}{x}\right). \qquad (1.68)$$

Соответственно можно записать

$$P\left(\frac{1}{\xi} \leqslant \frac{1}{x}\right) = 1 - P\left(\frac{1}{\xi} \geqslant \frac{1}{x}\right) = 1 - \lambda(x; \mu, \delta^2). \qquad (1.69)$$

Таким образом,

$$\lambda\left(\frac{1}{x}; -\mu, \delta^2\right) = 1 - \lambda(x, \mu, \delta^2), \qquad (1.70)$$

откуда

$$\lambda(x;\ \mu,\ \delta^2) = 1 - \lambda\left(\frac{1}{x};\ -\mu,\ \delta^2\right). \tag{1.71}$$

В геологической практике иногда приходится использовать в качестве модели случайную величину, представляющую собой сумму константы и случайной величины, т. е.

$$\eta = a + \xi, \tag{1.72}$$

где a — константа, а ξ — случайная величина. Если ξ распределена логарифмически-нормально с параметрами μ, σ^2, тогда

$$P(\eta \leqslant z) = P(\eta - a \leqslant z - a) = P(\xi \leqslant z - a) = \lambda(z - a;\ \mu,\ \delta^2) \tag{1.73}$$

Функция плотности вероятности этого распределения будет определена выражением

$$f_\eta(z) = \frac{1}{(z-a)\,\delta\sqrt{2\pi}}\, e^{-\frac{[\ln(z-a)-\mu]^2}{2\delta^2}}, \quad z > a. \tag{1.74}$$

Эта кривая достигает максимума в точке

$$z = a + e^{\mu - \delta^2} \tag{1.75}$$

и обращается в нуль в точке $z = a$. Используя формулу (1.74), нетрудно вычислить, что

$$M\eta = a + e^{\mu + \frac{1}{2}\delta^2} = a + M\xi, \tag{1.76}$$

$$D\eta = e^{2\mu + \delta^2}\left(e^{\delta^2} - 1\right) = D\xi. \tag{1.77}$$

Список рассматриваемых нами распределений, связанных с логнормальным, будет неполным без распределения случайной величины $\eta = c - \xi$, где c — константа, а ξ — логнормально распределенная случайная величина с параметрами μ, σ^2. Плотность вероятности этой величины определена выражением

$$f_\eta(y) = \frac{1}{\delta(c-y)\sqrt{2\pi}}\, e^{-\frac{[\ln(c-y)-\mu]^2}{2\delta^2}}, \quad y < c. \tag{1.78}$$

Нетрудно определить, что

$$M\eta = c - e^{\mu + \frac{1}{2}\delta^2}, \tag{1.79}$$

$$D\eta = e^{2\mu + \delta^2}\left(e^{\delta^2} - 1\right) = D\xi. \tag{1.80}$$

Обозначив $\gamma_3^{(\eta)}$, $\gamma_3^{(\xi)}$, $\gamma_4^{(\eta)}$, $\gamma_4^{(\xi)}$ асимметрию и эксцесс случайных величин η и ξ соответственно, можем записать

$$\gamma_3^{(\eta)} = -\gamma_3^{(\xi)}, \tag{1.81}$$

$$\gamma_4^{(\eta)} = \gamma_4^{(\xi)}. \tag{1.82}$$

На рис. 8 изображены плотности вероятности распределений случайных величин ξ и η, причем отчетливо видно, что эти кривые являются зеркальным отражением друг друга.

В заключение настоящего раздела необходимо сказать несколько слов об условиях возникновения логарифмически-нормальных распределений, которые, как и в случае нормального распределения, определены центральной предельной теоремой.

Пусть $\xi_1, \xi_2, \ldots, \xi_i, \ldots, \xi_n$ — последовательность независимых случайных величин, таких, что

$$0 < c' < \xi_i < c'' < \infty,$$

$$|\log \xi_i| < \max [|\log c'|, |\log c''|] = K, \tag{1.83}$$

где c', c'' и K являются константами. Кроме того, математические ожидания и дисперсии случайных величин $\log \xi_i$ существуют и являются конечными, т. е.

$$M \log \xi_i = \mu_i, \tag{1.84}$$

$$D \log \xi_i = \sigma_i^2. \tag{1.85}$$

Определим математическое ожидание и дисперсию случайной величины $\log \prod_{i=1}^{n} \xi_i$ с помощью выражений

$$M\left(\log \prod_{i=1}^{n} \xi_i\right) = \sum_{i=1}^{n} \mu_i = \mu_{(n)}, \tag{1.86}$$

$$D\left(\log \prod_{i=1}^{n} \xi_i\right) = \sum_{i=1}^{n} \sigma_i^2 = \sigma_{(n)}^2. \tag{1.87}$$

Теперь мы можем написать следующую цепочку равенств:

$$\left| P\left(\prod_{i=1}^{n} \xi_i \leqslant x\right) - \Lambda\left(x;\ \mu_{(n)},\ \sigma_{(n)}^2\right) \right| =$$

$$= \left| P\left(\prod_{i=1}^{n} \xi_i \leqslant e^y\right) - \Lambda\left(e^y;\ \mu_{(n)},\ \sigma_{(n)}^2\right) \right| =$$

$$= \left| P\left(\sum_{i=1}^{n} \log \xi_i \leqslant y\right) - N\left(y;\ \mu_{(n)},\ \sigma_{(n)}^2\right) \right|. \tag{1.88}$$

Таким образом, сходимость распределения случайной величины $\prod_{i=1}^{n} \xi_i$ к логарифмически-нормальному закону равносильна сходимости распределения случайной величины $\sum_{i=1}^{n} \log \xi_i$ к нормальному распределению

$$\left| P\left(\sum_{i=1}^{n} \log \xi_i \leqslant y\right) - N\left(y;\ \mu_{(n)},\ \sigma_{(n)}^2\right) \right| \leqslant c \frac{K}{\sigma_{(n)}}. \tag{1.89}$$

Из этого выражения следует, что распределение $P\left(\sum_{i=1}^{n} \log \xi_i \leqslant y\right)$ будет приближаться к нормальному, если константа K мала по сравнению с параметром $\sigma^2_{(n)}$, что определяет условие равномерной малости слагаемых $\log \xi_i$ в сумме $\sum_{i=1}^{n} \log \xi_i$ по сравнению с $\sigma_{(n)}$.

Распределение χ^2

Пусть $\xi_1, \xi_2, \ldots, \xi_i, \ldots, \xi_n$ — набор независимых одинаково нормально распределенных случайных величин с параметрами 0,1. Образуем новую случайную величину

$$\chi^2 = \sum_{i=1}^{n} \xi_i^2. \tag{1.90}$$

Функция распределения такой случайной величины будет определена выражением

$$P\left(\chi^2 \leqslant x\right) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \int_0^x v^{\frac{n}{2}-1} e^{-\frac{v}{2}} dv, \tag{1.91}$$

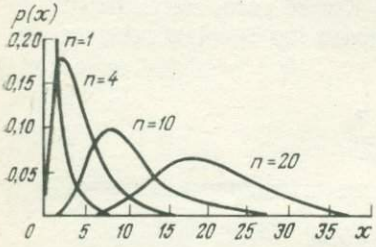

Рис. 9. Плотность вероятности χ^2-распределения при разных значениях числа степеней свободы

а соответствующая ей плотность вероятности будет

$$f(x) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, \quad 0 < x < \infty, \tag{1.92}$$

где $\Gamma\left(\frac{n}{2}\right)$ — гамма-функция, причем

$$\Gamma(a) = \int_0^\infty y^{a-1} e^{-y} dy, \tag{1.93}$$

и для любого действительного положительного a имеет место соотношение

$$\Gamma(a+1) = a\Gamma(A). \tag{1.94}$$

Если a целое число,

$$\Gamma(a) = (a-1)! \tag{1.95}$$

Из формул (1.91) и (1.92) видно, что распределение χ^2 определяется только одним параметром n, который называется числом степеней свободы. На рис. 9 приведены графики различных плотностей χ^2-распределения при разном значении n.

Необходимо отметить, что распределение случайной величины

$$\tau = \frac{\chi^2 - n}{\sqrt{2n}} \tag{1.96}$$

с ростом n приближается к нормальному с параметрами 0,1. Заметим также, что функция (1.91) табулирована, и таблицы можно найти в работе Л. Н. Большева и Н. В. Смирнова [6], сокращенный вариант которых приведен в конце книги (см. прил. 2).

Распределение Стьюдента

Пусть $\xi_1, \xi_2, \ldots, \xi_i, \ldots, \xi_n$ — набор независимых одинаково нормально распределенных случайных величин с параметрами 0,1 и пусть ξ_0 — также нормально распределенная величина, имеющая те же параметры. Образуем новую случайную величину

$$\tau = \frac{\xi_0}{\sqrt{\sum_{i=1}^{n} \xi_i^2}}. \tag{1.97}$$

Функция распределения этой величины имеет вид

$$P(\tau \leqslant t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} \int_{-\infty}^{t} \left(1 + \frac{v^2}{2}\right)^{-\frac{n+1}{2}} dv \tag{1.98}$$

и называется распределением Стьюдента, которое иногда называют также t-распределением. Нетрудно видеть, что это распределение, как и распределение χ^2, определяется только одним параметром n, который называется числом степеней свободы. Эта функция табулирована, и ее таблицы обычно имеются во многих учебниках математической статистики [6, 11, 19], сокращенный вариант которых приведен в конце книги (см. прил. 6).

Некоторые многомерные обобщения

Пусть $\xi_1, \xi_2, \ldots, \xi_i, \ldots, \xi_m$ — набор m случайных величин, который мы будем рассматривать как m-мерный вектор-столбец или вектор-строку, т. е.

$$\Xi = \{\xi_1, \xi_2, \ldots, \xi_i, \ldots, \xi_m\} \tag{1.99}$$

или

$$\Xi = \begin{vmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_i \\ \vdots \\ \xi_m \end{vmatrix}.$$

Пусть также $X = \{x_1, x_2, \ldots, x_i, \ldots, x_m\}$ — набор заданных значений $\{\xi_1, \xi_2, \ldots, \xi_i \ldots, \xi_m\}$. Тогда m-мерную функцию распределения случайной величины Ξ можно определить как вероятность совместного выполнения неравенств:

$$\xi_1 \leqslant x_1, \ \xi_2 \leqslant x_2, \ \ldots, \ \xi_i \leqslant x_i, \ \ldots, \ \xi_m \leqslant x_m, \quad (1.100)$$

т. е.

$$F(X) = P(\Xi \leqslant X) = P(\xi_1 \leqslant x_1, \ \ldots, \ \xi_i \leqslant x_i, \ \ldots, \ \xi_m \leqslant x_m). \quad (1.101)$$

Если все ξ_i независимы,

$$F(X) = \prod_{i=1}^{m} P(\xi_i \leqslant x_i). \quad (1.102)$$

Как и в одномерном случае, для m-мерного распределения можно определить m-мерную плотность вероятности. Если $F(X)$ непрерывна и дифференцируема, то для нее можно определить

$$f(X) = \frac{d}{dX} F(X), \quad (1.103)$$

где $dX = dx_1 dx_2, \ldots, dx_i, \ldots, dx_m$.

Таким образом,

$$F(X) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_m} f(v_1, v_2, \ldots, v_m)\, dv_1 dv_2, \ldots, dv_m. \quad (1.104)$$

В качестве примера непрерывного многомерного распределения m-мерной случайной величины $\Xi = \{\xi_1, \xi_2, \ldots, \xi_i, \ldots, \xi_m\}$ рассмотрим m-мерное нормальное распределение, функция плотности которого определена выражением

$$f(X) = (2\pi)^{\frac{-m}{2}} |\Sigma|^{\frac{-1}{2}} \exp\left[-\frac{1}{2}\{X-\underline{\mu}\}\Sigma^{-1}\{X-\underline{\mu}'\}\right], \quad (1.105)$$

где $X = \{x_1, x_2, \ldots, x_i, \ldots, x_m\}$ — m-мерный вектор-строка, $\underline{\mu} = \{\mu_1, \mu_2, \ldots, \mu_i, \ldots \mu_m\}$ — m-мерный вектор-строка, элементами которого являются математические ожидания случайных величин ξ_i, образующих m-мерную случайную величину Ξ, Σ — ковариационная матрица порядка $m \times m$, каждый элемент которой σ_{ij} определяется выражением

$$\sigma_{ij} = M[(\xi_i - M\xi_i)(\xi_j - M\xi_j)]. \quad (1.106)$$

Ясно, что диагональные элементы σ_{ii} матрицы Σ представляют собой дисперсии случайных величин ξ_i, так как

$$\sigma_{ii} = M(\xi_i - M\xi_i)^2 = \sigma_i^2. \quad (1.107)$$

Недиагональные же элементы σ_{ij} этой матрицы являются характеристиками силы линейной зависимости между случайными

величинами ξ_i и ξ_j и называются ковариациями. Чаще всего на практике в качестве меры линейной зависимости используется коэффициент корреляции ρ_{ij}, который определен формулой

$$\rho_{ij} = \sigma_{ij} / \sqrt{\sigma_{ii}\sigma_{jj}}. \tag{1.108}$$

Заметим, что в формуле (1.105) $|\Sigma|$ — детерминант матрицы Σ, Σ^{-1} — матрица, обратная матрице Σ, а $\{X - \underline{\mu}\}'$ — вектор-столбец, являющийся транспонированной вектор-строкой $\{X - \underline{\mu}\}$.

В качестве примера рассмотрим наиболее простой случай — двумерное нормальное распределение. Пусть $\Xi = \{\xi_1, \xi_2\}$ — двумерная нормальная случайная величина, $\underline{\mu} = \{\mu_1, \mu_2\}$ — ее математическое ожидание, а

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

— ее ковариационная матрица. Используя формулу (1.105), можем записать

$$f(x_1, x_2) = (2\pi)^{-1} \left| \begin{matrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{matrix} \right|^{-\frac{1}{2}} \exp \left[-\frac{1}{2} \{x_1 - \mu_1,\ x_2 - \mu_2\} \times \right.$$
$$\left. \times \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}^{-1} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \right]. \tag{1.109}$$

Вычислив детерминант и обратную матрицу с последующим выражением квадратичной формы, указанной в фигурных скобках, получим

$$f(x_1 x_2) = \frac{1}{2\pi\sigma_1\sigma_2 \left(\dfrac{\sigma_1^2\sigma_2^2 - \sigma_{12}}{\sigma_1^2\sigma_2^2} \right)^{\frac{1}{2}}} \times$$
$$\times \exp \left\{ -\frac{1}{2} \left[\frac{\left(\dfrac{x_1 - \mu_1}{\sigma_1}\right)^2 + \left(\dfrac{x_2 - \mu_2}{\sigma_2}\right)^2 - 2\dfrac{\sigma_{12}}{\sigma_1\sigma_2}\left(\dfrac{x_1 - \mu_1}{\sigma_1}\right)\left(\dfrac{x_2 - \mu_2}{\sigma_2}\right)}{\dfrac{\sigma_1^2\sigma_2^2 - \sigma_{12}}{\sigma_1^2\sigma_2^2}} \right] \right\}. \tag{1.110}$$

Обозначив $\sigma_{12}/\sigma_1\sigma_2 = \rho$, получим

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times$$
$$\times \exp \left\{ -\frac{1}{2(1-\rho)} \left[\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x_1 - \mu_1}{\sigma_1}\right)\left(\frac{x_2 - \mu_2}{\sigma_2}\right) \right] \right\}. \tag{1.111}$$

Если случайные величины ξ_1 и ξ_2 независимы, то безразмерный параметр $\rho = 0$ и приведенная выше формула примет вид

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left\{-\frac{1}{2}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\}. \quad (1.112)$$

В связи с тем, что логнормальное распределение нередко используется в геологических исследованиях в качестве модели реальных распределений различных геологических характеристик, будет полезно рассмотреть m-мерное логнормальное распределение. Пусть $\Xi = \{\xi_1, \xi_2, \ldots, \xi_i, \ldots, \xi_m\}$ m-мерная случайная величина, такая что новая случайная величина

$$\ln \Xi = \{\ln \xi_1, \ln \xi_2, \ldots, \ln \xi_i, \ldots, \ln \xi_m\}$$

распределена по m-мерному нормальному закону, т. е.

$$F(\ln X) = (2\pi)^{-\frac{m}{2}} |\Sigma|^{-\frac{1}{2}} \times$$

$$\times \int_{-\infty}^{\ln x_1} \ldots \int_{-\infty}^{\ln x_m} \exp\left[-\frac{1}{2}\{\ln X - \Theta\}\Sigma^{-1}\{\ln X - \Theta\}'\right] \times$$

$$\times d\ln x_1, \ldots, d\ln x_m, \quad (1.113)$$

где $X = \{x_1, x_2, \ldots, x_i, x_m\}$,

$\Theta = \{\theta_1, \theta_2, \ldots, \theta_i, \ldots, \theta_m\} = M\{\ln \xi_1, \ln \xi_2, \ldots, \ln \xi_m\}$,

Σ — ковариационная матрица случайной величины

$$\ln \Xi = \{\ln \xi_1, \ln \xi_2, \ldots, \ln \xi_m\}.$$

Так как значения X и $\ln X$ находятся в однозначном соответствии, можно записать

$$F(\ln X) = \Phi(X) = (2\pi)^{-\frac{m}{2}} |\Sigma|^{-\frac{1}{2}} \times$$

$$\times \int_0^{x_1} \ldots \int_0^{x_m} \exp\left[-\frac{1}{2}\{\ln X - \Theta\}\Sigma^{-1}\{\ln X - \Theta\}'\right] \times$$

$$\times \frac{dx_1 dv_2 \ldots dv_m}{x_1 x_2 \ldots v_m} = (2\pi)^{-\frac{m}{2}} |\Sigma|^{-\frac{1}{2}} \int_0^{x_1} \ldots \int_0^{x_m} \left(\prod_{i=1}^{m} v_i\right)^{-1} \times$$

$$\times \exp\left[-\frac{1}{2}\{\ln X - \Theta\}\Sigma^{-1}\{\ln X - \Theta\}'\right] dV. \quad (1.114)$$

Соответствующая $\Phi(X)$ функция m-мерной логнормальной плотности будет определена выражением

$$f(X) = (2\pi)^{-\frac{m}{2}} |\Sigma|^{-\frac{1}{2}} \left(\prod_{i=1}^{m} x_i\right)^{-1} \exp\left[-\frac{1}{2}\{\ln X - \Theta\}\Sigma^{-1}\{\ln X - \Theta\}'\right]. \quad (1.115)$$

В качестве примера для двумерного случая можем записать

$$f(x_1,\ x_2)=\frac{1}{2\pi\sigma_1\sigma_2 x_1 x_2\sqrt{1-\rho^2}}\times\exp\left\{-\frac{1}{2(1-\rho)}\left[\left(\frac{\ln x_1-\theta_1}{\sigma_1}\right)^2+\right.\right.$$
$$\left.\left.+\left(\frac{\ln x_2-\theta_2}{\sigma_2}\right)^2-2\rho\left(\frac{\ln x_1-\theta_1}{\sigma_1}\right)\left(\frac{\ln x_2-\theta_2}{\sigma_2}\right)\right]\right\}, \qquad (1.116)$$

где $\theta_1 = M \ln \xi_1$, $\theta_2 = M \ln \xi_2$, σ_1^2 и σ_2^2 — дисперсии случайных величин $\ln \xi_1$ и $\ln \xi_2$ соответственно, ρ — коэффициент корреляции для $\ln \xi_1$ и $\ln \xi_2$.

ГЛАВА 2

СТАТИСТИЧЕСКИЕ ОЦЕНКИ НЕИЗВЕСТНЫХ ПАРАМЕТРОВ

В главе 1 мы уже рассмотрели такие понятия как изучаемая, опробуемая и выборочная совокупности. Из этих определений следует, что изучаемая и опробуемая совокупности характеризуются некоторыми неизвестными нам значениями исследуемых свойств, например средними содержаниями различных компонентов, о которых исследователь вынужден судить по выборочным данным. Последние нередко бывают весьма ограничены, и поэтому вопрос об их использовании для суждения о неизвестных характеристиках стоит особенно остро.

В геологической практике для обоснования выводов, касающихся неизвестных значений изучаемых свойств, используются их приближенные характеристики, получаемые по выборочным данным, которые называются оценками. Так, например, в качестве приближенной характеристики неизвестного среднего значения нередко используется среднее арифметическое, хотя возможны и другие варианты оценок этого параметра.

Таким образом, возникает вопрос о выборе из набора возможных вариантов оценок тех из них, которые удовлетворяют некоторым требованиям качества, а это, в свою очередь, требует построения критериев качества статистических оценок.

2.1. ФОРМАЛЬНОЕ ОПРЕДЕЛЕНИЕ ОЦЕНОК И СПОСОБЫ ИХ ПОЛУЧЕНИЯ

Пусть ξ — случайная величина и $F(x; \Theta)$ — функция распределения этой величины, где $\Theta = \{\theta_1, \theta_2, \ldots, \theta_j, \ldots, \theta_k\}$ набор неизвестных параметров. Для простоты предположим, что $\Theta = \theta$. Обозначим через $X = \{x_1, x_2, \ldots, x_i, \ldots, x_n\}$ набор n наблюдений над случайной величиной ξ, т. е. выборку. Тогда оценку $\hat{\theta}$ для неизвестного параметра θ можно определить как некоторую функцию φ от набора наблюдений X, т. е.

$$\hat{\theta} = \varphi(X) = \varphi(x_1, x_2, \ldots, x_i, \ldots, x_n). \qquad (2.1)$$

Ясно, что вопрос построения оценки $\hat{\theta}$ неизвестного параметра θ сводится к определению вида функции φ, что можно сделать несколькими методами, которые могут дать различные результаты. В качестве одного из наиболее распространенных методов получения оценок рассмотрим метод максимального правдоподобия. Он заключается в следующем.

Как мы уже условились, рассматривается выборка $X = \{x_1, \ldots, x_n\}$, объем которой n и которая взята из совокупности значений непрерывной случайной величины ξ, функция плотности вероятности которой $f(x; \theta)$. Мы намеренно рассмотрим более сложный непрерывный случай, так как после этого приложение максимального правдоподобия к дискретным распределениям не составит труда.

Набор n значений $x_1, x_2, \ldots, x_n$ случайной величины ξ мы можем рассматривать как значения n независимых, одинаково распределенных случайных величин $\xi_1, \xi_2, \ldots, \xi_i, \ldots, \xi_n$, причем над каждой ξ_i проведено только одно наблюдение. В такой постановке каждому значению x_i будет соответствовать значение плотности $f(x_i, \theta)$, а совместная плотность вероятности случайных величин $\xi_1, \xi_2, \ldots, \xi_i, \ldots, \xi_n$ при заданном наборе $x_1, x_2, \ldots, x_i, \ldots, x_n$ будет определена выражением

$$L(x_1, x_2, \ldots, x_i, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \theta). \tag{2.2}$$

Так как $x_1, x_2, \ldots, x_i, \ldots, x_n$ известны, то L можно рассматривать как функцию неизвестного параметра θ и выбрать такое значение θ, при котором L достигает максимума, а это означает, что для любой другой оценки $\hat{\hat{\theta}}$ будет иметь место неравенство

$$\prod_{i=1}^{n} f(x_i, \hat{\theta}) > \prod_{i=1}^{n} f(x_i; \hat{\hat{\theta}}). \tag{2.3}$$

Вместо самой функции L обычно пользуются ее логарифмом. Максимум функции L выявляется обычным способом. Находится производная $dL(x_1, \ldots x_n)/d\theta$, приравнивается к нулю и решается относительно θ, т. е.

$$\frac{d \prod_{i=1}^{n} f(x_i; \theta)}{d\theta} = 0, \tag{2.4}$$

или, что равносильно

$$\frac{d \sum_{i=1}^{n} \ln f(x_i; \theta)}{d\theta} = 0. \tag{2.5}$$

В том случае, когда $\Theta = \{\theta_1, \theta_2, \ldots, \theta_j, \ldots, \theta_k\}$, выражения (2.4) и (2.5) примут вид

$$\frac{\partial \prod_{i=1}^{n} f(x_i; \theta_1, \theta_2, \ldots, \theta_i, \ldots, \theta_k}{\partial \theta_j} = 0. \tag{2.6}$$

$$\frac{\partial \sum_{i=1}^{n} \ln f(x_i; \theta_1, \theta_2, \ldots, \theta_j, \ldots, \theta_k)}{\partial \theta_j} = 0. \tag{2.7}$$

Полученное решение $\hat{\theta}_j = \varphi(x_1, \ldots, x_n)$ будет искомым при условии, что при проверке $\hat{\theta}_i$ дает максимум.

Проиллюстрируем метод максимального правдоподобия на примере нормального распределения. В данном случае для каждой случайной величины ξ_i

$$f(x_i;\ \mu,\ \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}, \tag{2.8}$$

где μ и σ^2 — параметры распределения, которые требуется оценить по выборке $x_1, x_2, \ldots, x, \ldots, x_n$. В соответствии с формулой (2.2) функция правдоподобия в данном случае будет иметь вид

$$L(x_1,\ x_2,\ \ldots,\ x_i,\ \ldots,\ x_n;\ \mu,\ \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] \tag{2.9}$$

или

$$\ln L(x_1,\ x_2,\ \ldots,\ x_i,\ \ldots,\ x_n;\ \mu,\ \sigma^2) = -\frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2. \tag{2.10}$$

Найдем вначале оценку $\hat{\mu}$ для μ. Для этого, используя выражения (2.9), возьмем частную производную по $\partial\mu$ и приравняем ее к нулю, т. е.

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0. \tag{2.11}$$

Нетрудно подсчитать, что

$$\sum_{i=1}^{n}(x_i-\mu) = 0, \qquad \sum_{i=1}^{n} x_i = n\mu, \tag{2.12}$$

откуда

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \hat{\mu}. \tag{2.13}$$

Таким образом, мы получили, что максимально правдоподобной оценкой среднего значения нормально распределенной случайной величины является среднее арифметическое. В статистике принято среднее арифметическое обозначать $\bar{x}$.

Аналогично получим оценку $\hat{\sigma}^2$ для σ^2

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0. \tag{2.14}$$

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

$$n\hat{\sigma}^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2, \tag{2.15}$$

откуда

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{2.16}$$

Заметим, что эта оценка несколько смещена (см. ниже о несмещенных оценках) относительно σ^2, что особенно сказывается в условиях малых выборок, и несмещенная оценка для σ^2 будет определена выражением

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2. \tag{2.17}$$

Другой весьма распространенный способ получения оценок неизвестных параметров — это метод моментов. Сущность его заключается в следующем.

Пусть, как и в предыдущем рассмотрении, из совокупности с плотностью вероятности $f(x; \theta_1, \theta_2, \ldots, \theta_k)$ взята выборка объема n и по выборочным данным $x_1, x_2, \ldots, x_n$ требуется получить оценки $\hat{\theta}_1, \ldots, \hat{\theta}_k$ для $\theta_1, \theta_2, \ldots, \theta_k$. Каждый h-й момент μ_h распределения $f(x; \theta_1, \ldots, \theta_k)$ можно рассматривать как некоторую функцию $v_h(\theta_1, \ldots, \theta_k)$ от значений оцениваемых параметров. Если значения μ_h заменить их выборочными значениями m_h, то можно получить систему из k уравнений, которая затем решается относительно неизвестных оценок $\hat{\theta}_i$. Так, например, в случае нормального распределения требуется оценить среднее μ и дисперсию σ^2. Положим $\mu = \theta_1$, $\sigma^2 = \theta_2$, тогда

$$m_1 = \hat{\theta}_1, \tag{2.18}$$

$$m_2 = \hat{\theta}_1^2 + \hat{\theta}_2, \tag{2.19}$$

где m_1 и m_2 — первый и второй начальные выборочные моменты, а $\hat{\theta}_1$ и $\hat{\theta}_2$ — искомые оценки для μ и σ^2. Легко подсчитать, что

$$m_1 = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{2.20}$$

$$s^2 = \hat{\theta}_2 = m_2 - m_1^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)^2 =$$

$$= \frac{1}{n} \left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2 \right]. \tag{2.21}$$

Нетрудно видеть, что в случае нормального распределения оценки, полученные с помощью метода моментов, совпадают с максимально правдоподобными оценками. Однако такое совпадение не всегда

обязательно и оценки по методу моментов могут существенно отличаться от оценок максимального правдоподобия. Так, например, в условиях логнормального распределения оценкой среднего значения рассматриваемой случайной величины по методу моментов будет среднее арифметическое, а оценка дисперсии, согласно Дж. Ачисону и Дж. Брауну [48], будет определена формулой (2.22). В данном случае максимально правдоподобные оценки среднего значения и дисперсии (обозначим их a и b^2 соответственно) имеют следующий вид:

$$a = e^{\overline{\ln x}} \Psi_n\left(\frac{1}{2} s^2_{\ln}\right), \tag{2.22}$$

$$b^2 = e^{2\overline{\ln x}} \left\{\Psi_n\left(2s^2_{\ln}\right) - \Psi_n\left(\frac{n-2}{n-1} s^2_{\ln}\right)\right\}, \tag{2.23}$$

где $\overline{\ln x}$ — среднее арифметическое логарифмов выборочных значений, $s^2_{\ln}$ — оценка дисперсии логарифмов, вычисляемая по формуле

$$s^2_{\ln} = \frac{1}{n-1} \sum_{i=1}^{n} (\ln x_i - \overline{\ln x})^2. \tag{2.24}$$

Функции $\Psi_n\left(\frac{1}{2} s^2_{\ln}\right)$, $\Psi_n\left(2s^2_{\ln}\right)$ и $\Psi_n\left(\frac{n-2}{n-1} s^2_{\ln}\right)$ вычисляются по приближенной формуле

$$\Psi_n(t) = e^t \left\{1 - \frac{t(t+1)}{n} + \frac{t^2(3t^2+22t+21)}{6n^2}\right\} + \Theta\left(\frac{1}{n^2}\right),$$

в которой t принимает значения $\frac{1}{2} s^2_{\ln}$, $2s^2_{\ln}$ и $\frac{n-2}{n-1} s^2_{\ln}$.

Сравнение свойств максимально правдоподобных оценок и оценок по методу моментов для логнормального распределения мы проведем ниже при рассмотрении критериев качества оценок.

2.2. РАСПРЕДЕЛЕНИЯ ВЫБОРОЧНЫХ ХАРАКТЕРИСТИК

Представим себе, что мы опробуем гранитный массив с последующим определением содержаний элементов в пробах, причем берем пробы наудачу независимыми сериями по n проб. Таким образом, по этим данным мы получим набор оценок средних значений содержания, например средних арифметических. Естественно, что эти оценки будут отличаться одна от другой, причем заранее нельзя точно предсказать, какое значение примет такая оценка. Все это дает основание рассматривать оценки неизвестных параметров как значения случайных величин, а это, в свою очередь, требует постановки вопроса о распределениях этих случайных величин.

2.2.1. Распределение среднего арифметического

Выше мы отмечали, что выборочные значения $x_1, x_2, \ldots, x_i, \ldots, x_n$ случайной величины ξ можно рассматривать как n значений n независимых одинаково распределенных случайных величин $\xi_1, \xi_2, \ldots, \xi_i, \ldots, \xi_n$. Следовательно, среднее арифметическое x, полученное по данным выборки, можно рассматривать как одно из значений случайной величины

$$\bar{\xi} = \frac{1}{n} \sum_{i=1}^{n} \xi_i. \tag{2.26}$$

Если для всех ξ_i существуют математические ожидания $\mu_i = \mu$ и дисперсии $\sigma_i^2 = \sigma^2$, которые конечны, то относительно суммы $\sum_{i=1}^{n} \xi_i$ будет выполнено неравенство Бэрри, рассмотренное в главе 1, а это дает основание полагать, что распределение случайной величины ξ будет приближаться к нормальному с увеличением n, причем математическое ожидание случайной величины $\bar{\xi}$ будет равно μ, а дисперсия σ^2/n, т. е.

$$M\bar{\xi} = \mu, \quad D\bar{\xi} = \frac{\sigma^2}{n}. \tag{2.27}$$

Таким образом, при весьма общих предположениях относительно изучаемой случайной величины ξ среднее арифметическое ее выборочных значений будет распределено приблизительно нормально с параметрами, определенными формулой (2.27). Это обстоятельство будет нами в дальнейшем широко использоваться при обосновании геологических выводов, связанных со сравнением средних значений изучаемых характеристик.

2.2.2. Распределение выборочной дисперсии

Рассмотрим выборку $x_1, x_2, \ldots, x_i, \ldots, x_n$ из совокупности значений нормально распределенной случайной величины ξ с параметрами μ, σ^2. Если значение μ известно, то оценка для σ^2 будет

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2. \tag{2.28}$$

Эту оценку можно рассматривать как одно значение случайной величины

$$\zeta^2 = \frac{1}{n} \sum_{i=1}^{n} (\xi_i - \mu)^2, \tag{2.29}$$

где ξ_i—n одинаково нормально распределенных с параметрами μ,

σ^2 независимых случайных величин. Образуем новую случайную величину

$$\chi_n^2 = \sum_{i=1}^{n} \left(\frac{\xi_i - \mu}{\sigma} \right)^2, \tag{2.30}$$

которая, как было показано в главе 1, представляет собой сумму квадратов независимых нормально распределенных с параметрами (0,1) случайных величин $(\xi_i - \mu)/\sigma$ и, следовательно, будет распределена как χ^2 с n степенями свободы. Случайную величину ζ^2 можно представить следующим образом:

$$\zeta^2 = \frac{\sigma^2}{n} \sum_{i=1}^{n} \left(\frac{\xi_i - \mu}{\sigma} \right)^2 = \frac{\sigma^2}{n} \chi_n^2. \tag{2.31}$$

Плотность вероятности этой случайной величины будет определена выражением

$$g(s^2) = \frac{1}{\left(\frac{n-2}{2}\right)! \, 2^{\frac{n}{2}}} \left(\frac{ns^2}{\sigma^2} \right)^{\frac{n-2}{2}} \frac{n}{\sigma^2} \exp\left[-\frac{n}{2} \left(\frac{s^2}{\sigma^2} \right) \right] \text{ при } s^2 > 0. \tag{2.32}$$

Так как математическое ожидание и дисперсия случайной величины χ^2 равны соответственно n и $2n$, то для случайной величины ζ^2 можно записать

$$M\zeta^2 = \frac{\sigma^2}{n} M\chi_n^2 = \frac{\sigma^2 n}{n} = \sigma^2, \tag{2.33}$$

$$D\zeta^2 = D\left(\frac{\sigma^2}{n} \chi_n^2 \right) = \frac{\sigma^4}{n^2} D\chi_n^2 = \frac{2\sigma^4}{n}. \tag{2.34}$$

Если же значение μ неизвестно, что чаще всего бывает на практике, то в качестве оценки для σ^2 используется число

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2, \tag{2.35}$$

которое представляет собой значение случайной величины

$$\zeta^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\xi_i - \bar{\xi})^2, \tag{2.36}$$

где $\bar{\xi} = \frac{1}{n} \sum_{i=1}^{n} \xi_i$.

Как и в предыдущем случае, можно записать

$$\zeta^2 = \frac{\sigma^2 \chi_{n-1}^2}{n-1}, \tag{2.37}$$

где χ_{n-1}^2 — случайная величина, распределенная как χ^2 с $n-1$

степенями свободы. Таким образом, функция плотности вероятности для ζ^2 будет иметь вид

$$g(s^2) = \frac{1}{\left(\frac{n-3}{2}\right)! \, 2^{\frac{n-1}{2}}} \left(\frac{n-1}{\sigma^2} s^2\right)^{\frac{n-3}{2}} \frac{n-1}{\sigma^2} \times$$

$$\times \exp\left[\frac{(n-1)s^2}{2\sigma^2}\right] \text{ при } s^2 > 0. \qquad (2.38)$$

Как и в предыдущем случае, легко подсчитать, что $M\zeta^2 = \sigma^2$ и $D\zeta^2 = \frac{2\sigma^4}{n-1}$.

2.2.3. Распределение выборочного размаха и оценки дисперсии

Весьма полезной выборочной характеристикой, особенно в условиях малых выборок ($n \leqslant 10$), является выборочный размах. Эта величина W_n представляет собой разность между максимальным и минимальным значениями выборки, т. е.

$$W_n = x_n - x_1 \quad \text{при } x_1 \leqslant x_2 \leqslant \ldots \leqslant x_n,$$

где $x_1, x_2, \ldots, x_i, \ldots, x_n$ — значения n одинаково распределенных случайных величин $\zeta_1, \zeta_2, \ldots, \zeta_i, \ldots, \zeta_n$ с функцией распределения $F(x)$. Наблюдаемое значение выборочного размаха W_n будем рассматривать как значение случайной величины ω_n, функция распределений которой определена выражением

$$P(\omega_n \leqslant W) = n \int_{-\infty}^{\infty} [F(x+W) - F(x)]^{n-1} dF(x). \qquad (2.40)$$

Различные варианты таблиц, связанных с распределением размаха для выборок из нормальной совокупности, приведены в книге Л. Н. Большева и Н. В. Смирнова [6]. В настоящей работе они помещены в виде прил. 3 и 4.

Если функция распределения случайных величин $\zeta_1, \zeta_2, \ldots, \zeta_i, \ldots, \zeta_n$ нормальная с параметрами μ, σ^2, то в качестве оценки для σ^2 можно использовать величину

$$s^2 = \frac{W_n}{d_n}, \qquad (2.41)$$

где

$$d_n = M\left(\frac{\omega_n}{\sigma}\right).$$

Значения величины d_n можно найти в приложениях к настоящей книге. Необходимо отметить, что эта оценка дисперсии особенно удобна и эффективна при $n \leqslant 10$, а при $n > 20$ ее применение приводит к значительной потере информации и поэтому приведенные таблицы составлены только для $n \leqslant 20$.

2.2.4. Распределение выборочного коэффициента корреляции

Рассмотрим еще один пример выборочного распределения — распределение оценки коэффициента корреляции двумерного нормального распределения.

Пусть (ξ, η) — двумерная нормально распределенная случайная величина и $(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_n, y_n)$ — n двумерных выборочных значений, которые мы рассматриваем как n значений n одинаково распределенных случайных величин

$$(\xi_1, \eta_1), (\xi_2, \eta_2), \ldots, (\xi_i, \eta_i), \ldots, (\xi_n, \eta_n).$$

Неизвестный коэффициент корреляции ρ, который мы рассматривали в главе 1, обычно оценивают с помощью величины

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}, \tag{2.42}$$

которая представляет собой значение случайной величины R, распределение которой при $n \geqslant 3$ в интервале $-1 < r < 1$ определяется плотностью вероятности

$$f_n(r, \rho) = \frac{2^{n-3}}{\pi \Gamma(n-2)} (1-\rho^2)^{\frac{n-1}{2}} (1-r^2)^{\frac{n-4}{2}} \times$$

$$\times \sum_{m=0}^{\infty} \left[\Gamma\left(\frac{n+m-1}{2} \right) \right]^2 \frac{(2\rho r)^m}{m!}. \tag{2.43}$$

Особое значение имеет случай, когда $\rho = 0$, тогда формула (2.43) примет вид

$$f_n(r, 0) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi}\ \Gamma\left(\frac{n-2}{2}\right)} (1-e^2)^{\frac{n-4}{2}} \text{ при } (|r| < 1,\ n \geqslant 3). \tag{2.44}$$

Из этой формулы следует, что новая случайная величина

$$\tau = R \sqrt{\frac{n-2}{1-R^2}} \tag{2.45}$$

при условии, что $\rho = 0$, будет обладать плотностью вероятности, определяемой выражением

$$\varphi(t) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi(n-2)}\ \Gamma\left(\frac{n-2}{2}\right)} \left(1 + \frac{t^2}{n-2}\right)^{-\frac{n-1}{2}}, \tag{2.46}$$

т. е. величина τ будет распределена по закону Стьюдента (см. главу 1) с n—2 степенями свободы. Этот результат будет нами использован в дальнейшем при статистическом обосновании геологических выводов (см. прил. 5 в конце книги).

2.3. КРИТЕРИИ КАЧЕСТВА ОЦЕНОК

В предыдущем изложении мы показали, что для одного и того же неизвестного параметра могут существовать различные варианты оценок, и для того чтобы обоснованно подходить к выбору той или иной из них, необходимо рассмотреть их критерии качества.

2.3.1. Несмещенность

Пусть $x_1, x_2, \ldots, x_i, \ldots, x_n$ — выборка объема n и θ — неизвестный оцениваемый параметр. Обозначим через $\hat{\theta}(x_1, x_2, \ldots x_i, \ldots, x_n)$ оценку для θ. Напомним, что оценку $\hat{\theta}(x_1, x_2, \ldots x_n)$ можно рассматривать как случайную величину.

Если при фиксированном n для оценки $\hat{\theta}(x_1, \ldots, x_n)$ неизвестного параметра θ выполнено условие

$$M\,\hat{\theta}(x_1, x_2, \ldots, x_n) = \theta, \tag{2.47}$$

то такая оценка называется несмещенной, т. е. не содержащей систематической ошибки.

Однако, если требование несмещенности не выполняется, этот недостаток обычно бывает легко устраним путем введения соответствующей поправки. Так, например, математическое ожидание оценки дисперсии

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, \tag{2.48}$$

особенно при небольших n, будет несколько заниженным по сравнению с σ^2, что исправляется выражением

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{2.49}$$

Оценка (2.49) является несмещенной.

2.3.2. Состоятельность

Пусть $\hat{\theta}(x_1)$, $\hat{\theta}(x_1, x_2)$, $\hat{\theta}(x_1, x_2, x_3), \ldots, \hat{\theta}(x_1, \ldots, x_n)$ — последовательность оценок, полученных по выборкам объема $k = 1, 2, 3, \ldots, n$. Такую последовательность называют состоятельной, если

$$\lim_{k\to\infty} P\{|\hat{\theta}(x_1, x_2 \ldots, x_k) - \theta| < \varepsilon\} = 1, \tag{2.50}$$

где ε — сколь угодно малое заданное число. Иными словами, со-

стоятельной называется такая последовательность оценок $\hat{\theta}(x_1, x_2, \ldots, x_k)$, для которой вероятность события, заключающегося в том, что $\hat{\theta}(x_1, x_2, \ldots, x_k)$ отличается от θ на величину, не превышающую сколь угодно малое заданное число ε, стремится к 1 при неограниченном возрастании k.

2.3.3. Эффективность

Оценка $\hat{\theta}(x_1, x_2, \ldots, x_n)$, обладающая минимальной дисперсией из всех возможных оценок, полученных по выборке объема n, называется эффективной. Естественно, что такая оценка, при условии что она не смещена, предпочтительнее любой другой, так как обеспечивает более тесную группировку результатов около истинного значения неизвестного оцениваемого параметра θ.

2.3.4. Достаточность

Примем, как и в предыдущих разделах, что $f(x_i, \theta)$ — это плотность вероятности случайной величины в точке x_i. Тогда для выборки объема n функция правдоподобия будет определена выражением

$$\prod_{i=1}^{n} f(x_i, \theta). \tag{2.51}$$

Оценка $\hat{\theta}(x_1, x_2, \ldots x_n)$ называется достаточной оценкой неизвестного параметра θ, если существует такая функция $h(x_1, x_2, \ldots, x_n)$, не зависящая от θ, для которой имеет место равенство

$$\prod_{i=1}^{n} f(x_i, \theta) = g[\hat{\theta}(x_1, x_2, \ldots, x_n), \theta]\, h(x_1, \ldots x_n). \tag{2.52}$$

Необходимо отметить, что достаточная оценка включает всю информацию, которую можно получить о неизвестном параметре по выборке объема n.

2.4. ИНТЕРВАЛЬНЫЕ ОЦЕНКИ

До сих пор мы рассматривали так называемые точечные оценки, которые выражены одним числом. Однако они не дают представления о том, какие отклонения от истинного значения оцениваемого параметра возможны для вычисленной оценки. От этого недостатка свободны интервальные оценки, которые представляют собой интервалы, каждому из которых поставлено в соответствие число, характеризующее вероятность события, заключающегося в том, что при бросании наудачу этот интервал накроет неизвестное значение оцениваемого параметра. Так, если θ — неизвестный параметр, а $\hat{\theta}$ — соответствующая статистическая оценка, то можно записать

$$P(\hat{\theta} - c \leqslant \theta \leqslant \hat{\theta} + c) = 1 - \alpha, \tag{2.53}$$

где α — заданное малое число, а c — число, которое соответствует

заданному значению α. Вероятность того, что интервал $(\hat{\theta}-c, \hat{\theta}+c)$ накроет θ равна $1-\alpha$ и называется надежностью, а величина c — точностью оценки $\hat{\theta}$ при заданной надежности $1-\alpha$. Интервал $(\hat{\theta}-c, \hat{\theta}+c)$ называется доверительным интервалом.

Рассмотрим построение интервальной оценки неизвестного среднего значения μ для среднего арифметического $\bar{x}$ при условии, что значение дисперсии σ^2 известно. Как уже отмечалось, при весьма общих предположениях относительно распределения изучаемой случайной величины оценку $\bar{x}$ среднего μ можно рассматривать как значение нормально распределенной случайной величины с математическим ожиданием, равным μ, и дисперсией σ^2/n. Выбрав соответствующее значение α, можно записать

$$P\left(-t \leqslant \frac{x-\mu}{\sigma/\sqrt{n}} \leqslant t\right) = 1-\alpha, \tag{2.54}$$

причем

$$N(-t;\ 0,1) = \frac{\alpha}{2}, \tag{2.55}$$

$$N(t;\ 0,1) = 1-\frac{\alpha}{2} \tag{2.56}$$

являются значениями стандартной нормальной функции в точках $-t$ и t. Выражение (2.54) можно записать следующим образом:

$$P\left(\bar{x}-t\frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \bar{x}+t\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha. \tag{2.57}$$

Таким образом, интервал, который с вероятностью $1-\alpha$ накрывает неизвестное значение μ, определяется следующими значениями $\left(\bar{x}-t_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\ \bar{x}+t_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$. Если выбрать $\alpha = 0{,}05$, что вполне достаточно для практических задач, то получим

$$N(-1{,}96) = 0{,}025,$$
$$N(1{,}96) = 0{,}975,$$

из чего следует, что границы 95 %-ного доверительного интервала среднего арифметического определяется как

$$\bar{x}-1{,}96\,\sigma/\sqrt{n};\ \bar{x}+1{,}96\sigma/\sqrt{n}.$$

Если значение σ неизвестно, то его можно оценить по выборке и использовать при построении доверительного интервала соответствующую оценку s, т. е.

$$\bar{x}-t_{\frac{\alpha}{2}}s/\sqrt{n};\quad \bar{x}+t_{1-\frac{\alpha}{2}}s/\sqrt{n}.$$

В том случае, когда выборка взята из нормальной совокупности, и число n не очень велико, можно воспользоваться тем, что величина

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

будет представлять собой значения случайной величины, распределенной по закону Стьюдента с $n-1$ степенями свободы. Тогда можно записать

$$P\left\{-t_{n-1,\,1-\frac{\alpha}{2}} \leqslant \frac{\bar{x} - \mu}{s/\sqrt{n}} \leqslant t_{n-1,\,1-\frac{\alpha}{2}}\right\} = 1 - \alpha,$$

откуда

$$P\left(\bar{x} - t_{n-1,\,1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \leqslant \mu \leqslant \bar{x} + t_{n-1,\,1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right) = 1 - \alpha. \quad (2.58)$$

Таким образом, в условиях нормального распределения и малых выборок границы 100 $(1-\alpha)$ %-ного доверительного интервала определяются как $\bar{x} - t_{n-1,\,1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$, $\bar{x} + t_{n-1,\,1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$, где $t_{n-1,\,1-\frac{\alpha}{2}}$ — значение аргумента функции Стьюдента, соответствующее значению вероятности $1 - \frac{\alpha}{2}$. Эти значения берутся из таблицы, приведенной в конце книги (см. прил. 6).

Кроме доверительного интервала для среднего значения полезно рассмотреть процедуру построения интервальной оценки для дисперсии. Пусть из нормальной совокупности с неизвестными параметрами μ и σ^2 взята выборка объема n и по ней вычислены соответствующие оценки $\bar{x}$ и s^2. Выше было показано, что отношение $s^2(n-1)/\sigma^2$ представляет значение случайной величины, распределенной как χ^2 с $n-1$ степенями свободы. В связи с этим можно записать

$$P\left[\chi^2_{n-1,\,1-\frac{\alpha}{2}} \geqslant \frac{s^2(n-1)}{\sigma^2} \geqslant \chi^2_{n-1,\,\frac{\alpha}{2}}\right] = 1 - \alpha. \quad (2.59)$$

Разделив все члены неравенства на $s^2(n-1)$, получим

$$P\left[\frac{\chi^2_{n-1,\,1-\frac{\alpha}{2}}}{s^2(n-1)} \geqslant \frac{1}{\sigma^2} \geqslant \frac{\chi^2_{n-1,\,\frac{\alpha}{2}}}{s^2(n-1)}\right] = 1 - \alpha, \quad (2.60)$$

что равносильно

$$P\left[\frac{s^2(n-1)}{\chi^2_{n-1,\,1-\frac{\alpha}{2}}} \leqslant \sigma^2 \leqslant \frac{s^2(n-1)}{\chi^2_{n-1,\,\frac{\alpha}{2}}}\right] = 1 - \alpha. \quad (2.61)$$

Таким образом, границы 100 $(1-\alpha)$ %-ного доверительного интервала для σ^2 определяются как $s^2(n-1)\chi^2_{n-1,\,1-\frac{\alpha}{2}}$, $s^2(n-1)/\chi^2_{n-1,\,\frac{\alpha}{2}}$, где значения $\chi^2_{n-1,\,1-\frac{\alpha}{2}}$ и $\chi^2_{n-1,\,\frac{\alpha}{2}}$ берутся из таблицы, приведенной в конце книги (см. прил. 2).

В качестве еще одного полезного примера построения интервальной оценки для решения непосредственно геологической задачи рассмотрим предложенный Р. И. Коганом метод интервального оценивания запасов твердых полезных ископаемых [25]. Аналогичный метод построен Б. В. Смирновым [40] для месторождений нефти и газа. Основная идея этих методов заключается в следующем.

Пусть Q_v — значение запасов полезного ископаемого, находящихся в заданной трехмерной области v. Величину Q_v требуется оценить по результатам измерения набора характеристик x_1, $x_2, \ldots, x_t, \ldots, x_n$ в n точках t области v, например таких, как содержания полезного компонента в пробах, мощность рудных тел, коэффициент рудоносности и др. В общем случае, по этим замерам вычисляется приближенная характеристика $\hat{Q}_v$ неизвестных запасов Q_v области v, т. е.

$$\hat{Q}_v = \varphi(n,\ x_1,\ x_2, \ldots, x_t, \ldots, x_n). \tag{2.62}$$

В результате мы получаем точечную оценку $\hat{Q}_v$ неизвестного значения запасов Q_v, для которой невозможно построить доверительный интервал в связи с тем, что мы располагаем только одним ее значением.

Однако, если имеющиеся n наблюдений разбить на m групп по k наблюдений в группе так, чтобы каждая из этих групп представляла более редкую, чем при n пробах сеть опробования всего объема v, то по каждой группе можно вычислить оценку

$$\hat{Q}_i = \varphi(k;\ x_1,\ x_2, \ldots, x_h). \tag{2.63}$$

Конечно, каждая такая оценка будет значительно менее точной чем $\hat{Q}_v$, но тем не менее мы будем располагать m измерениями $\hat{Q}_1$, $\hat{Q}_2, \ldots, \hat{Q}_i, \ldots, \hat{Q}_m$ одной и той же неизвестной нам величины. Таким образом, как и в предыдущих случаях, мы можем последовательность значений $\hat{Q}_1$, $\hat{Q}_2, \ldots, \hat{Q}_i, \ldots, \hat{Q}_m$ рассматривать как m значений m независимых одинаково распределенных случайных величин, имеющих математические ожидания, равные Q, и дисперсии, равные σ^2. В результате относительно величины

$$\overline{Q} = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i \tag{2.64}$$

будут выполнены условия неравенства Бэрри, рассмотренного в главе 1, из чего следует, что новая оценка $\overline{Q}$ для Q будет распреде-

лена приблизительно нормально с математическим ожиданием Q и дисперсией σ^2/m, оценкой для которой будет величина

$$s^2_{\overline{Q}} = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{Q}_i - Q)^2. \qquad (2.65)$$

Таким образом, мы можем записать

$$P\left(\overline{Q} - t_{1-\frac{\alpha}{2}}\, s_{\overline{Q}} \leqslant Q \leqslant \overline{Q} + t_{1-\frac{\alpha}{2}}\, s_{\overline{Q}}\right) = 1-\alpha, \qquad (2.66)$$

где $1-\alpha$ — заданный уровень надежности, $t_{1-\frac{\alpha}{2}}$ — значение аргумента стандартной нормальной функции для вероятности $1-\frac{\alpha}{2}$.

В результате границы доверительного интервала для $\overline{Q}$ при надежности $1-\alpha$ будут

$$\left(\overline{Q} - t_{1-\frac{\alpha}{2}}\, s_{\overline{Q}},\ \overline{Q} + t_{1-\frac{\alpha}{2}}\, s_{\overline{Q}}\right). \qquad (2.67)$$

Естественно, что здесь приведена только принципиальная схема получения интервальной оценки для неизвестного значения запасов. В данном виде она может быть применена только к весьма ограниченной части месторождения, например к блоку. Процедура же вычисления интервальной оценки для всего месторождения в целом значительно сложнее и подробно рассмотрена в уже упоминавшихся работах Р. И. Когана [25] и Б. В. Смирнова [40].

2.5. ПРИМЕРЫ

Вычислительные процедуры получения различных вариантов оценок для неизвестных параметров мы рассмотрим на примере двух выборок, представляющих собой пробы воды, взятые с поверхности залива Кара-Богаз-Гол (данные предоставлены автору В. П. Фединым). В первой выборке приведены определения содержания B_2O_3 и Br (табл. 3), тогда как во второй определялись SO_4 и Ca (табл. 4).

Таблица 3

Содержания B_2O_3 и Br в пробах воды с поверхности залива Кара-Богаз-Гол, %

Номер п/п	B_2O_3	Br	Номер п/п	B_2O_3	Br	Номер п/п	B_2O_3	Br
1	0,045	0,035	6	0,060	0,050	11	0,060	0,050
2	0,043	0,033	7	0,058	0,050	12	0,060	0,050
3	0,037	0,033	8	0,061	0,050	13	0,061	0,050
4	0,041	0,033	9	0,060	0,051	14	0,058	0,051
5	0,041	0,033	10	0,058	0,048			

Таблица 4

Содержания SO_4 и Ca в пробах воды с поверхности залива Кара-Богаз-Гол, %

Номер п/п	SO_4	Ca	Номер п/п	SO_4	Ca
1	3,634	0,018	4	3,514	0,018
2	3,359	0,008	5	3,555	0,009
3	3,950	0,008	6	3,567	0,009

По этим данным были вычислены средние арифметические и оценки дисперсий, причем для второй выборки, как содержащей менее 10 наблюдений для вычисления оценок дисперсий, использовались размахи.

Таким образом, для первой выборки были получены следующие оценки:

$$\bar{x} = 0{,}053, \quad s_x^2 = 0{,}8515 \cdot 10^{-4}, \quad s_x = 0{,}923 \cdot 10^{-2},$$

$$y = 0{,}0441, \quad s_y^2 = 0{,}6885 \cdot 10^{-4}, \quad s_y = 0{,}8297 \cdot 10^{-2},$$

где x — содержания B_2O_3, y — содержания Br.

Используя полученные результаты, построим доверительные интервалы для неизвестных средних значений μ_1 (для B_2O_3) и μ_2 (для Br), а также для соответствующих им дисперсий σ_1^2 и σ_2^2. Для μ_1 получим

$$\left(\bar{x} - t_{0,975}\frac{s_x}{\sqrt{n}}; \quad \bar{x} + t_{0,975}\frac{s_x}{\sqrt{n}}\right),$$

$$\left(0{,}053 - 1{,}96\frac{0{,}923 \cdot 10^{-2}}{\sqrt{14}}; \quad 0{,}053 + 1{,}96\frac{0{,}923 \cdot 10^{-2}}{\sqrt{14}}\right),$$

$$(0{,}035;\ 0{,}058).$$

Таким образом,

$$P\,(0{,}035 \leqslant \mu_1 \leqslant 0{,}058) = 0{,}95.$$

Для μ_2 получим

$$\left(\bar{y} - t_{0,975}\frac{s_y}{\sqrt{n}}; \quad \bar{y} + t_{0,975}\frac{s_y}{\sqrt{n}}\right),$$

$$\left(0{,}0441 - 1{,}96\frac{0{,}8297 \cdot 10^{-2}}{\sqrt{14}}; \quad 0{,}0441 + 1{,}96\frac{0{,}8297 \cdot 10^{-2}}{\sqrt{14}}\right),$$

$$(0{,}040;\ 0{,}048).$$

Теперь мы можем записать

$$P\,(0{,}040 \leqslant \mu_2 \leqslant 0{,}048) = 0{,}95.$$

Для построения доверительных интервалов дисперсий σ_1^2 и σ_2^2 воспользуемся выражением

$$P\left[\frac{s^2(n-1)}{\chi^2_{n-1,\,1-\frac{\alpha}{2}}} \leqslant \sigma^2 \leqslant \frac{s^2(n-1)}{\chi^2_{n-1,\,\frac{\alpha}{2}}}\right] = 1-\alpha.$$

Положив $\alpha = 0{,}05$, находим в таблицах соответствующие значения $\chi^2_{13;\,0.975} = 24{,}736$ и $\chi^2_{13;\,0.025} = 5{,}009$. В итоге получаем для σ_1^2

$$\left(\frac{0{,}8515\cdot 10^{-4}\cdot 13}{24{,}736};\quad \frac{0{,}8515\cdot 10^{-4}\cdot 13}{5{,}009}\right),$$

$$(0{,}4475\cdot 10^{-4};\quad 2{,}210\cdot 10^{-4}),$$

и в результате можно записать

$$P(0{,}448\cdot 10^{-4} \leqslant \sigma_1^2 \leqslant 2{,}210\cdot 10^{-4}) = 0{,}95.$$

Аналогично подсчитаем 95 %-ный доверительный интервал для σ_2^2

$$\left(\frac{0{,}6885\cdot 10^{-4}\cdot 13}{24.736};\quad \frac{0{,}6885\cdot 10^{-4}\cdot 13}{5{,}009}\right).$$

В итоге

$$P(0{,}362\cdot 10^{-4} \leqslant \sigma_2^2 \leqslant 1{,}787\cdot 10^{-4}) = 0{,}95.$$

Теперь воспользовавшись данными табл. 2, подсчитаем оценки для неизвестных дисперсий содержаний SO_4 и Ca. С этой целью находим для обоих компонентов минимальное и максимальное значение, а также разность между ними. Для SO_4 $\max x_i = 3{,}95$, $\min x_i = 3{,}359$ и $W_n = \max x_i - \min x_i = 0{,}591$.

Аналогично для Ca найдем

$$\max y_i = 0{,}018, \quad \min y_i = 0{,}008 \quad \text{и} \quad W_n = 0{,}010.$$

По формуле

$$s = \frac{W_n}{d_n},$$

где d_n для данного n находится из таблиц, получим $d_6 = 2{,}5344$, отсюда

$$s_1 = \frac{0{,}591}{2{,}5344} = 0{,}2284 \text{ для } SO_4,$$

$$s_2 = \frac{0{,}010}{2{,}5344} = 0{,}0039 \text{ для Ca}.$$

Нетрудно видеть, насколько просты вычисления, связанные с нахождением оценки для σ методом размахов.

ГЛАВА 3

ОСНОВНЫЕ ПРИНЦИПЫ ПОСТРОЕНИЯ СТАТИСТИЧЕСКИХ РЕШЕНИЙ

3.1. ОСОБЕННОСТИ ВЫВОДОВ ПО СТАТИСТИЧЕСКИМ ДАННЫМ

В предыдущей главе были рассмотрены различные способы получения статистических оценок неизвестных параметров. Естественно, что в геологии вычисление этих оценок не является самоцелью, а делается для последующего использования полученных характеристик при обосновании геологических выводов. Чаще всего эти выводы базируются на сравнении средних значений, в результате которого принимается одно из утверждений: неизвестные средние значения можно признать равными, т. е. пренебречь разницей между вычисленными оценками, или же неизвестные средние значения следует считать различными, т. е. разницу между оценками нужно признать существенной. Так, например, при опробовании двух разновидностей гранита, слагающих гранитный массив, из каждой разновидности было взято по 5 проб, в которых были определены содержания монацита, причем средние арифметические оказались равными 46 и 104 г/т. При обычном рассмотрении этих значений вполне возможен вывод об обогащении одной из разновидностей гранита монацитом и, как следствие, элементами группы редких земель. Однако, если учесть, что выборочные данные отличаются сильным разбросом, о чем свидетельствуют полученные значения оценок стандартных отклонений в выборках (52,28 и 54,69), то в обоснованности вывода о различных неизвестных средних содержаний можно усомниться. Дело в том, что если мы, пользуясь тем, что средние арифметические можно рассматривать как значения случайных величин, распределенных приблизительно нормально, построим соответствующие 95 %-ные доверительные интервалы, то получим довольно большую область их перекрытия, что свидетельствует о недостаточной обоснованности вывода о различиях истинных средних по имеющимся данным. Доверительные интервалы в нашем случае будут иметь следующие значения (0,17:91,82) и (56,34:151,66).

Аналогичный вопрос об обоснованности вывода о различных неизвестных средних может возникнуть и при изучении других статистических параметров — дисперсий, коэффициентов корреляции, регрессии, асимметрии и др. Выводы могут касаться не только пары выборок, но и одной выборки, а иногда и целого их набора.

Нетрудно видеть, что в силу ограниченности выборочных данных и их статистической природы при обосновании выводов по статистическим данным вполне возможно появление ошибочных за-

ключений, причем ошибки могут быть двух видов. Так, в рассмотренном выше случае со средними значениями монацита возможны два ошибочных исхода. Первый из них заключается в том, что неизвестные истинные значения средних содержаний монацита могут оказаться равными, а мы ошибочно примем решение, что они различны. Второй — в том, что будет принято ошибочное решение о равенстве средних, а на самом деле они существенно различаются. Вполне понятно, что события, заключающиеся в принятии ошибочного решения, следует рассматривать как случайные с соответствующими им вероятностями, которые, как правило, отличаются от нуля. Поэтому процедуру, на основании которой принимается решение, следует строить таким образом, чтобы упомянутые вероятности появления ошибочных решений были по возможности малыми. Этого можно достичь двумя путями. Первый путь заключается в увеличении числа наблюдений в выборках, за счет чего улучшается точность вычисляемых выборочных характеристик, что в свою очередь усиливает обоснованность выводов. Однако этот путь не всегда реален, так как в геологической практике нередки ситуации, когда число наблюдений ограничено и увеличить его не представляется возможным. Второй путь заключается в выборе такого решающего правила, которое бы минимизировало вероятности появления ошибочных решений или, по крайней мере, обеспечивало не очень большие их значения, а это требует четкой формальной постановки задачи о принятии статистических решений и выборе используемых для этого критериев.

3.2. СТАТИСТИЧЕСКИЕ ГИПОТЕЗЫ И КРИТЕРИИ ДЛЯ ИХ ПРОВЕРКИ

Вернемся к рассмотренной в предыдущем разделе ситуации, связанной с обоснованием вывода о равенстве или различии средних содержаний монацита в двух разновидностях гранитов. Обозначим через μ_1 неизвестное значение среднего содержания монацита в первой разновидности гранита и через μ_2 во второй соответственно. Пусть также $x_1', x_2', \ldots, x_{n_1}'$ — результаты определения содержаний монацита в n_1 пробах, взятых из первой разновидности гранита, а $x_1'', x_2'', \ldots, x_{n_2}''$ — аналогичные результаты для второй разновидности. Наше утверждение о равенстве неизвестных средних мы будем рассматривать как проверяемую гипотезу (иногда ее называют нулевой гипотезой) и обозначать

$$H_0 : \mu_1 = \mu_2 \tag{3.1}$$

или, что равносильно,

$$H_0 : \mu_1 - \mu_2 = 0. \tag{3.2}$$

Альтернативная гипотеза представляет собой утверждение о неравенстве μ_1 и μ_2, или, что равносильно, $\mu_1 - \mu_2 \neq 0$. Эту гипотезу мы будем обозначать

$$H_1 : \mu_1 - \mu_2 \neq 0. \tag{3.3}$$

Заметим, что проверяемая гипотеза содержит только одну точку на числовой оси, т. е. $\mu_1 - \mu_2 = 0$. Такие гипотезы мы будем называть простыми, в отличие от сложных гипотез, содержащих более чем одно значение изучаемого параметра. Таким образом, альтернатива $H_1 : \mu_1 - \mu_2 \neq 0$ будет представлять собой бесчисленное множество простых гипотез.

Ограничим множество альтернатив H_1 только одной простой гипотезой $\mu_1 - \mu_2 = a$. Тогда наша задача будет сведена к проверке гипотезы

$$H_0 : \mu_1 - \mu_2 = 0 \tag{3.4}$$

при простой альтернативе

$$H_1 : \mu_1 - \mu_2 = a, \; a > 0. \tag{3.5}$$

Для решения этой задачи в нашем распоряжении имеются две выборки $x_1', x_2', \ldots, x_{n_1}'$ и $x_1'', x_2'', \ldots, x_{n_2}''$, которые мы можем рассматривать как два набора значений независимых и одинаково распределенных (внутри каждого набора) случайных величин $\xi_1', \xi_2', \ldots, \xi_{n_1}', \xi_1'', \xi_2'', \ldots, \xi_{n_2}''$. Образуем новую случайную величину

$$\tau = \varphi(\xi_1', \ldots, \xi_{n_1}', \xi_1'', \ldots, \xi_{n_2}''; \; \mu_1 - \mu_2 = 0) \tag{3.6}$$

и выберем φ так, чтобы при условии $\mu_1 - \mu_2 = 0$ распределение случайной величины τ было точно известно. Пусть это распределение будет

$$P(\tau \leqslant t \mid \mu_1 - \mu_2 = 0) = \Phi_0(t), \tag{3.7}$$

и соответствующая ему плотность вероятности будет $f_0(t)$. Если в области значений T функции τ установить критическое значение t_q, которое делит область T на область принятия гипотезы T_0 и область ее отклонения T_1, то вычислив по результатам выборочных данных значение

$$t = \varphi(x_1', \ldots, x_{n_1}', x_1'', \ldots, x_{n_2}''; \; \mu_1 - \mu_2 = 0), \tag{3.8}$$

мы будем принимать H_0, если $t \in T_0$, и отклонять, если $t \in T_1$. Случайную величину τ, определенную формулой (3.6), мы будем называть критерием для проверки гипотезы H_0.

На рис. 10 показана плотность распределения $f_0(t)$ случайной величины τ, при условии что верна гипотеза H_0. Критическое значение t_q выбрано так, что вероятность, соответствующая области T_1 отклонения H_0, равна q, т. е.

$$P(\tau > t_{1-q} \mid H_0) = P(t \in T_1 \mid H_0) = \int_{T_1} f_0(t)\, dt = q. \tag{3.9}$$

Эта вероятность соответствует событию, заключающемуся в том, что проверяемая гипотеза H_0, будучи правильной, окажется ошибочно отклонена. Такая ошибка называется ошибкой первого рода, а вероятность ее появления называется уровнем значимости критерия τ относительно гипотезы H_0.

Отсюда следует, что вероятность события, заключающегося в том, что вычисленное значение t случайной величины τ при условии, что H_0 верна, окажется принадлежащим области T_0, определится выражением

$$P(t \in T_0 \mid H_0) = \int_{T_0} f_0(t)\, dt = 1 - q. \qquad (3.10)$$

Иными словами, эта вероятность соответствует событию, что проверяемая гипотеза H_0, будучи правильной, будет принята, т. е. не будет допущена ошибка первого рода.

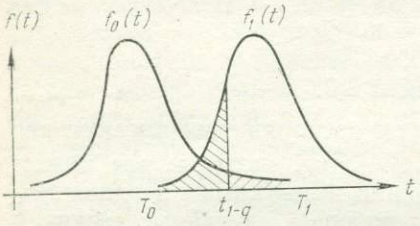

Рис. 10. Соотношение распределений критерия в условиях нулевой гипотезы и альтернативы

Вполне естественно желание выбрать t_q так, чтобы уровень значимости q был по возможности мал. Однако это может привести к нежелательным последствиям и чтобы понять их, нам нужно рассмотреть альтернативу H_1, которая как мы договорились, является в нашем случае простой гипотезой, т. е. $\mu_1 - \mu_2 = a > 0$. Обозначим через $f_1(t)$ плотность вероятности распределения критерия τ при условии, что верна альтернатива $H_1 : \mu_1 - \mu_2 = a$. На рис. 10 показано соотношение кривой $f_0(t)$ и $f_1(t)$. Нетрудно видеть, что событие $t \in T_0$, при условии, что гипотеза H_1 верна, будет соответствовать ошибочному решению, т. е. принятию гипотезы H_0, тогда как в действительности она неверна. Эта ошибка называется ошибкой второго рода, а ее вероятность будет определена выражением

$$P(\tau \leqslant t_{1-q} \mid H_1) = P(t \in T_0 \mid H_1) = \int_{T_0} f_1(t)\, dt = \beta. \qquad (3.11)$$

Вероятность противоположного события, заключающегося в том, что критерий τ примет значение $t \in T_1$ при условии, что верна альтернатива H_1, т. е. что будет принято правильное решение и ошибка второго рода не будет допущена, определится выражением

$$P(\tau > t_{1-q}) = P(t \in T_1 \mid H_1) = \int_{T_1} f_1(t)\, dt = 1 - \beta. \qquad (3.12)$$

Эта вероятность называется мощностью критерия относительно альтернативы H_1. Таким образом, при проверке гипотезы H_0, при фиксированной простой альтернативе H_1, возможны четыре собы-

тия, которые с соответствующими им вероятностями приведены в табл. 5.

Таблица 5

Вероятности риска, связанного с принятием ошибочных решений

Событие	Условие	Вероятность
$t \in T_0$	Верна H_0	$P(t \in T_0 \mid H_0) = \int_{T_0} f_0(t)\,dt = 1 - q$
$t \in T_1$	Верна H_1	$P(t \in T_1 \mid H_1) = \int_{T_1} f_1(t)\,dt = 1 - \beta$
$t \in T_1$	Верна H_0	$P(t \in T_1 \mid H_0) = \int_{T_1} f_0(t)\,dt = q$
$t \in T_0$	Верна H_1	$P(t \in T_0 \mid H_1) = \int_{T_0} f_1(t)\,dt = \beta$

Эти вероятности и их соотношение со статистическими решениями и истинным положением вещей можно представить в виде табл. 6.

Таблица 6

Вероятности ошибок и правильных решений

Применяемое решение	Истинное состояние	
	$H_0: \mu_1 = \mu_2$	$H_1: \mu_1 - \mu_2 = a > 0$
$H_1: \mu_1 - \mu_2 = a > 0$	q	$1 - \beta$
$H_0: \mu_1 = \mu_2$	$1 - q$	β

В обеих приведенных таблицах q и β соответствуют ошибкам первого и второго рода.

Необходимо отметить, что в условиях сложной альтернативы мощность критерия становится функцией от значений параметра. И поэтому при проверке гипотез очень важно учитывать риск, связанный с появлением как ошибок первого, так и второго рода, а для этого необходима некоторая общая процедура построения статистических критериев и выбора критических областей, которую мы рассмотрим в следующем разделе.

3.3. ПОСТРОЕНИЕ СТАТИСТИЧЕСКИХ КРИТЕРИЕВ И ВЫБОР КРИТИЧЕСКОЙ ОБЛАСТИ

Пусть ξ — случайная величина и $f(x)$ соответствующая ей плотность вероятности. Рассмотрим выборку $x_1, x_2, \ldots, x_n$, представляющую собой n независимых наблюдений над случайной величиной ξ. По этим данным требуется проверить нулевую гипотезу

$$H_0 : f(x) = f_0(x) \tag{3.13}$$

при альтернативе

$$H_1 : f(x) = f_1(x). \tag{3.14}$$

В данном случае мы используем более общую форму записи гипотезы, чем пользовались в предыдущих разделах. Любые частные случаи укладываются в это обобщение.

Как это уже делалось в главе 2, будем рассматривать выборку $x_1, x_2, \ldots, x_n$ как n значений n независимых одинаково распределенных случайных величин $\xi_1, \xi_2, \ldots, \xi_n$. Тогда для фиксированного набора значений $x_1, x_2, \ldots, x_n$ можно построить две функции правдоподобия

$$L_0(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_0(x_i) = L_0(x), \tag{3.15}$$

$$L_1(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_1(x_i) = L_1(x), \tag{3.16}$$

из которых первая учитывает условие, что верна проверяемая гипотеза H_0, а вторая — что верна альтернатива H_1. Отношение

$$\frac{L_1(x_1, \ldots, x_n)}{L_0(x_1, \ldots, x_n)} = \frac{\prod_{i=1}^{n} f_1(x_i)}{\prod_{i=1}^{n} f_0(x_i)} \tag{3.17}$$

называется отношением правдоподобия, и критическая область T_1 для отклонения H_0 и принятия H_1 будет определена неравенством

$$\frac{\prod_{i=1}^{n} f_1(x_i)}{\prod_{i=1}^{n} f_0(x_i)} \geqslant C, \tag{3.18}$$

где C — константа. Таким образом, вопрос заключается в том, как выбрать константу C. Общее правило для этого определено леммой Неймана—Пирсона [32], которая приведена ниже без доказательства. Она заключается в следующем: если критическая область T_1 такова, что

$$\int_{T_1} \prod_{i=1}^{n} f_0(x_i)\, dx_i = q = \int_{T_1} L_0(x)\, dx, \tag{3.19}$$

а C — такая константа, что $L_1(x) \geqslant CL_0(x)$ для всех $x \in T_1$ и $L_1(x) \leqslant CL_0(x)$ для всех $x \in T_0$, то не существует более мощной критической области уровня q, чем T_1.

Рассмотрим, как используется эта лемма на следующем простом примере. Допустим, что выборка $x_1, \ldots, x_n$ взята из нормальной совокупности, причем неизвестен только один параметр — среднее

значение μ, а дисперсия σ^2 известна. Ограничимся наиболее простым случаем, когда требуется проверить гипотезу

$$H_0 : \mu = a_1 \tag{3.20}$$

при простой альтернативе

$$H_1 : \mu = a_2, \tag{3.21}$$

где a_1 и a_2 заданные константы. Заметим, что такая задача нередко встречается при разведке рудных месторождений, когда изучаемый блок по данным опробования в зависимости от среднего содержания полезного компонента или бракуется, или учитывается при подсчете запасов. До сих пор в практике принятия этих решений статистические методы не применяются, хотя соответствующие критерии, учитывающие риск, связанный с принятием ошибочных решений, были бы весьма полезны.

Таким образом, в заданных условиях требуется построить решающее правило для принятия или отклонения $H_0 : \mu = a_1$ при альтернативе $H_1 : \mu = a_2$, уровне значимости q и мощности $1-\beta$. Задача сводится к определению критической области T_1 и области T_0 принятия нулевой гипотезы H_0.

Используя лемму Неймана—Пирсона, найдем такую константу C, для которой

$$\frac{\prod_{i=1}^{n} f_1(x_i)}{\prod_{i=1}^{n} f_0(x_i)} \geqslant C \text{ для всех } x \in T_1 \tag{3.22}$$

и

$$\frac{\prod_{i=1}^{n} f_1(x_1)}{\prod_{i=1}^{n} f_0(x_i)} \leqslant C \text{ для всех } x \in T_0, \tag{3.23}$$

где $f_0(x)$ — нормальная плотность с параметрами $\mu = a_1$, σ^2, а $f_1(x)$ — аналогичная функция с параметрами $\mu = a_2$, σ^2. Следовательно, можно записать

$$\frac{\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-a_2)^2\right]}{\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-a_1)^2\right]} \geqslant C \tag{3.24}$$

или

$$\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-a_2)^2-\sum_{=1}^{n}(x_i-a_1)^2\right]\right\} \geqslant C, \tag{3.25}$$

откуда

$$\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-a_2)^2-\sum_{i=1}^{n}(x_i-a_1)^2\right]\right\}\geqslant \ln C, \tag{3.26}$$

$$\sum_{i=1}^{n}(x_i-a_2)^2-\sum_{i=1}^{n}(x_i-a_1)^2\geqslant -2\sigma^2\ln C. \tag{3.27}$$

После несложных преобразований получим

$$\sum_{i=1}^{n}x_i(-2a_2+2a_1)>-2\sigma^2\ln C+n(a_1^2-a_2^2), \tag{3.28}$$

откуда

$$\sum_{i=1}^{n}x_i>\frac{n(a_1^2-a_2^2)-2\sigma^2\ln C}{2a_1-2a_2}. \tag{3.29}$$

Если правую часть этого неравенства, являющуюся константой, обозначить nb, т. е.

$$nb=\frac{n(a_1^2-a_2^2)-2\sigma^2\ln C}{2a_1-2a_2}, \tag{3.30}$$

то наше решающее правило примет вид

$$\sum_{i=1}^{n}x_i>nb \tag{3.31}$$

или, что равносильно,

$$\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{x}>b. \tag{3.32}$$

Рассматривая $\bar{x}$ как значение случайной величины $\bar{\xi}$, распределенной нормально с параметрами μ и σ^2/n, можем записать

$$P(\bar{\xi}\geqslant b)=q=P\left\{\frac{\bar{\xi}-a_1}{\sigma/\sqrt{n}}\geqslant\frac{b-a_1}{\sigma/\sqrt{n}}\right\}, \tag{3.33}$$

откуда

$$b=a_1+\frac{t_q\sigma}{\sqrt{n}}, \tag{3.34}$$

где t_q — значение аргумента стандартной нормальной функции, соответствующее значению q.

Вычислив величину b, можно определить и мощность критерия с помощью выражения

$$1-\beta=P\left(\frac{\bar{\xi}-a_2}{\sigma/\sqrt{n}}\geqslant\frac{b-a_2}{\sigma/\sqrt{n}}\right), \tag{3.35}$$

т. е., вычислив значение

$$t_{1-\beta} = \frac{b - a_2}{\sigma / \sqrt{n}}, \tag{3.36}$$

найти в таблице значение нормальной функции 1—β, соответствующее $t_{1-\beta}$.

Вполне естественно в качестве критерия для проверки гипотезы $H_0 : \mu = a_1$ при альтернативе $H_1 : \mu = a_2$ в условиях выборки объема n, взятой из нормальной совокупности с известной дисперсией σ^2, использовать величину

$$t = \frac{(\bar{x} - a_1)\sqrt{n}}{\sigma}. \tag{3.37}$$

Гипотеза H_0 будет приниматься, если вычисленное значение t окажется меньше, чем критическое t_q. Значение же t_q и уровень значимости q следует выбирать с учетом формулы (3.35), определяющей мощность критерия относительно альтернативы.

Для иллюстрации изложенного рассмотрим следующий гипотетический пример. Пусть по данным опробования породы (10 проб) установлено, что среднее арифметическое значение содержаний некоторого компонента равно 39 г/т. Предполагается, что распределение изучаемых содержаний близко к нормальному с неизвестным средним значением μ и дисперсией $\sigma^2 = 2809$, т. е. $\sigma = 53$. По этим данным требуется проверить гипотезу $H_0 : \mu = 100$ г/т при альтернативе $H_1 : \mu = 50$ г/т. Используя формулу (3.34), вычислим величину

$$b = a_1 + \frac{t_q \sigma}{\sqrt{n}}.$$

Положим $q = 0{,}01$, чему в нашем случае будет соответствовать $t_{0.01} = -2{,}33$. Тогда

$$b = 100 - \frac{2{,}33 \cdot 53}{\sqrt{10}} = 60{,}949 \simeq 61.$$

Таким образом, гипотеза $H_0 : \mu = 100$ г/т принимается, если $\bar{x} \geqslant 61$, и отклоняется при уровне значимости 0,01, если $\bar{x} < 61$.

Подсчитаем мощность критерия относительно альтернативы $H_1 : \mu = 50$ г/т

$$t_{1-\beta} = \frac{b - a_2}{\sigma\sqrt{n}} = \frac{(61 - 50)\sqrt{10}}{53} = 0{,}5967 \simeq 0{,}60.$$

Из таблицы нормального распределения находим

$$1 - \beta = 0{,}73, \text{ т. е. } \beta = 0{,}27.$$

Таким образом, вероятность появления ошибки второго рода при заданном уровне значимости довольно высока и равна 0,27.

В связи с этим вычислим новое значение b, соответствующее уровню значимости 0,05; при $t_{0,05} = -1,65$

$$b = 100 - \frac{1,65 \cdot 53}{\sqrt{10}} = 72,35 \simeq 72.$$

Соответствующее значение

$$t_{1-\beta} = \frac{(72-50)\sqrt{10}}{53} = 1,31,$$

откуда $1-\beta = 0,905$ и $\beta = 0,095$. Такие вероятности уже приемлемы и поэтому при проверке H_0 лучше использовать уровень значимости 0,05, чем 0,01. Так как наше значение $\bar{x} = 39$, что значительно меньше и 72 и 61, то проверяемая гипотеза уверенно отклоняется и принимается альтернатива H_1. Более того, если бы значение $\bar{x}$ принадлежало интервалу от 61 до 72, все же целесообразнее было бы отклонить проверяемую гипотезу, так как такое решение обеспечивало бы меньшую вероятность появления ошибки второго рода, наряду с небольшим значением уровня значимости.

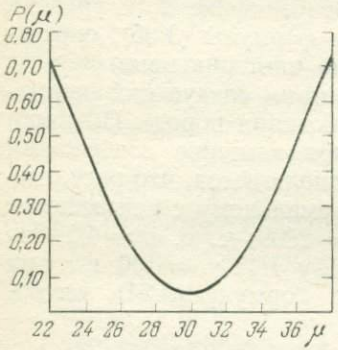

Рис. 11. Функции мощности по данным табл. 7

До сих пор мы рассматривали мощность критерия только относительно альтернативы. На практике же чаще всего приходится иметь дело со сложными альтернативами, например $\mu \neq a_1$ или $\mu_1 - \mu_2 \neq 0$ и т. п.

Во всех этих ситуациях сложная альтернатива представляет собой множество простых гипотез, каждой из которых можно поставить в соответствие значение мощности критерия.

Таким образом, в случае сложной конкурирующей гипотезы мощность будет представлять собой функцию, заданную на множестве альтернатив. Рассмотрим построение такой функции на следующем примере.

Пусть из нормальной совокупности с неизвестным средним μ и дисперсией, равной 100, взята выборка объемом $n = 10$. По этой выборке вычислена оценка для μ в виде среднего арифметического $\bar{x} = 31$. По этим данным требуется проверить гипотезу $H_0 : \mu = \mu_0 = 30$ при множестве альтернатив $H_1 : \mu \neq 30$. При проверке этой гипотезы мы имеем дело с так называемым двусторонним критерием. В соответствии с данными табл. 5 можем записать

$$q = P(t \in T_1 \mid H_0) = P(\bar{\xi} > b_2 \mid H_0) + P(\bar{\xi} < b_1 \mid H_0) \qquad (3.38)$$

или, что равносильно,

$$1-\Phi\left(\frac{b_2-\mu_0}{\sigma/\sqrt{n}}\right)+\Phi\left(\frac{b_1-\mu_0}{\sigma\sqrt{n}}\right)=q, \tag{3.39}$$

где Φ — стандартная нормальная функция, причем

$$\Phi\left(\frac{b_2-\mu_0}{\sigma/\sqrt{n}}\right)=1-\frac{q}{2}, \tag{3.40}$$

$$\Phi\left(\frac{b_1-\mu_0}{\sigma/\sqrt{n}}\right)=\frac{q}{2}. \tag{3.41}$$

Для $\frac{q}{2}=0{,}025$

$$\frac{b_2-\mu_0}{\sigma/\sqrt{n}}=1{,}96, \quad \frac{b_1-\mu_0}{\sigma/\sqrt{n}}=-1{,}96,$$

откуда

$$b_1=30-1{,}96\sqrt{10}=23{,}802, \qquad b_2=30+1{,}96\sqrt{10}=36{,}198.$$

Функция мощности (обозначим ее $P(\mu)$ будет определена в случае двустороннего критерия следующим выражением:

$$P(\mu)=1-\Phi\left(\frac{b_2-\mu}{\sigma/\sqrt{n}}\right)+\Phi\left(\frac{b_1-\mu}{\sigma/\sqrt{n}}\right). \tag{3.42}$$

Заметим, что при $\mu=\mu_0$, $P(\mu_0)=q$.
В нашем примере

$$P(\mu)=1-\Phi\left(\frac{36{,}198-\mu}{10/\sqrt{10}}\right)+\Phi\left(\frac{23{,}802-\mu}{10/\sqrt{10}}\right).$$

Ниже, в табл. 7, приведены значения этой функции, вычисленные для $12\leqslant\mu\leqslant 38$.

Таблица 7

Значения функции мощности

μ	$P(\mu)$	μ	$P(\mu)$	μ	$P(\mu)$
22	0,71567	29	0,06182	33	0,15806
24	0,46614	30	0,05000	34	0,24262
26	0,24262	31	0,06182	36	0,46614
27	0,15806	32	0,09656	38	0,71567
28	0,09656				

Из этой табл. 7 и рис. 11 видно, что функция $P(\mu)$ симметрична и достигает минимума при $\mu = \mu_0$, а это означает, что альтернативы, значения которых расположены близко к μ_0, наиболее трудноразличимы.

Необходимо отметить, что на функцию мощности в значительной степени влияет число наблюдений в выборке. Если в нашем примере увеличить число наблюдений с 10 до 100, оставив по-прежнему $\sigma^2 = 100$, а проверяемую гипотезу $H : \mu = \mu_0 = 30$, то получим

$$b_1 = 30 - 1{,}96\,\frac{\sqrt{100}}{10} = 28{,}04, \qquad b_2 = 30 + 1{,}96\,\frac{\sqrt{100}}{10} = 31{,}96.$$

ПРОВЕРКА НЕКОТОРЫХ ТИПОВЫХ ГИПОТЕЗ

В этой главе мы рассмотрим некоторые простейшие ситуации, с которыми приходится сталкиваться в геологических исследованиях, связанных с обоснованием выводов по статистическим данным. К ним прежде всего относятся такие задачи, как выбор математической модели, описывающей распределение значений изучаемой характеристики, проверка различных предположений о средних значениях, о дисперсиях как мерах рассеяния результатов наблюдений и о корреляционных характеристиках при изучении зависимостей между рассматриваемыми признаками.

В данной главе мы ограничимся наиболее простыми постановками с тем, чтобы в дальнейшем перейти к более сложным ситуациям. Следует отметить, что несмотря на простоту описанных в этой главе ситуаций, они очень часто наблюдаются в геологической практике и приведенные здесь приемы по обработке данных будут весьма полезны в работе геолога.

4.1. ПРОВЕРКА ГИПОТЕЗ О ФУНКЦИЯХ РАСПРЕДЕЛЕНИЯ

Во многих геологических исследованиях требуется дать сводное описание совокупности значений изучаемой характеристики, например содержания элемента в породе и т. п. Для этих целей одного среднего значения, или среднего и дисперсии, бывает недостаточно. Требуется еще указать вид функции, которая описывает данное эмпирическое распределение. Знание этого вида необходимо еще и для того, чтобы сознательно выбирать по возможности эффективные критерии и статистические оценки параметров, соответствующие установленному распределению при решении геологических задач.

4.1.1. Общая постановка задачи о виде функции распределения и ее проверке

Допустим, что в нашем распоряжении имеется выборка объема n, элементы которой мы обозначим $x_1, x_2, \ldots, x_n$, и пусть $F(x)$ — неизвестная функция распределения, оцениваемая по выборке. Обозначим через $F_0(x)$ заданную функцию распределения, которую предполагается использовать в качестве модели изучаемого распределения. Таким образом, задача заключается в проверке гипотезы $H_0 : F(x) = F_0(x)$ при альтернативе $H_1 : F(x) \neq F_0(x)$.

Наиболее общий способ проверки гипотезы H_0 заключается в следующем. Область выборочных значений $x_1, x_2, \ldots, x_n$ раз-

бивается на k интервалов, необязательно равных, и подсчитываются частости n_l попадания значений выборки в эти интервалы, $l = 1, 2, \ldots, k$. Если проверяемая гипотеза H_0 верна, то число

$$\chi^2_{k-m} = \sum_{l=1}^{k} \frac{(N_l - n_l)^2}{N_l} \tag{4.1}$$

будет представлять собой значение случайной величины, распределенной как χ^2 с $k—m$ степенями свободы, где k — число интервалов, а m — число наложенных связей. Значения N_l представляют собой теоретические частоты попадания выборочных значений изучаемой случайной величины в интервалы с номерами $l = 1, 2, \ldots, k$.

Таким образом, гипотеза H_0 принимается, если вычисленное значение $\chi^2_{k-m} \leqslant \chi^2_{q,\,k-m}$, где $\chi^2_{q,\,k-m}$ допустимое значение, соответствующее уровню значимости q и $k—m$ степеням свободы.

Теоретические частоты N_l вычисляются следующим образом. Обозначим через a_l границы интервалов, на которые разделена область выборочных значений, т. е. $a_1, a_2, \ldots, a_l, \ldots, a_{k+1}$. Тогда для каждого значения a_l можно вычислить функцию $F_0(a_l)$, а для всех интервалов определить разности

$$P_l = F_0(a_{l+1}) - F(a_l), \tag{4.2}$$

которые представляют собой вероятности событий, заключающихся в том, что значение изучаемой случайной величины будет принадлежать интервалу от a_l до a_{l+l}. Тогда теоретические частоты n_l можно вычислить по формуле

$$N_l = nP_l, \tag{4.3}$$

где n — объем выборки.

Необходимо отметить, что описанный критерий является практически универсальным, так как не зависит от вида функции распределения, выбранной для H_0. Неудобство его заключается в том, что он неприменим в условиях малых выборок, которые дают очень ненадежную информацию о частотах попадания значений выборки в заданные интервалы.

Для иллюстрации вычислений, связанных с этим критерием, рассмотрим следующий пример.

В результате проведения количественно-минералогических подсчетов в 50 шлифах песчаников одного возраста, отобранных наудачу с заданной территории было установлено, что общее содержание темноцветных минералов колеблется в них от 1 до 6 %. В табл. 8 показано, как результаты этих 50 подсчетов распределились по пяти интервалам, на которые разбита область от 1 до 6 %.

Полученное эмпирическое распределение можно изобразить в виде гистограммы, Форма гистограммы значительно отличается от какого-либо одновершинного распределения и поэтому в каче-

Таблица 8

Проверка гипотезы о равномерном распределении

a_l	$F(a_l)$	P_l	N_l	n_l	$\frac{(N_l - n_l)^2}{N_l}$
1	0,0000				
2	0,2000	0,20	10	7	0,9
3	0,4000	0,20	10	12	0,4
4	0,6000	0,20	10	13	0,3
5	0,8000	0,20	10	6	1,6
6	1,0000	0,20	10	12	0,4
					$\chi^2_{k-m} = 3{,}6$

стве $F_0(x)$, т. е. проверяемой гипотезы, была выбрана функция равномерного распределения. Таким образом,

$$H_0 : F(x) = F_0(x) = \int_1^x \frac{dv}{6-1} = \frac{1}{5}\int_1^x dv \qquad (4.4)$$

при альтернативе $H_1 : F(x) \neq F_0(x)$.

В этой таблице показана и процедура вычислений значения χ^2_{k-m}, которое оказалось равным 3,6. Число степеней свободы в данном случае равно 4, так как существует только одно наложенное условие $\sum_{l=1}^{k} n_l/n = 1$. Критическое значение $\chi^2_{0,05;\,4} = 9{,}488$, а так как $\chi^2_4 = 3{,}6$, то проверяемую гипотезу о равномерном распределении следует принять, как не противоречащую выборочным данным.

4.1.2. Проверка гипотезы о нормальном распределении

Описанный выше критерий можно с успехом использовать для проверки гипотезы о нормальном распределении. Однако вычисление теоретических частот N_l в данном случае производится несколько сложнее, чем при проверке гипотезы о равномерном распределении. Эти вычисления сводятся к следующему. Пусть взята выборка объема n, которая распределена по соответствующим интервалам (a_l, a_{l+1}). Обозначим, как мы это уже делали, число наблюдений, попавших в интервал с номером l, через n_l а середину этого интервала через x_l, т. е.

$$x_l = \frac{a_l + a_{l+1}}{2}. \qquad (4.5)$$

Если значения среднего μ и дисперсии σ^2 неизвестны (а это обычная практическая ситуация), то их оценки в случае группиро-

ванных данных вычисляются по формулам

$$\bar{x} = \frac{1}{n} \sum_{l=1}^{k} n_l x_l, \tag{4.6}$$

$$s^2 = \frac{1}{n-1} \sum_{l=1}^{k} n_l (x_l - \bar{x})^2. \tag{4.7}$$

После того как эти оценки вычислены, для каждой границы интервала a_l вычисляется величина

$$t_l = \frac{a_l - \bar{x}}{s} \tag{4.8}$$

и для всех t_l находятся значения нормальной стандартной функции $\Phi(t_l)$, которые берутся из таблиц. Таким образом, как и в случае равномерного распределения, можно подсчитать теоретические частоты N_l

$$N_l = n[\Phi(t_{l+1}) - \Phi(t_l)]. \tag{4.9}$$

Эти частоты показывают, как бы распределились наши n наблюдений, если бы выборка была взята из нормальной совокупности со средним $\mu = \bar{x}$ и дисперсией $\sigma^2 = s^2$. Таким образом, при проверке нашей гипотезы используются три ограничения

$$\sum_{l=1}^{k} n_l = n, \quad \text{или} \quad \frac{1}{n} \sum_{l=1}^{k} n_l = 1, \tag{4.10}$$

$$\mu = x, \tag{4.11}$$

$$\sigma^2 = s^2, \tag{4.12}$$

и, следовательно, число степеней свободы критерия

$$\chi^2_{k-m} = \sum_{l=1}^{k} \frac{(N_l - n_l)^2}{N_l} \tag{4.13}$$

будет равно $k-3$.

Если $\chi^2_{k-3} \leqslant \chi^2_{q,\,k-3}$, то гипотеза о нормальном распределении принимается как не противоречащая выборочным данным, тогда как если $\chi^2_{k-3} > \chi^2_{q,\,k-3}$, она отклоняется при уровне значимости q.

В качестве примера проверки гипотезы о нормальном распределении можно рассмотреть данные по содержаниям Al_2O_3 в гранитах, заимствованные из книги Е. А. Струве [41]. Всего $n = 313$ результатов анализов следующим образом распределились по 12 интервалам через 1% (табл. 9).

Из полученных результатов видно, что гипотеза о нормальном распределении не противоречит выборочным данным, так как вычисленное значение критерия, равное 13,36, меньше, чем допустимое при 5 %-ном уровне значимости и 7 степенях свободы, которое равно 14,067.

Таблица 9

Проверка гипотезы о нормальном распределении

a_l	x_l	n_l	t_l	$\Phi(t_l)$	N_l		$\frac{(N_l - n_l)^2}{N_l}$
8			—3,33	0,00048342			
	8,5	2			0,58	3,78	2,74
9			—2,83	0,00232740	3,20		
	9,5	5					
10			—2,24	0,0125455	10,02		0,91
	10,5	7					
11			—1,70	0,0445655	25,20		1,53
	11,5	19					
12			—1,15	0,125072	45,65		0,88
	12,5	52					
13			—0,61	0,270931	62,97		0,57
	13,5	57					
14			—0,07	0,472097	66,45		0,46
	14,5	72					
15			0,48	0,684386	50,63		2,12
	15,5	61					
16			1,02	0,846136	29,94		4,00
	16,5	19					
17			1,57	0,9417924	12,76		0,12
	17,5	14					
18			2,11	0,9825708	4,20	5,43	0,03
	18,5	4					
19			2,65	0,99597541	1,23		
	20,0	1					
21			3,74	0,999907990			
							$\chi_7^2 = 13{,}36$

Другой удобный способ проверки гипотезы о нормальном распределении основан на рассмотрении оценок асимметрии $\hat{\gamma}_1$ и эксцесса $\hat{\gamma}_2$. Напомним, что

$$\hat{\gamma}_1 = \frac{m_3}{s^3}, \tag{4.14}$$

$$\hat{\gamma}_2 = \frac{m_4}{s^4} - 3, \tag{4.15}$$

где m_3 и m_4 — третий и четвертый центральные выборочные моменты, а s^2 — оценка дисперсия.

В условиях нормального распределения случайные величины, значения которых $\hat{\gamma}_1$ и $\hat{\gamma}_2$ мы наблюдаем, распределены приблизительно нормально со средними значениями, равными нулю и приближенными дисперсиями $\sqrt{\frac{6}{n}}$ и $\sqrt{\frac{24}{n}}$ соответственно.

Таким образом, числа

$$t_1 = \frac{\hat{\gamma}_1}{\sqrt{\frac{6}{n}}}, \quad t_2 = \frac{\hat{\gamma}_2}{\sqrt{\frac{24}{n}}} \tag{4.16}$$

в случае нормального распределения, будут представлять собой значения случайных величин, распределенных приблизительно нормально с параметрами 0 и 1. Поэтому гипотезу о нормальном распределении следует отклонить, если хотя бы одно значение t_1 или t_2 по абсолютной величине превысит допустимое t_q, при заданном уровне значимости q.

В качестве примера проверки гипотезы о нормальном распределении с помощью этого критерия рассмотрим данные по содержанию MgO в пробах из дунитов (данные представлены Ю. Е. Молдаванцевым), приведенные в табл. 10.

Таблица 10

Содержание MgO в дунитах Войкаро-Сыньинского гипербазитового массива (в %)

i	MgO	i	MgO	i	MgO	i	MgO
1	43,36	9	44,33	17	24,87	25	41,66
2	43,16	10	44,62	18	36,10	26	41,79
3	49,61	11	44,61	19	41,43	27	42,83
4	40,53	12	44,61	20	45,60	28	36,64
5	48,57	13	36,67	21	36,96	29	42,07
6	43,75	14	38,93	22	43,65	30	37,39
7	39,92	15	11,97	23	39,55	31	39,52
8	43,15	16	30,14	24	39,01	32	31,42

По этим данным были вычислены $\bar{x} = 39{,}64$; $s^2 = 51{,}49$; $s^3 = 370{,}14$; $s^4 = 2657{,}6$; $\hat{\gamma}_1 = -1{,}96$ и $\hat{\gamma}_2 = 5{,}12$, а также значения двух критериев

$$t_1 = \frac{|\hat{\gamma}_1|}{\sqrt{\frac{6}{n}}} = \frac{1{,}96}{\sqrt{\frac{6}{32}}} = -4{,}53,$$

$$t_2 = \frac{|\hat{\gamma}_2|}{\sqrt{\frac{24}{n}}} = \frac{5{,}12}{\sqrt{\frac{24}{n}}} = 5{,}88.$$

Нетрудно видеть, что гипотезу о нормальном распределении следует отклонить, так как $t_1 < -3{,}00$ и $t_2 > 3{,}00$.

4.1.3. Проверка гипотезы о логнормальном распределении

В геологии впервые близость статистических кривых распределения содержаний некоторых элементов в породах и рудах к логнормальной кривой была отмечена Н. К. Разумовским [35, 36].

Проверка гипотезы о логнормальном распределении не вызывает каких-либо затруднений и сводится к проверке гипотезы о нормальном распределении логарифмов изучаемой случайной величины, которую можно провести одним из рассмотренных выше способов. В качестве примера рассмотрим результаты проверки гипотезы о логнормальном распределении содержаний циркона в гранитоидах послениженекарбонового возраста на Урале. Наиболее распространенной породой среди них является биотитовый гранит, иногда переходящий в биотит-роговообманковый, и менее распространены гранодиориты, наблюдающиеся в основном как краевые фации крупных массивов. В значительном подчинении находятся пегматоидные и аплитовидные разновидности гранитов. Всего на акцессорные минералы было опробовано пять массивов — Сысертский, Шилово-Коневский, Рефтинский, Мурзинский и Магнитогорский, в которых в общей сложности было собрано 119 проб. Ниже, в табл. 11, приведены результаты проверки гипотезы о логнормальном распределении содержаний (в г/т) циркона по данным упомянутых 119 проб.

Таблица 11

Проверка гипотезы о логнормальном распределении

a_l	$\lg a_l$	y_l	n_l	t_l	$\Phi(t_l)$	N_l	$\frac{(N_l - n_l)^2}{N_l}$
0,06	—1,200			—2,33			
		—0,950	5			4,37	0,09
0,20	—0,7000			—1,79	0,03673		
		—0,450	8			8,20	0,00
0,63	—0,200			—1,25	0,10565		
		0,050	18			11,95	3,06
2,00	0,300			—0,82	0,20611		
		0,550	19			26,94	2,34
6,30	0,800			—0,17	0,43251		
		1,050	27			24,76	0,20
20,00	1,300			0,36	0,64058		
		1,550	15			20,87	1,65
63,0	1,800			0,90	0,81594		
		2,050	18			12,99	1,93
200,0	2,300			1,44	0,92507		
		2,550	9			8,92	0,0
630,0	2,800			1,98			
							$\chi^2_5 = 9,27$

Так как число интервалов группировки равно 8, то число степеней свободы будет 8—3 = 5, чему при 5 %-ном уровне значи-

мости соответствует значение 11,07. Таким образом, так как вычисленное значение $\chi_5^2 = 9{,}27$ меньше, чем допустимое $\chi^2_{0{,}05;5} = 11{,}07$, гипотезу о логнормальном распределении следует принять как непротиворечащую выборочным данным.

В главе 1 нами были рассмотрены также распределения, связанные с логнормальным, в частности распределение вида $\Lambda(x—a; \mu, \sigma^2)$, когда изучаемая случайная величина представляет собой сумму константы a и логнормально распределенной случайной величины, и распределение вида $1—\Lambda(a—x; \mu, \sigma^2)$, описывающего поведение случайной величины, являющейся разностью константы a и логнормально распределенной случайной величины.

Для проверки гипотезы о согласованности выборочного распределения с распределением вида $\Lambda(x—a; \mu, \sigma^2)$ нужно выборочные значения x_i заменить на $y_i = x_i—a$. Показанием к такой проверке является наличие существенно выраженной положительной асимметрии в распределении логарифмов изучаемой случайной величины, т. е. случай, когда гипотеза о нормальном распределении логарифмов отклоняется за счет значительной положительной асимметрии.

Технически сама проверка гипотезы о нормальным распределении по значениям $\log y_i = \log(x_i—a)$ не представляет собой каких-либо затруднений, за исключением процедуры выбора константы a. Для этого пока не существует никаких приемов кроме перебора различных вариантов значений константы до тех пор, пока не будет найден тот, при котором гипотеза о нормальном распределении будет принята для выборочных данных $\log(x_i—a)$.

В качестве примера проверки гипотезы об этом распределении можно привести данные табл. 12 по содержаниям Ta_2O_5 в амазонитовых гранитах, предоставленные автору А. М. Гребенниковым.

Таблица 12

Содержание Ta_2O_5 в амазонитовых гранитах (в %)

x_i	n_i	x_i	n_i	x_i	n_i	x_i	n_i
0,005	3	0,014	13	0,023	3	0,042	1
0,006	7	0,015	8	0,025	5	0,045	2
0,007	13	0,016	5	0,026	1	0,047	1
0,008	11	0,017	4	0,027	4	0,050	2
0,009	23	0,018	6	0,028	2	0,055	1
0,010	19	0,019	4	0,030	5	0,057	1
0,011	11	0,020	4	0,033	1	0,060	1
0,012	18	0,021	3	0,037	1	0,064	1
0,013	4	0,022	5	0,041	1	0,080	1

В этой таблице n_i означает наблюдаемую частоту соответствующего значения x_i — содержания Ta_2O_5 (в %).

Проверка гипотезы о логнормальном распределении содержаний Ta_2O_5 дала отрицательный результат. Расхождение наблюдаемого эмпирического распределения логарифмов содержаний и нормальной модели проявляется в отчетливо выраженной положительной асимметрии и положительном эксцессе. Вычисления показали, что оценки асимметрии распределения логарифмов содержаний Ta_2O_5 и эксцесса равны соответственно 1,094 и 1,560, тогда как допустимые их значения при числе наблюдений 195 и 1 %-ном уровне значимости составляют 0,525 и 1,050. Нетрудно видеть, что вычисленные оценки асимметрии и эксцесса распределения логарифмов содержаний Ta_2O_5 сильно превышают допустимые значения и поэтому гипотезу о логнормальном распределении можно уверенно отклонить.

После того как выборочные значения x_i содержаний Ta_2O_5 были заменены величинами $y_i = x_i - 0{,}0045$ и по значениям $\log y_i$ были вычислены значения оценок асимметрии и эксцесса, оказалось, что проверяемая гипотеза не противоречит выборочным данным. Вычисленные оценки асимметрии и эксцесса оказались равными соответственно 0,164 и 0,480, т. е. меньшими, чем допустимые значения 0,521 и 1,050, что позволяет уверенно принять гипотезу о нормальном распределении для величин $\log y_i$ и, следовательно, гипотезу о распределении $\Lambda(x-a;\ \mu,\ \sigma^2)$ для содержаний Ta_2O_5.

Вопрос о генетическом истолковании полученного результата мы рассмотрим в следующем разделе, а до этого остановимся на примере проверки гипотезы о распределении вида $1-\Lambda(a-x;\ \mu,\ \sigma^2)$. Показанием к такой проверке является наличие в изучаемом выборочном распределении существенно выраженной отрицательной асимметрии и положительного эксцесса. Именно с такой ситуацией мы столкнулись в предыдущем разделе при проверке гипотезы о нормальном распределении по содержаниям MgO в дунитах Войкаро-Сыньинского гипербазитового массива (см. табл. 10).

После того как выборочные значения x_i содержаний MgO были заменены величинами $y_i = 60 - x_i$, для которых была проведена проверка гипотезы о логнормальном распределении, оказалось, что это распределение является вполне подходящей моделью. Полученные значения отношений оценок асимметрии $\hat{\gamma}_1$ и эксцесса $\hat{\gamma}_2$ к их стандартным отклонениям $\sqrt{\frac{6}{n}}$ и $\sqrt{\frac{24}{n}}$ оказались равными

$$|\hat{\gamma}_1| \Big/ \sqrt{\frac{6}{n}} = 1{,}898, \quad |\hat{\gamma}_2| \Big/ \sqrt{\frac{24}{n}} = 1{,}532.$$

Так как эти отношения значительно меньше 3, гипотеза о распределении вида $1-\Lambda(a-x;\ \mu,\ \sigma^2)$ для содержаний MgO в дунитах может быть уверенно принята, как непротиворечащая выборочным данным.

Геохимическое объяснение полученного результата будет дано в следующем разделе.

4.1.4. Роль моделей распределений в геологии

Вид функций распределения различных числовых геологических характеристик давно интересовал многих исследователей. Несмотря на то, что эти исследования нередко ограничивались только визуальным изучением формы эмпирической кривой распределения, они не только сыграли существенную роль в методологическом отношении, но и позволили сделать ряд важных геологических выводов. Необходимо отметить, что первоначально особенно широко эти вопросы затрагивались при подсчете запасов полезных ископаемых, примерами чего могут служить работы Н. В. Барышева [3], Д. А. Зенкова [20] и др.

Еще в 1925 г. Ф. Ю. Левинсон-Лессинг [30, 31] использовал статистические методы для обоснования геохимических выводов. Основываясь на построении статистических аналогов функций распределения содержаний некоторых окислов в излившихся породах, он провел разграничение семейств андезитов-базальтов, трахитов-базальтов и др. Следует отметить, что большинство изученных Ф. Ю. Левинсоном-Лессингом породообразующих окислов в пределах заданного узкого типа пород характеризовалось содержаниями, распределения которых были близки к нормальному. Значительно позднее, в 1940 и 1941 гг., Н. И. Разумовским была отмечена близость статистических кривых распределения содержаний некоторых элементов в породах и рудах к логнормальному закону. В 1954 г. Л. Г. Аренс [45, 46, 47] на основании рассмотрения гистограмм распределения результатов анализов 20 элементов в различных изверженных породах высказал утверждение, что содержания элементов в изверженных горных породах распределены логнормально. Однако он не проводил статистическую проверку согласованности эмпирических распределений с теоретическим логнормальным, и этот весьма категоричный вывод Л. Г. Аренса был основан только на визуальном рассмотрении гистограмм.

Несколько позднее Р. Л. Миллер и Е. Д. Гольдберг [54] провели проверку данных Аренса при помощи критерия χ^2, в результате чего установили, что из 20 приведенных Аренсом примеров с логнормальным распределением согласуются только шесть, 12 элементов характеризуются промежуточным положением распределения содержаний между нормальным и логнормальным законами, а содержания двух элементов распределены нормально.

Критике работ Л. Г. Аренса посвящен ряд статей К. В. Обри [49], С. Дюровича [51], А. В. Вистелиуса [57]. Последняя работа посвящена рассмотрению таких понятий, как «локальное» и «обобщенное» распределение, причем предполагается, что «локальное» распределение (в небольшой пространственно ограниченной части породы) близко к нормальному, тогда как «обобщенное» (связано с совокупностями более широкого плана) является асимметричным.

В начале 60-х годов автором было проведено систематическое исследование по выявлению типичных моделей распределения содержаний элементов и минералов в изверженных горных породах [37]. На основании большого фактического материала, как собственного, так и заимствованного из разнообразных литературных источников, в этой работе было изучено 109 различных распределений. Этот материал был разделен на три группы, из которых первая (42 примера) посвящена изучению распределений содержаний минералов в изверженных горных породах, вторая (35 примеров)— распределениям содержаний элементов в минералах и, наконец, третья содержит 32 примера изучения распределений содержаний элементов в изверженных породах. Существенное внимание было уделено в этой работе условиям возникновения того или иного вида распределения.

В результате проведенных исследований было установлено, что распределения содержаний минералов (как породообразующих, так и акцессорных) очень часто согласуются с логнормальной функцией, причем весьма характерна устойчивость этого распределения как для «локальных», так и «обобщенных» совокупностей.

Те немногие случаи отклонения распределений содержаний минералов от логнормального распределения при детальном петрографическом исследовании удалось объяснить наличием более чем одной независимой генерации минерала, что позволяло рассматривать содержание изучаемого минерала как сумму двух или более независимых случайных величин. Так, например, при изучении локальных распределений рассматривалась выборка из 31 определения содержаний калиевого полевого шпата, плагиоклаза, кварца и биотита в шлифах из одного штуфа гранодиорита (массив Огузтау в Бетпак-Дале). Оказалось, что распределения всех изученных минералов, кроме кварца, хорошо согласуются с логнормальным законом, тогда как распределение логарифмов содержаний кварца характеризуется существенной отрицательной асимметрией и положительным эксцессом. Детальное петрографическое исследование показало, что кварц в гранодиоритах представлен двумя независимыми генерациями, из которых первая связана с кристаллизацией породы, а вторая — с процессом более позднего окварцевания.

Кроме того, на примере пироксенов, замещенных роговой обманкой и оливином, в пироксенитах Баранчинского массива на Среднем Урале было установлено, что исходное распределение содержаний минерала до замещения его другими минералами не противоречит логнормальному закону.

Исследование распределений содержаний элементов в различных минералах показало, что случаи несогласованности этих распределений с логнормальным законом встречаются крайне редко. Характерно, что в редких случаях несогласованности распределений с логнормальным законом изучаемый минерал оказался представленным несколькими генерациями. Так, например, при изуче-

нии распределения содержаний галлия в сфалеритах Такобского месторождения [12] оказалось, что это распределение хорошо согласуется с нормальным законом, а детальное минералогическое исследование руд показало, что сфалерит представлен четырьмя независимыми генерациями. Таким образом, содержание галлия в сфалеритах этого месторождения можно рассматривать как взвешенную сумму четырех случайных величин.

Изучение распределений содержаний элементов в различных изверженных породах показало, что наблюдаются случаи согласованности этих распределений с логнормальным и нормальным распределением, а также с распределениями вида $\Lambda(x—a;\ \mu,\ \sigma^2)$ и $1—\Lambda(a—x;\ \mu,\ \sigma^2)$. Примерами распределений, согласующихся с логнормальным законом, могут служить распределения содержания кальция в гранитах (по данным К. К. Турекьяна и Дж. Л. Калпа [42]), мышьяка в гранитах (по данным Х. Ониси и Е. Б. Санделла [33]), титана в долеритах Карру (по данным, Ф. Уокера и А. Польдерварта [43]) и ряд других примеров. Отличительной особенностью этих элементов является то, что в изученных породах они в подавляющей массе находятся в одном минерале. Так, приблизительно 90 % всего кальция в гранитах содержится в плагиоклазе, мышьяк в основном концентрируется в акцессорном арсенопирите. Что же касается титана в долеритах Карру, то около 93 % от всей его массы приходится на ильменит, и только около 7 % на долю пироксена. Подробно эти вопросы рассмотрены в одной из работ автора [37].

С примером нормального распределения содержаний элемента в породе нам уже приходилось встречаться в разделе 4.1.2 при рассмотрении распределения содержаний Al_2O_3 в гранитах по данным Е. А. Струве [41]. Другими примерами согласованности распределений содержаний элементов с нормальным законом могут служить распределения содержаний SiO_2 в гранитах [41], Na_2O в долеритах Карру и др. Весьма характерно, что SiO_2 и Al_2O_3 в гранитах входят в состав всех главных породообразующих минералов — плагиоклазов, калиевых полевых шпатов, биотитов, амфиболов и др., причем количества SiO_2, приходящиеся на долю кварца, плагиоклазов и калиевых полевых шпатов, приблизительно равны. Аналогично, Na_2O в долеритах Карру содержится не менее чем в четырех минералах приблизительно равными порциями — в плагиоклазах ранней стадии кристаллизации, амфиболах, пироксенах и позднем альбите.

Таким образом, во всех этих примерах изучаемые содержания можно рассматривать как суммы нескольких случайных величин, которые приблизительно равновелики.

Что же касается распределений вида $\Lambda(x—a;\ \mu,\ \sigma^2)$ и вида $1—\Lambda(x—a;\ \mu,\ \sigma^2)$, то связанные с ними примеры были рассмотрены в предыдущем разделе. Здесь следует только отметить, что из факта согласованности распределений Ta_2O_5 с распределением вида $\Lambda(x—a;\ \mu,\ \sigma^2)$ следует, что содержания Ta_2O_5 в амазонит-альби-

товых границах можно рассматривать как сумму константы и логнормально распределенной случайной величины. Петрографические исследования показали, что амазонит-альбитовые граниты являются продуктом метасоматического изменения порфировидных биотитовых гранитов, которое выразилось в их микроклинизации и альбитизации. Это замещение сопровождалось привносом тантала, вошедшего в состав танталита и в меньшей степени микролита. Однако некоторое количество тантала содержалось в исходных биотитовых гранитах, так как в тех их участках, которые не затронуты процессами метасоматоза, содержится около 0,0045 % Ta_2O_5 с очень незначительными колебаниями. Эту величину и можно рассматривать как константу, к которой добавлена логнормальная случайная величина. Таким образом, и этот вид распределения находит свое генетическое объяснение.

Тот факт, что распределение MgO в дунитах Войкаро-Сыньинского гипербазитового массива согласуется с распределением $1—\Lambda\ (a—x;\ \mu,\ \sigma^2)$, объясняется тем, что в оливине, которым практически полностью сложены дуниты, двухвалентные ионы магния и железа изоморфно замещают друг друга, а сумма содержаний MgO и FeO практически представляет собой константу. Изучение же функции распределения содержаний FeO показало, что она практически не отличается от логнормальной, а поэтому содержание MgO следует рассматривать как разность константы и логнормально распределенной случайной величины.

Все приведенные примеры свидетельствуют о том, что вид функции распределения изучаемых геологических характеристик может содержать важную геологическую информацию. Однако следует помнить, что сам по себе факт пригодности в качестве модели того или иного вида распределения свидетельствует только о том, что по отношению к изучаемой величине выполнены те или иные формальные условия. Например, в случае нормального распределения выполнены условия центральной предельной теоремы и случайную величину можно рассматривать как сумму независимых приблизительно равновеликих слагаемых. На вопрос о том, что собой представляют эти слагаемые, можно ответить только после специального геологического исследования и дать единого рецепта получения ответа на этот вопрос невозможно.

4.2. ПРОВЕРКА ГИПОТЕЗ О РАВЕНСТВЕ СРЕДНИХ ЗНАЧЕНИЙ

Практически не существует такого раздела геологии, где бы не приходилось сравнивать средние значения изучаемых характеристик и по результатам этих сравнений делать геологические выводы. Геологические задачи, требующие сравнения средних значений, столь многообразны, что их невозможно перечислить, и каждый геолог обязательно сталкивался с этим вопросом в своей работе.

В главе 3 мы уже встречались с некоторыми примерами проверки

гипотез о средних, в частности о равенстве неизвестного среднего заданному значению в условиях простой и сложной альтернативы для выборок из нормальной совокупности с известной дисперсией. Однако нам все же придется вернуться к этим ситуациям для того, чтобы рассмотреть критерии, рекомендуемые для малых выборок.

Все дальнейшее изложение способов проверки гипотез о средних построено в соответствии с тремя типичными для геологических задач ситуациями, а именно для нормального распределения, для логнормального распределения и для случая, когда распределение априори неизвестно или же отличается как от нормального, так и логнормального. Кроме того, в геологических исследованиях практически не встречаются случаи, когда значение дисперсии точно известно. Поэтому во всех дальнейших задачах этой главы мы будем рассматривать только те ситуации, когда дисперсия оценивается по выборке.

4.2.1. Проверка гипотезы о равенстве неизвестного среднего заданному значению

Формальная постановка этой задачи сводится к следующему. Пусть ξ — изучаемая случайная величина, над которой проведено n наблюдений $X = (x_1, x_2, \ldots, x_i, \ldots, x_n)$, а $M\,\xi$ — математическое ожидание случайной величины ξ. Требуется построить некоторый критерий t, с помощью которого можно было бы проверить гипотезу $H_0 : \mu = \mu_0$, если μ_0 — заданное значение.

В условиях нормального распределения случайной величины ξ в качестве такого критерия можно использовать выражение

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \tag{4.17}$$

где

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{4.18}$$

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}^2)}, \tag{4.19}$$

т. е. $\bar{x}$ и s являются средним арифметическим и оценкой стандартного отклонения соответственно.

Если проверяемая гипотеза верна, то при достаточно большом числе n (больше 20) t будет представлять собой значение нормально распределенной случайной величины со средним, равным 0, и дисперсией, равной 1. Таким образом, в зависимости от альтернативы с помощью таблиц нормального распределения выбирается критическое t_q, соответствующее заданному уровню значимости q. Так, при альтернативе $H_1 : \mu < \mu_0$ гипотеза H_0 отклоняется, если, $t < t_q$, при альтернативе $H_1 : \mu > \mu_0$ H_0 отклоняется, если $t > t_q$,

а при альтернативе $H_1 : \mu \neq \mu_0$ H_0 отклоняется, если $|t| > t_{1-\frac{q}{2}}$.

Так, при уровне значимости 0,05 для двух первых односторонних альтернатив критическое значение $t_{0,05}$ будет равно — 1,63 и 1,63 соответственно, тогда как для двусторонней альтернативы $t_{1-0,025} = t_{0,975} = 1,96$.

Выше было отмечено, что распределение критерия (4.17) будет практически совпадать с нормальным, если число наблюдений в выборке больше 20. Если же n невелико, то в условиях нормального распределения изучаемой случайной величины ξ для критерия (4.17) можно построить точное распределение, которое, если H_0 верна, описывается распределением Стьюдента, рассмотренным нами в главе 1. Действительно, так как s^2 представляет собой значение случайной величины $\sigma^2\chi^2/(n-1)$, то после несложных преобразований критерий t можно рассматривать как значение случайной величины, являющейся частным от деления нормальной случайной величины с параметрами 0,1 на корень квадратный из χ^2-распределенной случайной величины. Выполнение этих условий, как было показано в главе 1, приводит к распределению Стьюдента, в данном случае с $n-1$ степенями свободы. Таким образом, для принятия H_0 или ее отклонения вычисленное значение t нужно сравнить с допустимым значением $t_{q, n-1}$, соответствующим заданной альтернативе при уровне значимости q и $n-1$ степенями свободы. Эти значения приведены в прил. 6 в конце книги. Заметим, что при $n-1$, превышающем 20, распределение Стьюдента практически совпадает с нормальным.

В условиях логнормального распределения изучаемой случайной величины ξ проверка гипотезы $H_0 : M\xi = a_0$, т. е. о равенстве математического ожидания ξ заданному значению, производится следующим образом.

Допустим, что над логнормально распределенной с параметрами μ, σ^2 случайной величиной проведено n наблюдений $x_1, x_2, \ldots, x_i, \ldots, x_n$. По этим данным нетрудно получить статистические оценки для μ и σ^2, которые будут соответственно равны

$$\overline{\ln x} = \frac{1}{n} \sum_{i=1}^{n} \ln x_i, \tag{4.20}$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\ln x_i - \overline{\ln x})^2, \tag{4.21}$$

$\mu + \frac{1}{2}\sigma^2$.

В связи с тем, что $M\xi = e$ гипотезу $H_0 : M\xi = a_0$ можно заменить равносильной

$$H_0' : \mu + \frac{1}{2}\sigma^2 = \ln a_0. \tag{4.22}$$

Оценкой для суммы $\mu + \frac{1}{2}\sigma^2$ является число $\overline{\ln x} + \frac{1}{2}s^2$, являющееся значением случайной величины, распределение которой при достаточно больших n будет близко к нормальному с параметрами $\mu + \frac{1}{2}\sigma^2$, $\frac{\sigma^2}{n} + \frac{\sigma^4}{2(n-1)}$. Таким образом, число

$$t = \frac{\overline{\ln x} + \frac{1}{2}s^2 - \ln a_0}{\sqrt{\frac{s^2}{n} + \frac{s^4}{2(n-1)}}} \tag{4.23}$$

будет при условии, что верна гипотеза H_0, представлять собой значение случайной величины, распределенной приблизительно нормально со средним значением, равным 0, и дисперсией, равной 1. Следовательно, критическое значение t_q, соответствующее уровню значимости q и заданной альтернативе H_1, можно взять из таблицы нормального распределения (см. прил. 1), которая приведена в конце книги. Так, при уровне значимости 0,05 принимается альтернатива $H_1 : M\xi < a_0$, если $t < t_{0,05} = -1{,}63$, или альтернатива $H : M\xi > a_0$, если $t > t_{0,95} = 1{,}63$, или же альтернатива $H_1 : M\xi \neq a_0$, если $|t| > t_{0,975} = 1{,}96$. Следует отметить, что на практике проводить вычисления удобнее с десятичными логарифмами и поэтому критерий (4.23) можно записать в виде

$$t = \frac{\overline{\lg x} + 1{,}153\,s^2 - \lg a_0}{\sqrt{\frac{s^2}{n} + \frac{2{,}650\,s^4}{n-1}}}. \tag{4.24}$$

4.2.2. Проверка гипотезы о равенстве двух неизвестных средних

Эта задача является одной из самых распространенных в геологических исследованиях, так как утверждение о равенстве или различии средних значений характеристик двух геологических объектов лежит в основе многих геологических выводов. В общем случае ситуация сводится к следующему. Рассматриваются две случайные величины ξ_1 и ξ_2, над которыми проведено по n_1 и n_2 наблюдений $X_1 = (x_{11}, x_{12}, \ldots, x_{1i}, \ldots, x_{1n_1})$ и $X_2 = (x_{21}, x_{22}, \ldots, x_{2i}, \ldots, x_{2n_2})$ соответственно. По двум выборкам X_1 и X_2 требуется проверить гипотезу $H_0 : M\xi_1 = M\xi_2$ при одной из следующих возможных альтернатив $H_1 : M\xi_1 < M\xi_2$, $H_1 : M\xi_1 > M\xi_2$ и $H_1 : M\xi_1 \neq M\xi_2$ и при уровне значимости q.

В условиях нормального распределения случайных величин ξ_1 и ξ_2 математические ожидания $M\xi_1$ и $M\xi_2$ равны соответствующим параметрам распределения μ_1 и μ_2, что позволяет нам гипотезу H_0 записать как $H_0 : \mu_1 = \mu_2$, а три перечисленные выше альтерна-

тивы представить в следующем виде: $H_1 : \mu_1 < \mu_2$, $H_1 : \mu_1 > \mu_2$ и $H_1 : \mu_1 \neq \mu_2$. Рассмотрим общий случай, когда неизвестные дисперсии распределений σ_1^2 и σ_2^2 неравны. По выборочным данным X_1 и X_2 можно вычислить следующие статистические оценки:

$$\bar{x}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_{1\,i}, \quad \bar{x}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} x_{2\,i}, \tag{4.25}$$

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_{1\,i} - \bar{x}_1)^2, \tag{4.26}$$

$$s_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (x_{2\,i} - \bar{x}_2)^2. \tag{4.27}$$

Нетрудно видеть, что разность $\bar{x}_1 - \bar{x}_2$ будет представлять собой значение случайной величины, распределенной нормально со средним $\mu_1 - \mu_2$ и дисперсией $\sigma_1^2 / n_1 + \sigma_2^2 / n_2$ и, следовательно, число

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{4.28}$$

также будет значением нормально распределенной случайной величины, параметры которой при условии $\mu_1 - \mu_2$ будут 0 и 1. Таким образом, гипотезу $H_0 : \mu_1 = \mu_2$ следует принять, если вычисленное значение t не попадет в критическую область, соответствующую уровню значимости q и одной из трех упомянутых альтернатив. Гипотеза H_0 отклоняется при первой альтернативе и уровне значимости q, если $t < t_q$, при второй альтернативе, если $t > t_{1-q}$, и при третьей, если $|t| > t_{1-\frac{q}{2}}$. Значения t_q, соответствующие уровню значимости q, берутся из таблицы нормального распределения, приведенной в конце книги (см. прил. 1). Так, например, для уровня значимости 0,05 $t_{0,05} = -1{,}63$, $t_{0,95} = 1{,}63$ и $t_{0,975} = 1{,}96$.

Необходимо отметить, что весьма распространенная ошибка практических приложений критерия (4.28) в геологических исследованиях заключается в том, что распределение этого критерия в условиях H_0 принимают совпадающим с распределением Стьюдента. Это неверно, так как подкоренное выражение $s_1^2 / n_1 + s_2^2 / n_2$ не является величиной, распределенной как

$$\frac{\sigma^2 \chi^2}{\text{число степеней свободы}}.$$

Подкоренное выражение $s_1^2 / n_1 + s_2^2 / n_2$ вычислено в предположении, что дисперсии σ_1^2 и σ_2^2 неравны и применять этот критерий следует в условиях нормального распределения в тех случаях, когда суммарный объем выборок превышает 20, или же, когда

предварительная проверка гипотезы о равенстве неизвестных дисперсий σ_1^2 и σ_2^2 (см. раздел 4.3) покажет, что следует принять альтернативу $\sigma_1^2 \neq \sigma_2^2$.

Если же выборки, взятые из нормальных совокупностей, невелики, то прежде чем проверять гипотезу о равенстве средних значений, следует проверить гипотезу о равенстве дисперсий с помощью критерия Фишера, описанного в разделе 4.3. Если окажется что гипотезу о равенстве дисперсий следует отклонить, то для проверки гипотезы о равенстве средних нужно воспользоваться критерием (4.28). Если же гипотеза о равенстве дисперсий не противоречит выборочным данным, то для проверки гипотезы о равенстве средних можно воспользоваться следующим критерием:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{n_1 + n_2}{n_1 n_2}}}, \tag{4.29}$$

где

$$s^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}. \tag{4.30}$$

Если проверяемая гипотеза $H_0 : \mu_1 = \mu_2$ верна, то число t, определенное выражением (4.29), будет представлять собой значение случайной величины, распределенной по закону Стьюдента с $n_1 + n_2 - 2$ степенями свободы, и, следовательно, критические значения для принятия или отклонения проверяемой гипотезы при соответствующей альтернативе и заданном уровне значимости можно найти в таблице, приведенной в конце книги (см. прил. 6).

В условиях логнормального распределения проверка аналогичной гипотезы, т. е. гипотезы о равенстве математических ожиданий $M\xi_1$ и $M\xi_2$ двух логнормально распределенных случайных величин ξ_1 и ξ_2, может быть проведена следующим образом. Предварительно заметим, что гипотезе $H_0 : M\xi_1 = M\xi_2$ в данном случае равносильна гипотеза $H_0 : \mu_1 + \frac{1}{2}\sigma_1^2 = \mu_2 + \frac{1}{2}\sigma_2^2$, где $\mu_1 = M \ln \xi_1$, $\mu_2 = M \ln \xi_2$, $\sigma_1^2 = D \ln \xi_1$, $\sigma_2^2 = D \ln \xi_2$.

По выборкам, объем которых n_1 и n_2, получены следующие оценки этих параметров:

$$\overline{\ln x_1} = \frac{1}{n_1} \sum_{i=1}^{n_1} \ln x_{1i}, \tag{4.31}$$

$$\overline{\ln x_2} = \frac{1}{n_2} \sum_{i=1}^{n_2} \ln x_{2i}, \tag{4.32}$$

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (\ln x_{1i} - \overline{\ln x_1})^2, \tag{4.33}$$

$$s_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (\ln x_{2i} - \overline{\ln x_2})^2. \tag{4.34}$$

Используя тот факт, что оценки для сумм $\mu_1 + \frac{1}{2}\sigma_1^2$ и $\mu_2 + \frac{1}{2}\sigma_2^2$, которыми являются суммы $\overline{\ln x_1} + \frac{1}{2}s_1^2$ и $\overline{\ln x_2} + \frac{1}{2}s_2^2$, распределены приблизительно нормально со средними $\mu_1 + \frac{1}{2}\sigma_1^2$, $\mu_2 + \frac{1}{2}\sigma_2^2$ и дисперсиями $\sigma_1^2/n_1 + \sigma_1^4/2(n_1 - 1)$, $\sigma_2^2/n_2 + \sigma_2^4/2(n_2 - 1)$ соответственно, можно построить следующий критерий, предложенный автором ранее [37]:

$$t = \frac{\overline{\ln x_1} - \overline{\ln x_2} + \frac{1}{2}\left(s_1^2 - s_2^2\right)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} + \frac{1}{2}\left(\frac{s_1^4}{n_1 - 1} + \frac{s_2^4}{n_2 - 1}\right)}} \tag{4.35}$$

или, что удобнее на практике,

$$t = \frac{\overline{\lg x_1} - \overline{\lg x_2} + 1{,}153\left(s_1^2 - s_2^2\right)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} + 2{,}650\left(\frac{s_1^4}{n_1 - 1} + \frac{s_2^4}{n_2 - 1}\right)}}, \tag{4.36}$$

т. е. тот же критерий, только использующий оценки параметров распределения, вычисленные по десятичным логарифмам.

Если проверяемая гипотеза верна, то вычисленное значение критерия будет значением случайной величины, распределенной приблизительно нормально с математическим ожиданием, равным нулю, и дисперсией, равной 1. В связи с этим проверяемая гипотеза о равенстве средних значений двух логнормально распределенных случайных величин отклоняется, если вычисленное значение t попадет в критическую область, определяемую значением t_q, соответствующим выбранному уровню значимости q и выбранной альтернативе. Так, при $H_1 : M\xi_1 < M\xi_2$ гипотеза H_0 отклоняется, если $t < t_q$, при $H_1 : M\xi_1 > M\xi_2$ проверяемая гипотеза отклоняется, если $t > t_{1-q}$, и при альтернативе $H_1 : M\xi_1 \neq M\xi_2$ критическая область определяется неравенством $|t| > t_{1-\frac{q}{2}}$.

В практической работе, прежде чем приступить к проверке гипотезы $H_0 : M\xi_1 = M\xi_2$ с помощью описанного критерия, целесообразно провести проверку гипотезы $\sigma_1^2 = \sigma_2^2$ с помощью критерия Фишера, описанного в разделе 4.3. Если эта гипотеза будет принята, то дальнейшую проверку гипотезы $M\xi_1 = M\xi_2$ можно провести с помощью критерия Стьюдента, проверяя гипотезу $\mu_1 = \mu_2$, так как при условии $\sigma_1^2 = \sigma_2^2$ она равносильна гипотезе

$M\xi_1 = M\xi_2$. Критерий Стьюдента в данном случае выглядит следующим образом:

$$t = \frac{\overline{\lg x_1} - \overline{\lg x_2}}{s\sqrt{\frac{n_1 + n_2}{n_1 n_2}}}, \tag{4.37}$$

где

$$s = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}. \tag{4.38}$$

Правила применения этого критерия описаны выше.

Если же гипотеза $\sigma_1^2 = \sigma_2^2$ отклонена и принята альтернатива $\sigma_1^2 \neq \sigma_2^2$, то из этого следует, что в данных условиях гипотеза $\mu_1 = \mu_2$ не будет равносильной гипотезе $M\xi_1 = M\xi_2$. Однако при $\sigma_1^2 \neq \sigma_2^2$ равенство $\mu_1 + \frac{1}{2}\sigma_1^2 = \mu_2 + \frac{1}{2}\sigma_2^2$ может быть выполнено, и поэтому следует перейти к проверке гипотезы $M\xi_1 = M\xi_2$ с помощью критерия (4.36).

4.2.3. Проверка гипотезы о равенстве k неизвестных средних

В реальных геологических задачах нередко приходится сталкиваться с ситуациями, когда требуется высказать суждение о равенстве или различии более чем двух неизвестных средних значений. Иногда такую задачу пытаются решить путем множества попарных сравнений изучаемых групп наблюдений, но такой подход нельзя считать удовлетворительным для решения поставленной задачи. Таким образом, мы имеем дело с наиболее общим случаем проверки гипотез о равенстве средних, когда задача сводится к следующему. Пусть $\xi_1, \xi_2, \ldots, \xi_i, \ldots, \xi_k$ набор k независимых случайных величин и над каждой величиной с номером $i = 1, 2, \ldots, k$ проведено по n_i наблюдений, которые мы обозначим $x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{in_i}$. Обозначим также математические ожидания случайных величин $\xi_1, \xi_2, \ldots, \xi_i, \ldots, \xi_k$ через $\mu_1, \mu_2, \ldots, \mu_i, \ldots, \mu_k$, а дисперсии через $\sigma_1^2, \sigma_2^2, \ldots, \sigma_i^2, \ldots, \sigma_k^2$. Таким образом, наша проверяемая гипотеза будет иметь следующий вид: $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k = \mu_0$, а множество альтернатив представим как $H_1 : \mu_i \neq \mu_0$ хотя бы для одного $i = 1, 2, \ldots, k$.

В условиях нормального распределения при условии, что дисперсии σ_i^2 равны между собой, т. е. $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$, эту гипотезу можно проверить с помощью критерия, аналогичного критерию Стьюдента, который представляет собой набор k отношений

$$t_i = \frac{y_i\sqrt{n_i(N-2)}}{\sqrt{N - n_i - n_i y_i^2}}, \tag{4.39}$$

где

$$N = \sum_{i=1}^{k} n_i, \tag{4.40}$$

$$y_i = \frac{\bar{x}_i - \bar{x}}{s}, \tag{4.41}$$

где в свою очередь

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, \tag{4.42}$$

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{k} n_i \bar{x}_i = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}, \tag{4.43}$$

$$s^2 = \frac{1}{N-k} \sum_{i=1}^{k} (n_i - 1)\, s_i^2, \tag{4.44}$$

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2. \tag{4.45}$$

Величины $\bar{x}_i$ и s_i^2 представляют собой оценки среднего значения и дисперсии в группе наблюдений с номером i, а величины $\bar{x}$ и s^2 — обобщенные оценки среднего и дисперсии, вычисленные в предположении, что все k средних и все k дисперсий равны.

Естественно, что прежде чем применить этот критерий для проверки гипотезы $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k = \mu_0$, необходимо проверить гипотезу о равенстве дисперсий $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 = \sigma_0^2$ с помощью критериев, описанных в следующем разделе этой главы. Если гипотеза о равенстве дисперсий не отклоняется, это служит основанием для уверенного применения описанного критерия. Принятие или отклонение гипотезы $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k = \mu_0$ основано на том, что полученные числа t_i, при условии что проверяемая гипотеза верна, представляют собой значения случайных величин, распределенных по закону Стьюдента с $N-2$ степенями свободы. Таким образом, проверяемую гипотезу о равенстве средних следует отклонить, если хотя бы одно из вычисленных значений t_i превысит по абсолютной величине допустимое при уровне значимости q и $N-2$ степенях свободы. Проверяемая гипотеза принимается, если все значения t_i по абсолютной величине окажутся меньше критического.

Теперь рассмотрим случай, когда предварительная проверка гипотезы о равенстве дисперсий дала отрицательный результат, т. е. неизвестные значения k дисперсий нельзя признать равными. В подобной ситуации можно воспользоваться критерием

$$V = \sum_{i=1}^{k} \frac{(\bar{x}_i - \bar{x})^2\, n_i}{s_i^2}, \tag{4.46}$$

где $\bar{x} = \frac{1}{N} \sum_{i=1}^{k} n_i \bar{x}_i$, s_i^2 — оценка дисперсии в группе с номером i, x_i — соответствующее этой группе среднее арифметическое.

Если проверяемая гипотеза о равенстве средних верна, то числа

$$\frac{\bar{x}_i - \bar{x}}{s_i / \sqrt{n_i}} \tag{4.47}$$

будут представлять собой значения случайных величин, распределенных приблизительно нормально со средним значением, равным нулю, и дисперсией, равной 1. Следовательно, число V, определенное формулой (4.46), представляет собой сумму значений квадратов k независимых, одинаково нормально распределенных случайных величин с параметрами 0, 1, что позволяет нам рассматривать V как значение случайной величины, распределенной по закону χ^2 с $k-1$ степенями свободы. Число степеней свободы уменьшено на 1 в связи с тем, что в данном случае наложено одно ограничение $\mu_0 = \bar{x}$. Таким образом, проверяемая гипотеза о равенстве k средних отклоняется в том случае, если вычисленное значение V превысит допустимое $\chi^2_{q,\ k-1}$, соответствующее уровню значимости q и $k-1$ степени свободы. Если же $V \leqslant \chi^2_{q,\ k-1}$, гипотеза $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k = \mu_0$ принимается, как непротиворечащая выборочным данным. Следует также отметить, что этим критерием можно пользоваться и при условии, что предварительная проверка гипотезы о равенстве дисперсий в группах дала положительный результат, т. е. гипотеза была принята.

В условиях логнормального распределения гипотезу о равенстве математических ожиданий k случайных величин $\xi_1, \xi_2, \ldots, \xi_i, \ldots, \xi_k$ можно свести к равносильной гипотезе о равенстве математических ожиданий случайных величин $\ln \xi_1, \ln \xi_2, \ldots, \ln \xi_i, \ldots, \ln \xi_k$, т. е. $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k = \mu_0$ при условии, что дисперсии $\sigma_i^2 = D \ln \xi$ равны между собой. В этих условиях проверка гипотезы $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k = \mu_0$ при альтернативе $H_1 : \mu_i \neq \mu_0$ хотя бы для одного $i = 1, 2, \ldots, k$ будет равносильна проверке гипотезы $M \xi_1 = M \xi_2 = \ldots = M \xi_k = a_0$ и может быть проведена одним из описанных выше критериев.

Однако, если гипотеза о равенстве дисперсий логарифмов изучаемых случайных величин отклоняется, то дальнейшую проверку следует прекратить, так как не существует специального критерия для проверки гипотезы о равенстве k математических ожиданий логнормально распределенных случайных величин при условии неравенства дисперсий их логарифмов. В подобной ситуации следует воспользоваться более общим критерием, который будет рассмотрен ниже для случая, когда сведения о виде распределения ограничены.

4.2.4. Проверка гипотез о равенстве средних значений в условиях распределений, отличающихся от нормального

В геологической практике нередки случаи, когда распределения изучаемых геологических характеристик существенно отличаются как от нормального, так и логнормального закона, или же сведения об этих распределениях столь ограничены, что не позволяют сделать какие-либо заключения о виде функции распределения. Однако несмотря на столь неблагоприятные условия, геолог все же вынужден принимать решения на основе сравнения средних значений, и задача заключается в том, чтобы выбрать такие статистические методы, которые бы позволили даже в таких неблагоприятных ситуациях делать статистически обоснованные выводы.

Весьма распространенное заблуждение среди геологов-практиков заключается в том, что если распределения изучаемых случайных величин отличаются от нормального, то выводы, базирующиеся на сравнении средних значений, необоснованы. Другим типом ошибки является применение статистических критериев, требующих строгого выполнения ряда ограничений в ситуациях, когда эти ограничения не выполняются. Примером может служить применение критерия Стьюдента в условиях распределений, отличающихся от нормального, или же когда оценки дисперсий сравниваемых совокупностей существенно отличаются.

Для того чтобы сделать какие-либо рекомендации по применению статистических методов в описанных выше условиях, необходимо еще раз вернуться к вопросу о риске, связанном с принятием статистических решений. Как уже было показано в главе 3, в результате статистической проверки гипотезы может возникнуть два вида ошибок — первого и второго рода. Вероятность появления ошибки первого рода, заключающейся в ложном отклонении проверяемой гипотезы, когда она в действительности верна, представляет собой уровень значимости критерия. Вероятность не допустить ошибку второго рода, заключающуюся в ошибочном принятии проверяемой гипотезы (тогда как она неверна, а верна альтернатива), представляет собой мощность критерия, которая характеризует чувствительность критерия к альтернативе.

Б. Л. Ван дер Варденом [11] было показано, что отклонение распределения от нормального закона мало влияет на уровень значимости критерия, приспособленного для нормального распределения, и если заданный уровень значимости равен q, то в условиях распределения, отличающегося от нормального, он не превзойдет $2q$. Таким образом, если q выбрать достаточно малым, вероятность появления ошибки первого рода будет соответственно невелика.

Совершенно иначе обстоит дело с мощностью критериев, предназначенных для нормального распределения, в условиях отклонения распределений изучаемых случайных величин от нормального. В этих условиях мощность критерия может существенно умень-

шаться, что приводит к увеличению вероятности появления ошибки второго рода, иными словами, к уменьшению чувствительности критерия по отношению к альтернативе.

В связи с этим, если в силу необходимости приходится использовать критерий, соответствующий нормальному распределению, в описанных выше условиях, да еще и при малых выборках, то следует обязательно учитывать риск, связанный с принятием ошибочных решений. Если в подобной ситуации проверяемая гипотеза отклоняется, такой вывод можно признать достаточно обоснованным, так как вероятность появления ошибки первого рода, которой в этом случае следует опасаться, мала и не превосходит $2q$. Отклонение проверяемой гипотезы в подобной ситуации свидетельствует о том, что несмотря на плохие условия, критерий все же оказался чувствительным к альтернативе, и ее можно уверенно принять.

Серьезные сомнения в выводе могут возникнуть тогда, когда в описанных условиях, особенно при малых выборках, проверяемая гипотеза не отклоняется. В данном случае следует помнить, что вероятность появления ошибки второго рода может оказаться весьма большой, и в подобной ситуации целесообразно обратить внимание на последствия возможного ошибочного принятия проверяемой гипотезы. Если эти последствия существенны, то в подобной ситуации целесообразно увеличить число наблюдений для получения более обоснованного решения.

Естественно, что для проверки гипотез в упомянутых выше осложненных условиях целесообразно выбирать такие критерии, на которые мало влияет отклонение распределений от нормального. В частности, для проверки гипотезы о равенстве двух неизвестных средних значений μ_1 и μ_2 случайных величин ξ_1 и ξ_2, о которых известно только, что они имеют конечные математические ожидания и дисперсии, по выборкам $x_{11}, x_{12}, \ldots, x_{1i}, \ldots, x_{1n_1}, x_{21}, x_{22}, \ldots, x_{2n_2}$ объема n_1 и n_2, удобно пользоваться критерием, описанным в разделе 4.2.2. Этот критерий определяется выражением

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \tag{4.48}$$

где $\bar{x}_1$ и $\bar{x}_2$ — средние арифметические, вычисленные по данным соответствующих выборок, а s_1^2 и s_2^2 — оценки дисперсий. Несмотря на то что распределения ξ_1 и ξ_2 отличаются от нормального, средние арифметические $\bar{x}_1$ и $\bar{x}_2$ при достаточно больших значениях n_1 и n_2 можно рассматривать как значения нормально распределенных случайных величин с математическими ожиданиями μ_1 и μ_2. Таким образом, если гипотеза $H_0 : \mu_1 = \mu_2$ верна, то, число t будет представлять собой значение нормально распределенной слу-

чайной величины со средним 0 и дисперсией, равной 1, что позволяет, пользуясь таблицей нормальной функции, выбирать критическое значение t_q, соответствующее уровню значимости q, при той или иной альтернативе.

Точно так же для случая k выборок, взятых из совокупностей отличающихся от нормальных, для проверки гипотезы $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k = \mu_0$ о равенстве k средних значений, можно воспользоваться критерием, описанным в разделе 4.2.3, который определяется выражением

$$V = \sum_{i=1}^{k} \frac{(\bar{x}_i - \bar{x})^2 n_i}{s_i^2}, \tag{4.49}$$

где $\bar{x}_i$ — среднее арифметическое в выборке с номером i, $\bar{x}$ — среднее арифметическое, вычисленное в предположении, что проверяемая гипотеза верна, s_i^2 — оценка дисперсии в выборке с номером i. Несмотря на то что распределения в выборках предполагаются отличающимися от нормального, средние арифметические $\bar{x}_i$ можно рассматривать как значения нормально распределенных случайных величин, что в условиях проверяемой гипотезы о равенстве средних приводит к χ^2-распределению критерия с k—1 степенями свободы. Таким образом, гипотеза $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k = \mu_0$ принимается, если $V \leqslant \chi^2_{q,\ k-1}$, и отклоняется, если $V > \chi^2_{q,\ k-1}$, где $\chi^2_{q,\ k-1}$ — критическое значение, соответствующее уровню значимости q и k—1 степенями свободы.

Если в распоряжении исследователя нет таблиц χ^2-распределения, то можно воспользоваться тем, что при не очень малых k распределение χ^2 близко к нормальному со средним значением, равным числу степеней свободы, и дисперсией, равной удвоенному числу степеней свободы. Тогда величина

$$t = \frac{V - (k-1)}{\sqrt{2(k-1)}} \tag{4.50}$$

при условии, что проверяемая гипотеза верна, будет значением нормально распределенной случайной величины со средним, равным 0, и дисперсией, равной 1, и проверяемую гипотезу можно отклонить, если $t > t_{1-q}$. Следует подчеркнуть, что в данном случае критерий односторонний.

4.3. ПРОВЕРКА ГИПОТЕЗ О ДИСПЕРСИЯХ

Такая характеристика, как дисперсия, исключительно важна в геологических исследованиях, так как будучи мерой рассеяния результатов наблюдений относительно среднего значения, она может использоваться для описания изменчивости свойств геологических объектов. Наряду со средними значениями, такие меры из-

менчивости могут служить предметом сравнения, результаты которого используются для обоснования геологических выводов.

Кроме того, в разделе 4.2 мы уже неоднократно сталкивались с необходимостью проверки гипотез о равенстве двух или более дисперсий с целью определения условий, в которых будет применяться критерий для проверки гипотез о равенстве средних. Так, например, применение критерия Стьюдента предполагает, что гипотеза $H_0 : \sigma_1^2 = \sigma_2^2$ уже проверена и не противоречит выборочным данным.

4.3.1. Проверка гипотезы о равенстве двух дисперсий

Формально задача проверки гипотезы о равенстве двух неизвестных дисперсий ставится следующим образом. Пусть ξ_1 и ξ_2 — две независимые, нормально распределенные случайные величины, имеющие дисперсии σ_1^2 и σ_2^2. Над случайными величинами ξ_1 и ξ_2 проведено по n_1 и n_2 наблюдений $x_{11}, x_{12}, \ldots, x_{1i}, \ldots, x_{1n_1}$ и $x_{21}, x_{22}, \ldots, x_{2i}, \ldots, x_{2n_i}$. По этим данным требуется проверить гипотезу $H_0 : \sigma_1^2 = \sigma_2^2$ при альтернативе $H_1 : \sigma_1^2 > \sigma_2^2$.

В главе 2 было показано, что величины

$$\frac{1}{\sigma_1^2}\sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2 = \frac{s_1^2(n_1-1)}{\sigma_1^2} \tag{4.51}$$

и

$$\frac{1}{\sigma_2^2}\sum_{i=1}^{n_2}(x_{2i}-\bar{x}_2)^2 = \frac{s_2^2(n_2-1)}{\sigma_2^2} \tag{4.52}$$

являются значениями случайных величин, распределенных как χ^2 с n_1-1 и n_2-1 степенями свободы соответственно. Кроме того, было показано [6,11], что распределение отношения двух χ^2-распределенных случайных величин зависит только от n_1-1 и n_2-1 и называется распределением Фишера или F-распределением. Таким образом, отношение

$$F\binom{n_1-1}{n_2-1} = \frac{s_1^2\sigma_1^2}{s_2^2\sigma_2^2} \tag{4.53}$$

будет распределено по закону Фишера с n_1-1 и n_2-1 степенями свободы. Если гипотеза $H_0 : \sigma_1^2 = \sigma_2^2$ верна, то этому же закону будет подчинено и распределение отношения

$$F\binom{n_1-1}{n_2-1} = \frac{s_1^2}{s_2^2}, \tag{4.54}$$

что можно использовать в качестве критерия для проверки гипотезы $\sigma_1^2 = \sigma_2^2$. F-распределение табулировано (см. прил. 7 в конце книги), причем указанная таблица составлена для уровней значимости

0,05 и 0,025 при альтернативе $H_1: \sigma_1^2 > \sigma_2^2$. В связи с этим в формуле (4.54) в числитель всегда ставится большее значение оценки дисперсии, т. е. $s_1^2 > s_2^2$. Гипотеза H_0 принимается, если $F\binom{n_1-1}{n_2-1} < F_{(n_1-1,\ n_2-1),\ 1-q}$ и отклоняется, если $F\binom{n_1-1}{n_2-1} > F_{(n_1-1,\ n_2-1)\ 1-q}$.

4.3.2. Проверка гипотезы о равенстве более чем двух дисперсий

Пусть $\xi_1, \xi_2, \ldots, \xi_i, \ldots, \xi_k$ — набор k независимых, нормально распределенных случайных величин с неизвестными дисперсиями $\sigma_1^2, \sigma_2^2, \ldots, \sigma_i^2, \ldots, \sigma_k^2$. Над каждой величиной произведено n_i наблюдений $x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{in_i}$, по которым вычислены соответствующие оценки неизвестных параметров распределения $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_i, \ldots, \bar{x}_k$ и $s_1^2, s_2^2, \ldots, s_i^2, \ldots, s_k^2$.

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, \tag{4.55}$$

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2. \tag{4.56}$$

По этим данным требуется проверить гипотезу $H_0: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_i^2 = \ldots = \sigma_k^2 = \sigma_0^2$ при альтернативе $H_1: \sigma_i^2 \neq \sigma_0^2$, хотя бы для одного $i = 1, 2, \ldots, k$.

Описываемый ниже критерий, предназначенный для проверки этой гипотезы, был предложен Бартлетом в 1937 г. и поэтому называется критерием Бартлета. Этот исследователь показал, что если H_0 верна, то величина

$$B = \frac{1}{C}\left[\sum_{i=1}^{k} (n_i - 1) \ln s^2 - \sum_{i=1}^{k} (n_i - 1) \ln s_i^2\right] = \frac{2{,}3056}{C} \times$$

$$\times \left[\sum_{i=1}^{k} (n_i - 1) \lg s^2 - \sum_{i=1}^{k} (n_i - 1) \lg s_i^2\right] \tag{4.57}$$

будет распределена как χ^2 с $k-1$ степенями свободы. В выражении (4.57)

$$C = 1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k} \frac{1}{(n_i - 1)} - \frac{1}{\sum_{i=1}^{k} (n_i - 1)}\right], \tag{4.58}$$

$$s^2 = \frac{1}{N - 1} \sum_{i=1}^{k} (n_i - 1)\, s_i^2,$$

т. е. величина s^2 представляет собой обобщенную оценку диспер-

сии, вычисленную в предположении, что проверяемая гипотеза верна.

Таким образом, гипотеза H_0 принимается, если вычисленное значение B окажется меньше допустимого $\chi^2_{q,\,k-1}$ соответствующего уровню значимости q и k—1 степеням свободы. Наоборот, проверяемая гипотеза H_0 отклоняется и принимается альтернатива H_1, если $B > \chi^2_{q,\,k-1}$.

4.4. ПРОВЕРКА ГИПОТЕЗ О КОЭФФИЦИЕНТЕ КОРРЕЛЯЦИИ

Как было показано в главе 1 при рассмотрении двумерного нормального распределения, коэффициент корреляции ρ представляет собой меру линейной зависимости двух случайных величин ξ_1, ξ_2 и определяется выражением

$$\rho = \frac{M[(\xi_1 - M\xi_1)(\xi_2 - M\xi_2)]}{\sqrt{D\xi_1 D\xi_2}}. \tag{4.60}$$

Напомним, что в качестве оценки r для ρ обычно используется величина

$$r = \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\left[\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2 \sum_{i=1}^{n}(x_{2i} - \bar{x}_2)^2\right]^{\frac{1}{2}}}, \tag{4.61}$$

где x_{1i} и x_{2i} — наблюдения с номером i над случайными величинами ξ_1 и ξ_2 соответственно, $\bar{x}_1$ и $\bar{x}_2$ — средние арифметические, полученные по наблюдениям x_{1i} и x_{2i}, n — число наблюдений.

Естественно, что утверждение о наличии линейной зависимости двух изучаемых геологических характеристик или, наоборот, о ее отсутствии требует соответствующего статистического обоснования, задачу которого можно сформулировать в виде проверки гипотезы $H_0 : \rho = 0$, заключающейся в утверждении о равенстве нулю неизвестного значения коэффициента корреляции при одной из альтернатив $\rho < 0$, $\rho > 0$, $\rho \neq 0$.

Необходимо отметить, что распределение оценки коэффициента корреляции в условиях H_0 имеет довольно сложный вид, и его неудобно использовать для построения критерия проверяющего гипотезу $H_0 : \rho = 0$. С этой целью удобно пользоваться выражением

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \tag{4.62}$$

которое, как было показано Р. Фишером [52], при условии, что H_0 верна, представляет собой значение случайной величины, распределенной по закону Стьюдента с n—2 степенями свободы. Таким

образом, гипотеза $H_0 : \rho = 0$ отклоняется, если вычисленное значение t превысит соответствующие критические значения $t_{1-q,\, n-2}$ и $t_{1-\frac{q}{2},\, n-2}$ для двух последних альтернатив или же окажется меньше $t_{q,\, n-2}$ для первой альтернативы. Критические значения, соответствующие заданному уровню значимости q и числу степеней свободы $n-2$, берутся из таблиц распределения Стьюдента.

Второй простой способ проверки гипотезы $H_0 : \rho = 0$, также предложенный Р. Фишером [52], основан на том, что отношение

$$z = \frac{1}{2} \ln \frac{1+r}{1-r} \tag{4.63}$$

представляет собой значение случайной величины, распределенной в условиях H_0 приблизительно нормально со средним значением, равным нулю, и дисперсией, равной $\frac{1}{n-3}$. Из этого следует, что число

$$t = z\sqrt{n-3} \tag{4.64}$$

при условии, что H_0 верна, будет значением нормальной случайной величины с параметрами 0 и 1.

Таким образом, соответствующие критические значения t_q, при заданном уровне значимости q можно взять из таблиц нормального распределения.

Аналогичный критерий можно построить и для проверки гипотезы о равенстве двух коэффициентов корреляции, что представляет собой задачу, нередко встречающуюся в геологической практике. Такой вопрос возникает при геологических исследованиях, когда требуется сделать выводы о том, одинакова ли зависимость между двумя изучаемыми геологическими характеристиками двух геологических объектов или же различна. Таким образом, требуется проверить гипотезу $H_0 : \rho_1 = \rho_2$, например, при альтернативе $H_1 : \rho_1 \neq \rho_2$. В нашем распоряжении имеются две двухмерные выборки, объемы которых n_1 и n_2, и по ним вычислены две оценки r_1 и r_2 неизвестных коэффициентов корреляции ρ_1 и ρ_2. Используя описанное выше z — преобразование Фишера, получим два числа z_1 и z_2, которые будут распределены приблизительно нормально со средними значениями ρ_1, ρ_2 и дисперсиями $\frac{1}{n_1-3}$, $\frac{1}{n_2-3}$. Теперь мы можем записать критерий для проверки гипотезы $H_0 : \rho_1 = \rho_2$, который был предложен К. Р. Рао [55]

$$t = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}}. \tag{4.65}$$

Если проверяемая гипотеза верна, то число t будет значением случайной величины, распределенной нормально с параметрами 0 и

1, и гипотезу $H_0 : \rho_1 = \rho_2$ следует отклонить и принять альтернативу $H_1 : \rho_1 \neq \rho_2$, если $|t| > t_{1-\frac{q}{2}}$, где $t_{1-\frac{q}{2}}$ можно взять из таблиц нормальной функции.

Вполне естественно возникает также вопрос о критерии для проверки гипотезы о равенстве k коэффициентов корреляции, т. е. $H_0 : \rho_1 = \rho_2 = \ldots = \rho_i = \ldots = \rho_k = \rho_0$, при альтернативе $H_1 : \rho_i \neq \rho_0$ при $k > 2$ хотя бы для одного $i = 1, 2, \ldots, k$.

В распоряжении исследователя для решения поставленной задачи имеется k выборок из двухмерных нормальных совокупностей, объемы которых $n_1, n_2, \ldots, n_i, \ldots, n_k$. По этим выборкам вычислены оценки $r_1, r_2, \ldots, r_i, \ldots, r_k$ для неизвестных коэффициентов корреляции $\rho_1, \rho_2, \ldots, \rho_i, \ldots, \rho_k$. Удобный критерий, предназначенный для проверки H_0 по перечисленным данным, был предложен К. Р. Рао [55]. Этот критерий определен выражением

$$V = \sum_{i=1}^{k} (n_i - 3)(\operatorname{arctg} r_i)^2 - \frac{\left[\sum_{i=1}^{k} (n_i - 3) \operatorname{arctg} r_i\right]^2}{\sum_{i=1}^{k} (n_i - 3)}. \qquad (4.66)$$

К. Р. Рао показал, что этот критерий при условии, что проверяемая гипотеза верна, распределен как χ^2 с $k-1$ степенями свободы и поэтому H_0 отклоняется в тех случаях, когда $V > \chi^2_{q,\,k-1}$, т. е. превышает допустимое значение при уровне значимости q и $k-1$ степенях свободы.

Необходимо отметить, что задачи, связанные с рассмотрением коэффициентов корреляции, весьма разнообразны в геологических исследованиях, и в данной книге нет возможности рассмотреть их все. Поэтому читателю можно рекомендовать работы В. Н. Бондаренко [7, 8], С. А. Айвазяна [1], А. Б. Вистелиуса [13], где эти вопросы рассмотрены достаточно подробно.

4.5. ПРИМЕРЫ

Переходя к примерам, необходимо отметить, что большинство описанных критериев для проверки гипотез, связанных со средними значениями, дисперсиями и коэффициентами корреляции столь просты, что нет надобности в их специальной иллюстрации. Поэтому мы остановимся только на наиболее сложных критериях, а именно на предназначенных для проверки гипотезы о равенстве набора дисперсий, о равенстве нескольких средних и некоторых других. Кроме того, иногда нет надобности отдельно рассматривать нормальное и логнормальное распределение, так как в случае последнего те же действия, что и в условиях нормального распределения производятся с логарифмами.

Пример 1. В качестве примера, включающего проверку нескольких гипотез, мы воспользуемся результатами изучения со-

держаний индия в халькопиритах из различных типов месторождений [22].

Изучаемый халькопирит был представлен образцами из пяти генетических типов месторождений, которые перечислены в табл. 13. При этом наименьшим числом анализов n_i были охарактеризованы халькопириты из медно-никелевых и медно-молибденовых месторождений (7 и 10 соответственно). Однако для этих типов месторождений по данным анализа образцов, взятых из печенгских, мончегорских и норильских медно-никелевых руд и из армянских и среднеазиатских медно-молибденовых, характерен наименьший разброс содержаний (от 0,5 до 10 и от 1,3 до 40 г/т).

По халькопиритам касситерит-силикат-сульфидных месторождений материал был более представительным при колебаниях содержаний от 6 до 1500 г/т. Проанализированные образцы были отобраны из руд месторождений Якутии, Северо-Востока, Дальнего Востока и Восточного Забайкалья. Образцы из руд свинцово-цинковых месторождений были взяты из месторождений Центрального Казахстана, Средней Азии, Мексики, ГДР и ФРГ. Интервал колебаний содержаний индия в этих халькопиритах от 0,04 до 180 г/т.

Наибольшим числом анализов оказались охарактеризованы колчеданные месторождения медно-цинкового и свинцово-цинкового рядов. Первые из них представлены главным образом уральскими и в меньшей степени северокавказскими и алтайскими халькопиритами, а вторые в основном образцами из алтайских месторождений.

Проверка гипотезы о нормальном распределении логарифмов содержаний индия в халькопиритах с помощью отношений оценок асимметрии и эксцесса к их стандартным отлонениям показала, что логнормальную функцию можно использовать в качестве модели реальных распределений содержаний. Результаты этой проверки приведены в табл. 13.

Таблица 13

Оценки параметров логнормального распределения содержаний индия в халькопиритах

Тип месторождения	n_i	$\overline{\lg x_i}$	s_i^2	$\frac{\lvert\hat{\gamma}_1\rvert}{\sigma_{\hat{\gamma}_1}}$	$\frac{\lvert\hat{\gamma}_2\rvert}{\sigma_{\gamma_2}}$
1. Медно-никелевые	7	0,123	0,201	0,15	0,43
2. Медно-цинково-колчеданные, колчеданно-свинцово-цинковые, медно-цинково-пирротиновые	40	0,871	0,350	0,38	0,62
3. Медно-молибденовые	10	0,702	0,314	0,64	0,85
4. Касситерит-силикат-сульфидные	35	2,405	0,224	2,43	2,96
5. Свинцово-цинковые	21	0,689	0,844	1,15	0,06

Таким образом, гипотезу о равенстве набора средних значений содержания индия в халькопиритах можно представить как последовательную проверку двух гипотез, из которых первая является утверждением о равенстве дисперсий логарифмов содержаний при альтернативе, что хотя бы одна дисперсия выпадает из однородного ряда. Вторая гипотеза, при условии, что дисперсии логарифмов равны, заключается в утверждении о равенстве средних значений логарифмов.

По данным, приведенным в табл. 13, с помощью описанного выше критерия Бартлета была проведена проверка гипотезы о равенстве дисперсий логарифмов содержаний индия для всех пяти изучаемых типов месторождений. В результате было получено значение $B = 14{,}48$, что превышает допустимое значение χ^2, равное 13,3, которое соответствует уровню значимости 0,01 и четырем степеням свободы. Таким образом, проверяемую гипотезу о равенстве дисперсий следует отклонить.

Однако после того, как из набора дисперсий была исключена дисперсия пятого, свинцово-цинкового типа, как максимально отличающаяся от обобщенной оценки дисперсии, а для оставшихся четырех дисперсий повторена проверка нулевой гипотезы с помощью критерия Бартлета, оказалось что вычисленное значение $B = 3{,}69$, тогда как допустимое значение при уровне значимости 0,01 и трех степенях свободы равно 11,3. Таким образом, гипотеза о равенстве дисперсий логарифмов содержаний индия в халькопиритах первых четырех типов месторождений не противоречит выборочным данным и ее можно принять. Следовательно, для этих четырех типов месторождений можно провести проверку гипотезы о равенстве средних значений логарифмов с помощью критерия (4.39), аналогичного критерию Стьюдента. Эта проверка (вычисления мы предоставляем провести читателю) показала, что нулевая гипотеза не отклоняется только для второго и третьего типов месторождений, тогда как для первого и четвертого типов средние значения логарифмов существенно отличаются друг от друга, а также от среднего третьего и четвертого типов.

Из этого, в свою очередь, следует (при однородности ряда статистических оценок дисперсий логарифмов), что халькопириты из руд первого и четвертого типов обладают своими, присущими только каждому из них, средними содержаниями индия. Однако проверка гипотезы о равенстве средних значений содержания в условиях неравных дисперсий логарифмов, для второго, третьего и пятого типов с помощью критерия (4.28) показала, что расхождение между статистическими ожиданиями этих средних следует рассматривать как несущественное, так как вычисленные значения критерия $t = 0{,}34$ (второй и пятый) и $t = 1{,}38$ (третий и пятый) значительно меньше, чем допустимое 1,96 при уровне значимости 0,05.

Таким образом, проведенные исследования показали, что средние значения содержаний индия можно рассматривать как равные для второго, третьего и пятого типов месторождений, тогда как от

них существенно отличаются по средним содержаниям халькопириты первого и четвертого типов. Соответствующие максимально правдоподобные оценки (a) средних содержаний, вычисленные по формуле (2.22), приведены в табл. 14.

Таблица 14

Максимально правдоподобные оценки (a) средних содержаний индия в халькопиритах

Номер группы	Номер типа	Оценка, г/т	Точность ($\pm\lambda$) при надежности 0,95
1	1 2	2	0,72
2	3 5	22	4,00
3	4	445	17,00

Заметим, что значение точности (λ) при надежности 0,95 оценивалось по формуле

$$\lambda = \pm \frac{1{,}96a}{\sqrt{n}} \sqrt{s^2 - \frac{s^4}{2}},$$

где s^2 — оценка дисперсии логарифмов.

Пример 2. Для иллюстрации применения критерия (4.66), предназначенного для проверки гипотезы о равенстве k коэффициентов корреляции, приведем следующий пример, заимствованный из книги Р. Миллера и Дж. Кана [32].

Из коллекции Spirifer pennatus были взяты наудачу четыре выборки объемом 50, 50, 30 и 25 особей, на которых измерялись значения двух признаков. Для этих признаков были вычислены четыре оценки коэффициента корреляции соответственно каждой выборке. Эти данные приведены в табл. 15.

Таблица 15

Проверка гипотезы о равенстве коэффициентов корреляции

Номер выборки	n_i	r_i	arctg r_i	$(n_i - 3)$ arctg r_i	$(n_i - 3)$ (arctg $r_i)^2$
1	50	0,364	0,38	17,86	6,7868
2	50	0,312	0,32	15,04	4,8128
3	30	0,405	0,43	11,61	4,9923
4	25	0,327	0,34	7,48	2,5432
				51,99	19,1351

Возникает вопрос, можно ли рассматривать набор полученных оценок коэффициентов корреляции как однородный, или же, нао-

борот, различия между ними столь велики, что они превышают выборочные погрешности в одной популяции.

В табл. 15 приведены промежуточные результаты вычислений критерия V, определенного равенством (4.66). Окончательное его значение

$$V = 19{,}1351 - \frac{2702{,}96}{143} = 0{,}2333,$$

что значительно меньше критического значения χ^2, которое при уровне значимости 0,05 и трех степенях свободы равно 7,81. Из этого следует, что гипотезу о равенстве неизвестных значений четырех коэффициентов корреляции следует принять, как не противоречащую выборочным данным.

НЕКОТОРЫЕ МНОГОМЕРНЫЕ ГИПОТЕЗЫ

В главе 1 мы уже сталкивались с понятием многомерных данных и распределений многомерных случайных величин, которые, как было отмечено, играют в геологии исключительно важную роль. Дело в том, что большинство геологических задач требует совместного рассмотрения комплекса характеристик изучаемых объектов с последующими заключениями, ложащимися в основу геологических выводов. Как и в одномерных случаях, рассмотренных в предыдущей главе, в распоряжении исследователя обычно имеются выборочные многомерные данные, на основании которых принимаются те или иные решения, сопровождающиеся риском, определяемым как ошибки первого и второго рода. Так, например, многомерным обобщением гипотезы о равенстве средних значений является гипотеза о равенстве многомерных средних, проверка которой лежит в основе последующих утверждений о сходстве или различии целого комплекса изучаемых геологических характеристик.

В этой главе мы рассмотрим методы решения наиболее распространенных многомерных задач, таких, как проверка гипотез о равенстве многомерных средних и о равенстве ковариационных матриц, являющихся многомерными аналогами дисперсий. Однако прежде чем перейти к рассмотрению этих методов, мы кратко остановимся на некоторых понятиях матричной алгебры, которые нам понадобятся в дальнейшем.

5.1. МАТРИЦЫ И ВЕКТОРЫ

Матрицей порядка $n \times m$ называется прямоугольная таблица, состоящая из n строк и m столбцов. Так, например, таблица

$$\begin{vmatrix} 4 & 8 & 3 \\ 1 & 7 & 5 \\ 0 & \lg 3 & \operatorname{arctg} 2 \\ 5 & 6 & 7 \end{vmatrix}$$

будет матрицей порядка 4×3.

Обычно принято обозначать матрицы большими жирными буквами, а их элементы малыми с нижними индексами, указывающими

номер строки и номер столбца. Например,

$$\mathbf{A}=\begin{vmatrix} a_{11} & a_{12} & \dots & a_{1j} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2j} & \dots & a_{2m} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{t1} & a_{t2} & \dots & a_{tj} & \dots & a_{tm} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nj} & \dots & a_{nm} \end{vmatrix} \qquad (5.1)$$

будет матрицей порядка $n \times m$, где a_{tj} — элемент этой матрицы, стоящий в строке с номером t и столбце с номером j, причем $t = 1, 2, \dots, n$, $j = 1, 2, \dots, m$. Матрица, число строк которой равно числу столбцов, называется квадратной. Например,

$$\mathbf{A}=\begin{pmatrix} 3 & 0 & -1 \\ 4 & 5 & 2 \\ 7 & 1 & 3 \end{pmatrix}.$$

Заметим, что элементы a_{ii} такой матрицы, т. е. те которые лежат на главной диагонали, называются диагональными элементами.

Квадратная матрица, для всех элементов которой $a_{tj} = a_{jt}$ при $t \neq j$, называется симметричной. Иными словами, элементы этой матрицы симметричны относительно главной диагонали. Например,

$$\mathbf{A}=\begin{pmatrix} 6 & -2 & 3 \\ -2 & 5 & 0 \\ 3 & 0 & 4 \end{pmatrix}.$$

Квадратная матрица, все элементы которой, не лежащие на главной диагонали, равны нулю, называется диагональной матрицей. Примером такой матрицы может служить

$$\mathbf{A}=\begin{vmatrix} 3 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 4 \end{vmatrix}.$$

Диагональная матрица, все диагональные элементы которой равны 1, называется единичной матрицей. Обычно ее принято обозначать буквой **I**. Например,

$$\mathbf{I}=\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Эта матрица имеет очень большое значение в матричной алгебре и обычно играет роль 1.

И, наконец, введем еще одно определение нулевой матрицы, все элементы которой равны нулю. Эта матрица очень часто играет ту же роль, что и 0 при действиях с обычными числами.

Матрица порядка $n \times 1$ или $1 \times n$, т. е. состоящая только из одного столбца или одной строки, называется вектором-столбцом или вектором-строкой соответственно. Примерами могут служить

$$X = \begin{vmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{vmatrix}, \quad Y = \{y_1,\ y_2,\ y_3,\ y_4,\ y_5\} \tag{5.2}$$

Две матрицы $\mathbf{A} = \{a_{ij}\}$ и $\mathbf{B} = \{B_{ij}\}$ равны, если они имеют один и тот же порядок $n \times m$ и если $a_{ij} = b_{ij}$ для всех $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$, т. е. если все элементы одной матрицы равны соответствующим элементам другой матрицы. Так, например, матрицы

$$\mathrm{A} = \begin{pmatrix} 5 & 7 \\ 4 & 3 \end{pmatrix} \text{ и } \mathbf{B} = \begin{pmatrix} 5 & 7 \\ 4 & 3 \end{pmatrix}$$

равны, тогда как обе эти матрицы не равны матрице

$$\mathbf{C} = \begin{pmatrix} 5 & 7 & 0 \\ 4 & 3 & 0 \end{pmatrix},$$

как имеющей другой порядок, и не равны матрице

$$\mathbf{D} = \begin{pmatrix} 5 & 7 \\ 3 & 3 \end{pmatrix},$$

так как один элемент в них не совпадает.

Если строки матрицы порядка $n \times m$ преобразовать в столбцы, то получим новую матрицу порядка $m \times n$, которая называется транспозицией первой матрицы. Транспонированная матрица обычно обозначается $\mathbf{A}'$, где $\mathbf{A}$ исходная матрица. Так, например,

$$\mathbf{A} = \begin{pmatrix} 3 & 7 \\ 2 & 4 \\ 2 & 6 \end{pmatrix} \qquad \mathbf{A}' = \begin{pmatrix} 3 & 2 & 2 \\ 7 & 4 & 6 \end{pmatrix}.$$

Естественно, что транспонированный вектор-столбец — это вектор-строка, и наоборот. Например,

$$X = \begin{vmatrix} 1 \\ 0 \\ 3 \\ 4 \end{vmatrix} \quad x' = \{1\ 0\ 3\ 4\}.$$

Операцию сложения (вычитания) можно определить только для матриц, имеющих один и тот же порядок. Так, если **A** и **B** матрицы порядка $n \times m$, то сложив их, получим новую матрицу **D** того же порядка, т. е.

$$\mathbf{A}+\mathbf{B}=\mathbf{D}, \tag{5.3}$$

где каждый элемент d_{ij} матрицы **D** получается из соответствующих элементов a_{ij}, b_{ij} матриц **A** и **B**, т. е.

$$d_{ij}=a_{ij}+b_{ij}. \tag{5.4}$$

Заметим, что $\mathbf{A}+\mathbf{B}=\mathbf{B}+\mathbf{A}$. Все это относится и к вычитанию матриц. Естественно, что сложение и вычитание векторов представляет собой частный случай для матриц порядка $n \times 1$ или $1 \times n$.

Необходимо также отметить, что для двух матриц **A** и **B** одного порядка имеет место равенство

$$(\mathbf{A}+\mathbf{B})'=\mathbf{A}'+\mathbf{B}'. \tag{5.5}$$

Теперь мы определим операцию умножения матрицы на скаляр, предварительно заметив, что скаляр — это значение, выраженное одним действительным числом. В матричной алгебре скаляр противопоставляют вектору. Таким образом, любая матрица может быть умножена на скаляр, причем результат умножения не зависит от порядка сомножителей. Для того чтобы умножить матрицу $\mathbf{A}=\{a_{ij}\}$ на скаляр c, нужно каждый элемент этой матрицы a_{ij} умножить на c, т. е.

$$c\mathbf{A}=\mathbf{A}c=\{ca_{ij}\} \tag{5.6}$$

Например, $c = 15$,

$$\mathbf{A}=\begin{pmatrix} 2 & 1 \\ 3 & 1 \end{pmatrix}$$

$$c\mathbf{A}=\begin{pmatrix} 30 & 15 \\ 45 & 15 \end{pmatrix}.$$

Следует отметить, что операция деления матрицы на скаляр эквивалентна умножению этой матрицы на величину, обратную делителю.

Две матрицы **A** и **B** могут быть перемножены, т. е. **A** будет умножена на **B**, если порядок матрицы **A** будет $n \times m$, а матрицы **B** $m \times k$. Произведение **AB** будет иметь порядок $n \times k$. Это означает, что умножение двух матриц возможно только тогда, когда число столбцов матрицы **A**, стоящей слева, равно числу строк матрицы **B**, стоящей справа. Пусть

$$\mathbf{AB}=\mathbf{P}. \tag{5.7}$$

Процедура получения элементов p_{ij} матрицы **P**, порядок которой $n \times k$, определена через элементы a_{ij} и b_{ij} матриц **A** и **B** следующим выражением:

$$\mathbf{P}_{ij} = \sum_{s=1}^{m} a_{is} b_{sj}. \tag{5.8}$$

Иными словами, каждый элемент p_{ij} матрицы **P** вычисляется путем умножения строки с номером i матрицы **A** на столбец с номером j матрицы **B**, причем результат такого умножения представляет собой сумму произведений соответствующих пар элементов с номером $s = 1, 2, \ldots, m$. Так, например, перемножим вектор-строку $\{2, 4, 5, 7\}$ на вектор-столбец

$$\begin{vmatrix} 3 \\ 7 \\ 2 \\ 1 \end{vmatrix},$$

т. е.

$$(2,\ 4,\ 5,\ 7) \begin{vmatrix} 3 \\ 7 \\ 2 \\ -1 \end{vmatrix} = (2 \cdot 3) + (4 \cdot 7) + (5 \cdot 2) -$$

$$-(7 \cdot 1) = 6 + 28 + 10 - 7 = 37.$$

Рассмотрим пример умножения матриц

$$\mathbf{AB} = \begin{pmatrix} 3 & -1 \\ 0 & 4 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} 4 & 0 & 2 \\ 1 & 2 & 6 \end{pmatrix} = \begin{pmatrix} 11 & -2 & 0 \\ 4 & 8 & 24 \\ 14 & 12 & 40 \end{pmatrix} = \mathbf{P}.$$

Первый элемент p_{11} матрицы **P** получен следующим образом:

$$p_{11} = (3 - 1) \begin{pmatrix} 4 \\ 1 \end{pmatrix} = 12 - 1 = 11,$$

следующий

$$p_{12} = (3 - 1) \begin{pmatrix} 0 \\ 2 \end{pmatrix} = 0 - 2 = -2$$

и т. д.

Последний элемент

$$p_{33} = (2\ \ 6) \begin{pmatrix} 2 \\ 6 \end{pmatrix} = 4 + 36 = 40.$$

Степенью n квадратной матрицы $\mathbf{A}$ называется n-кратное умножение матрицы $\mathbf{A}$ саму на себя, т. е.

$$\mathbf{A} = \mathbf{A} \cdot \mathbf{A} \cdot \ldots \cdot \mathbf{A}. \tag{5.9}$$

Так, например, если $\mathbf{A} = \begin{pmatrix} 3 & 4 \\ 1 & 2 \end{pmatrix}$, то

$$\mathbf{A}^2 = \begin{pmatrix} 3 & 4 \\ 1 & 2 \end{pmatrix}^2 = \begin{pmatrix} 3 & 4 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 3 & 4 \\ 1 & 2 \end{pmatrix} =$$

$$= \begin{pmatrix} 3 \cdot 3 + 4 \cdot 1 & 3 \cdot 4 + 4 \cdot 2 \\ 1 \cdot 3 + 2 \cdot 1 & 1 \cdot 4 + 2 \cdot 2 \end{pmatrix} = \begin{pmatrix} 13 & 20 \\ 5 & 8 \end{pmatrix}.$$

Заметим также, что следом $\mathrm{tr}\,\mathbf{A}$ квадратной матрицы $\mathbf{A}$ называется сумма ее диагональных элементов, т. е.

$$\mathrm{tr}\,\mathbf{A} = \sum_{i=1}^{m} a_{ii}. \tag{5.10}$$

Например,

$$\mathrm{tr} \begin{pmatrix} 13 & 20 \\ 5 & 8 \end{pmatrix} = 13 + 8 = 21.$$

Одной из весьма сложных операций, которая тем не менее широко используется в прикладных задачах, связанных с матричной алгеброй, является обращение матрицы. Заметим, что эта операция определена только для квадратных матриц.

Пусть $\mathbf{A}$ квадратная матрица. Если произведение $\mathbf{AB} = \mathbf{I}$, т. е. равно единичной матрице, то матрица $\mathbf{B}$ называется матрицей, обратной по отношению к $\mathbf{A}$ и обозначается $\mathbf{A}^{-1}$. Следовательно, можно записать $\mathbf{AA}^{-1} = \mathbf{I}$, а также $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$.

Процедура нахождения обратной матрицы довольно сложна, и существуют различные методы ее реализации, например метод Дулитла, который подробно рассмотрен в книге К. Крамбейна и Ф. Грейбилла [27]. Здесь мы на этих процедурах останавливаться не будем, так как в настоящее время все ЭВМ снабжены стандартными программами реализации этой процедуры.

В дальнейшем рассмотрении нам потребуется такое понятие, как скалярное произведение c, которое представляет собой результат умножения вектора-строки X порядка $1 \times m$ на матрицу $\mathbf{S}$ порядка $m \times m$ и на вектор-столбец Y порядка $m \times 1$, т. е.

$$c = XSY = \{x_1, x_2, \ldots, x_m\} \begin{pmatrix} s_{11} & \ldots & s_{1m} \\ \ldots & \ldots & \ldots \\ s_{m1} & \ldots & s_{mm} \end{pmatrix} \begin{vmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_m \end{vmatrix}. \tag{5.11}$$

Например,

$$X = (2, \quad 4),$$

$$\mathbf{S} = \begin{pmatrix} 2 & 3 \\ 4 & 2 \end{pmatrix},$$

$$Y = \begin{pmatrix} 3 \\ 3 \end{pmatrix},$$

$$c = (2, 4) \begin{pmatrix} 2 & 3 \\ 4 & 2 \end{pmatrix} \begin{pmatrix} 3 \\ 3 \end{pmatrix} = 102.$$

Перемножив вектор-строку (2, 4) на матрицу $\begin{pmatrix} 2 & 3 \\ 4 & 2 \end{pmatrix}$, получим вектор-строку (20, 14). В результате умножения этого результата на $\begin{pmatrix} 3 \\ 3 \end{pmatrix}$ получим

$$(20, 14) \begin{pmatrix} 3 \\ 3 \end{pmatrix} = 60 + 42 = 102.$$

В ходе дальнейшего рассмотрения многомерных статистических критериев нам нередко придется сталкиваться с таким понятием, как детерминант, или определитель квадратной матрицы. Пусть

$$\mathbf{A} = \begin{vmatrix} a_{11} & a_{12} & \cdot & \cdot & \cdot & a_{1m} \\ a_{21} & a_{22} & \cdot & \cdot & \cdot & a_{2m} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{m1} & a_{m2} & \cdot & \cdot & \cdot & a_{mm} \end{vmatrix} \tag{5.12}$$

квадратная матрица порядка $m \times m$. Определителем матрицы $\mathbf{A}$, который мы будем обозначать $|\mathbf{A}|$, называется многочлен вида

$$\sum \pm a_{1\alpha} a_{2\beta} \cdot \quad \cdot a_{m\gamma}, \tag{5.13}$$

где $\alpha, \beta, \ldots, \gamma$ — есть произвольная перестановка чисел $1, 2, \ldots, m$. Суммирование ведется по всем перестановкам и поэтому определитель содержит $n!$ членов, из которых $\frac{1}{2} n!$ четные и $\frac{1}{2} n!$ нечетные. Примером простейшего определителя матрицы

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \tag{5.14}$$

будет число

$$|\mathbf{A}| = a_{11} \quad a_{22} - a_{21} a_{12}. \tag{5.15}$$

Современные ЭВМ, как правило, оснащены стандартными программами для вычисления определителей.

5.2. ПРОВЕРКА ГИПОТЕЗ О РАВЕНСТВЕ МНОГОМЕРНЫХ СРЕДНИХ

Постановка этой задачи по существу не отличается от рассмотренного нами в предыдущей главе одномерного варианта, за исключением того, что в данном случае нам приходится иметь дело с m-мерными данными. Для простоты сначала рассмотрим задачу проверки гипотезы о равенстве двух многомерных средних, а затем дадим ее обобщенное решение для случая k выборок. Конечно, нет необходимости еще раз подчеркивать значение этих задач в геологии, где, как правило, приходится принимать практические решения на основании сходства или различия комплекса признаков.

5.2.1. Проверка гипотезы о равенстве двух многомерных средних

Пусть $\Xi_1 = \{\xi_1', \xi_2', \ldots, \xi_j', \ldots, \xi_m'\}$ и $\Xi_2 = \{\xi_1'', \xi_2'', \ldots, \xi_j'', \ldots, \xi_m''\}$ две независимые m-мерные нормально распределенные случайные величины с математическими ожиданиями

$$M\Xi_1 = \underline{\mu}_1 = \{\mu_1', \mu_2', \ldots, \mu_j', \ldots, \mu_m'\}, \tag{5.16}$$

$$M\Xi_2 = \underline{\mu}_2 = \{\mu_1'', \mu_2'', \ldots, \mu_j'', \ldots, \mu_m''\} \tag{5.17}$$

и ковариационными матрицами

$$\Sigma_1 = \begin{vmatrix} \sigma_{21}' & \sigma_{12}' & \ldots & \sigma_{1j}' & \ldots & \sigma_{1m}' \\ \sigma_{21}' & \sigma_{22}' & \ldots & \sigma_{2j}' & \ldots & \sigma_{2m}' \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \sigma_{i1}' & \sigma_{i2}' & \ldots & \sigma_{ij}' & \ldots & \sigma_{im}' \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \sigma_{m1}' & \sigma_{m2}' & \ldots & \sigma_{mj}' & \ldots & \sigma_{mm}' \end{vmatrix}, \tag{5.18}$$

$$\Sigma_2 = \begin{vmatrix} \sigma_{11}'' & \sigma_{12}'' & \ldots & \sigma_{1j}'' & \ldots & \sigma_{1m}'' \\ \sigma_{21}'' & \sigma_{22}'' & \ldots & \sigma_{2j}'' & \ldots & \sigma_{2m}'' \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \sigma_{i1}'' & \sigma_{i2}'' & \ldots & \sigma_{ij}'' & \ldots & \sigma_{im}'' \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \sigma_{m}'' & \sigma_{m2}'' & \ldots & \sigma_{mj}'' & \ldots & \sigma_{mm}'' \end{vmatrix}. \tag{5.19}$$

Над каждой случайной величиной Ξ_1 и Ξ_2 проведено по n_1 и n_2 наблюдений $X_1, X_2, \ldots, X_t, \ldots, X_{n_1}$ и $Y_1, Y_2, \ldots, Y_t, \ldots, Y_{n_2}$, где

$$X_t = \{x_{t1}, x_{t2}, \ldots, x_{tj}, \ldots, x_{tm}\},$$
$$Y_t = \{y_{t1}, y_{t2}, \ldots, y_{tj}, \ldots, y_{tm}\}.$$

По этим данным требуется проверить гипотезу $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ при альтернативе $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$. Сначала мы рассмотрим критерий, предназначенный для проверки этой гипотезы, при условии, что $\Sigma_1 = \Sigma_2$. Естественно, что выполнение такого условия требует такой статистической проверки, соответствующие критерии которой будут рассмотрены в разделе 5.3.

С целью иллюстрации мы покажем как для проверки гипотезы $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ строится критерий отношения правдоподобия. Обозначим $L(X, Y, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma)$ функцию правдоподобия в условиях сделанных ограничений, которая будет иметь вид

$$L(X_1, X_2, \ldots, X_{n_1}, Y_1, Y_2, \ldots, Y_{n_2}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma) =$$
$$= (2\pi)^{-\frac{m(n_1+n_2)}{2}} |\Sigma|^{-\frac{(n_1+n_2)}{2}} \exp\left[-\frac{1}{2}\sum_{t=1}^{n_1}\{X_t-\boldsymbol{\mu}_1\}\times\right.$$
$$\left.\times\Sigma^{-1}\{X_t-\boldsymbol{\mu}_1\}' + \sum_{t=1}^{n_2}\{Y_t-\boldsymbol{\mu}_2\}\Sigma^{-1}\{Y_t-\boldsymbol{\mu}_2\}'\right], \qquad (5.20)$$

где $\Sigma = \Sigma_1 = \Sigma_2$, Σ^{-1} — матрица, обратная Σ.

Максимуму этой функции по $\boldsymbol{\mu}$ в условиях H_0 будет соответствовать выражение

$$\max_{\boldsymbol{\mu}} L(H_0) = (2\pi)^{-\frac{m(n_1+n_2)}{2}} |\mathbf{S}_0|^{-\frac{n_1+n_2}{2}} \exp\left[-\frac{1}{2}m(n_1+n_2)\right], \qquad (5.21)$$

где $\mathbf{S}_0$ — оценка для Σ, вычисленная при условии, что H_0 верна.

Максимум функции правдоподобия в условиях альтернативы определяется выражением

$$\max_{\boldsymbol{\mu}} L(H_1) = (2\pi)^{-\frac{m(n_1+n_2)}{2}} |\mathbf{S}|^{-\frac{n_1+n_2}{2}} \exp\left[-\frac{1}{2}m(n_1+n_2)\right], \qquad (5.22)$$

где $\mathbf{S}_1$ — оценка ковариационной матрицы Σ при условии, что верна альтернатива H_1.

В соответствии с работой Т. Андерсона [2], критерий для проверки гипотезы H_0, основанный на отношении правдоподобия, будет представлен выражением

$$\lambda = \frac{\max\limits_{\boldsymbol{\mu}} L(H_0)}{\max\limits_{\boldsymbol{\mu}} L(H_1)} = \left(\frac{|\mathbf{S}_1|}{|\mathbf{S}_0|}\right)^{\frac{n_1+n_2}{2}}. \qquad (5.23)$$

Однако при проверке гипотезы H_0 обычно не используется непосредственно отношение λ, а применяется число

$$V = -\left(n_1 + n_2 - 2 - \frac{m}{2}\right) \ln \frac{|S_1|}{|S_0|}, \tag{5.24}$$

которое, если H_0 верна, представляет собой значение случайной величины, распределенной как χ^2 с m степенями свободы. Таким образом, гипотеза H_0 отвергается, если вычисленное значение V превысит допустимое $\chi^2_{q,m}$, соответствующее уровню значимости q и m степеням свободы.

Необходимо отметить, что элементы s^0_{ij} и s^1_{ij} матриц S_0 и S_1 соответственно вычисляются по следующим формулам:

$$s^0_{ij} = \frac{1}{n_1 + n_2 - 2}\left[\sum_{t=1}^{n_1} x_{ti}x_{tj} + \sum_{t=1}^{n_2} y_{ti}y_{tj} - \right.$$

$$\left. - \frac{1}{n_1 + n_2}\left(\sum_{t=1}^{n_1} x_{ti} + \sum_{t=1}^{n_2} y_{ti}\right)\left(\sum_{t=1}^{n_1} x_{tj} + \sum_{t=1}^{n_2} y_{tj}\right)\right], \tag{5.25}$$

$$s^1_{ij} = \frac{1}{n_1 + n_2 - 2}\left[\sum_{t=1}^{n_1} x_{ti}x_{tj} + \sum_{t=1}^{n_2} y_{ti}y_{tj} - \right.$$

$$\left. - \frac{1}{n_1}\left(\sum_{t=1}^{n_1} x_{ti}\right)\left(\sum_{t=1}^{n_1} x_{tj}\right) - \frac{1}{n_2}\left(\sum_{t=1}^{n_2} y_{ti}\right)\left(\sum_{t=1}^{n_2} y_{tj}\right)\right]. \tag{5.26}$$

Нетрудно видеть, что элементы s^0_{ij} выборочной ковариационной матрицы S_0 вычисляются так, что обе выборки объединены в одну, что и учитывает предположение о равенстве средних. Элементы же s^0_{ij} матрицы S_1, наоборот, вычисляются так, что возможная разница между средними исключается, чем достигается соответствие этих оценок гипотезе H_1. Необходимо отметить, что для проверки гипотезы о равенстве двух многомерных средних существуют и другие критерии, например критерий T^2 Готелинга, описание которого можно найти в книгах Т. Андерсона [2], а также Р. Миллера и Дж. Кана [22].

5.2.2. Проверка гипотезы о равенстве k многомерных средних

Для проверки гипотезы о равенстве k m-мерных средних при условии, что все k ковариационных матриц изучаемых совокупностей равны между собой, можно воспользоваться методом, предложенным К. Р. Рао [55], известным под названием обобщенного дисперсионного анализа.

Формально данная задача выглядит следующим образом. Пусть $\Xi_1, \Xi_2, \ldots, \Xi_l, \ldots, \Xi_k$ — набор k независимых m-мерных нормально распределенных случайных величин с математическими ожиданиями $\mu_1, \mu_2, \ldots, \mu_k$ и ковариационными матрицами $\Sigma_1 =$

$= \Sigma_2 = \ldots = \Sigma_k = \Sigma_0$. По результатам наблюдений $X_{lt} = \{x_{lt1}, x_{lt2}, \ldots, x_{ltj}, \ldots, x_{ltm}\}$, где $l = 1, 2, \ldots, k$ (номер выборки), $t = 1, 2, \ldots, n_l$ (номер наблюдения в выборке с номером l) и $j = 1, 2, \ldots, m$ (номер признака) требуется проверить гипотезу

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_l = \ldots = \mu_k = \mu_0. \qquad (5.27)$$

при альтернативе $H_1 : \mu_l \neq \mu_0$ хотя бы для одного $l = 1, 2, \ldots, k$.

Не останавливаясь на процедуре вывода критерия, предложенного К. Р. Рао [55], рассмотрим конечный результат, который и используется в практических вычислениях

$$V = -\left(\sum_{l=1}^{k} n_l - 1 - \frac{m+k}{2}\right) \ln \frac{|\mathbf{S}_1|}{|\mathbf{S}_0|}, \qquad (5.29)$$

где $\mathbf{S}_0$ — выборочная ковариационная матрица, элементы которой s_{ij}^0 вычислены в предположении, что верна проверяемая гипотеза H_0, а $\mathbf{S}_1$ оценка ковариационной матрицы Σ_0, элементы которой s_{ij}^1 вычислены в предположении, что верна альтернатива H_1. $|\mathbf{S}_0|$ и $|\mathbf{S}_1|$ — детерминанты этих матриц. Матрицы $\mathbf{S}_0$ и $\mathbf{S}_1$ вычисляются по следующим формулам (X_{lt} — вектор-строка):

$$\mathbf{S}_0 = \frac{1}{N-1} \sum_{l=1}^{k} \sum_{t=1}^{n_l} \{X_{lt} - \overline{X}\}' \{X_{lt} - \overline{X}\}, \qquad (5.30)$$

где

$$N = \sum_{l=1}^{k} n_l; \qquad \overline{X} = \frac{1}{N} \sum_{l=1}^{k} \sum_{t=1}^{n_l} X_{lt} \qquad (5.31)$$

обобщенный вектор средних арифметических для всех k групп наблюдений

$$\mathbf{S}_1 = \frac{1}{N-1} \sum_{l=1}^{k} \sum_{t=1}^{n_l} \{X_{lt} - \overline{X}_l\}' \{X_{lt} - \overline{X}_l\}, \qquad (5.32)$$

где

$$\overline{X}_l = \frac{1}{n_l} \sum_{t=1}^{n_l} X_{lt} \qquad (5.33)$$

вектор средних арифметических для группы с номером l.

Если эту запись сделать более детальной, т. е. не в матричной форме, то можно показать, что элементы s_{ij}^0 и s_{ij}^1 выборочных ковариационных матриц $\mathbf{S}_1$ и $\mathbf{S}_0$ вычисляются по следующим формулам:

$$s_{ij}^0 = \frac{1}{N-1} \left[\sum_{l=1}^{k} \sum_{t=1}^{n_l} x_{lti} x_{ltj} - \frac{1}{N} \left(\sum_{l=1}^{k} \sum_{t=1}^{n_l} x_{lti} \right) \left(\sum_{l=1}^{k} \sum_{t=1}^{n_l} x_{ltj} \right) \right], \qquad (5.34)$$

$$s_{ij}^1 = \frac{1}{N-1}\left[\sum_{l=1}^{k}\sum_{t=1}^{n_l} x_{lti}x_{ltj} - \sum_{l=1}^{k}\frac{1}{n_l}\left(\sum_{t=1}^{n_l} x_{lti}\right)\left(\sum_{t=1}^{n_l} x_{ltj}\right)\right], \quad (5.35)$$

где $N = \sum_{l=1}^{k} n_l$, т. е. общее число наблюдений во всех k выборках.

Если проверяемая гипотеза верна, то вычисленное значение критерия V, определенного выражением (5.29), будет представлять собой значение случайной величины, распределенной как χ^2 с $m\,(k-1)$ степенями свободы. Таким образом, проверяемая гипотеза H_0, заключающаяся в утверждении о равенстве k m-мерных средних, отклоняется и принимается альтернатива $H_1 : \mu_l \neq \mu_0$, если $V > \chi^2_{q,\, m\,(k-1)}$ при уровне значимости q и $m\,(k-1)$ степенях свободы. Гипотеза H_0 принимается, как не противоречащая выборочным данным, если $V \leqslant \chi^2_{q,\, m\,(k-1)}$.

5.2.3. Проверка гипотезы о равенстве k m-мерных средних при неравных ковариационных матрицах

Требование равенства ковариационных матриц при реальных геологических исследованиях не всегда выполняется. В связи с эгим необходимо рассмотреть критерий, позволяющий проверять гипотезу о равенстве многомерных средних, при условии, что ковариационные матрицы неравны. В данном случае нет надобности рассматривать отдельно частный случай при $k = 2$, а следует перейти сразу к более общему для любого $k \geqslant 2$. Заметим также, что ниже мы рассмотрим критерий, известный под названием критерия Джеймса [28, 53].

Пусть, как и в предыдущем разделе, $\Xi_1, \Xi_2, \ldots, \Xi_l, \ldots, \Xi_k$ — набор m-мерных независимых нормально распределенных случайных величин, каждой из которых соответствует m-мерный вектор-строка средних значений $\mu_1, \mu_2, \ldots, \mu_l, \ldots, \mu_k$ и ковариационная матрица $\Sigma_1, \Sigma_2, \ldots, \Sigma_l, \ldots, \Sigma_k$.

Предполагается, что матрицы Σ_l неравны между собой. Над каждой случайной величиной Ξ_l проведено по n_l наблюдений $X_{lt} = \{x_{lt1}, x_{lt2}, \ldots, x_{ltj}, \ldots, x_{ltm}\}$, представляющих собой векторы-строки. По этим данным требуется проверить гипотезу

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k = \mu_0 \quad (5.36)$$

при альтернативе $H_1 : \mu_1 \neq \mu_0$ хотя бы для одного $l = 1, 2, \ldots, k$. Проверку требуется провести при условии, что ковариационные матрицы Σ_l неравны между собой.

Критерий, с помощью которого можно провести такую проверку, определяется выражением

$$V = \sum_{l=1}^{k} n_l \{\overline{X}_l - \overline{X}\}\, S_l^{-1} \{\overline{X}_l - \overline{X}\}', \quad (5.37)$$

где

$$\overline{X}_l = \frac{1}{n_l} \sum_{t=1}^{n_l} X_{lt}, \tag{5.38}$$

т. е. вектор-строка, образованная средними арифметическими в выборке с номером l, $\mathbf{S}_l^{-1}$ — матрица, обратная ковариационной матрице $\mathbf{S}_l$, соответствующей выборке с номером l,

$$\mathbf{S}_l = \frac{1}{n_l - 1} \sum_{t=1}^{n_l} \{X_{lt} - \overline{X}_l\}' \{X_{lt} - \overline{X}_l\}. \tag{5.39}$$

Обобщенное выборочное m-мерное среднее $\overline{X}$, в данной ситуации вычисляется по формуле

$$\overline{X} = \left(\sum_{l=1}^{k} n_l \mathbf{S}^{-1} \right)^{-1} \left(\sum_{l=1}^{k} n_l \mathbf{S}_l^{-1} \overline{X}_l' \right) \tag{5.40}$$

Если проверяемая гипотеза верна, то число V, определенное выражением (5.37), будет представлять собой значение случайной величины, распределенной асимптотически как χ^2 с $m\,(k-1)$ степенями свободы. Однако Г. Джеймс [53] показал, что распределение величины V несколько смещено и в качестве критического значения, соответствующего уровню значимости q и $m\,(k-1)$ степеням свободы, следует брать число

$$V_{q,\, m\,(k-1)} = \chi^2_{q,\, m\,(k-1)} \left(A + B\chi^2_{q,\, m\,(k-1)} \right), \tag{5.41}$$

где

$$A = 1 + \frac{1}{2m\,(k-1)} \sum_{l=1}^{k} \frac{1}{(n_l - 1)} \left\{ \mathrm{tr} \left[\mathbf{I}_m - \left(\sum_{l=1}^{k} n_l \mathbf{S}_l^{-1} \right)^{-1} n_l \mathbf{S}_l^{-1} \right] \right\}^2, \tag{5.42}$$

$$B = \frac{1}{m\,(k-1)\,[m\,(k-1)+2]} \left\{ \sum_{l=1}^{k} \frac{1}{n_l - 1} \mathrm{tr} \times \right.$$

$$\left. \times \left[\mathbf{I}_m - \left(\sum_{l=1}^{k} n_l \mathbf{S}_l^{-1} \right)^{-1} n_l \mathbf{S}_l^{-1} \right]^2 + (A-1)\, m\,(k-1) \right\}, \tag{5.43}$$

где $\mathbf{I}_m$ — единичная матрица порядка $m \times m$, tr — знак следа матрицы (см. раздел 5.1). Таким образом, проверяемая гипотеза отклоняется, если вычисленное значение V превысит допустимое при уровне значимости q и $m\,(k-1)$ степенях свободы, т. е.

$$V > \chi^2_{q,\, m\,(k-1)} \left(A + B\chi^2_{q,\, m\,(k-1)} \right), \tag{5.44}$$

и принимается, если это неравенство имеет обратный знак.

5.3. ПРОВЕРКА ГИПОТЕЗ О КОВАРИАЦИОННЫХ МАТРИЦАХ

В предыдущих разделах этой главы мы уже сталкивались с условием равенства ковариационных матриц, которое, естественно, требует соответствующей статистической проверки. Более того, в геологических исследованиях вопрос о равенстве или различии ковариационных матриц представляет собой самостоятельный интерес, так как ковариационные матрицы, помимо векторов средних значений, несут дополнительную информацию о сходстве или различии изучаемых геологических объектов, причем с учетом связей между отдельными характеристиками.

Не нарушая общности, сразу же рассмотрим случай k выборок, так как ситуация с двумя выборками представляет собой только частный вариант. Таким образом, наша задача сводится к следующему.

Пусть $\Xi_1, \Xi_2, \ldots, \Xi_l, \ldots, \Xi_k$ — набор k m-мерных случайных величин, которые нормально распределены с ковариационными матрицами $\Sigma_1, \Sigma_2, \ldots, \Sigma_l, \ldots, \Sigma_k$ соответственно. По результатам наблюдения $X_{lt} = \{x_{lt\,1}, x_{lt\,2}, \ldots, x_{ltj}, \ldots, x_{ltm}\}$, которые представляют собой m-мерные вектор-строки, требуется проверить гипотезу

$$H_0: \Sigma_1 = \Sigma_2 = \ldots = \Sigma_k = \Sigma_0 \tag{5.45}$$

при альтернативе $H_1: \Sigma_l \neq \Sigma_0$ хотя бы для одного $l = 1, 2, \ldots, k$. (5.46)

Следует отметить, что для проверки этой гипотезы существуют различные критерии, например описанные в книге Т. Андерсона [2]. Здесь мы рассмотрим только один из них, описанный С. Кульбаком [28]. Этот критерий определен выражением

$$V = \sum_{l=1}^{k} \frac{(n_l - 1)}{2} \ln \frac{|\mathbf{S}|}{|\mathbf{S}_l|}, \tag{5.47}$$

где

$$\mathbf{S}_l = \frac{1}{n_l - 1} \sum_{t=1}^{n_l} \{X_{lt} - \overline{X}_l\}' \{X_{lt} - \overline{X}_l\}, \tag{5.48}$$

$$\overline{X}_l = \frac{1}{n_l} \sum_{t=1}^{n_l} X_{lt}, \tag{5.49}$$

$$\mathbf{S} = \frac{1}{N - k} \sum_{l=1}^{k} (n_l - 1)\, \mathbf{S}_l, \tag{5.50}$$

$$N = \sum_{l=1}^{k} n_l. \tag{5.51}$$

Таким образом, $\mathbf{S}_l$ — оценка ковариационной матрицы в выборке с номером l, $\mathbf{S}$ — обобщенная выборочная матрица, вычисленная в предположении, что верна проверяемая гипотеза H_0. Если проверяемая гипотеза верна, то число V будет представлять собой значение случайной величины, распределенной асимптотически как χ^2 с $(k-1)\,m\,(m+1)/2$ степенями свободы. Однако лучшим приближением является нецентральное χ^2-распределение с параметром нецентральности [28]

$$\beta = \frac{2m^3 + 3m^2 - m}{12}\left(\sum_{l=1}^{k}\frac{1}{n_l - 1} - \frac{1}{N-m}\right) \qquad (5.52)$$

и $f = (k-1)\,m\,(m+1)/2$ степенями свободы. При числе степеней свободы, превышающем 7, это распределение следует рассматривать как χ^2 с $\frac{(k-1)\,m\,(m+1)}{2}$ степенями свободы. Таким образом, если при $f > 7$, $V > \chi^2_{q,f}$, то проверяемую гипотезу следует отклонить.

В том случае, когда $f \leqslant 7$, критические значения, соответствующие уровню значимости 0,05, можно взять из таблицы, приведенной в конце книги (см. прил. 8).

5.4. ПРИМЕРЫ

Пример 1. В качестве простейшего примера проверки гипотезы о равенстве многомерных средних воспользуемся результатами обоснования различий средних значений четырех измеряемых признаков двух близких видов фораминифер Valvulineria turcmenica и Valvulineria iphigenia по двум выборкам, состоящим из 15 наблюдений каждая. Эти данные опубликованы автором ранее [37]. В выборках у каждой особи измерялись следующие характеристики: 1) длина раковины, 2) ширина раковины, 3) длина последней камеры, 4) ширина последней камеры. Средние арифметические каждого признака (в мм) приведены ниже.

Признак	V. iphigenia	V. turcmenica
1	0,293	0,316
2	0,217	0,245
3	0,183	0,167
4	0,205	0,91

Статистические оценки ковариационных матриц $\mathbf{S}_0$ и $\mathbf{S}_1$, вычисленных в предположении, что верна гипотеза H_0 и H_1, соответственно умноженные на 1000, имеют вид

$$\mathbf{S}_0 = \begin{vmatrix} 4{,}6 & 3{,}2 & 2{,}4 & 2{,}6 \\ 3{,}2 & 2{,}7 & 1{,}6 & 1{,}8 \\ 2{,}4 & 1{,}6 & 1{,}7 & 1{,}5 \\ 2{,}6 & 1{,}8 & 1{,}5 & 1{,}8 \end{vmatrix},$$

$$S_1 = \begin{vmatrix} 4{,}6 & 3{,}1 & 2{,}5 & 2{,}6 \\ 3{,}1 & 2{,}5 & 1{,}7 & 1{,}8 \\ 2{,}5 & 1{,}7 & 1{,}6 & 1{,}4 \\ 2{,}6 & 1{,}8 & 1{,}4 & 1{,}8 \end{vmatrix}.$$

Детерминанты этих матриц соответственно равны

$$|S_0| = 0{,}271,$$

$$|S_1| = 0{,}147.$$

Подставив их значения в формулу (5.24), т. е. в

$$V = -\left(n_1 + n_2 - 2 - \frac{m}{2}\right) \ln \frac{|S_1|}{|S_0|},$$

получим

$$V = -\left(15 + 15 - 2 - \frac{4}{2}\right) \ln 0{,}542 = 15{,}18.$$

Допустимое значение χ^2, соответствующее уровню значимости 0,01 при 4 степенях свободы, $\chi^2_{0.01;\,4} = 13{,}28$. Так как вычисленное значение $V = 15{,}18$ значительно превышает 13,28, гипотезу о равенстве четырех мерных средних значений для этих двух видов фораминифер можно уверенно отклонить и различия между ними считать доказанными.

Пример 2. Для иллюстрации процедуры вычисления критерия Джеймса, определенного формулой (5.37), воспользуемся его собственными данными [53].

Даны три двухмерные выборки, объемы которых $n_1 = 16$, $n_2 = 11$ и $n_3 = 11$. По этим выборкам вычислены три вектора-строки средних арифметических

$$\overline{X}_1 = \{9{,}8 \quad 15{,}06\},$$

$$\overline{X}_2 = \{13{,}05 \quad 22{,}57\},$$

$$\overline{X}_3 = \{14{,}67 \quad 25{,}17\},$$

являющиеся оценками трех неизвестных двухмерных средних μ_1, μ_2, μ_3. В геологическом плане под этими средними значениями можно понимать, например, содержания двух акцессорных минералов, выраженных в граммах на тонну в трех сериях проб, взятых с одного и того же участка породы различными методами, и требуется выяснить, различаются ли эти методы по своим результатам или нет. Можно найти и другие геологические аналоги этой задачи. Иными словами, под имеющимися данными можно иметь в виду

самые различные геологические показатели, замеренные на трех изучаемых геологических объектах или тремя способами на одном объекте.

По этим же трем выборкам вычислены оценки $\mathbf{S}_1$, $\mathbf{S}_2$, $\mathbf{S}_3$ соответствующих неизвестных ковариационных матриц Σ_1, Σ_2, Σ_3

$$\mathbf{S}_1 = \begin{pmatrix} 120{,}0 & -16{,}3 \\ -16{,}3 & 17{,}8 \end{pmatrix},$$

$$\mathbf{S}_2 = \begin{pmatrix} 81{,}8 & 32{,}1 \\ 32{,}1 & 53{,}8 \end{pmatrix},$$

$$\mathbf{S}_3 = \begin{pmatrix} 100{,}3 & 23{,}2 \\ 23{,}2 & 97{,}1 \end{pmatrix}.$$

По имеющимся данным требуется проверить гипотезу H_0; $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3 = \boldsymbol{\mu}_0$ при альтернативе $H_1 : \boldsymbol{\mu}_l \neq \boldsymbol{\mu}_0$ хотя бы для одного $l = 1, 2, 3$.

Для того чтобы вычислить значение критерия

$$V = \sum_{l=1}^{k} n_l \{\overline{X}_l - \overline{X}\} \mathbf{S}_l^{-1} \{\overline{X}_l - \overline{X}\},$$

нам нужно вычислить обобщенный средний вектор-строку X, который определен выражением

$$\overline{X} = \left(\sum_{l=1}^{k} n_l \mathbf{S}^{-1} \right)^{-1} \left(\sum_{l=1}^{k} n_l \mathbf{S}_l^{-1} \overline{X}_l' \right),$$

что в свою очередь требует вычисления обратных матриц

$$(n_l \mathbf{S}_l^{-1}) \quad \text{и} \quad (n_l \mathbf{S}_l^{-1} \overline{X}_l').$$

Проведем эти вычисления:

$$(n_1 \mathbf{S}_1^{-1}) = \begin{pmatrix} 0{,}1523 & 0{,}1396 \\ 0{,}1396 & 1{,}0272 \end{pmatrix},$$

$$(n_2 \mathbf{S}_2^{-1}) = \begin{pmatrix} 0{,}1756 & -0{,}1048 \\ -0{,}1048 & 0{,}2670 \end{pmatrix},$$

$$(n_3 \mathbf{S}_3^{-1}) = \begin{pmatrix} 0{,}1161 & -0{,}0277 \\ -0{,}0277 & 0{,}1199 \end{pmatrix}.$$

Сложив эти матрицы, получим

$$\sum_{l=1}^{3} n_l \mathbf{S}_l^{-1} = \begin{pmatrix} 0{,}4440 & 0{,}0071 \\ 0{,}0071 & 1{,}4141 \end{pmatrix}.$$

Далее находим матрицу, обратную только что полученной

$$\left(\sum_{l=1}^{3} n_l \mathbf{S}_l^{-1}\right)^{-1} = \begin{pmatrix} 2{,}2524 & -0{,}0113 \\ -0{,}0113 & 0{,}7072 \end{pmatrix},$$

теперь нам нужно найти двухмерные векторы-столбцы $n_l \mathbf{S}_l^{-1} \overline{X}_l'$

$$n_1 \mathbf{S}_1^{-1} \overline{X}_1' = \begin{pmatrix} 3{,}5980 \\ 16{,}8405 \end{pmatrix},$$

$$n_2 \mathbf{S}_2^{-1} \overline{X}_2' = \begin{pmatrix} -0{,}0738 \\ 4{,}6586 \end{pmatrix},$$

$$n_3 \mathbf{S}_3^{-1} \overline{X}_3' = \begin{pmatrix} 1{,}0060 \\ 2{,}6165 \end{pmatrix}.$$

Найдя сумму полученных трех векторов-столбцов

$$\sum_{l=1}^{3} n_l \mathbf{S}_l^{-1} \overline{X}_l' = \begin{pmatrix} 4{,}5302 \\ 24{,}1106 \end{pmatrix},$$

можем подсчитать искомое значение $\overline{X}$:

$$\overline{X}' = \begin{pmatrix} 2{,}2524, & -0{,}0113 \\ -0{,}0113 & 0{,}7072 \end{pmatrix} \begin{pmatrix} 4{,}5302 \\ 24{,}1106 \end{pmatrix} = \begin{pmatrix} 9{,}9314 \\ 16{,}9998 \end{pmatrix},$$

т. е. $\overline{X} = \{9{,}9314 \quad 16{,}9998\}$.

Теперь, используя полученные данные, вычислим значение критерия, учитывая, что

$$\{\overline{X}_1 - X\}' = \begin{pmatrix} 9{,}82 & -9{,}93 \\ 15{,}06 & -17{,}00 \end{pmatrix} = \begin{pmatrix} -0{,}11 \\ -1{,}94 \end{pmatrix},$$

$$\{\overline{X}_2 - \overline{X}\}' = \begin{pmatrix} 13{,}05 & -9{,}93 \\ 22{,}57 & -17{,}00 \end{pmatrix} = \begin{pmatrix} 3{,}12 \\ 5{,}57 \end{pmatrix},$$

$$\{\overline{X}_3 - \overline{X}\}' = \begin{pmatrix} 14{,}67 & -9{,}93 \\ 25{,}17 & -17{,}00 \end{pmatrix} = \begin{pmatrix} 4{,}74 \\ 8{,}17 \end{pmatrix}.$$

Подставив эти данные в формулу (5.37), получим

$$V = \{-0{,}11, \quad -1{,}94\} \begin{pmatrix} 0{,}1523 & 0{,}1396 \\ 0{,}1396 & 1{,}0272 \end{pmatrix} \begin{pmatrix} -0{,}11 \\ -1{,}94 \end{pmatrix} +$$

$$+ \{3{,}12, \quad 5{,}57\} \begin{pmatrix} 0{,}1756 & -0{,}1048 \\ -0{,}1048 & 0{,}2670 \end{pmatrix} \begin{pmatrix} 3{,}12 \\ 5{,}57 \end{pmatrix} +$$

$$+\{4{,}74,\quad 8{,}17\}\begin{pmatrix} 0{,}1161 & -0{,}0277 \\ -0{,}0277 & 0{,}1199 \end{pmatrix}\begin{pmatrix} 4{,}74 \\ 8{,}17 \end{pmatrix}=$$

$$=3{,}9274+6{,}3503=8{,}4664=18{,}7441.$$

Это число превышает допустимое значение χ^2, соответствующее уровню значимости 0,01 и (3—1) 2 = 4 степеням свободы, которое равно 13,277. Однако, учитывая результаты Г. Джеймса, определенные формулами (5.42) и (5.43), можно улучшить наше приближение, вычислив числа A и B с последующим определением исправленного критического значения. Процедура этих вычислений сводится к следующему:

$$\mathbf{I}_2-\left(\sum_{l=1}^{3} n_l\mathbf{S}_l^{-1}\right)^{-1} n_1\mathbf{S}_1^{-1}=$$

$$=\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}-\begin{pmatrix} 2{,}2524 & -0{,}0113 \\ -0{,}0113 & 0{,}7072 \end{pmatrix}\begin{pmatrix} 0{,}1523 & 0{,}1396 \\ 0{,}1396 & 1{,}0272 \end{pmatrix}=$$

$$=\begin{pmatrix} 0{,}6585 & -0{,}3028 \\ -0{,}0970 & 0{,}2751 \end{pmatrix},$$

$$\mathbf{I}_2-\left(\sum_{l=1}^{3} n_l\mathbf{S}_l^{-1}\right)^{-1} n_2\mathbf{S}_2^{-1}=$$

$$=\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}-\begin{pmatrix} 2{,}2524 & -0{,}0113 \\ -0{,}0113 & 0{,}7072 \end{pmatrix}\begin{pmatrix} 0{,}1756 & -0{,}1048 \\ -0{,}1048 & 0{,}2760 \end{pmatrix}=$$

$$=\begin{pmatrix} 0{,}6033 & 0{,}2391 \\ 0{,}0761 & 0{,}8100 \end{pmatrix},$$

$$\mathbf{I}_2-\left(\sum_{l=1}^{3} n_l\mathbf{S}_l^{-1}\right)^{-1} n_3\mathbf{S}_3^{-1}=$$

$$=\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}-\begin{pmatrix} 2{,}2524 & -0{,}0113 \\ -0{,}0113 & 0{,}7072 \end{pmatrix}\begin{pmatrix} 0{,}1161 & -0{,}0277 \\ -0{,}0277 & 0{,}1199 \end{pmatrix}=$$

$$=\begin{pmatrix} 0{,}7382 & 0{,}0637 \\ 0{,}0209 & 0{,}9149 \end{pmatrix}.$$

Далее нужно вычислить три следа квадратов матриц и три квадрата следов матриц.

$$\mathrm{tr}\begin{pmatrix} 0{,}6585 & -0{,}3028 \\ -0{,}0970 & 0{,}2751 \end{pmatrix}^2=0{,}5680,$$

$$\operatorname{tr}\begin{pmatrix} 0{,}6033 & 0{,}2391 \\ 0{,}0761 & 0{,}8100 \end{pmatrix}^2 = 1{,}0565,$$

$$\operatorname{tr}\begin{pmatrix} 0{,}7382 & 0{,}0637 \\ 0{,}0209 & 0{,}9149 \end{pmatrix}^2 = 1{,}3846,$$

$$\left[\operatorname{tr}\begin{pmatrix} 0{,}6585 & -0{,}3028 \\ -0{,}0970 & 0{,}2751 \end{pmatrix}\right]^2 = 0{,}8716,$$

$$\left[\operatorname{tr}\begin{pmatrix} 0{,}6033 & 0{,}2391 \\ 0{,}0761 & 0{,}8100 \end{pmatrix}\right]^2 = 1{,}9974,$$

$$\left[\operatorname{tr}\begin{pmatrix} 0{,}7382 & 0{,}0637 \\ 0{,}0209 & 0{,}9149 \end{pmatrix}\right]^2 = 2{,}7327.$$

Откуда, в соответствии с формулами (5.42) и (4.43),

$$A = 1 + \frac{1}{8}\left(\frac{0{,}8716}{15} + \frac{1{,}9972}{10} + \frac{2{,}7327}{10}\right) = 1{,}0664,$$

$$B = \frac{1}{24}\left(\frac{0{,}5680}{15} + \frac{1{,}0565}{10} + \frac{1{,}3846}{10} + \frac{0{,}5311}{2}\right) = 0{,}02281.$$

Таким образом, критическое значение $V_{0,01;\,4}$, соответствующее уровню значимости 0,01 и четырем степеням свободы, будет

$$V_{0,01;4} = \chi^2_{0,01;\,4}\left(A + B\chi^2_{0,01;\,4}\right) = 13{,}277\,(1{,}0664 + 0{,}02281 \cdot 13{,}277) =$$
$$= 13{,}277 \cdot 1{,}360 = 18{,}18.$$

Вычисленное нами значение $V = 18{,}74$, что позволяет уверенно отклонить проверяемую гипотезу о равенстве трех двухмерных средних. Необходимо отметить, что приведенный пример является наиболее простым по своим вычислениям, так как в нем использовались только двухмерные данные. Объем вычислений резко возрастает даже при незначительном увеличении m, что делает практически невозможным ручной счет в подобных задачах, что требует обязательного применения ЭВМ.

Пример 3. С целью иллюстрации метода проверки гипотезы о равенстве k ковариационных матриц воспользуемся данными, приведенными в книге Р. Миллера и Дж. Кана [32], представляющими собой результаты измерения четырех характеристик черепной коробки семи ископаемых видов млекопитающих ореодонтов. В упомянутой книге приведены результаты измерения следующих четырех характеристик: 1) ширина черепной коробки в области теменного шва (в см), 2) максимальная длина ряда коренных зубов (в мм), 3) максимальная длина черепной коробки (в см) и 4) максимальная глубина черепной коробки, измеренная от затылочного основания на уровне затылочного бугра (в см).

Так как вычисления, связанные с анализом четырехмерных данных, весьма трудоемки, для целей иллюстрации можно воспользоваться трехмерной задачей, исключив из рассмотрения, например, третий признак — максимальную длину черепной коробки. Эти трехмерные результаты измерений черепов семи видов ореодонтов приведены ниже.

Номер наблюдения в выборке	Номера признаков 1	2	4
Subdesmatochoerus			
1	47	99	15
2	42	93	16
3	40	90	13
4	46	100	11
5	46	96	16
6	42	88	15
7	43	89	14
8	44	78	13
9	44	90	11
10	47	99	15
$n_1 = 11$	47	92	13
Mogoreodon gigas loomisi			
1	78	165	18
2	77	165	19
3	65	148	20
4	74	163	15
5	65	169	16
6	70	176	23
7	69	161	13
8	67	178	14
9	65	174	18
10	64	168	13
$n_2 = 11$	68	166	15
Oreodontes osborni			
1	42	81	8
2	48	83	8,6
3	45	87	9
4	48	83	8
5	46	84	6,1
6	51	87	7,9
7	46	80	7
8	50	90	8,1
9	46	85	6,5
10	48	85	7,2
11	47	85	8
12	49	83	7,7
13	43	79	7,1
14	47	87	7,5
$n_3 = 15$	46	87	8
Desmatochoerus hatcheri			
1	58	129	16
2	52	126	18
3	50	122	22
4	52	123	18
5	60	138	17
6	61	122	17
7	54	132	17
8	65	131	18
9	55	130	17
10	64	125	16
$n_4 = 11$	56	124	16
Pseudodesmatochoerus			
1	60	114	20
2	60	118	19
3	60	111	21
4	58	102	20
5	55	116	20
6	59	117	17
7	59	114	17
$n_5 = 8$	60	121	19
Merychoidodon culbertsoni			
1	45	91	7,5
2	46	93	6,5
3	48	92	5,0
4	46	91	6,0
5	45	86	6,5
6	51	93	7,5
7	47	92	5,0
8	48	89	6,5
9	47	91	6,0
10	50	91	7,2
11	48	91	7,6
12	49	93	7,0
13	49	87	6,5
$n_6 = 14$	49	91	7,7
Prodesmatochoerus meeni			
1	37	88	3,9
2	43	79	4,0
3	43	84	4,2
4	42	80	5,2
5	39	83	4,5
6	39	87	4,5
7	40	86	4,5
8	34	77	4,8
9	35	82	4,6
10	45	88	4,9
11	33	80	3,9
$n_7 = 12$	42	85	4,0

Таким образом, мы располагаем семью трехмерными выборками, объемы которых $n_1 = 11$, $n_2 = 11$, $n_3 = 15$, $n_4 = 11$, $n_5 = 8$, $n_6 = 14$, $n_7 = 12$.

По этим данным требуется проверить гипотезу о равенстве семи неизвестных ковариационных матриц, т. е.

$$H_0 : \Sigma_1 = \Sigma_2 = \ldots = \Sigma_7 = \Sigma_0$$

при альтернативе

$H_1 : \Sigma_l \neq \Sigma_0$ хотя бы для одного $l = 1, 2, \ldots, 7$.

Для проверки H_0 мы воспользуемся критерием (5.47), описанным в разделе 5.2, т. е.

$$V = \sum_{l=1}^{k} \frac{(n_l - 1)}{2} \ln \frac{|S|}{|S_l|}.$$

Для этого необходимо по имеющимся выборочным данным вычислить оценки S_l ковариационных матриц Σ_l для каждой выборки, а также оценку S для Σ_0, которая вычисляется в предположении, что H_0 верна. Эти оценки следующие:

$$S_1 = \begin{pmatrix} 5{,}86 & 8{,}23 & -0{,}03 \\ 8{,}23 & 40{,}76 & 1{,}84 \\ -0{,}03 & 1{,}84 & 3{,}16 \end{pmatrix},$$

$$S_2 = \begin{pmatrix} 24{,}82 & -2{,}99 & 3{,}78 \\ -2{,}99 & 67{,}65 & -0{,}81 \\ 3{,}78 & -0{,}81 & 10{,}02 \end{pmatrix},$$

$$S_3 = \begin{pmatrix} 5{,}74 & 4{,}09 & 0{,}29 \\ 4{,}09 & 8{,}97 & 0{,}70 \\ 0{,}29 & 0{,}70 & 0{,}58 \end{pmatrix},$$

$$S_4 = \begin{pmatrix} 25{,}20 & 7{,}50 & -4{,}40 \\ 7{,}50 & 25{,}27 & 2{,}53 \\ -4{,}40 & 2{,}53 & 2{,}87 \end{pmatrix},$$

$$S_5 = \begin{pmatrix} 2{,}98 & 1{,}73 & -0{,}27 \\ 1{,}73 & 32{,}98 & -3{,}16 \\ -0{,}27 & -3{,}16 & 2{,}12 \end{pmatrix},$$

$$S_6 = \begin{pmatrix} 3{,}30 & 1{,}01 & 0{,}59 \\ 1{,}01 & 4{,}49 & 0{,}02 \\ 0{,}50 & 0{,}02 & 0{,}79 \end{pmatrix},$$

$$S_7 = \begin{pmatrix} 15{,}15 & 5{,}36 & 0{,}28 \\ 5{,}36 & 13{,}66 & -0{,}20 \\ 0{,}28 & -0{,}20 & 0{,}18 \end{pmatrix}.$$

Обобщенная матрица S легко вычисляется по матрицам с помощью формулы

$$S=\frac{1}{\left(\sum_{l=1}^{7} n_l\right)-7}\sum_{l=1}^{7}\frac{n_l-1}{2}S_l.$$

В итоге было получено

$$S=\begin{pmatrix}11{,}61 & 3{,}58 & 0{,}07\\ 3{,}58 & 25{,}36 & 0{,}28\\ 0{,}07 & 0{,}28 & 2{,}61\end{pmatrix}.$$

Определители всех семи матриц, соответствующих отдельным выборкам, равны:

$$|S_1|=519{,}96,\quad |S_2|=17603{,}90,\quad |S_3|=18{,}26,$$
$$|S_4|=848{,}67,\quad |S_5|=172{,}58,\quad |S_6|=9{,}79,$$
$$|S_7|=29{,}80,\ \text{а определитель}\ |S|=734{,}12.$$

Используя формулу (5.47) и учитывая переход от десятичных логарифмов к натуральным, можем вычислить значение критерия

$$V=2{,}3056\left(5\lg\frac{734{,}12}{519{,}96}+5\lg\frac{734{,}12}{17603{,}9}+7\lg\frac{734{,}12}{18{,}26}+\right.$$
$$+5\lg\frac{734{,}12}{848{,}67}+3{,}5\lg\frac{734{,}12}{172{,}58}+6{,}5\lg\frac{734{,}12}{9{,}79}+$$
$$\left.+5{,}5\lg\frac{734{,}12}{29{,}80}\right)=61{,}80.$$

Как уже отмечалось в разделе 5.2. число степеней свободы критерия V равно

$$\frac{(k-1)\,m\,(m+1)}{2}=\frac{(7-1)\,3\,(3+1)}{2}=12.$$

Так как число степеней свободы значительно больше 7, то можно воспользоваться для выбора критического значения таблицами χ^2-распределения. Уровню значимости 0,01 при 12 степенях свободы соответствует значение 58,62, что позволяет уверенно отклонить проверяемую гипотезу о равенстве ковариационных матриц и принять альтернативу, так как вычисленное значение $V=61{,}80$ превышает 58,62.

Таким образом, набор выборочных ковариационных матриц семи видов ореодонтов нельзя признать однородным, что означает, что различия видов сказываются не только в средних значениях изученных характеристик черепа, но и в характере зависимостей между этими показателями. Это также означает, что ковариационные матрицы могут служить диагностическим средством, так как несут в данном случае соответствующую информацию.

ИССЛЕДОВАНИЕ РАЗЛИЧИЙ МЕЖДУ ГЕОЛОГИЧЕСКИМИ ОБЪЕКТАМИ

Весьма часто геологический вывод базируется на утверждении об одинаковой или разной величине различий между парами геологических объектов. Так, например, на основании анализа микроклинов из гранитов, слагающих основную часть интрузива, и анализа микроклинов пегматитовой жилы были получены средние арифметические содержания Rb_2O: в гранитах $\bar{x}_1 = 0{,}18$, в пегматитах $\bar{x}_2 = 0{,}236$ %; причем статистическая проверка показала, что эти оценки неизвестных средних значений существенно различаются. Аналогичное опробование с последующими определениями Rb_2O в микроклинах было проведено для гранитов и пегматитов другого массива, где были получены следующие средние арифметические: $\bar{x}_3 = 0{,}226$ для гранитов и $\bar{x}_4 = 0{,}309$ для пегматитов. Как и в предыдущем случае, $\bar{x}_3$ и $\bar{x}_4$ различаются существенно.

Таким образом, разность между средними арифметическими в первом случае $|\bar{x}_2 - \bar{x}_1| = 0{,}053$, а во втором $|\bar{x}_4 - \bar{x}_3| = 0{,}083$. Вполне естественно, может возникнуть вопрос — можно ли рассматривать эти расхождения практически одинаковыми или же они существенно различны. На принятом решении могут базироваться самые разнообразные геологические выводы, и поэтому обоснованность подобных утверждений представляет в геологии значительный методический интерес.

С вопросами подобного типа особенно часто приходится сталкиваться при стратиграфических исследованиях, когда требуется различать границы между крупными стратиграфическими подразделениями, например ярусами, и более мелкими, детализирующими, например между свитами. Интуитивно ясно, что границы между более крупными стратиграфическими подразделениями должны обладать большими различиями, чем между более мелкими.

Таким образом, возникает задача не только количественного выражения различий между геологическими объектами, но и сравнения самих различий, причем, как и во всех предыдущих задачах, с учетом риска, связанного с принятием ошибочных решений.

Прежде чем перейти к статистическим методам таких сравнений, целесообразно рассмотреть возможные меры таких различий и их статистические оценки.

6.1. МЕРЫ РАЗЛИЧИЙ И ИХ ОЦЕНКИ

Рассмотрим простейшую ситуацию. Пусть ξ_1 и ξ_2 две случайные величины с математическими ожиданиями μ_1 и μ_2 и пусть также η_1 и η_2 — другая пара случайных величин с математическими ожи-

даниями θ_1 и θ_2 соответственно. Если $\mu_1 \neq \mu_2$ и $\theta_1 \neq \theta_2$, то в качестве меры различий между μ_1 и μ_2, а также θ_1 и θ_2 естественно рассматривать расстояния между ними, т. е.

$$\delta_1 = |\mu_1 - \mu_2| \tag{6.1}$$

и

$$\delta_2 = |\theta_1 - \theta_2|. \tag{6.2}$$

Естественно, что множество возможных мер различий не ограничивается только расстояниями, а включает и некоторые функции расстояния. Примером может служить величина

$$\delta_1' = \frac{|\mu_1 - \mu_2|}{\sigma},$$

где σ — константа.

В том случае, если Ξ_1, Ξ_2, H_1 и H_2 — m-мерные случайные величины, то расхождения между объектами, моделями которых они являются, можно представить как расстояния между $\boldsymbol{\mu}_1$ и $\boldsymbol{\mu}_2$, а также между $\boldsymbol{\theta}_1$ и $\boldsymbol{\theta}_2$ в m-мерном пространстве, где $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ — m-мерные векторы-строки математических ожиданий упомянутых случайных величин. В данном случае эти расстояния выразятся следующим образом:

$$\begin{aligned} \delta_1 &= \{\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\}\{\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\}', \\ \delta_2 &= \{\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\}\{\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\}'. \end{aligned} \tag{6.4}$$

В обычной практической геологической ситуации точные значения δ_1 и δ_2, так же как $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, неизвестны и оцениваются по выборке. Таким образом, возникает задача проверки гипотезы $H_0 : \delta_1 = \delta_2$ при альтернативе $H_1 : \delta_1 \neq \delta_2$ (альтернатива может быть и другой, например $\delta_1 < \delta_2$), по имеющимся статистическим оценкам $\hat{\delta}_1$ и $\hat{\delta}_2$. Если же в качестве меры различий выбрана некоторая функция δ_1 и δ_2, то аналогичные гипотезы можно сформулировать и для этого случая. Рассмотрим сначала простейший случай проверки гипотезы о равенстве двух расхождений.

6.2. СРАВНЕНИЕ ДВУХ РАСХОЖДЕНИЙ

Начнем с наиболее простого случая, когда даны две одномерные случайные величины ξ_1 и ξ_2 с неизвестными средними μ_1 и μ_2 и дисперсиями σ_1^2 и σ_2^2. Даны также две случайные величины ξ_3 и ξ_4 со средними μ_3 и μ_4 и дисперсиями σ_3^2 и σ_4^2. Пусть над каждой случайной величиной проведено по n_1, n_2, n_3 и n_4 наблюдений, по которым получены средние арифметические $\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$, $\bar{x}_4$, являющиеся оценками для μ_1, μ_2, μ_3 и μ_4, а также статистические оценки дисперсий S_1^2, S_2^2, S_3^2 и S_4^2.

Допустим, что для проверки гипотез $H_1' : \mu_1 - \mu_2 = 0$ и $H_1'' : \mu_3 - \mu_4 = 0$ использовался критерий

$$t_1 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}, \tag{6.6}$$

$$t_2 = \frac{|\bar{x}_3 - \bar{x}_4|}{\sqrt{\frac{S_3^2}{n_3} + \frac{S_4^2}{n_4}}}, \tag{6.7}$$

и в обоих случаях оказалось $t_1 > t_{1-\frac{q}{2}}$ и $t_2 > t_{1-\frac{q}{2}}$, что позволило отклонить проверяемые нулевые гипотезы.

Только наличие такого результата означает правомерность постановки задачи о равенстве расхождений между двумя парами средних значений. Так как гипотезы H_0' и H_0'' отклонены, то это значит, что числа t_1 и t_2 можно рассматривать как значения двух случайных величин τ_1 и τ_2, которые распределены приблизительно нормально со средними значениями $a_1 \neq 0$ и $a_2 \neq 0$ и дисперсиями, равными 1. В этих случаях новая случайная величина

$$\tau = \frac{\tau_1 - \tau_2}{\sqrt{2}} \tag{6.8}$$

при условии $a_1 = a_2$ будет распределена приблизительно нормально со средним значением 0 и дисперсией, равной 1. Числа a_1 и a_2 являются функциями разностей $\mu_1 - \mu_2$ и $\mu_3 - \mu_4$, так что их можно рассматривать в качестве мер различий между средними. Таким образом, мы будем считать расхождения между средними значениями μ_1, μ_2 и μ_3, μ_4 равноценными, если будет принята гипотеза $H_0 : a_1 = a_2$ и, наоборот, неравноценными, если будет принята альтернатива $H_1 : a_1 \neq a_2$. Критерием для проверки H_0 будет служить число

$$t = \frac{t_1 - t_2}{\sqrt{2}}, \tag{6.9}$$

которое в условиях H_0 будет значением случайной величины τ, распределенной приблизительно нормально со средним, равным нулю и дисперсией, равной 1. Следовательно, гипотеза H_0 отвергается, если $|t| > t_{1-\frac{q}{2}}$, где $t_{1-\frac{q}{2}}$ берется из таблиц нормального распределения в соответствии с заданным уровнем значимости q.

Несколько иначе обстоит дело с многомерными задачами. Заметим, что очень многие критерии проверки многомерных гипотез о равенстве тех или иных параметров в условиях нулевой гипотезы распределены приблизительно как χ^2 с соответствующим числом

степеней свободы f. Так, например, при проверке гипотезы $H_0 : M\,\Xi_1 = M\,\Xi_2$ можно воспользоваться критерием (5.24), описанным в предыдущей главе, который в условиях H_0 распределен как χ^2 с f степенями свободы. Точно так же критерий с χ^2-распределением можно подобрать и для проверки гипотезы о равенстве ковариационных матриц. Если гипотеза о равенстве тех или иных многомерных параметров отклоняется, то это значит, что распределение критерия не согласуется с χ^2-распределением, а подчинено так называемому распределению нецентрального χ с параметром нецентральности a и f степенями свободы. Этот параметр a можно рассматривать как некоторую функцию расхождения между сравниваемыми параметрами и отражающую обобщенный результат различий по всем m рассматриваемым характеристикам. Таким образом, вполне естественно принять параметр нецентральности a как меру различий между сравниваемыми значениями многомерного параметра. Допустим, что в результате применения некоторого критерия V для проверки гипотезы H_0, эта гипотеза отклонена и мы считаем V значением случайной величины, распределенной как нецентральный χ^2 с f степенями свободы. Известно [5], что при достаточно большом значении f распределение величины $\chi^2(a, f)$ можно рассматривать как близкое к нормальному со средним значением $a + f$ и дисперсией $2(f + 2a)$, т. е.

$$P[\chi^2(a, f) \leqslant x] \simeq N[x; f+a, 2(f+2a)]. \tag{6.10}$$

Следовательно, как статистическую оценку для a, можно использовать величину $\hat{a} = V - f$, дисперсия которой будет $2(f + 2a)$. Эту дисперсию, в свою очередь, можно оценить с помощью формулы

$$s^2_{\hat{a}} = 4V - 2f. \tag{6.11}$$

Так как оценка $\hat{a}$ параметра нецентральности a является значением случайной величины $\chi^2(f, a) - f$, которая распределена приблизительно нормально с математическим ожиданием, равным a, и дисперсией $2(f + 2a)$, то легко построить статистический критерий для проверки гипотезы о равенстве двух неизвестных параметров нецентральности по их статистическим оценкам.

Пусть a_1 и a_2 — неизвестные параметры нецентральности двух случайных величин $\chi^2(f, a_1)$ и $\chi^2(f, a_2)$, которые распределены как нецентральный χ^2 с f степенями свободы. Обозначим статистические оценки для a_1 и a_2 через $\hat{a}_1$ и $\hat{a}_2$ соответственно. Число

$$t = \frac{\hat{a}_1 - \hat{a}_2}{\sqrt{\hat{D}(\hat{a}_1 - \hat{a}_2)}} \tag{6.12}$$

будет представлять собой значение случайной величины

$$\tau = \frac{\chi^2(f, a_1) - \chi^2(f, a_2)}{2\sqrt{\chi^2(f, a_1) + \chi^2(f, a_2) - f}}, \tag{6.13}$$

которая при условии $a_1 = a_2$ будет распределена приблизительно нормально с параметрами 0, 1.

Таким образом, если V_1 и V_2 — два значения некоторого критерия, распределенного в условиях нулевой гипотезы как χ^2 с f степенями свободы, и эта гипотеза на основании значений V_1 и V_2 в обоих случаях отвергнута, то для проверки гипотезы о равноценности выявленных различий, которой равносильна гипотеза о равенстве параметров нецентральности a_1 и a_2, достаточно вычислить величину

$$t = \frac{V_1 - V_2}{2\sqrt{V_1 + V_2 - f}} \tag{6.14}$$

и сравнить ее с допустимым значением t_q при уровне значимости q. Необходимо отметить, что выбор критического значения t_q зависит от альтернативы. В данном случае нулевой гипотезе $H_0 : a_1 = a_2$ можно противопоставить три альтернативы:

$$H_1 : a_1 < a_2, \tag{6.15}$$

$$H_2 : a_1 > a_2, \tag{6.16}$$

$$H_3 : a_1 \neq a_2. \tag{6.17}$$

Выбор критических значений, как и в ряде рассмотренных в предыдущих главах случаев, производится исходя из следующих соотношений:

$$P(\tau < t_q \mid H_0) = \Phi(t_q) = q, \tag{6.18}$$

$$P(\tau > t_{1-q} \mid H_0) = 1 - \Phi(t_{1-q}) = q, \tag{6.19}$$

$$P\left(\tau < t_{\frac{q}{2}} \mid H_0\right) + P\left(\tau > t_{1-\frac{q}{2}} \mid H_0\right) = q, \tag{6.20}$$

где Φ — нормальная функция с параметрами 0 и 1.

6.3. СРАВНЕНИЕ БОЛЕЕ ЧЕМ ДВУХ РАСХОЖДЕНИЙ

В геологической практике нередко бывает недостаточно сравнения только двух расхождений. В большинстве ситуаций, возникающих при геологических исследованиях, требуется высказать суждение о наборе более чем двух расхождений между парами сравниваемых геологических объектов. Примером может служить стратиграфический разрез, в котором установлены по комплексу m признаков k границ и требуется ответить на вопрос — можно ли рассматривать все эти границы как равноценные или же, наоборот, среди них можно выделить главные границы, характеризующиеся наиболее сильными расхождениями, и второстепенные, играющие роль детализирующих разграничений.

Допустим, что каждое из k выявленных расхождений обосновано значением критерия V_l, $l = 1, 2, \ldots, k$, превысившего допустимое значение $\chi^2_{q,f}$. (Предполагается что в условиях нулевой

гипотезы критерий V_l распределен как χ^2 с f степенями свободы.) В результате получено k чисел $V_1, V_2, \ldots, V_l, \ldots, V_k$, которые можно рассматривать как значения случайных величин, распределенных как нецентральный χ^2 с параметрами нецентральности $a_1, a_2, \ldots, a_l, \ldots, a_k$ соответственно и с f степенями свободы. Утверждению, что все k расхождений равноценны, равносильно следующее равенство, которое мы будем рассматривать в качестве нулевой гипотезы

$$H_0 : a_1 = a_2 = \ldots = a_k = a_0.$$

Альтернатива, соответствующая утверждению, что набор k распределений неравноценен, формально выражена неравенством $H_1 : a_l \neq a_0$ хотя бы для одного $l = 1, 2, \ldots, k$. Автором в одной из своих работ [38] подробно рассмотрена процедура вывода критерия для проверки гипотезы H_0. Здесь мы на ней останавливаться не будем, а только отметим, что она основана на том, что набор k случайных величин, $\chi^2(a_1), \chi^2(a_2), \ldots, \chi^2(a_k)$, распределенных как нецентральный χ^2 с параметрами нецентральности $a_1, a_2, \ldots, a_k$, можно рассматривать как k приблизительно нормально распределенных случайных величин с математическими ожиданиями $a_1 + f, a_2 + f, \ldots, a_k + f$ и дисперсиями $2(f + 2a_1)$, $2(f + 2a_2), \ldots, 2(f + 2a_k)$. С помощью функции правдоподобия был получен следующий критерий для проверки гипотезы H_0:

$$\chi^2_{k-2} = \frac{1}{2(2\overline{V} - f)} \sum_{l=1}^{k} (V_l - \overline{V})^2, \tag{6.22}$$

где

$$\overline{V} = \frac{1}{k} \sum_{l=1}^{k} V_l. \tag{6.23}$$

Если проверяемая гипотеза верна, то χ^2_{k-2} будет представлять собой значение случайной величины, распределенной как χ^2 с $k-2$ степенями свободы и поэтому гипотезу H_0 следует отклонить и принять альтернативу H_1, если вычисленное значение χ^2_{k-2} окажется больше, чем допустимое [значение $\chi^2_{q,\,k-2}$, соответствующее уровню значимости q и $k-2$ степеням свободы. Если же $\chi^2_{k-2} \leq \chi^2_{q,\,k-2}$, то все рассматриваемые k расхождений следует признать равноценными.

Необходимо отметить, что результат применения описанного выше критерия, заключающийся в принятии или, наоборот, отклонении гипотезы о равноценности k различий, зачастую не может удовлетворить исследователя. Вернее, если гипотеза принимается, то этот результат является окончательным. Если же гипотеза H_0 отклонена, то возникает новая задача, заключающаяся в выделении среди набора k различий равнозначных групп. Подойти к решению этой задачи можно следующим образом. Пусть $\chi^2(a_1)$, $\chi^2(a_2), \ldots, \chi^2(a_k)$ — набор k случайных величин, распределен-

ных как нецентральный χ^2 с параметрами нецентральности a_1, a_2, . . . , a_k и f степенями свободы. Заметим, что предшествующая проверка гипотезы $H_0 : a_1 = a_2 = \ldots = a_k = a_0$ о равенстве параметров нецентральности с помощью описанного выше критерия (6.22) по эмпирическим значениям V_1, V_2, . . . , V_k дала отрицательный результат. Как и раньше, будем рассматривать V_l как значения приблизительно нормально распределенных случайных величин со средними значениями $a_l + f$ и дисперсиями $2(f + 2a_l)$. Тогда для любой пары a_h, a_l из всех $k(k-1)/2$ возможных пар, получаемых в наборе из k элементов, можно сформулировать проверяемую гипотезу

$$H_0^{(hl)}: \; a_h = a_l, \text{ при } (h \neq l) \tag{6.24}$$

и соответствующую альтернативу

$$H_1^{(hl)}: \; a_h \neq a_l, \text{ при } (h \neq l). \tag{6.25}$$

В условиях H_0 разность $V_h - V_l$ будет представлять собой значение случайной величины, распределенной приблизительно нормально с математическим ожиданием, равным нулю, и дисперсией, равной $4(f + a_h + a_l)$. Функция правдоподобия при условии, что верна гипотеза $H_0^{(hl)}$: $a_h = a_l$, для этой разности будет иметь вид

$$L[V_h - V_l;\; 0;\; 4(f + a_h + a_l)] = 2[(f + a_h + a_l)2\pi]^{-\frac{1}{2}} \times$$
$$\times \exp\left\{-\frac{(V_h - V_l)^2}{8(f + a_h + a_l)}\right\}. \tag{6.26}$$

Эта функция максимальна для той пары значений h, l при $h \neq l$, для которой

$$\frac{(V_h - V_l)^2}{8(f + a_h + a_l)} = \min_{h,\, l}. \tag{6.27}$$

Заменив a_h и l_l соответствующими оценками $V_h - f$ и $V_l - f$, получим число

$$\omega_{hl}^2 = \frac{(V_h - V_l)^2}{2(V_h + V_l - f)}, \text{ при } h \neq l, \tag{6.28}$$

которое, если проверяемая гипотеза верна, будет представлять собой значение случайной величины, распределенной как χ^2 с одной степенью свободы. Таким образом, набор k расхождений можно признать полностью неравнозначным, т. е. таким, в котором все сравниваемые пары расхождений существенно различаются, если

$$\min_{h,\, l} \omega_{hl}^2 > \chi_{q,1}^2, \tag{6.29}$$

и, наоборот, имеющим не менее чем одну пару одинаковых расхождений, если

$$\min_{h,\,l} \omega_{hl}^2 \leqslant \chi_{q,1}^2. \tag{6.30}$$

Если в результате вычислений окажется, что это неравенство выполнено, то расхождения, соответствующие минимуму ω_{hl}^2, следует объединить как наиболее близкие из всех возможных пар расхождений и вычислить для них обобщенную оценку

$$V_{h_0 l_0} = \frac{V_{h_0} + V_{l_0}}{2}, \tag{6.31}$$

где h_0, l_0, номера расхождений, для которых значение ω_{hl}^2 минимально. Дисперсия величины $V_{h_0 l_0}$ определяется выражением $V_{h_0} + V_{l_0} - f$. Если расхождения с номерами h_0 и l_0 объединены в одну группу, то весь набор расхождений будет рассчитывать $k-1$ элементов, для которых нужно проверить $k-2$ гипотезы $H_0: a_{h_0 l_0} = a_{hr}$, где $r \neq h_0$, l_0. Критерий для проверки этих гипотез будет представлять собой выражение

$$\omega_{h_0 l_0 h_r}^2 = \frac{(V_{h_r} - V_{h_0 l_0})^2}{\hat{D}V_{h_r} + \hat{D}V_{h_0 l_0}}, \tag{6.32}$$

где $\hat{D}V_{h_r}$ и $\hat{D}V_{h_0 l_0}$ оценки дисперсий величин V_{h_r} и $V_{h_0 l_0}$.

Формулу (6.32) можно также записать в виде

$$\omega_{h_0 l_0 h_r} = \frac{(V_{h_r} - V_{h_0 l_0})}{4V_{h_r} + V_{h_0} + V_{l_0} - 3f}. \tag{6.33}$$

Таким образом, $k-2$ значения критерия, вычисленные с помощью формулы (6.28), и оставшиеся после нахождения $\min \omega_{hl}$ $(k-2)^2/2$ значений в сумме составят $1/2$ $(k-1)$ $(k-2)$ значений, из которых снова отыскивается $\min \omega^2$ и сравнивается с критическим $\chi_{q,1}^2$. Если окажется, что

$$\min_{r,\,d} \omega^2 > \chi_{q,1}^2 \begin{pmatrix} r = 1,\ 2,\ \ldots,\ k-1 \\ d = 1,\ 2,\ \ldots,\ k-1 \end{pmatrix}, \tag{6.34}$$

то процедура прекращается, так как уже на этом этапе все группы расхождений неравнозначны. Если же $\min\limits_{r,\,d} \omega^2 \leqslant \chi_{q,1}^2$, то процедура объединения равнозначных расхождений повторяется до тех пор, пока будет достигнуто неравенство

$$\min \omega^2 > \chi_{q,1}^2. \tag{6.35}$$

Таким образом, в итоге будут получены группы равноценных расхождений, для каждой из которых необходимо вычислить обобщенную оценку параметра нецентральности $\overline{V}_i$—f, где

$$\overline{V}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} V_{ij}, \tag{6.36}$$

где i — номер группы, n_i — число расхождений в группе с номером i, V_{ij} — значения критерия (6.28), попавшие в i-ю группу. Если все значения $\overline{V}_i$ расположить в убывающем порядке

$$\overline{V}_{i1} > \overline{V}_{i2} > \ldots > \overline{V}_{ip}, \tag{6.37}$$

то при одном и том же числе степеней свободы f для всех групп получим последовательность групп расхождений, упорядоченную по их роли в различении геологических объектов.

6.4. СРАВНЕНИЕ РАСХОЖДЕНИЙ ПРИ РАЗНОМ ЧИСЛЕ СТЕПЕНЕЙ СВОБОДЫ

Рассмотренные способы сравнения величин расхождений при фиксированном числе степеней свободы не всегда могут удовлетворить исследователя. В практической работе геологу нередко приходится высказывать суждения о равнозначности или, наоборот, различии расхождений между комплексами характеристик геологических объектов, причем число этих характеристик может различаться, что, в свою очередь, приводит к разному числу степеней свободы. Примером может служить задача сравнения расхождений между подразделениями больших стратиграфических разрезов, когда комплекс признаков, по которому устанавливаются границы, меняется по разрезу. В подобной ситуации по-прежнему требуется указать главные и второстепенные границы, но описанные выше критерии для этой цели непригодны. Для того чтобы подойти к решению этой задачи, необходимо ввести сопоставимую количественную меру различий, соответствующих установленным расхождениям.

В предыдущем разделе при сравнении расхождений, охарактеризованных одним и тем же числом степеней свободы, мы рассматривали в качестве такой меры параметр нецентральности нецентрального χ^2-распределения. Однако сравнение параметров нецентральности при различном числе степеней свободы не имеет смысла и потому такая мера различий сопоставляемых объектов в подобной ситуации непригодна. В данном случае в качестве меры различий можно воспользоваться частным от деления параметра нецентральности на соответствующее число степеней свободы. Полученная величина будет характеризовать различия, приходящиеся в среднем на одну степень свободы.

Обозначим, как это делалось выше, значение критерия, соответствующего двум сопоставляемым геологическим объектам, че-

рез V и будем рассматривать его как значение случайной величины $\chi^2(a, f)$, распределенной как нецентральный χ^2 с параметром нецентральности a и f степенями свободы. Как это делалось в разделе 6.2, будем рассматривать случайную величину $\chi^2(a, f)$ как распределенную асимптотически нормально со средним значением $f + a$ и дисперсией $2(f + 2a)$. Образуем новую случайную величину

$$Z = \frac{\chi^2(a,f)}{f}, \tag{6.38}$$

которая, так же как и $\chi^2(a, f)$, будет распределена асимптотически нормально со средним значением, равным $1 + \frac{a}{f}$, и дисперсией $2\left(\frac{1}{f} + \frac{2a}{f^2}\right)$, т. е.

$$P(Z \leqslant z) \simeq N\left[z;\ 1 + \frac{a}{f},\ \frac{2}{f}\left(1 + \frac{2a}{f}\right)\right]. \tag{6.39}$$

Для простоты записи положим $\frac{a}{f} = b$. Тогда в качестве статистической оценки для b можно использовать величину

$$b = \frac{a}{f} = \frac{V - f}{f} = \frac{V}{f} - 1. \tag{6.40}$$

Дисперсия этой оценки будет

$$\hat{D}(\hat{b}) = \frac{2}{f}\left(1 + \frac{2a}{f}\right) = \frac{2}{f}(1 + 2b). \tag{6.41}$$

Эту дисперсию можно оценить по формуле

$$s^2_{\hat{b}} = \frac{2}{f}\left(\frac{2N}{f} - 1\right) = \frac{4V}{f^2} - \frac{2}{f}. \tag{6.42}$$

Так как оценка $\hat{b}$ распределена асимптотически нормально с математическим ожиданием, равным b, и дисперсией $\frac{2}{f}(1 + 2b)$, то нетрудно построить критерий для проверки гипотезы о равенстве двух неизвестных параметров b_1 и b_2 по их статистическим оценкам.

Пусть V_1 и V_2 значения некоторого критерия, предназначенного для сравнения геологических объектов и распределенного в условиях нулевой гипотезы как χ^2, а в условиях альтернативы как нецентральный χ^2 с f_1 и f_2 степенями свободы соответственно. Обозначим через a_1 и a_2 неизвестные параметры нецентральности. Так как в данном случае $f_1 \neq f_2$, то параметры нецентральности a_1 и a_2 не позволяют судить о величине наблюдаемых расхождений, вернее не позволяют их сравнивать. Однако расхождения можно считать равноценными, если величины $b_1 = a_1/f_1$ и $b_2 = a_2/f_2$ равны, как обладающие одинаковыми различиями, приходящимися в среднем на одну степень свободы. Равенство $b_1 = b_2$ будем рас-

сматривать как нулевую гипотезу H_0, которая проверяется при множестве альтернатив $H_1: b_1 \neq b_2$. Учитывая, что оценки b_1 и b_2 для b_1 и b_2 распределены асимптотически нормально со средними значениями b_1 и b_2, то в качестве критерия для проверки $H_0: b_1 = b_2$ можно воспользоваться выражением

$$t = \frac{|\hat{b_1} - \hat{b_2}|}{\sqrt{S^2_{\hat{b_1}} + S^2_{\hat{b_2}}}} = \frac{|V_1 f_2 - V_2 f_1|}{2 f_1 f_2 \sqrt{\frac{V_1}{f_1} + \frac{V_2}{f_2} - \frac{1}{2f_1} - \frac{1}{2f_2}}}. \quad (6.43)$$

Этот критерий представляет собой значение случайной величины, которая при условии что верна H_0, распределена приблизительно нормально с математическим ожиданием, равным 0, и дисперсией, равной 1. В связи с этим гипотеза H_0 отклоняется и принимается альтернатива H_1, если вычисленное значение t окажется меньше допустимого $t_{1-\frac{q}{2}}$, соответствующего уровню значимости q для двустороннего критерия. Значение $t_{1-\frac{q}{2}}$ берется из таблиц нормальной функции. Если же $t \leqslant t_{1-\frac{q}{2}}$, то рассматриваемые расхождения следует считать равноценными.

Нетрудно видеть, что при $f_1 = f_2 = f$ критерий (6.43) совпадает с критерием для проверки гипотезы о равноценности двух расхождений, определенным формулой (6.14), т. е.

$$t = \frac{|V_1 - V_2|}{2\sqrt{V_1 + V_2 - f}}. \quad (6.44)$$

В связи с тем что в геологических задачах нельзя ограничиться сравнением только двух различий, необходимо построить критерий для сравнения любого их числа. Аналогичный критерий был рассмотрен выше для фиксированного числа степеней свободы.

Обозначим $V_1, V_2, \ldots, V_l, \ldots, V_k$ значения некоторого критерия, устанавливающего k существенных различий между геологическими объектами. Как и раньше, будем считать, что в условиях альтернативы критерий распределен как нецентральный χ^2. Пусть $f_1, f_2, \ldots, f_l, \ldots, f_k$ значения числа степеней свободы, соответствующие значениям критерия. Обозначим $b_1, b_2, \ldots, b_l, \ldots, b_k$ неизвестные значения отношений a_l/f_l, где a_l — параметр нецентральности, а статистические оценки отношений b_l обозначим как $\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_k$, где

$$b_l = \frac{V_l}{f_l} - 1 \quad (6.45)$$

По имеющимся данным требуется проверить гипотезу $H_0: b_1 = b_2 = \ldots = b_k = b_0$ при альтернативе $H_1: b_l \neq b_0$ хотя бы для одного $l = 1, 2, \ldots, k$.

В связи с тем что оценки $\hat{b}_1$, $\hat{b_2}$, . . . , $\hat{b_k}$ можно рассматривать как значения случайных величин, распределенных приблизительно нормально со средними значениями b_1, b_2, . . . , b_k, то число

$$\sum_{l=1}^{k} \frac{(\hat{b_l} - \hat{b}_0)^2}{D(b_l)} \tag{6.46}$$

будет представлять собой значение случайной величины, которая в условиях нулевой гипотезы распределена как χ^2 с k—2 степенями свободы. Оценка $\hat{b}_0$ константы b_0 в формуле (6.45) определена выражением

$$\hat{b}_0 = \frac{1}{\sum_{l=1}^{k} f_l} \sum_{l=1}^{k} (V_l - f_l), \tag{6.47}$$

а $D(\hat{b}_l)$ выразится как

$$D(\hat{b}_l) = \frac{2}{f_l}(1 + 2\hat{b}_0). \tag{6.48}$$

Используя эти значения, получим

$$\chi^2_{k-2} = \sum_{l=1}^{k} \frac{\left(\frac{V_l}{f_l} - 1 - \hat{b}_0\right)}{\frac{2}{f_l}(1 + 2b_0)}. \tag{6.49}$$

В развернутом виде эта формула будет иметь вид

$$\chi^2_{k-2} = \frac{1}{\frac{4\sum_{l=1}^{k} V_l}{\sum_{l=1}^{k} f_l} - 2} \sum_{l=1}^{k} f_l \left(\frac{V_l}{f_l} - \frac{\sum_{l=1}^{k} V_l}{\sum_{l=1}^{k} f_l} \right)^2. \tag{6.50}$$

Таким образом, гипотеза H_0 отклоняется, если вычисленное значение χ^2_{k-2} превысит допустимое $\chi^2_{q,\,k-2}$, соответствующее уровню значимости q и $k-1$ степеням свободы, и, наоборот, гипотеза H_0, заключающаяся в утверждении о равноценности рассматриваемого набора расхождений принимается, как непротиворечащая эмпирическим данным, если будет выполнено неравенство $\chi^2_{k-2} \leqslant \chi^2_{q,\,k-2}$. Если гипотеза о равноценности расхождений отклонена, то последующее разделение всего набора расхождений на однородные группы можно провести путем попарного сравнения

оценок b_h с последующим объединением по минимальному значению критерия (6.43), как это делалось для постоянного значения f.

Следует отметить, что в случае равенства $f_1 = f_2 = \ldots = f_k$ критерий (6.49) совпадает с критерием (6.22).

6.5. ПРИМЕРЫ

Пример на сравнение двух расхождений. При изучении ультракислых гранитоидов Центрального Казахстана (материалы В. И. Серых) была построена классификация этих пород по следующим четырем признакам в зависимости от содержаний плагиоклаза: содержание калиевого полевого шпата, кварца, биотита и суммарное содержание акцессорных минералов. Исходные данные, по которым производилось построение классификации, были опубликованы ранее [38]. В результате проведенных исследований среди этих гранитоидов были выделены четыре группы. Так как выделенные группы представляли собой последовательность, соответствующую увеличению содержаний плагиоклаза, то для обоснования их выделения проводились сравнения только смежных групп по четырем перечисленным выше признакам. Для этого использовался многомерный критерий [38], который в условиях нулевой гипотезы, заключающейся в утверждении равенства многомерных средних, распределен как χ^2 с числом степеней свободы, равным числу признаков.

Таким образом, каждым двум смежным в последовательности группам соответствовало значение критерия, которое, в связи с тем, что оно превысило критическое, можно рассматривать как значение случайной величины, распределенной как нецентральный χ^2 с четырьмя степенями свободы. Эти значения критерия следующие:

$$V_1 = 51{,}37,\ V_2 = 20{,}20,\ V_3 = 16{,}10.$$

В данном случае вполне уместен вопрос, можно ли рассматривать установленные расхождения между группами гранитоидов как равноценные, или же они играют различную роль, т. е. среди них есть главные и второстепенные.

С помощью критерия

$$t = \frac{V_1 - V_2}{2\sqrt{V_1 + V_2 - f}},$$

предназначенного для сопоставления двух расхождений с одинаковым числом степеней свободы, проведем попарное сравнение приведенных выше трех значений критерия, т. е. проверим гипотезу о равенстве параметров нецентральности. Соответствующие значения упомянутого критерия имеют вид

$$t_{1,2} = \frac{V_1 - V_2}{2\sqrt{V_1 + V_2 - f}} = \frac{51{,}37 - 20{,}20}{2\sqrt{51{,}37 + 20{,}20 - 4}} = 1{,}90,$$

$$t_{1,3} = \frac{51{,}37 - 16{,}11}{2\sqrt{51{,}37 + 16{,}11 - 4}} = 2{,}21,$$

$$t_{2,3} = \frac{20{,}20 - 16{,}11}{2\sqrt{20{,}20 + 16{,}11 - 4}} = 0{,}36.$$

Из полученных результатов только один $t_{1,3} = 2{,}21$ превышает допустимое значение $t_{0,05} = 1{,}96$, соответствующее 5 %-ному уровню значимости при двусторонней альтернативе. Значение $t_{2,3} = 0{,}36$, соответствующее второму и третьему расхождениям, значительно меньше $t_{0,05} = 1{,}96$, что позволяет уверенно принять гипотезу о равноценности этих расхождений.

Сомнительный результат получен при сопоставлении первого и второго расхождений, когда $t_{1,2} = 1{,}90$ очень незначительно отличается от критического (1,96) в меньшую сторону, что не позволяет отклонить проверяемую гипотезу при заданном уровне значимости, но и не дает большой уверенности при ее принятии, как не противоречащей выборочным данным.

Для получения однозначного решения объединим второе и третье расхождения, как наиболее близкие и несущественно различающиеся, вычислив для них единую оценку

$$V_{(2,3)} = \frac{V_2 + V_3}{2} = \frac{20{,}20 + 16{,}11}{2} = 18{,}15.$$

Затем с помощью того же критерия проверим гипотезу о равноценности первого расхождения и усредненного для второго и третьего расхождений

$$t = \frac{V_1 - V_{(2,3)}}{2\sqrt{V_1 - V_{(2,3)} - 4}} = \frac{51{,}37 - 18{,}15}{2\sqrt{51{,}37 + 18{,}15 - 4}} = 2{,}05.$$

Так как полученное значение t превышает критическое при уровне значимости 0,05, то гипотезу о равенстве параметров нецентральности двух нецентральных χ^2-распределений, а следовательно, и предположение о равноценности расхождений отклонить, как противоречащее выборочным данным.

Таким образом, установлено, что из трех выявленных расхождений одно (первое) играет главную роль, а остальные два носят характер детализации. Следовательно, имеющиеся гранитоиды разделяются на две крупные группы, причем первая группа больше не делится, а вторая разделена на три подгруппы, чему соответствуют два равноценных расхождения.

Пример на сравнение k расхождений. В связи с тем что критерий для проверки гипотезы о равенстве более чем двух расхождений при постоянном значении числа степеней свободы является частным случаем более общего критерия, предназначенного для проверки упомянутой нулевой гипотезы при неравном числе степеней свободы, мы ограничимся примером, связанным с применением последнего. При статистическом расчлене-

нии (см. главу 9) стратиграфического разреза палеогеновых отложений по комплексу микрофауны, включающему 59 видов фораминифер, были выявлены 11 границ между 12 установленными стратиграфическими подразделениями. В связи с тем что комплекс микрофауны по разрезу варьирует, число видов фораминифер, по которым доказывалась граница между смежными подразделениями, менялось, а следовательно, менялось и число степеней свободы применявшегося для этого обоснования критерия, предназначенного для проверки гипотезы о равенстве многомерных средних. Упомянутый критерий в условиях проверяемой гипотезы распределен как χ^2 с числом степеней свободы, равным числу изучаемых признаков. Таким образом, все значения критерия V_l, соответствующие границам с номером l, можно рассматривать как значения случайной величины, распределенной как нецентральной χ^2 с параметром нецентральности a_l и f_l степенями свободы, так как вычисленные числа V_l превысили допустимые χ^2_{q,f_l}. В табл. 16 приведены все 11 значений критерия V_l с соответствующим числом степеней свободы f_l.

Таблица 16

Значения критерия V_l и числа степеней свободы

l	V_l	f_l	$\frac{V_l}{f_l}$	l	V_l	f_l	$\frac{V_l}{f_l}$
1	41,13	8	5,16	7	42,53	23	1,85
2	12,55	5	2,51	8	42,87	23	1,86
3	15,76	6	2,64	9	47,70	26	1,83
4	15,26	6	2,54	10	52,17	25	2,09
5	42,05	19	2,21	11	74,95	25	2,99
6	114,61	31	3,69				

Для проверки гипотезы о равенстве одиннадцати отношений $a_l/f_l = b_l$, что соответствует равноценности всех рассматриваемых расхождений, характеризующих границы, мы воспользуемся критерием 6.50, описанным в разделе 6.4,

$$\chi^2_{k-2} = \frac{1}{4\alpha - 2} \sum_{l=1}^{k} f_l \left(\frac{V_l}{f_l} - \alpha \right)^2,$$

где

$$\alpha = \frac{\sum_{l=1}^{k} V_l}{\sum_{l=1}^{k} f_l}.$$

Процедура вычисления этого критерия, по данным табл. 14, приведена в табл. 17.

Таблица 17

Проверка гипотезы о равенстве расхождений

l	V_l	f_l	$\frac{V_l}{f_l}$	$\frac{V_l}{f_l}-\alpha$	$f_l\left(\frac{V_l}{f_l}-\alpha\right)^2$
1	41,18	8	5,15	2,60	54,08
2	12,55	6	2,51	—0,04	0,01
3	15,76	6	2,63	0,08	0,04
4	15,26	6	2,54	0,00	0,00
5	42,05	19	2,21	—0,34	2,20
6	114,61	31	3,70	1,15	41,00
7	42,53	23	1,85	—0,70	11,27
8	42,87	23	1,86	—0,69	10,95
9	47,70	26	1,84	—0,71	13,11
10	52,17	25	2,09	—0,46	5,29
11	74,95	25	3,00	0,45	5,06

$$\sum_{l=1}^{11} V_l = 501{,}63, \qquad \alpha = \frac{501{,}66}{197} = 2{,}55,$$

$$\sum_{l=1}^{11} f_l = 197, \qquad \sum_{l=1}^{11} f_l\left(\frac{V_l}{f_l}-\alpha\right)^2 = 143{,}01,$$

$$4\alpha - 2 = 4 \cdot 2{,}55 - 2 = 8{,}20,$$

$$\chi^2_{k-2} = \frac{143{,}01}{8{,}20} = 17{,}44.$$

Так как вычисленное значение $\chi^2_9 = 17{,}44$ превышает допустимое $\chi^2_{0,\,05;\,9} = 16{,}92$, то из этого следует, что набор рассматриваемых расхождений, а значит и выделенных стратиграфических границ нельзя рассматривать как равноценный. Однако, если из рассматриваемого набора расхождений удалить первое значение $V_1 = 41{,}18$, как максимально уклоняющееся от значения $\alpha = 2{,}55$, и для оставшихся десяти значений V_l повторить всю процедуру проверки гипотезы о равноценности расхождений, получим

$$\alpha = 2{,}43, \quad \sum_{l=2}^{11} f_l\left(\frac{V_l}{f_l}-\alpha\right)^2 = 86{,}03,$$

$$4\alpha - 2 = 7{,}22, \quad \chi^2_{10-2} = \chi^2_8 = \frac{86{,}03}{7{,}22} = 11{,}14.$$

Так как полученное значение $\chi^2_8 = 11{,}14$ меньше, чем $\chi^2_{0{,}05;\,8} = 15{,}51$, то оставшиеся 10 границ можно считать равноценными, а первую границу рассматривать как выраженную более ярко, чем все остальные, т. е. соответствующую более сильным изменениям в комплексе фауны, по сравнению с остальными десятью границами.

ГЛАВА 7

ВЫБОР ИНФОРМАТИВНЫХ КОМБИНАЦИЙ ПРИЗНАКОВ

7.1. ПОСТАНОВКА ВОПРОСА

При изучении геологических объектов по комплексу признаков число последних может оказаться весьма большим. При этом вполне естественно желание выделить из всего комплекса рассматриваемых характеристик такую их комбинацию, которая бы обладала наибольшими различиями по сравнению с другими комбинациями признаков. Так, например, при изучении пород, слагающих рудоносные и безрудные гранитные массивы, обычно проводятся измерения самых разнообразных характеристик (содержаний различных элементов как петрогенных, так и редких, содержаний породообразующих и акцессорных минералов и т. п.). В подобной ситуации, особенно когда число изучаемых признаков велико, и возникает задача о сокращении без ущерба для дела числа изучаемых характеристик, т. е. выявления такой их комбинации, которая при меньшем числе входящих в нее характеристик обеспечивала бы ту же или большую надежность различения сравниваемых объектов. Поиск такой комбинации особенно важен для последующего применения дискриминантного анализа (см. главу 8), который нередко используется для построения решающих правил при геологическом прогнозировании. Естественно, что для построения дискриминантной функции желательно найти такую комбинацию признаков, которая бы обеспечивала наилучшее распознавание изучаемых объектов.

Следует отметить, что при геологических исследованиях представляют интерес не только признаки, которые несут информацию о различиях между сравниваемыми объектами, но также и комплексы признаков, которые такими различиями не обладают. Такие неинформативные комплексы признаков, характеризующие сходство сравниваемых объектов, нередко можно объяснить с генетических позиций, что может оказаться весьма важным при выяснении генезиса изучаемых геологических образований.

Таким образом, в геологических исследованиях представляет интерес выбор как информативных, так и неинформативных комбинаций. Однако прежде чем перейти к изложению методов этого выбора, необходимо ввести некоторые формальные определения.

7.1.1. Формальные определения

Пусть $\Xi = \{\xi_1, \xi_2, \ldots, \xi_j, \ldots, \xi_m\}$ и $H = \{\eta_1, \eta_2, \ldots, \eta_j, \ldots, \eta_m\}$ — m-мерные случайные величины, а θ_1 и θ_2 — соответствующие им многомерные параметры. Под θ_1 и θ_2 можно

иметь в виду m-мерные векторы средних значений или же ковариационные матрицы порядка $m \times m$, а можно также θ_1 и θ_2 рассматривать как соответствующие векторы средних и ковариационных матриц, т. е.

$$\theta_1 = (\mu_1, \Sigma_1), \quad \theta_2 = (\mu_2, \Sigma_2), \tag{7.1}$$

где μ_1 и μ_2 — m-мерные векторы средних значений, а Σ_1 и Σ_2 — ковариационные матрицы порядка $m \times m$.

Обозначим множество всех значений индекса j через M и пусть J_k — произвольное множество в M, содержащее k элементов, которое мы в дальнейшем будем называть комбинацией k признаков. Дополнение множества J_k до M обозначим J_{m-k}, которое, естественно, представляет собой комбинацию m—k признаков.

Обозначим через $\theta_1(J_k)$ и $\theta_2(J_k)$ наборы элементов многомерных параметров θ_1 и θ_2, соответствующие комбинации k признаков J_k и дадим определение неинформативной комбинации.

Неинформативной комбинацией J_k^0 относительно многомерного параметра θ будем называть любую комбинацию k признаков J_k, для которой $\theta_1(J_k) = \theta_2(J_k)$. Таким образом, если под θ понимается среднее значение, т. е. $\theta_1 = \mu_1$, $\theta_2 = \mu_2$, то неинформативная комбинация будет определена относительно k-мерного вектора средних значений, если же $\theta_1 = \Sigma_1$, $\theta_2 = \Sigma_2$, то неинформативная комбинация будет определена относительно подматриц Σ_1^k и Σ_2^k порядка $k \times k$ матриц Σ_1 и Σ_2 соответственно.

Наоборот, любую комбинацию k признаков J_k, для которой

$$\theta_1(J_k) \neq \theta_2(J_k), \tag{7.2}$$

будем называть информативной комбинацией относительно многомерного параметра θ и обозначать J*.

7.2. ПРОВЕРКА ГИПОТЕЗЫ О НЕИНФОРМАТИВНОСТИ ЗАДАННОЙ КОМБИНАЦИИ ПРИЗНАКОВ

При геологических исследованиях вполне реальна постановка вопроса об исключении из дальнейших исследований некоторого заданного набора $J_k k$ признаков. Это может быть вызвано тем, что такие признаки изучались только на части подлежащих исследованию объектов, или же тем, что продолжение наблюдений над этими характеристиками невозможно в связи с большой стоимостью, или же не может быть выполнено в силу каких-либо других причин. При этом желательно знать, содержит ли исключаемый из рассмотрения комплекс признаков какую-либо информацию о различении двух типов объектов. Естественно, что неинформативный набор признаков может быть исключен без существенных потерь для решаемой геологической задачи, тогда как исключение информативного комплекса может привести к нежелательным последствиям.

Задачу об отнесении данного комплекса k признаков J_k к не-

информативной комбинации J_k^0 или, наоборот, к информативной J_k^* относительно параметра θ можно сформулировать как проверку гипотезы $H_0 : \theta_1 (J_k) = \theta_2 (J_k)$ при альтернативе $H_1 : \theta_1 (J_k) \neq \theta_2 (J_k)$. В зависимости от того, как задан многомерный параметр θ, и выбирается соответствующий критерий $V (J_k)$ для проверки гипотезы H_0. Так, например, если $\theta_1 = \mu_1 = \{\mu_1', \mu_2', \ldots, \mu_j', \ldots, \mu_m'\}$ и $\theta_2 = \mu_2 = \{\mu_1'', \mu_2'', \ldots, \mu_j'', \ldots, \mu_m''\}$, т. е. в качестве параметра θ рассматриваются многомерные средние значения, а ковариационные матрицы Σ_1 и Σ_2 предполагаются равными, то в качестве $V (J_k)$ можно воспользоваться описанным в главе 5 критерием (5.24), основанным на отношении правдоподобия

$$V (J_k) = -\left(n_1 + n_2 - 2 - \frac{k}{2}\right) \ln \frac{|S_1|}{|S_0|}, \qquad (7.3)$$

где n_1 и n_2 — числа наблюдений в выборках, соответствующих случайным величинам Ξ и H, S_0 — оценка ковариационной матрицы Σ, вычисленная в предположении, что проверяемая гипотеза верна, S — оценка для Σ в предположении, что верна альтернатива H_1. Как следует из текста главы 5, гипотеза о неинформативности заданной комбинации признаков относительно средних значений может быть принята, если вычисленное значение $V (J_k)$ окажется меньше допустимого значения $\chi^2_{q,k}$, соответствующего уровню значимости q и k степеням свободы. Если же $V (J_k) > \chi^2_{q,k}$, то из этого следует, что заданная комбинация J_k признаков информативна относительно вектора средних значений.

Если для ковариационных матриц Σ_1 и Σ_2 нельзя ввести допущения о их равенстве, то в качестве критерия $V (J_k)$ для проверки предположения о неинформативности комбинации признаков J_k относительно k-мерных средних можно воспользоваться описанным также в главе 5 критерием (5.38), известным под названием критерия Джеймса.

Аналогично, если параметр θ задан как ковариационные матрицы $\theta_1 = \Sigma_1$ и $\theta_2 = \Sigma_2$, то гипотеза о неинформативности заданного набора признаков J_k будет равносильна гипотезе $H_0 : \Sigma_1^k = \Sigma_2^k$, где Σ_1^k и Σ_2^k подматрицы порядка $k \times k$ матриц Σ_1 и Σ_2, соответствующие комплексу признаков J_k. Проверить эту гипотезу также нетрудно с помощью критерия (5.47), описанного в главе 5, который в случае двух выборок объема n_1 и n_2 примет вид

$$V (J_k) = \frac{(n_1 - 1)}{2} \ln \frac{|S^k|}{|S_1^k|} + \frac{(n_2 - 1)}{2} \ln \frac{|S^k|}{|S_2^k|}, \qquad (7.3)$$

где S_1^k и S_2^k — оценки ковариационных матриц Σ_1^k и Σ_2^k, а матрица

S^k вычислена в предположении, что $\Sigma_1^k = \Sigma_2^k$ и определена выражением

$$S^k = \frac{1}{n_1 + n_2 - 2}\left[(n_1 - 1)\right] S_1^k + (n_2 - 1) S_2^k]. \qquad (7.4)$$

Методика определения критического значения $V_{q,f}$ для принятия или отклонения проверяемой гипотезы при уровне значимости q и f степенях свободы описана в разделе 5.3. Таким образом, если $V(J_k) > V_{q,f}$, то заданную комбинацию признаков J_k следует рассматривать как информативную относительно ковариационных матриц.

В результате нетрудно видеть, что для любой комбинации признаков можно проверить предположение о ее неинформативности относительно заданного параметра.

Описанные выше приемы проверки предположения о неинформативности заданной комбинации признаков нетрудно обобщить на более чем две выборки, т. е. на случай, когда имеется более двух типов изучаемых геологических объектов. Для этой цели можно воспользоваться критериями (5.38) и (5.47), предназначенными для проверки гипотез о равенстве набора многомерных средних и о равенстве нескольких ковариационных матриц. Точно так же, как и в случае двух выборок, некоторой заданной комбинации признаков J_k будет поставлено в соответствие одно значение критерия $V(J_k)$, в результате сравнения которого с критическими значениями принимается решение о неинформативности или, наоборот, информативности изучаемой комбинации признаков.

7.3. ПОЛНАЯ НЕИНФОРМАТИВНАЯ КОМБИНАЦИЯ

Выше мы определили, что комбинация k признаков J_k, для которой выполнено равенство $\theta_1(J_k) = \theta_2(J_k)$, является неинформативной относительно многомерного параметра θ. Допустим, что k задано и тогда этому заданному значению будет соответствовать множество I_k^0 неинформативных комбинаций J_{kl}, где l — номер комбинации. Обозначим через L множество всех значений l, соответствующих I_k^0. Тогда полную неинформативную комбинацию J^0 можно определить как объединение всех неинформативных комбинаций при заданном k, т. е.

$$I_0 = \bigcup_{k=1}^{m} \bigcup_{l \in L} J_{kl} = \sup_k \sup_{l \in L} J_{kl}^0. \qquad (7.5)$$

Естественно, что J^0 нам остается неизвестна, и по имеющимся эмпирическим данным требуется построить оценку $\hat{J}^0$ для J^0.

Подробно этот вопрос был рассмотрен автором ранее [38], а здесь мы приведем только наиболее простой способ, заключающийся в следующем.

Как уже отмечалось в предыдущем разделе, комбинация J_k, образованная k признаками, рассматривается как неинформативная относительно параметра θ, если для нее имеет место равенство

$$\theta_1(J_k) = \theta_2(J_k). \tag{7.6}$$

Таким образом, оценкой $\hat{J}^0_{kj}$ для J^0_{kl} будет такая комбинация признаков J_k, для которой гипотеза $H_0: \theta_1(J_k) = \theta_2(J_k)$ не отклоняется, т. е. значение некоторого критерия $V(J_k)$, предназначенного для проверки гипотезы H_0, окажется меньше критического V_q, соответствующего уровню значимости q. Следовательно, мы можем записать

$$\hat{J}^0_{kl} = J_k : V(J_k) \leqslant V_q. \tag{7.7}$$

В результате оценка $\hat{J}^0$ полной неинформативной комбинации J^0 будет определена выражением

$$\hat{J}^0 = \sup_k \sup_{l \in L} \hat{J}^0_{kl}. \tag{7.8}$$

Естественно, что дополнение $\hat{J}^0$ до M будет представлять собой оценку $\hat{J}^{**}$ полной информативной комбинации J^{**}.

7.4. СОВМЕСТНЫЙ ПОИСК ПОЛНОЙ ИНФОРМАТИВНОЙ И ПОЛНОЙ НЕИНФОРМАТИВНОЙ КОМБИНАЦИИ ПРИЗНАКОВ

Построение описанной выше оценки $\hat{J}^{**}$ для полной информативной комбинации J^{**} основано на проверке гипотез о неинформативности и не включает рассмотрения информативных признаков. Следует отметить, что подобный подход обеспечивает малый риск, связанный с появлением ошибки, заключающейся в ошибочном включении неинформативных признаков в оценку полной неинформативной комбинации, тогда как от ошибки, заключающейся в невключении некоторых информативных признаков в оценку полной информативной комбинации, этот способ слабо застрахован. Избежать этого недостатка можно путем совместного рассмотрения информативной и неинформативной комбинаций.

Ниже мы рассмотрим построение совместного метода поиска информативной и неинформативной комбинаций признаков относительно многомерного среднего значения. Обозначим, как мы уже это делали, изучаемые m-мерные случайные величины через

$$\Xi = \{\xi_1, \xi_2, \ldots, \xi_j, \ldots, \xi_m\} \text{ и } \mathrm{H} = \{\eta_1, \eta_2, \ldots, \eta_i, \ldots, \eta_m\},$$

а соответствующие им математические ожидания через μ' и μ''. Представим каждую из случайных величин Ξ и H в виде двух векторов Ξ_k, Ξ_{m-k} и H_k, H_{m-k} порядка $1 \times k$ и $1 \times (m-k)$ соответственно, т. е.

$$\Xi = \{\Xi_k, \Xi_{m-k}\}, \tag{7.9}$$

$$\mathrm{H} = \{\mathrm{H}_k, \mathrm{H}_{m-k}\}, \tag{7.10}$$

а соответствующие им математические ожидания будут выглядеть следующим образом:

$$\mu' = \{\mu'_k,\ \mu'_{m-k}\}, \tag{7.11}$$

$$\mu'' = \{\mu''_k,\ \mu''_{m-k}\}. \tag{7.12}$$

Такому произвольному разбиению m-мерных случайных величин на k-мерный и $(m-k)$-мерный подвекторы соответствуют две комбинации признаков J_k и J_{m-k}, относительно которых можно сформулировать нулевую гипотезу $H_0 : J_k = J_k^0,\ J_{m-k} = J_{m-k}^0$, т. е. $H_0 : \mu' = \mu''$. Этой нулевой гипотезе можно противопоставить множество альтернатив $H_1 : \mu' \neq \mu''$, которое содержит три подмножества:

$$H_1^1 : \mu'_k \neq \mu''_k,\ \mu'_{m-k} = \mu''_{m-k}, \tag{7.13}$$

$$H_1^2 : \mu'_k = \mu''_k,\ \mu'_{m-k} \neq \mu''_{m-k}, \tag{7.14}$$

$$H_1^3 : \mu'_k \neq \mu''_k,\ \mu'_{m-k} \neq \mu''_{m-k}. \tag{7.15}$$

Множество альтернатив H_1^1 представляет собой утверждение, что комбинация k признаков J_k является информативной, тогда как J_{m-k} представляет собой неинформативную комбинацию. Гипотеза H_1^2 — это обратное утверждение относительно комбинаций признаков J_k и J_{m-k}, тогда как гипотеза H_1^3 утверждает, что обе комбинации как J_k, так и J_{m-k}, содержат информативные признаки. Таким образом, если в результате проверки нулевой гипотезы $H_0 : \mu' = \mu''$ оказалось, что ее следует отклонить и принять множество альтернатив $H_1 : \mu' \neq \mu''$, то любую из гипотез H_1^1, H_1^2, H_1^3 можно выбрать при заданном k как проверяемую, а любую из оставшихся двух гипотез как альтернативу.

При поиске полной информативной комбинации признаков и ее дополнения до полного множества признаков, которое является полной неинформативной комбинацией, в качестве нулевой гипотезы естественно выбрать H_1^1 или H_1^2. Пусть это будет H_1^1. Если этой гипотезе противопоставить H_1^3, то задача сведется просто к проверке гипотезы $J_{m-k} = J_{m-k}^0$ при альтернативе $J_{m-k} \neq J_{m-k}^0$, а комбинация J_k будет автоматически исключена из рассмотрения. В связи с этим целесообразно проверяемой гипотезе H_1^1 противопоставить в качестве альтернативы H_1^2.

Если через $\Lambda\left(H_1^1,\ J_{kl},\ J_{(m-k)l}\right)$ обозначить функцию правдоподобия в условиях гипотезы H_1^1 для фиксированного значения k и конкретного варианта с номером l разбиения множества признаков M на J_{kl}, $J_{(m-k)\,l}$, а через $\Lambda\left(H_1^2,\ J_{kl},\ J_{(m-k)l}\right)$ обозначить аналогичную функцию правдоподобия в условиях альтернативы H_1^2, то допустив, что распределения изучаемых случайных величин Ξ и H описываются m-мерными нормальными функциями с раз-

ными m-мерными средними и общей неизвестной ковариационной матрицей Σ, нетрудно построить критерий отношения правдоподобия

$$\lambda(J_{kl},\ J_{(m-k)\,l}) = \frac{\max\limits_{\mu} \Lambda\left(H_1^1,\ J_{kl},\ J_{(m-k)\,l}\right)}{\max\limits_{\mu} \Lambda\left(H_1^2,\ J_{kl},\ J_{(m-k)l}\right)}, \tag{7.16}$$

который будет представлять собой функцию, заданную на множестве разбиений набора признаков M на J_{kl} и $J_{(m-k)\,l}$.

Здесь мы не будем рассматривать процедуру построения критерия, которая уже была детально рассмотрена [38], а приведем только конечный результат, который в условиях сделанных ограничений будет представлен выражением

$$\lambda(J_{kl},\ J_{(m-k)l}) = \frac{\begin{vmatrix} S^{(0)}_{lkk} & S^{(0,1)}_{l(m-k)k} \\ S^{(0,1)}_{lk(m-k)} & S^{(1)}_{l(m-k)(m-k)} \end{vmatrix}^{\frac{n_1+n_2}{2}}}{\begin{vmatrix} S^{(1)}_{lkk} & S^{(1,0)}_{l(m-k)k} \\ S^{(1,0)}_{lk(m-k)} & S^{(0)}_{l(m-k)(m-k)} \end{vmatrix}^{\frac{n_1+n_2}{2}}}, \tag{7.17}$$

где $S^{(0)}_{lkk}$ — оценка ковариационной матрицы варианта с номером lk признаков, вычисленная в предположении, что эти k признаков неинформативны, $S^{(1)}_{lkk}$ — оценка ковариационной матрицы для этой же комбинации признаков, вычисленная в предположении, что данная комбинация k признаков информативна. Аналогично определены и выборочные матрицы $S^{(0)}_{l(m-k)(m-k)}$ и $S^{(1)}_{l(m-k)(m-k)}$ только по отношению к оставшимся признакам. Оценки $S^{(0,1)}_{l(m-k)k}$ и $S^{(0,1)}_{lk(m-k)}$ вычисляются в предположении, что оба соотношения $\mu^1_k = \mu_k$, $\mu'_{m-k} \neq \mu''_{m-k}$ выполнены, тогда как оценки $S^{(1,0)}_{l(m-k)k}$ и $S^{(1,0)}_{lk(m-k)}$ соответствуют обратному соотношению $\mu^1_k \neq \mu_k$, $\mu_{m-k} = \mu''_{m-k}$. Оценка искомого разбиения всего набора изучаемых признаков на полную информативную и полную неинформативную комбинации соответствует тому варианту, для которого выражение $\lambda(J_{kl}, J_{(m-k)l})$ достигает максимума. Естественно, что при поиске максимума можно ограничиться только отношением детерминантов в выражении (7.17), без возведения в степень $(n_1 + n_2)/2$.

Ниже мы рассмотрим алгоритм, реализующий описанный метод на практике.

7.4.1. Алгоритм выбора полной комбинации признаков, информативной относительно многомерных средних

Приведенный ниже алгоритм полезен при решении разнообразных геологических задач, требующих выделения из большого набора признаков такой комбинации, которая обладает наиболее

сильными различиями для сравниваемых объектов, например рудоносными и безрудными гранитными массивами.

Алгоритм представлен в виде трех частей, из которых первая (предложенная А. В. Гараниным), связанная с ранжированием признаков по их информативности, предназначена для сокращения рассматриваемых вариантов разбиения. Вторая часть содержит процедуры вычисления элементов выборочных матриц сумм смешанных произведений в условиях проверяемых предположений, тогда как третья часть посвящена описанию процедур поиска полной информативной комбинации признаков и проверке гипотез о равенстве многомерных средних по всему рассматриваемому набору признаков и по выбранной комбинации.

Часть I

1. Даны две m-мерные выборки объемом n_1 и n_2

$$X_1,\ X_2,\ \ldots,\ X_t,\ \ldots,\ X_{n_1} \text{ и } Y_1,\ Y_2,\ \ldots,\ Y_t,\ \ldots,\ Y_{n_2}, \quad (7.18)$$

где

$$X_t = \{x_{t1},\ x_{t2},\ \ldots,\ x_{tj},\ \ldots.\ x_{tm}\}, \quad (7.19)$$

$$Y_t = \{y_{t1},\ y_{t2},\ \ldots,\ y_{tj},\ \ldots,\ y_{tm}\} \quad (7.20)$$

2. Вычисляются

$$\bar{X} = \{\bar{x}_1,\ \bar{x}_2 \ldots,\ x_j,\ \ldots\ \bar{x}_m\} = \frac{1}{n_1}\sum_{t=1}^{n_1} X_t \quad (7.21)$$

и

$$\bar{Y} = \{\bar{y}_1,\ \bar{y}_2,\ \ldots,\ \bar{y}_j,\ \ldots,\ \bar{y}_m\} = \frac{1}{n_2}\sum_{t=1}^{n_2} Y_t, \quad (7.22)$$

где

$$\bar{x}_j = \frac{1}{n_1}\sum_{t=1}^{n_1} x_{tj}, \quad (7.23)$$

$$\bar{y}_j = \frac{1}{n_2}\sum_{t=1}^{n_2} y_{tj}. \quad (7.24)$$

3. Вычисляются элементы двух матриц

$$S_X = \{s_{Xij}\},\ i,\ j = 1,\ 2,\ \ldots,\ m, \quad (7.25)$$

$$S_Y = \{s_{Yij}\},\ i,\ j = 1,\ 2,\ \ldots,\ m, \quad (7.26)$$

где

$$s_{Xij} = \frac{1}{n_1 - 1}\sum_{t=1}^{n_1} (x_{ti} - \bar{x}_i)(x_{tj} - \bar{x}_j) =$$

$$= \frac{1}{n_1 - 1}\left[\sum_{t=1}^{n_1} x_{ti}x_{tj} - \frac{1}{n_1}\left(\sum_{t=1}^{n_1} x_{ti}\right)\left(\sum_{t=1}^{n_1} x_{tj}\right)\right], \quad (7.27)$$

$$s_{Yij} = \frac{1}{n_2 - 1} \sum_{t=1}^{n_2} (y_{ti} - \bar{y}_i)(y_{tj} - \bar{y}_j) =$$

$$= \frac{1}{n_2 - 1} \left[\sum_{t=1}^{n_2} y_{ti} y_{tj} - \frac{1}{n_2} \left(\sum_{t=1}^{n_2} y_{ti} \right) \left(\sum_{t=1}^{n_2} y_{tj} \right) \right]. \qquad (7.28)$$

В матричной записи

$$\mathbf{S}_X = \frac{1}{n_1 - 1} \sum_{t=1}^{n_1} \{X_t - \bar{X}\}' \{X_t - \bar{X}\}, \qquad (7.29)$$

$$\mathbf{S}_Y = \frac{1}{n_1 - 1} \sum_{t=1}^{n_2} \{Y_t - \bar{Y}\}' \{Y_t - \bar{Y}\}, \qquad (7.30)$$

где X_t и Y_t — векторы-строки.

4. Вычисляются m величин

$$D_{1j} = (\bar{x}_j - \bar{y}_j)^2 \Big/ \left(\frac{1}{n_1} s_{Xjj} + \frac{1}{n_2} s_{Yjj} \right) \qquad (7.31)$$

и выбирается то значение j_1, для которого $D_{1j} = \max$.

5. Вычисляется $m-1$ величин

$$D_{2j} = \{\bar{X}_j^{(2)} - \bar{Y}_j^{(2)}\} \left(\frac{1}{n_1} \mathbf{S}_{Xj}^{(2)} + \frac{1}{n_2} \mathbf{S}_{Yj}^{(2)} \right)^{-1} \{\bar{X}_j^{(2)} - \bar{Y}_j^{(2)}\}', \quad (7.32)$$

где $\bar{X}_j^{(2)}$ и $\bar{Y}_j^{(2)}$ — подвекторы порядка 1×2 векторов X и $\bar{Y}$, образованные теми $\bar{x}_{j_1}$ и $\bar{y}_{j_1}$, для которых достигнут максимум D_{1j}, и одним из оставшихся $\bar{x}_j$ и $\bar{y}_j$, для которых $j \neq j_1$. Величины D_{2j} — обобщенные расстояния между всеми парами двумерных векторов $\bar{X}_j^{(2)}$ и $\bar{Y}_j^{(2)}$, содержащими $\bar{x}_{j_1}$ и $\bar{y}_{j_1}$. $\mathbf{S}_{Xj}^{(2)}$ и $\mathbf{S}_{Yj}^{(2)}$ — подматрицы порядка 2×2 матриц $\mathbf{S}_X$ и $\mathbf{S}_Y$, соответствующие векторам $\bar{X}_j^{(2)}$ и $\bar{Y}_j^{(2)}$.

Выбирается то значение j_2, для которого $D_{2j} = \max$.

6. Процедура повторяется для всех $j = 1, 2, \ldots, m$. Так, для $j = k \leqslant m$ вычисляется $m-k$ значений

$$D_{kj} = \{\bar{X}_j^{(k)} - \bar{Y}_j^{(k)}\} \left(\frac{1}{n_1} \mathbf{S}_{Xj}^{(k)} + \frac{1}{n_2} \mathbf{S}_{Yj}^{(k)} \right)^{-1} \{\bar{X}_j^{(k)} - \bar{Y}_j^{(k)}\}' \quad (7.33)$$

и выбирается j_k, для которого $D_{kj} = \max_{j \neq j_1, j_2, \ldots, j_k}$ $\bar{X}_j^{(k)}$ и $Y_j^{(k)}$ векторы-строки порядка $1 \times k$, являющиеся подвекторами векторов $\bar{X}$ и $\bar{Y}$ и представляющие собой комбинации векторов $\bar{X}_j^{(k-1)}$ и $\bar{Y}_j^{(k-1)}$, для которых достигнут $\max D_{(k-1)j}$ на предыдущей стадии и одной из оставшихся пар $\bar{x}_j$, $\bar{y}_j$ ($j \neq j_1, j_2, \ldots, j_{k-1}$). При $k = m$ полученная последовательность номеров признаков $j_1, j_2, \ldots, j_m$ будет соответствовать их расположению от наилучшего к наихудшему.

7. В процессе всей дальнейшей работы признаки расположены в полученном порядке $j_1, j_2, \ldots, j_m$. Далее вычисляются элементы четырех матриц порядка $m \times m$

$$\mathbf{S}^{(0)} = \{s_{ij}^{(0)},\ i,\ j=1,\ 2,\ \ldots\ m\},\quad \mathbf{S}^{(1)} = \{s_{ij}^{(1)},\ i,\ j=1,\ 2,\ \ldots,\ m\}, \tag{7.34}$$

$$\mathbf{S}^{(01)} = \{s_{ij}^{(01)},\ i,\ j=1,\ 2,\ \ldots,\ m\},\quad \mathbf{S}^{(10)} = \{s_{ij}^{(10)},\ i,\ j=1,\ 2,\ \ldots,\ m\}. \tag{7.35}$$

Матрицы $\mathbf{S}^{(01)}$ и $\mathbf{S}^{(10)}$ — треугольные, т. е. $s_{ij}^{(01)} = 0$ при $i \geqslant j$ и $S_{ij}^{(10)} = 0$ при $i \geqslant j$. Вычисления $s_{ij}^{(0)}$, $s_{ij}^{(1)}$, $s_{ij}^{(01)}$, $s_{ij}^{(10)}$ производятся по формулам

$$s_{ij}^{(0)} = \sum_{t=1}^{n_1} (x_{ti} - \hat{\mu}_i)(x_{tj} - \hat{\mu}_j) + \sum_{t=1}^{n_2} (y_{ti} - \hat{\mu}_i)(y_{ti} - \hat{\mu}_j), \tag{7.36}$$

$$s_{ij}^{(1)} = \sum_{t=1}^{n_1} (x_{ti} - \bar{x}_i)(x_{tj} - \bar{x}_j) + \sum_{t=1}^{n_2} (y_{ti} - \bar{y}_i)(y_{tj} - \bar{y}_j), \tag{7.37}$$

$$s_{ij}^{(01)} = \sum_{t=1}^{n_1} (x_{ti} - \hat{\mu}_i)(x_{tj} - \bar{x}_j) + \sum_{t=1}^{n_2} (y_{ti} - \hat{\mu}_i)(y_{tj} - \bar{y}_j), \tag{7.38}$$

$$s_{ij}^{(10)} = \sum_{t=1}^{n_1} (x_{ti} - \bar{x}_i)(x_{tj} - \hat{\mu}_j) + \sum_{t=1}^{n_2} (y_{ti} - \bar{y}_i)(y_{tj} - \hat{\mu}_j). \tag{7.39}$$

В этих выражениях величины $\bar{x}_i$, $\bar{y}_i$, $\hat{\mu}_i$ подсчитываются по формулам

$$\bar{x}_i = \frac{1}{n_1} \sum_{t=1}^{n_1} x_{ti}, \tag{7.40}$$

$$\bar{y}_i = \frac{1}{n_2} \sum_{t=1}^{n_2} y_{ti}, \tag{7.41}$$

$$\hat{\mu}_i = \frac{1}{n_1 + n_2} \left(\sum_{t=1}^{n_1} x_{ti} + \sum_{t=1}^{n_2} y_{ti} \right). \tag{7.42}$$

8. Разбиению векторов $\bar{X}$ и $\bar{Y}$ на два подвектора $\bar{X}^{(k)}$ $\bar{X}^{(m-k)}$ и $\bar{Y}^{(k)}, \bar{Y}^{(m-k)}$, т. е.

$$\bar{X} = \{\bar{X}^{(k)},\ \bar{X}^{(m-k)}\}, \tag{7.43}$$

$$\bar{Y} = \{\bar{Y}^{(k)},\ \bar{Y}^{(m-k)}\}, \tag{7.44}$$

$$\bar{X}^{(k)} = \{\bar{x}_1,\ \bar{x}_2,\ \ldots,\ \bar{x}_k\}, \tag{7.45}$$

$$\bar{X}^{(m-k)} = \{\bar{x}_{k+1},\ \bar{x}_{k+2},\ \ldots,\ \bar{x}_m\}, \tag{7.46}$$

$$\bar{Y}^{(k)} = \{\bar{y}_1,\ \bar{y}_2,\ \ldots,\ \bar{y}_k\}, \tag{7.47}$$

$$\bar{Y}^{(m-k)} = \{\bar{y}_{k+1},\ \bar{y}_{k+2},\ \ldots,\ \bar{y}_m\}, \tag{7.48}$$

соответствует разбиение матриц $S^{(0)}$, $S^{(1)}$, $S^{(01)}$ и $S^{(10)}$ на следующие блоки:

$$S^{(0)} = \begin{pmatrix} S^{(0)}_{kk} & S^{(0)}_{k(m-k)} \\ S^{(0)}_{(m-k)k} & S^{(0)}_{(m-k)(m-k)} \end{pmatrix}, \tag{7.49}$$

$$S^{(1)} = \begin{pmatrix} S^{(1)}_{kk} & S^{(1)}_{k(m-k)} \\ S^{(1)}_{(m-k)k} & S^{(1)}_{(m-k)(m-k)} \end{pmatrix}, \tag{7.50}$$

$$S^{(01)} = \begin{pmatrix} S^{(01)}_{kk} & S^{(01)}_{k(m-k)} \\ 0 & S^{(01)}_{(m-k)(m-k)} \end{pmatrix}, \tag{7.51}$$

$$S^{(10)} = \begin{pmatrix} S^{(10)}_{kk} & S^{(10)}_{k(m-k)} \\ 0 & S^{(10)}_{(m-k)(m-k)} \end{pmatrix}, \tag{7.52}$$

где $S^{(0)}_{kk}$, $S^{(1)}_{kk}$ — квадратные симметричные матрицы порядка $k \times k$, $S^{(01)}_{kk}$, $S^{(10)}_{kk}$ — треугольные матрицы порядка $k \times k$, диагональные элементы которых равны 0, $S^{(01)}_{k(m-k)}$, $S^{(10)}_{k(m-k)}$ $S^{(01)}_{(m-k)k}$, $S^{(10)}_{(m-k)k}$ — матрицы порядка $k \times (m-k)$ и $(m-k) \times k$, причем $\left[S^{(01)}_{k(m-k)}\right]' = S^{(01)}_{(m-k)k}$, $\left[S^{(10)}_{k(m-k)}\right]' = S^{(10)}_{(m-k)k}$, где штрих означает операцию транспонирования; $S^{(0)}_{(m-k)(m-k)}$, $S^{(1)}_{(m-k)(m-k)}$ — квадратные симметричные матрицы порядка $(m-k) \times (m-k)$; $S^{(01)}_{(m-k)(m-k)}$, $S^{(10)}_{(m-k)(m-k)}$ — треугольные матрицы порядка $(m-k) \times (m-k)$, диагональные элементы которых равны 0; **0** — нулевые матрицы порядка $(m-k)\ k$.

Часть III

9. Для каждого заданного k строятся матрицы S_k и S_k^* порядка $m \times m$

$$S_k = \begin{pmatrix} S^{(1)}_{kk} & S^{(10)}_{k(m-k)} \\ \left[S^{(10)}_{k(m-k)}\right]' & S^{(0)}_{(m-k)(m-k)} \end{pmatrix}, \tag{7.53}$$

$$S_k^* = \begin{pmatrix} S^{(0)}_{kk} & S^{(01)}_{k(m-k)} \\ \left[S^{(01)}_{k(m-k)}\right]' & S^{(0)}_{(m-k)(m-k)} \end{pmatrix}, \tag{7.54}$$

где $\left[S^{(10)}_{k(m-k)}\right]'$ и $\left[S^{(01)}_{k(m-k)}\right]'$ — транспонированные матрицы $S^{(10)}_{k(m-1)}$ и $S^{(01)}_{k(m-k)}$ соответственно, причем $k = 1, 2, \ldots, m$.

10. Для каждого k вычисляется значение отношения детерминантов $|S_k^*|$ к $|S_k|$, т. е.

$$\lambda_k = \frac{|S_k^*|}{|S_k|}, \tag{7.55}$$

или

$$\lambda_k = \frac{\left| S^{(0)}_{kk} - S^{(01)}_{k(m-k)} \left(S^{(1)}_{(m-k)(m-k)} \right)^{-1} \left[S^{(01)}_{k(m-k)} \right]' \right| \cdot \left| S^{(1)}_{(m-k)(m-k)} \right|}{\left| S^{(1)}_{kk} - S^{(10)}_{k(m-k)} \left(S^{(0)}_{(m-k)(m-k)} \right)^{-1} \left[S^{(10)}_{k(m-k)} \right]' \right| \cdot \left| S^{(0)}_{(m-k)(m-k)} \right|}, \tag{7.56}$$

где $\left(S^{(1)}_{(m-k)(m-k)}\right)^{-1}$ и $\left(S^{(0)}_{(m-k)(m-k)}\right)^{-1}$ — матрицы, обратные матрицам $S^{(1)}_{(m-k)(m-k)}$ и $S^{(0)}_{(m-k)(m-k)}$ соответственно.

11. Значения λ_k или $\log \lambda_k$ вычисляются для всех $k = 1, 2, \ldots, m$ в упорядоченной последовательности $j_1, j_2, \ldots, j_k, \ldots, j_m$ и выбирается та комбинация $j \leqslant k_{\max}$, которой соответствует $\max \lambda_k$.

12. Для полученного значения k_0 вычисляется величина

$$V_0 = -\left(n_1 + n_2 - \frac{k_0}{2}\right) \ln \frac{\left| S^{(1)}_{k_0 k_0} \right|}{\left| S^{(0)}_{k_0 k_0} \right|}, \tag{7.57}$$

которая сравнивается с критическим значением $\chi^2_{q,\,k_0}$.

7.5. ВЫБОР НАИЛУЧШЕЙ ИНФОРМАТИВНОЙ КОМБИНАЦИИ

Полная информативная комбинация, определенная по отношению к некоторому заданному параметру, может включать такие признаки, которые хотя и являются информативными, но вносят столь малый вклад в величину, определяющую расхождение между значениями сравниваемых многомерных характеристик, что практически ухудшает надежность распознавания.

В связи с этим возникает задача выбора такой комбинации, которая была бы наилучшей для характеристик расхождения между значениями изучаемого многомерного параметра.

Для определения наилучшей информативной комбинации можно воспользоваться подходом, описанным в разделе 6.4 при сравнении расхождений, охарактеризованных разным числом признаков. В этом разделе в качестве меры расхождения между сравниваемыми объектами предлагалось использовать параметр $b = \frac{a}{f}$, представляющий собой частное от деления параметра нецентральности a нецентрального χ^2, распределения на соответствующее число степеней свободы f.

Необходимо отметить, что при проверке весьма многих многомерных гипотез используются критерии, которые в условиях альтернативы распределены как нецентральный χ^2. Таким образом, каждому набору k признаков J_k можно поставить в соответствие значение критерия $V(J_k)$ и число степеней свободы f_k. В результате частное от деления $V(J_k)/f_k$ можно рассматривать как функцию, заданную на множестве всех возможных вариантов разбиения множества признаков M на две части J_k и J_{m-k}, причем эта функция будет характеризовать средний вклад, который обеспечивает в различиях данная комбинация признаков на одну степень сво-

боды. Очевидным решением будет тот вариант разбиения всего множества признаков M на две части, которому соответствует максимум отношения $V(J_k)/f_k$.

Ниже приведены конкретные алгоритмы поиска наилучшей информативной комбинации относительно многомерных средних и относительно ковариационных матриц.

7.5.1. Алгоритм выбора наилучшей информативной комбинации относительно многомерных средних

Даны две m-мерные выборки, объемы которых n_1 и n_2

$$X_1, X_2, \ldots, X_t, \ldots, X_{n_1} \quad \text{и} \quad Y_1, Y_2, \ldots, Y_t, \ldots, Y_{n_2}, \tag{7.58}$$

где

$$X_t = \{x_{t1}, x_{t2}, \ldots, x_{tj}, \ldots, x_{tm}\}, \tag{7.59}$$

$$Y_t = \{y_{t1}, y_{t2}, \ldots, y_{tj}, \ldots, y_{tm}\}. \tag{7.60}$$

1. Проводится процедура ранжирования признаков, описанная в первой части раздела 7.4.1, т. е. для каждого $k = 1, 2, \ldots, m$, начиная с $k = 1$, вычисляется $m-k$ значений обобщенного k-мерного расстояния

$$D_{kj} = \{\overline{X}_j^{(k)} - \overline{Y}_j^{(k)}\} \left(\frac{1}{n_1} \mathbf{S}_{Xj}^{(k)} + \frac{1}{n_2} \mathbf{S}_{Yj}^{(k)}\right)^{-1} \{\overline{X}_j^{(k)} - \overline{Y}_j^{(k)}\}' \tag{7.61}$$

выбирается j_k, для которого $D_{kj} = \max$. В данном случае $\overline{X}_j^{(k)}$ и $\overline{Y}_j^{(k)}$ — векторы-строки порядка $1 \times k$, представляющие собой комбинации $(k-1)$-мерных векторов $\overline{X}_{j\,(k-1)}^{(k-1)}$, $\overline{Y}_{j\,(k-1)}^{(k-1)}$, для которых достигнут $\max D_{k-1}$ на предыдущей стадии. В результате будет получена новая последовательность номеров признаков $j_1, j_2, \ldots, j_m$, которая соответствует их расположению от наилучшего к наихудшему в смысле информативности.

2. Полученную новую последовательность признаков, не меняя расположения, можно разделить $m-1$ способами на две части, состоящие из k и $m-k$ признаков. При $k = m$ получим полный набор m признаков без разделения.

Если относительно ковариационных матриц Σ_1 и Σ_2 сравниваемых выборок справедливо предположение $\Sigma_1 = \Sigma_2 = \Sigma$, то для последующего поиска наилучшей информативной комбинации можно воспользоваться критерием (5.24), который для любого $k = 1, 2, \ldots, m$ будет иметь вид

$$V(J_k) = -\left(n_1 + n_2 - 2 - \frac{k}{2}\right) \ln \frac{|\mathbf{S}_k^{(1)}|}{|\mathbf{S}_k^{(0)}|}, \tag{7.62}$$

где k — число признаков, вошедших в рассматриваемую комбинацию, $\mathbf{S}_k^{(0)}$ — выборочная ковариационная матрица, соответствующая k признакам, элементы которой вычислены в предположении,

что верна гипотеза о равенстве k-мерных средних, $\mathbf{S}_k^{(1)}$ — аналогичная ковариационная матрица, вычисленная в предположении, что k-мерные средние различны, т. е. данная комбинация признаков J_k является информативной. Элементы этих матриц вычисляются по формулам

$$s_{ij}^{(0)} = \sum_{t=1}^{n_1} (x_{ti} - \hat{\mu}_i)(x_{tj} - \hat{\mu}_j) + \sum_{t=1}^{n_2} (y_{ti} - \hat{\mu}_i)(y_{tj} - \hat{\mu}_j), \quad (7.63)$$

$$s_{ij}^{(1)} = \sum_{t=1}^{n_1} (x_{ti} - \bar{x}_i)(x_{tj} - \bar{x}_j) + \sum_{t=1}^{n_2} (y_{ti} - \bar{y}_i)(y_{tj} - \bar{y}_j), \quad (7.64)$$

где

$$\bar{x}_i = \frac{1}{n_1} \sum_{t=1}^{n_1} x_{ti}, \quad (7.65)$$

$$\bar{y}_i = \frac{1}{n_2} \sum_{t=1}^{n_2} y_{ti}, \quad (7.66)$$

$$\hat{\mu}_i = \frac{1}{n_1 - n_2} \left(\sum_{t=1}^{n_1} x_{ti} + \sum_{t=1}^{n_2} y_{ti} \right). \quad (7.67)$$

Естественно, что элементы $s_{ij}^{(0)}$ и $s_{ij}^{(1)}$ для полного множества рассматриваемых признаков образуют две матрицы $\mathbf{S}^{(0)}$ и $\mathbf{S}^{(1)}$ порядка $m \times m$. Заметим, что элементы этих матриц для последующих вычислений удобно расположить в новом порядке, полученном в результате процедуры ранжирования.

Обозначим через v_k отношения $V(J_k)/f_k$, где f_k число степеней свободы. В результате получим последовательность

$$v_1, v_2, \ldots, v_k, \ldots, v_m. \quad (7.68)$$

В данном случае число степеней свободы $f_k = k$, и для получения последовательности (7.68) достаточно каждое значение критерия (7.62) поделить на k. Искомая наилучшая комбинация признаков будет соответствовать тому значению k_0, для которого $v_{k_0} = \max_k$. Естественно, что в эту комбинацию входят все признаки с номерами, меньшими или равными k_0. Для полученного значения k_0 уже вычислено значение

$$V(J_{k_0}) = -\left(n_1 + n_2 - 2 - \frac{k_0}{2} \right) \ln \frac{|\mathbf{S}_{k_0}^{(1)}|}{|\mathbf{S}_{k_0}^{(0)}|}, \quad (7.69)$$

которое в условиях гипотезы о равенстве k_0-мерных средних является значением случайной величины, распределенной как χ^2 с k_0 степенями свободы. Таким образом, если окажется, что $V(J_{k_0}) \leqslant \chi^2_{q,\,k_0}$, то из этого следует, что даже наилучшая комбинация признаков не содержит какой-либо информации о различиях изучаемых геологических объектов и никакой комбинацией из

всех рассматриваемых m признаков нельзя пользоваться для последующего распознавания. Если же $V(J_{k_0}) > \chi^2_{q,k_0}$, то полученную комбинацию J_{k_0} можно считать наилучшей информативной комбинацией признаков, относительно многомерных средних в условиях равенства ковариационных матриц.

3. Если же относительно ковариационных матриц двух изучаемых случайных величин Ξ и H нельзя ввести допущение о равенстве и их следует рассматривать как неравные, то вместо критерия (7.62) следует воспользоваться критерием Джеймса, определенным выражением (5.38) в главе 5. Нетрудно видеть, что для случая двух выборок этот критерий совпадает со значением обобщенного расстояния D_{kj}, используемого для ранжирования и определенного формулой (7.61).

Таким образом, полученная упорядоченная последовательность $\max\limits_{j} D_{kj}$ будет представлять собой упорядоченный набор значений критерия Джеймса, который, как отмечалось в главе 5, в условиях нулевой гипотезы распределен как нецентральный χ^2 с числом степеней свободы, равным числу признаков в случае двух выборок.

Следовательно, для выбора наилучшей информативной комбинации относительно многомерных средних, в условиях неравных ковариационных матриц нужно получить последовательность отношений $v_k = \max\limits_{j} D_{kj,k}$.

Таким образом, для решения поставленной задачи в последовательности

$$v_1,\ v_2,\ \ldots,\ v_k,\ \ldots,\ v_m \tag{7.70}$$

нужно найти то значение k_0, которому соответствует $\max\limits_{k} v_k$.

Если окажется, что

$$k_0 v_{k_0} \leqslant \chi^2_{q,k_0}\left(A + B\chi^2_{q,k_0}\right), \tag{7.71}$$

то из этого следует, что даже наилучшая комбинация не несет информации относительно k_0-мерных средних. Если же $k_0 v_{k_0} > \chi^2_{q,k_0}\left(A + B\chi^2_{q,k_0}\right)$, то выявленную комбинацию k_0 признаков можно рассматривать как наилучшую информативную относительно средних значений в условиях неравных ковариационных матриц. Числа A и B определены формулами (5.42) и (5.43).

7.5.2. Алгоритм выбора наилучшей информативной комбинации относительно ковариационных матриц

Задача выбора наилучшей информативной комбинации признаков относительно ковариационных матриц возникает в тех случаях, когда многомерные средние значения не несут существенной информации о различиях между сравниваемыми геологическими объектами и для последующего распознавания приходится строить не

линейные, а квадратичные решающие правила. Вполне естественно в подобной ситуации возникновение задачи о выборе тех признаков, по которым это решающее правило следует строить. Используя общие принципы выбора наилучших информативных комбинаций, изложенные в начале раздела 7.5, построим следующий алгоритм решения данной задачи.

1. Даны две m-мерные выборки объемом n_1 и n_2

$$X_1,\ X_2,\ \ldots,\ X_t,\ \ldots,\ X_{n_1} \text{ и } Y_1,\ Y_2,\ \ldots,\ Y_t,\ \ldots,\ X_{n_2}, \tag{7.72}$$

где

$$X_t = \{x_{t1},\ x_{t2},\ \ldots,\ x_{tj},\ \ldots,\ x_{tm}\}, \tag{7.73}$$

$$Y_t = \{y_{t1},\ y_{t2},\ \ldots,\ y_{tj},\ \ldots,\ y_{tm}\}. \tag{7.74}$$

2. Вычисляются элементы трех матриц $\mathbf{S}_X$, $\mathbf{S}_Y$ и $\mathbf{S}$, каждая из которых имеет порядок $m \times m$

$$\mathbf{S}_X = \frac{1}{n_1 - 1} \sum_{t=1}^{n_1} \{X_t - \overline{X}\}' \{X_t - \overline{X}\}, \tag{7.75}$$

$$\mathbf{S}_Y = \frac{1}{n_2 - 1} \sum_{t=1}^{n_2} \{Y_t - \overline{Y}\}' \{Y_t - \overline{Y}\}, \tag{7.76}$$

$$\mathbf{S} = \frac{1}{n_1 + n_2 - 2} [(n_1 - 1)\,\mathbf{S}_X + (n_2 - 1)\,\mathbf{S}_Y], \tag{7.77}$$

где

$$\overline{X} = \frac{1}{n_1} \sum_{t=1}^{n_1} X_t, \qquad \overline{Y} = \frac{1}{n_2} \sum_{t=1}^{n_2} Y_t. \tag{7.78}$$

3. Для каждого из m признаков вычисляются значения

$$V_j^{(1)} = (n_1 - 1) \ln \frac{s_{jj}}{s_{Xjj}} + (n_2 - 1) \ln \frac{s_{jj}}{s_{Yjj}} \tag{7.79}$$

$$(j = 1, 2,\ \ldots,\ m)$$

и выбирается то значение $j = j_1$, для которого $V_j^{(1)} = \max\limits_j$.

4. Рассматриваются m — 1 пар, которые образуют признак j_1 с оставшимися m—1 признаками и для каждой пары вычисляется величина

$$V_j^{(2)} = (n_2 - 1)\,\mathrm{en}\,\frac{|\mathbf{S}_j^{(2)}|}{|\mathbf{S}_{Xj}^{(2)}|} + (n_2 - 1) \ln \frac{|\mathbf{S}_j^{(2)}|}{|\mathbf{S}_{Yj}^{(2)}|}, \tag{7.80}$$

где $j \neq j_1$, $\mathbf{S}_j^{(2)}$, $\mathbf{S}_{Xj}^{(2)}$, $\mathbf{S}_{Yj}^{(2)}$ — подматрицы порядка 2×2 ковариационных матриц $\mathbf{S}$, $\mathbf{S}_X$, $\mathbf{S}_Y$, соответствующие рассматриваемой паре признаков $j_1 j$.

Выбирается такая пара признаков $j_1 j = j_1 j_2$, для которой, $V_j^{(2)} = \max\limits_{j \neq j_1}$.

5. Для k-того шага процедуры вычисляется $m-k$ значений

$$V_j^{(k)} = (n_1 - 1) \ln \frac{|\mathbf{S}_j^{(k)}|}{|\mathbf{S}_{Xj}^{(k)}|} + (n_2 - 1) \ln \frac{|\mathbf{S}_j^{(k)}|}{|\mathbf{S}_{Yj}^{(k)}|}, \qquad (7.81)$$

где $j \neq j_1, j_2, \ldots, j_{k-1}$, $\mathbf{S}_j^{(k)}$, $\mathbf{S}_{Xj}^{(k)}$, $\mathbf{S}_{Yj}^{(k)}$ — матрицы порядка $k \times k$ соответствующие комплексу признаков $j_1, j_2, \ldots, j_{k-1}, j$ являющиеся подматрицами матриц $\mathbf{S}$, $\mathbf{S}_X$, $\mathbf{S}_Y$.

6. Процедура повторяется для всех $k = 1, 2, \ldots, m$. В результате будет получена последовательность значений $\max\limits_j V_0$, которые для простоты обозначим V_k, т. е.

$$V_1, V_2, \ldots, V_k, \ldots, V_m. \qquad (7.82)$$

7. Для каждого k вычисляется значение $2V_k/k(k+1)$ и отыскивается такое значение k_0, для которого это отношение максимально. В данном случае V_k разделено на число степеней свободы критерия, которое в случае двух выборок равно $k(k+1)/2$ в соответствии с результатами, изложенными в главе 5.

8. Если величина V_{k_0}, соответствующая найденному значению k_0, окажется меньше допустимого $V_{q,\, k(k+1)/2}$, соответствующего уровню значимости q и $k(k+1)/2$ степеням свободы, то изучаемая комбинация признаков неинформативна относительно ковариационных матриц. Если же $V_{k_0} > V_{q,\, \frac{k(k+1)}{2}}$, то найденная комбинация J_{k_0} является наилучшей информативной комбинацией относительно ковариационных матриц. Критическое значение $V_{q,\, k(k+1)/2}$ определяется так, как это описано в главе 5.

7.6. ПРИМЕР

В качестве примера выбора полной информативной комбинации признаков относительно средних значений, при условии равенства ковариационных матриц, воспользуемся данными химического анализа пород сыверлинской и ергаланской свит трапповой формации Норильского плато, которые уже приводились ранее автором [38] в качестве примера иллюстрирующего метод выбора полной информативной комбинации признаков. Однако с целью сокращения объема вычислений в указанной работе была значительно упрощена процедура ранжирования признаков. Вместо алгоритма, описанного в первой части раздела 7.4.1, для ранжирования признаков вычислялись только отношения

$$d_j = (\bar{x}_j - \bar{y}_j)^2 / s_{jj}^{(1)}, \qquad (7.83)$$

где $\bar{x}_j$ и $\bar{y}_j$ — средние арифметические для признака с номером j, соответствующее двум сравниваемым выборкам, $s_{jj}^{(1)}$ — оценка дисперсии признака с номером j, которая совпадает с соответствую-

щим диагональным элементом матрицы $\mathbf{S}^{(1)}$, определенной выражением (7.37).

Расположение вычисленных значений d_j в ряд

$$d_{j_1} \geqslant d_{j_2} \geqslant \ldots \geqslant d_{j_i} \geqslant \ldots \geqslant d_{j_m} \qquad (7.84)$$

определяло новую последовательность признаков.

Однако такой метод хотя и является простым, обладает существенным недостатком — не учитывает зависимостей между признаками, т. е. основан на предположении, что признаки независимы.

Кроме того, в указанной работе [38] при вычислении отношения правдоподобия λ_k, определенного формулой (7.56), было допущено упрощение, заключающееся в том, что при разбиении полного набора признаков на две части J_k и J_{m-k} эти две части предполагались независимыми, что в значительной степени упрощало вычисления, придавая формуле (7.56) следующий более простой вид:

$$\lambda_k = \frac{\left|\mathbf{S}^{(0)}_{kk}\right| \cdot \left|\mathbf{S}^{(1)}_{(m-k)\,(m-k)}\right|}{\left|\mathbf{S}^{(1)}_{kk}\right| \cdot \left|\mathbf{S}^{(0)}_{(m-k)\,(m-k)}\right|}. \qquad (7.85)$$

В приводимом ниже примере эти упрощения устранены и все вычисления проводились по формулам, рекомендуемым в разделе 7.4.1 настоящей книги.

Аналитические данные были предоставлены автору Ю. А. Тумановской. Для сокращения объема вычислений вместо десяти обычно приводимых компонентов рассматривались следующие четыре суммы:

$$\xi_1 = SiO_2 + TiO_2,\ \xi_2 = Al_2O_3 + Fe_2O_3,$$
$$\xi_3 = MgO + FeO + CaO + MnO,\ \xi_4 = Na_2O + K_2O.$$

В нашей задаче необходимо выяснить, существует ли разница в среднем составе траппов, являющихся двумя последовательными порциями деятельности магматического очага, или же этой разницы нет и проведенная стратиграфическая граница не соответствует изменениям в среднем химическом составе выделенных свит. Объединение окислов в четыре перечисленные суммы может привести к тому, что разница между составами будет завуалирована и неустановлена. Однако, если даже при таком объединении разница окажется существенной, то при раздельном рассмотрении компонентов она тем более не может быть случайной. Если расхождение между четырехмерными средними составами рассматриваемых свит окажется существенным, то вполне естественна постановка задачи о выборе полной информативной и полной неинформативной комбинации признаков; первой, как обеспечивающей различия между свитами, а второй, как характеризующей сходство между ними.

В таблицах 18 и 19 приведены значения упомянутых сумм, полученные по данным опробования ергаланской и сыверлинской свит. Ранее автором [38] был проведен анализ этих данных с целью вы-

бора полной информативной и полной неинформативной комбинации признаков. При этом использовался упрощенный метод ранжирования признаков, основанный на рассмотрении разностей

Таблица 18

Значения сумм ξ_1, ξ_2, ξ_3, ξ_4 для пород ергаланской свиты

t	ξ_1	ξ_2	ξ_3	ξ_4	t	ξ_1	ξ_2	ξ_3	ξ_4
1	46,14	19,53	20,79	5,95	9	53,24	16,51	22,72	5,31
2	51,31	19,27	16,62	6,06	10	48,76	19,59	21,43	4,44
3	45,05	19,75	22,35	5,85	11	46,18	16,62	34,69	2,51
4	43,51	21,20	25,14	5,02	12	50,30	17,31	22,31	3,82
5	50,76	22,50	18,74	3,48	13	47,86	19,11	22,57	3,66
6	46,45	20,26	21,37	5,66	14	48,85	21,00	21,00	3,09
7	48,49	20,29	20,44	4,93	15	47,30	19,71	21,51	2,84
8	51,20	22,99	17,35	4,84	16	40,15	15,54	27,35	2,58

Таблица 19

Значения ξ_1, ξ_2, ξ_3 и ξ_4 для пород сыверлинской свиты

t	ξ_1	ξ_2	ξ_3	ξ_4	t	ξ_1	ξ_2	ξ_3	ξ_4
1	53,46	20,33	17,93	3,57	12	50,99	19,14	22,83	2,64
2	54,34	16,65	90,92	4,08	13	50,54	19,31	24,87	3,34
3	53,10	19,39	20,17	3,70	14	50,48	19,08	25,78	2,82
4	54,25	18,39	19,90	4,08	15	52,11	18,45	24,34	2,66
5	54,70	17,73	20,30	4,26	16	52,74	20,76	19,08	3,74
6	50,49	18,50	24,31	2,41	17	41,56	12,61	33,95	1,20
7	52,40	19,55	21,06	3,24	18	41,88	12,54	34,55	0,37
8	52,54	16,51	21,80	4,12	19	43,76	12,02	37,62	1,78
9	49,82	18,57	23,76	2,78	20	42,90	11,84	38,12	0,80
10	52,32	18,89	23,18	2,47	21	45,77	14,61	34,41	1,15
11	51,29	19,08	19,57	5,30	22	47,92	14,75	31,23	0,78

$$d_j = \frac{(\bar{x}_j - \bar{y}_j)^2}{s_{jj}^{(1)}}. \qquad (7.86)$$

Как уже было отмечено выше, этот упрощенный метод ранжирования не учитывает зависимостей между признаками. В результате признаки были расположены в следующем порядке: 4, 2, 1, 3. В этом же порядке расположены строки и столбцы приведенных ниже выборочных ковариационных матриц

$$\mathbf{S}^{(1)} = \begin{vmatrix} 1{,}75 & 1{,}97 & 3{,}64 & -4{,}69 \\ 1{,}97 & 6{,}53 & 3{,}58 & -10{,}42 \\ 3{,}64 & 3{,}58 & 16{,}03 & -15{,}53 \\ -4{,}69 & -10{,}42 & -15{,}53 & 31{,}97 \end{vmatrix},$$

$$S^{(0)} = \begin{vmatrix} 2{,}39 & 2{,}92 & 2{,}67 & -5{,}94 \\ 2{,}92 & 7{,}92 & 2{,}19 & -12{,}79 \\ 2{,}67 & 2{,}19 & 17{,}39 & -13{,}75 \\ -5{,}94 & -12{,}79 & -13{,}75 & 34{,}31 \end{vmatrix}.$$

С помощью упрощенной формулы были получены следующие значения критерия (7.85) в степени 2/38 для $k = 0, 1, 2, 3, 4$.

$$\lambda_0^{\frac{2}{38}} = \frac{1074}{789} = 1{,}360, \qquad \lambda_1^{\frac{2}{38}} = \frac{2{,}39 \cdot 783}{1{,}75 \cdot 1026} = 1{,}042,$$

$$\lambda_2^{\frac{2}{38}} = \frac{10{,}40 \cdot 270}{7{,}55 \cdot 407} = 0{,}912, \qquad \lambda_3^{\frac{2}{38}} = \frac{149 \cdot 31{,}97}{77 \cdot 34{,}31} = 1{,}820,$$

$$\lambda_4^{\frac{2}{38}} = \frac{789}{1074} = 0{,}735.$$

Таким образом, среди приведенных значений $\lambda_k^{\frac{2}{n_1+n_2}}$ максимальное, равное 1,820, соответствует $k = 3$, что дает возможность разделить набор признаков на две части, из которых первая (4, 2, 1) состоящая из трех признаков, является информативной, а вторая (3), содержащая только один третий признак, неинформативная.

После исключения из рассмотрения третьего признака, представляющего собой сумму содержаний двухвалентных окислов, была проведена проверка гипотезы о равенстве трехмерных средних, соответствующих полной информативной комбинации признаков, в которую вошли $Na_2O + K_2O$, $Al_2O_3 + Fe_2O_3$ и $SiO_2 + TiO_2$. Используя критерий отношения правдоподобия (7.62), получим

$$V = -\left(16 + 22 - 2 - \frac{3}{2}\right) \ln \frac{77}{149} =$$

$$= -34{,}5 \ln (0{,}516) = 34{,}5 \cdot 0{,}661 = 22{,}8.$$

Допустимое значение χ^2, соответствующее уровню значимости 0,05 и трем степеням свободы, равно 7,815, а вычисленное нами значение $V = 22{,}8$ значительно превышает 7,815, что позволяет уверенно отклонить гипотезу о равенстве трехмерных средних, т. е. сделать обоснованный вывод о их различии.

Как уже было отмечено, в данном примере был сделан ряд упрощений в процедуре ранжирования признаков и в самом алгоритме вычисления критерия, предназначенного для выбора искомых комбинаций. Эти упрощения сводились к недоучету зависимостей между признаками и некоторыми их комбинациями. Ниже мы проанализируем те же данные, но с помощью метода, свободного от перечисленных недостатков. Так, для ранжирования признаков

была использована процедура, описанная в первой части раздела 7.4.1 этой главы. В результате была получена следующая последовательность признаков: 4, 1, 2, 3. Нетрудно видеть, что от первого варианта такой последовательности, полученной упрощенным методом, эта последовательность отличается положением признаков с номерами 1 и 2.

Для выбора полной информативной и полной неинформативной комбинации был использован критерий, описанный во второй и третьей частях раздела 7.4.1, который учитывает зависимости между комбинациями рассматриваемых признаков. В результате была получена следующая последовательность значений $\ln \lambda_k^{2/(n_1+n_2)}$ для $k = 0, 1, 2, 3, 4$:

$$\overline{\lambda_0^{2/38}} = -0{,}306, \quad \overline{\lambda_1^{2/38}} = -0{,}050, \quad \overline{\lambda_2^{2/38}} = 0{,}406,$$

$$\overline{\lambda_3^{2/38}} = 0{,}601, \quad \overline{\lambda_4^{2/38}} = 0{,}306.$$

Нетрудно видеть, что максимальное значение критерия, равное 0,601, соответствует разбиению последовательности признаков на две комбинации (4, 1, 2) и (3), из которых первая является информативной, а вторая неинформативной. Таким образом, оказалось, что в данном случае в результате применения упрощенного и полного методов был получен почти один и тот же результат (расхождение в последовательности признаков), что дает основание в случае необходимости, особенно при недостаточных возможностях имеющейся в наличии ЭВМ, использовать упрощенный метод. Однако следует отметить, что несмотря на совпадение этих методов полной информативной и неинформативной комбинации различная ранжировка существенно влияет на результат выбора наилучшей информативной комбинации.

Ниже мы рассмотрим результаты применения описанного в разделе 7.51 метода выбора наилучшей информативной комбинации. В соответствии с разделом 7.51 критерий для выбора наилучшей информативной комбинации относительно средних значений в условиях равных ковариационных матриц определен выражением

$$v_k = \frac{V(J_k)}{f_k}, \tag{7.87}$$

где $k = 1, 2, \ldots, m$, f_k — число степеней свободы, а $V(J_k)$ — критерий отношения правдоподобия, определенный выражением (5.24). Таким образом v_k представляет собой функцию, заданную на множестве разбиений упорядоченной последовательности при-

знаков на две части. Используя формулу (7.87) и последовательность (4, 1, 2, 3), вычислим значения v_k для наших данных

$$\frac{V(4)}{1} = \frac{-\left(n_1 + n_2 - 2 - \frac{k}{2}\right)}{1} \ln \frac{|S^{(1)}|}{|S^{(0)}|} =$$

$$= -\left(16 + 22 - 2 - \frac{1}{2}\right) \ln \frac{1{,}75}{2{,}39} 11{,}131;$$

$$\frac{V(4,1)}{2} = \frac{-\left(16 + 22 - 2 - \frac{2}{2}\right)}{2} \ln \frac{\begin{vmatrix} 1{,}75 & 3{,}64 \\ 3{,}64 & 16{,}03 \end{vmatrix}}{\begin{vmatrix} 2{,}39 & 2{,}67 \\ 2{,}67 & 17{,}39 \end{vmatrix}} = 14{,}81;$$

$$\frac{V(4,\ 1,\ 2)}{3} = \frac{-\left(16 + 22 - 2 - \frac{3}{2}\right)}{3} \ln \frac{\begin{vmatrix} 1{,}75 & 1{,}97 & 3{,}64 \\ 1{,}97 & 6{,}53 & 3{,}58 \\ 3{,}64 & 3{,}58 & 16{,}03 \end{vmatrix}}{\begin{vmatrix} 2{,}39 & 2{,}92 & 2{,}67 \\ 2{,}92 & 7{,}92 & 2{,}19 \\ 2{,}67 & 2{,}19 & 17{,}39 \end{vmatrix}} = 9{,}76;$$

$$\frac{V(4,\ 1,\ 2,\ 3)}{4} = -\frac{\left(16 + 22 - 2 - \frac{4}{2}\right)}{4} \ln \frac{|S^{(0)}|}{|S^{(1)}|} = 2{,}63.$$

Таким образом, в результате проведенных вычислений получена последовательность значений критерия (табл. 20).

Таблица 20

Последовательность значений критерия

I_k	4	4, 1	4, 1, 2	4, 1, 2, 3
$V(I_k)/t_k$	11,13	14,81	9,76	2,63
f_k	1	2	3	4

Из этой таблицы видно, что v_k достигает максимума на втором шаге, что соответствует комбинации двух признаков — четвертого и первого, т. е. комбинации суммы щелочных окислов и суммы четырехвалентных окислов. Таким образом, второй признак хотя и входит в полную информативную комбинацию, в наилучшую информативную комбинацию не вошел. Установленная в результате вычислений комбинация двух признаков является наилучшей для различения сыверлинской и ергаланской свит.

ДИСКРИМИНАНТНЫЙ АНАЛИЗ

В предыдущей главе мы упоминали термин «дискриминантная функция», так как информативные признаки в подавляющем большинстве случаев используются для ее построения. В этой же главе мы рассмотрим типовые геологические задачи, связанные с дискриминантным анализом, и опишем процедуры построения дискриминантных функций.

8.1. ПОСТАНОВКА ЗАДАЧ ДИСКРИМИНАНТНОГО АНАЛИЗА

Почти в любом геологическом исследовании приходится сталкиваться с ситуацией, когда нужно сделать вывод о принадлежности одного объекта или их набора к одной из нескольких заранее заданных групп. Иногда эти выводы делаются по одному, а иногда по комплексу признаков, причем до последнего десятилетия все эти выводы были интуитивными, что приводило к существенному влиянию субъективных факторов.

Однако начиная с середины шестидесятых годов в геологии для решения упомянутой задачи различные исследователи начинают применять математические методы. Примерами могут служить работы Ш. А. Губермана и др. [22], А. Н. Бугайца [9, 10], А. А. Дорофеюка [18], Ю. А. Воронина [16] и др.

Необходимо отметить, что в проблеме применения математических методов в геологии при решении задач классификационного отнесения объектов к одной из заданных групп сразу же наметились два основных направления. Первое из них, называемое распознаванием образов, является чисто эвристическим и не содержит в основе своей каких-либо теорий. Существенным его достоинством является возможность использования для распознавания качественных характеристик геологических объектов, что в значительной степени расширяет возможности этого направления. В данной книге мы это направление рассматривать не будем, так как оно не является статистическим, и отсылаем читателя к другой работе [18].

Второе направление в решении задач классификационного отнесения изучаемых объектов к одной из заданных групп по комплексу признаков называется дискриминантным анализом. В отличие от методов первого направления, методы дискриминантного анализа требуют только количественных данных, что несколько сужает области их применения. Однако в основе этого анализа лежит хорошо развитая математическая теория, что позволяет учитывать риск, связанный с принятием ошибочных решений.

Формально задача дискриминантного анализа сводится к следующему. Пусть $A_1, A_2, \ldots, A_l, \ldots, A_k$ — k множеств объектов. Не нарушая общности, мы ограничимся для простоты рассмотрением только двух множеств A_1 и A_2. Каждому из множеств A_1 и A_2 поставим в соответствие m-мерную случайную величину $\Xi = \{\xi_1, \xi_2, \ldots, \xi_j, \ldots, \xi_m\}$ и $\mathrm{H} = \{\eta_1, \eta_2, \ldots, \eta_j, \ldots, \eta_m\}$, причем известно, что некоторые m-мерные параметры θ_1 и θ_2, присущие Ξ и H, различны, т. е. $\theta_1 \neq \theta_2$. В большинстве случаев под θ_1 и θ_2 понимаются многомерные средние, но не исключено и рассмотрение ковариационных матриц или же многомерных средних и ковариационных матриц совместно.

Допустим, что из каждой совокупности A_1 и A_2 объектов a взята выборка объемом n_1 и n_2 соответственно и по выборочным данным $X_t = \{x_{t1}, x_{t2}, \ldots, x_{tj}, \ldots, x_{tm}\}$ и $Y_t = \{y_{t1}, y_{t2}, \ldots, y_{tj}, \ldots, y_{tm}\}$ требуется построить решающее правило D, которое бы позволяло относить объекты из третьей совокупности A, представляющей собой смесь объектов из A_1 и A_2 к A_1 и к A_2. Обозначим результат m-мерного наблюдения из совокупности A, которая требует распознавания через $Z_t = \{z_{t1}, z_{t2}, \ldots, z_{tj}, \ldots, z_{tm}\}$. Таким образом, наше решающее правило должно заключаться в том, что рассматриваемое наблюдение Z_t относится к совокупности A_1 (например, к перспективным в отношении рудоносности образования), если оно характеризуется определенным множеством значений $\{z_1, z_2, \ldots, z_j, \ldots, z_m\}$, а при других значениях $\{z_1, z_2, \ldots, z_j, \ldots, z_m\}$ к совокупности A_2 (например, к бесперспективным в отношении рудоносности объектам). Такое условие приводит к тому, что все m-мерное пространство, в котором результаты наблюдений представлены m-мерными точками, будет разделено на две области R_1 и R_2, причем если результат наблюдения попал в R_1, мы принимаем решение о его принадлежности к группе A_1, а если он попадает в R_2, то мы относим его к совокупности A_2.

Естественно, что и то и другое решение не исключает возможности появления ошибок, которые заключаются в следующем. Решение о принадлежности классифицируемого объекта $a \in A$ о принадлежности его к A_1, т. е. $a \in A_1$, ошибочно и он в действительности принадлежит к A_2, т. е. $a \in A_2$. Вторая возможная ошибка заключается в том, что принимается решение $a \in A_2$, тогда как в действительности $a \in A_1$.

Каждой из этих ошибок можно приписать соответствующую цену, так как нередко появление этих ошибок приводит к тем или иным потерям. Так, например, ошибочное отнесение объекта к перспективно рудоносным, тогда как в действительности он бесперспективен, приведет к потерям, связанным с безрезультативным проведением поисковых работ на этом объекте. Наоборот, ошибочное отнесение перспективного объекта к бесперспективным приведет к потере месторождения, которое стоит, как правило, дороже, чем затраты на поисковые работы. Обозначим стоимости этих по-

терь через $C(A_1|A_2)$ и $C(A_2|A_1)$ соответственно. Ошибки и их стоимости сведены воедино в табл. 21.

Таблица 21

Ошибки и их стоимости

		Действительное состояние	
		$a \in A_1$	$a \in A_2$
Принимаемое решение	$a \in A_1$	Правильное решение $C(A_1 \mid A_1) = 0$	$a \in A_1 \mid a \in A_2$ $C(A_1 \mid A_2) > 0$
	$a \in A_2$	$a \in A_2 \mid a \in A_1$ $C(A_2 \mid A_1) > 0$	Правильное решение $C(A_2 \mid A_2) = 0$

Допустим, что в выборке, которую нужно подвергнуть разделению на объекты, принадлежащие совокупностям A_1 и A_2, эти объекты смешаны в определенных соотношениях и доля объектов $a \in A_1$ равна q_1, а объектов $a \in A_2$ — q_2 и $q_1 + q_2 = 1$. Тогда величину q_1 можно рассматривать как вероятность события, заключающегося в том, что взятый наудачу из изучаемой смешанной совокупности объект будет принадлежать к A_1. Аналогично интерпретируется и вероятность q_2.

Кроме того, не нарушая общности, будем считать, что вероятностные свойства совокупностей A_1 и A_2 описываются плотностями вероятности

$$f_1(X) = f_1(x_1, x_2, \ldots, x_j, \ldots, x_m), \quad f_2(Y) = f_2(y_1, \ldots, y_m).$$

Таким образом, если, как мы это сделали выше, область R значений X и Y разделена на две непересекающиеся области R_1 и R_2, то вероятности появления ошибочных решений будут определены следующим образом:

$$P(A_1 \mid A_2) = \int_{R_1} f_2(X)\, dX, \tag{8.1}$$

$$P(A_2 \mid A_1) = \int_{R_2} f_1(X)\, dX. \tag{8.2}$$

Теперь мы можем охарактеризовать и потери, связанные с неправильной классификацией, которые, если известны значения q_1 и q_2, будут определены следующим выражением:

$$W = C(A_1 \mid A_2)\, P(A_1 \mid A_2)\, q_2 + C(A_2 \mid A_1)\, P(A_2 \mid A_1)\, q_1. \tag{8.3}$$

Это выражение представляет собой математическое ожидание потерь классификации или, как его еще называют, средние потери. Таким образом, области принятия решений R_1 и R_2 нужно выбрать так, чтобы потери W были по возможности меньшими. Метод, ко-

торый обеспечивает минимум W при заданных q_1 и q_2 называется методом Бейеса. Подробно вопрос о риске, связанном с принятием решений в задачах классификации, рассмотрен в работах Т. Андерсона [2], а также Д. Блэкуела и М. Гиршика [4]. Здесь же мы ограничимся только общим правилом, заключающимся в том, что R_1 и R_2 выбираются следующим образом:

$$R_1 : \frac{f_1(X)}{f_2(X)} \geqslant \frac{C(A_1 \mid A_2)\, q_2}{C(A_2 \mid A_1)\, q_1}, \tag{8.4}$$

$$R_2 : \frac{f_1(X)}{f_2(X)} < \frac{C(A_1 \mid A_2)\, q_2}{C(A_2 \mid A_1)\, q_1}. \tag{8.5}$$

8.2. ПОСТРОЕНИЕ ЛИНЕЙНОГО РЕШАЮЩЕГО ПРАВИЛА ДЛЯ ДВУХ МНОГОМЕРНЫХ СОВОКУПНОСТЕЙ

8.2.1. Случай, когда $f_1(X)$ и $f_2(X)$ известны

Прежде чем перейти к рассмотрению реальных ситуаций, с которыми приходится сталкиваться в задачах классификационного отнесения объектов при геологических исследованиях, рассмотрим случай, когда функции $f_1(X)$ и $f_2(X)$ заранее известны. Более того, мы будем полагать, что $f_1(X)$ и $f_2(X)$ являются m-мерными нормальными плотностями с параметрами $\boldsymbol{\mu}_1$, Σ и $\boldsymbol{\mu}_2$, Σ соответственно. Таким образом, мы заранее вводим условие, что ковариационные матрицы распределений f_1 и f_2 равны Σ.

В соответствии с выражением (8.4) и положив

$$\frac{C(A_1 \mid A_2)\, q_2}{C(A_2 \mid A_1)\, q_1} = k, \tag{8.6}$$

можно определить область R_1 с помощью неравенства

$$R_1 : \frac{f_1(X)}{f_2(X)} = \frac{\exp\left[-\frac{1}{2}\{X-\boldsymbol{\mu}_1\}\,\Sigma^{-1}\{X-\boldsymbol{\mu}_1\}'\right]}{\exp\left[-\frac{1}{2}\{X-\boldsymbol{\mu}_2\}\,\Sigma^{-1}\{X-\boldsymbol{\mu}_2\}'\right]}. \tag{8.7}$$

Заметим, что $\boldsymbol{\mu}_1$ и $\boldsymbol{\mu}_2$ рассматриваются как m-мерные векторы-строки. После несложных преобразований (8.7) можно записать

$$R_1 : X\,\Sigma^{-1}\{\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\}' - \frac{1}{2}\{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2\}\,\Sigma^{-1}\{\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\}' \geqslant \ln k. \tag{8.8}$$

В подавляющем большинстве геологических ситуаций нет никаких данных, позволяющих судить о вероятностях q_1 и q_2, а нередко и цены потерь $C(A_2 \mid A_1)$ и $C(A_1 \mid A_2)$ невозможно бывает определить. В таких ситуациях ничего не остается делать, как допустить, что $C(A_1 \mid A_2) = C(A_2 \mid A_1)$ и $q_1 = q_2$. Тогда $\ln k = 0$ и R_1 определится неравенством

$$R_1 : X\Sigma^{-1}\{\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\}' \geqslant \frac{1}{2}\{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2\}\,\Sigma^{-1}\{\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\}'. \tag{8.9}$$

Область R_2 будет определена строгим неравенством имеющим обратный знак, т. е.

$$R_2 : X\Sigma^{-1}\{\mu_1-\mu_2\}' < \frac{1}{2}\{\mu_1+\mu_2\}\,\Sigma^{-1}\{\mu_1-\mu_2\}'. \qquad (8.10)$$

Выражение, стоящее в левой части этих неравенств, называется дискриминантной функцией, которая и служит критерием для отнесения рассматриваемого результата наблюдения Z_t в выборке к совокупности A_1 или A_2. Подставив значение Z_t на место X в формулу (8.7), получим в результате или неравенство (8.9), или (8.10), что даст возможность сделать вывод о принадлежности Z_t. Заметим также, что правая часть неравенств (8.9) и (8.10), равная

$$\frac{1}{2}\{\mu_1+\mu_2\}\Sigma^{-1}\{\mu_1-\mu_2\}', \qquad (8.11)$$

представляет собой константу, являющуюся пороговым значением в условиях сделанных ограничений.

Необходимо подчеркнуть, что описанный метод позволяет проводить классификационное отнесение только одного элемента выборки и не распространяется на выборку в целом. Более того, выводы, получаемые в результате применения такого метода относительно геологического объекта, могут оказаться неоднозначными.

8.2.2. Случай, когда μ_1, μ_2 и Σ оцениваются по выборке

Хотя ситуация, описанная в предыдущем разделе, и может встретиться в практике при геологических исследованиях, обычно параметры μ_1, μ_2 и Σ остаются неизвестными и оцениваются по выборке. Таким образом, в распоряжении исследователя имеется выборка объема n_1 наблюдений X_t ($t = 1, 2, \ldots, n_1$) над объектами совокупности A_1, и вторая выборка, объем которой n_2 наблюдений V_t ($t = 1, 2, \ldots, n_2$) над объектами совокупности A_2. По этим данным мы должны построить решающее правило, позволяющее относить некоторое наблюдение Z_t из третьей выборки к A_1 или A_2.

Как уже было показано в главе 3, наилучшими оценками для μ_1 и μ_2 будут векторы средних арифметических:

$$\overline{X} = \frac{1}{n_1}\sum_{t=1}^{n_1} X_t = \{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_j, \ldots, \bar{x}_m\}, \qquad (8.12)$$

$$\overline{Y} = \frac{1}{n_2}\sum_{t=1}^{n_2} Y_t = \{\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_i, \ldots, \bar{y}_m\}, \qquad (8.13)$$

а оценкой неизвестной ковариационной матрицы Σ будет матрица

$$S = \frac{1}{n_1+n_2-2}\left[\sum_{t=1}^{n_1}\{X_t-\overline{X}\}'\{X_t-\overline{X}\} + \sum_{t=}^{n_2}\{Y_t-\overline{Y}\}'\{Y_t-\overline{Y}\}\right]. \qquad (8.14)$$

Если мы подставим эти оценки в выражение (8.9) вместо μ_1, μ_2 и Σ, то получим решающее правило, определяющее область R_1, т. е.

$$R_1: ZS^{-1}\{\overline{X}-\overline{Y}\}' \geqslant \frac{1}{2}\{\overline{X}+\overline{Y}\}\,S^{-1}\{\overline{X}-\overline{Y}\}'. \qquad (8.15)$$

Необходимо отметить, что левая часть неравенства, предложенная Р. Фишером [52], представляет собой линейную дискриминантную функцию, обладающую наибольшей дисперсией между выборками относительно дисперсии внутри выборок.

Таким образом, m-мерное наблюдение Z_t подставляется на место Z в выражение (8.15), чем определяется его принадлежность к A_1 (если выполнено неравенство (8.15), и, наоборот, к A_2, если неравенство имеет обратный знак.

8.2.3. Случай, когда ковариационные матрицы известны и неравны

В предыдущих двух разделах мы рассматривали случай двух m-мерных нормальных распределений при условии, что соответствующие им ковариационные матрицы Σ_1 и Σ_2 равны. Однако это требование в большинстве геологических ситуаций может не выполняться, или же вообще могут отсутствовать какие-либо данные о соотношении ковариационных матриц. В связи с этим, мы сначала рассмотрим наиболее простой случай, когда μ_1 и μ_2, Σ_1 и Σ_2 известны и неравны, а затем перейдем к более сложной ситуации.

Как показано в работе Г. С. Лбова [29], в данной ситуации полезно применить квадратичное решающее правило, основанное на отношении правдоподобия, которое определит области R_1 и R_2 следующим образом:

$$R_1: \{X-\mu_2\}\Sigma_2^{-1}\{X-\mu_2\}' - \{X-\mu_1\}\Sigma_1^{-1}\{X-\mu_1\}' \geqslant$$

$$\geqslant 2\ln\frac{|\Sigma_1|^{\frac{1}{2}}}{|\Sigma_2|^{\frac{1}{2}}}, \qquad (8.16)$$

$$R_2: \{X-\mu_2\}\Sigma_2^{-1}\{X-\mu_2\}' - \{X-\mu_1\}\Sigma_1^{-1}\{X-\mu_1\}' <$$

$$<2\ln\frac{|\Sigma_1|^{\frac{1}{2}}}{|\Sigma_2|^{\frac{1}{2}}}. \qquad (8.17)$$

Таким образом, подлежащее классификации m-мерное наблюдение

$$Z_t = \{z_{t1}, z_{t2}, \ldots, z_{tj}, \ldots, z_{tm}\}$$

будет отнесено к первой группе, если в результате подстановки на

место X в формулу (8.16) будет получено значение, принадлежащее R_1, и, наоборот, ко второй группе, если вычисленное значение дискриминантной функции окажется в области R_2.

Вопрос о риске, связанном с принятием решений в данной ситуации, подробно рассмотрен в работах А. Н. Бугайца [9], Г. С. Лбова [29] и М. К. Камалова [24].

8.2.4. Случай, когда ковариационные матрицы неравны и μ_1, μ_2, Σ_1 и Σ_2 оцениваются по выборке

Наиболее часто наблюдаемая в геологии реальная ситуация заключается в том, что параметры многомерных распределений μ_1, μ_2, Σ_1 и Σ_2 остаются неизвестными и оцениваются по выборочным значениям

$$X_t = \{x_{t1}, x_{t2}, \ldots, x_{tj}, \ldots, x_{tm}\}, \quad t = 1, 2, \ldots, n_1 \tag{8.18}$$

и

$$Y_t = \{y_{t1}, y_{t2}, \ldots, y_{tj}, \ldots, y_{tm}\}, \quad t = 1, 2, \ldots, n_2, \tag{8.19}$$

взятым из двух совокупностей, для которых нужно построить дискриминантную функцию. Оценки для μ_1, μ_2, Σ_1 и Σ_2 будут определены следующими формулами:

$$\overline{X} = \frac{1}{n_1} \sum_{t=1}^{n_1} X_t = \{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_j, \ldots, \bar{x}_m\}, \tag{8.20}$$

$$\overline{Y} = \frac{1}{n_2} \sum_{t=1}^{n_2} Y_t = \{\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_j, \ldots, \bar{y}_m\}, \tag{8.21}$$

$$\mathbf{S}_1 = \frac{1}{n_1 - 1} \sum_{t=1}^{n_1} \{X_t - \overline{X}\}' \{X_t - \overline{X}\}, \tag{8.22}$$

$$\mathbf{S}_2 = \frac{1}{n_2 - 1} \sum_{t=1}^{n_2} \{Y_t - \overline{Y}\}' \{Y_t - \overline{Y}\}. \tag{8.23}$$

В данной ситуации возможны два решения. Первое из них — это построение линейной дискриминантной функции, которая будет определена выражением

$$Z\left(\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2\right)^{-1} \{\overline{X} - \overline{Y}\}'. \tag{8.24}$$

Критические области R_1 и R_2 для этой функции определятся следующими неравенствами:

$$R_1 : Z\left(\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2\right)^{-1} \{\overline{X} - \overline{Y}\}' \geqslant \frac{1}{2} \{\overline{X} + \overline{Y}\} \left(\frac{1}{n_1}\mathbf{S}_1 + \right.$$

$$\left. + \frac{1}{n_2} S_2\right)^{-1} \{\overline{X} - \overline{Y}\}', \tag{8.25}$$

$$R_2 : Z\left(\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2\right)^{-1} \{\overline{X} - \overline{Y}\}' < \frac{1}{2} \{\overline{X} + \overline{Y}\} \times$$

$$\times\left(\frac{1}{n_1}\mathbf{S}_1+\frac{1}{n_2}S_2\right)^{-1}\{\overline{X}-\overline{Y}\}'. \qquad (8.26)$$

Таким образом, результат наблюдения $Z_t=\{z_{t1}, z_{t2}, \ldots, z_{tj}, \ldots, z_{tm}\}$ относится к первой совокупности, если после его подстановки в формулу (8.24) на место Z окажется, что вычисленное значение попадет в область R_1. Наоборот, если это значение попадет в область R_2, Z_t следует отнести ко второй совокупности.

Квадратичная дискриминантная функция, построенная по этим данным, будет иметь вид

$$\{Z-\overline{Y}\}\,\mathbf{S}_2^{-1}\{Z-\overline{Y}\}'-\{Z-\overline{X}\}\,\mathbf{S}_1^{-1}\{Z-\overline{X}\}'. \qquad (8.27)$$

Критические области R_1 и R_2 на основе этой функции определяются следующим образом:

$$R_1:\{Z-\overline{Y}\}\,\mathbf{S}_2^{-1}\{Z-\overline{Y}\}'-\{Z-\overline{X}\}\,S_1^{-1}\{Z-\overline{X}\}'\geqslant$$

$$\geqslant 2\ln\frac{|\mathbf{S}_1|^{\frac{1}{2}}}{|\mathbf{S}_2|^{\frac{1}{2}}}, \qquad (8.28)$$

$$R_2:\{Z-\overline{Y}\}\,\mathbf{S}_2^{-1}\{Z-\overline{Y}\}'-\{Z-\overline{X}\}\,\mathbf{S}_1^{-1}\{Z-\overline{X}\}'<$$

$$<2\ln\frac{|\mathbf{S}_1|^{\frac{1}{2}}}{|\mathbf{S}_2|^{\frac{1}{2}}}. \qquad (8.29)$$

Таким образом, если в результате подстановки классифицируемого наблюдения $Z_t=\{z_{t1}, z_{t2}, \ldots, z_{tm}\}$ в формулу (8.27) на место Z будет выполнено неравенство (8.28), то Z_t относится к первой совокупности, если же окажется, что вычисленное значение функции (8.27) принадлежит R_2, то Z_t нужно отнести ко второй совокупности.

8.3. ОБ ЭФФЕКТИВНОСТИ ПРИМЕНЕНИЯ ЛИНЕЙНЫХ И КВАДРАТИЧНЫХ РЕШАЮЩИХ ПРАВИЛ

В данной главе вполне уместен вопрос о том, сколь эффективно применение того или иного решающего правила в различных реальных геологических ситуациях. Этот вопрос подробно изучался А. Н. Бугайцом [9], которым были получены следующие результаты.

Им было установлено, что линейные решающие правила оптимальны или близки к оптимальным не только в условиях многомерных нормальных распределений изучаемых совокупностей, но и при весьма более широких условиях. Оказалось, что линейные решающие правила являются оптимальными, если изучаемые распределе-

ния являются унимодальными и их плотности равномерно убывают с удалением от средних значений. Этот результат в значительной степени расширяет область применения линейных решающих правил, которые отличаются значительной простотой построения.

Квадратичные решающие правила, как отмечает А. Н. Бугаец [9], оказываются полезными в тех случаях, когда число классифицируемых совокупностей более двух.

8.4. ПРИМЕРЫ

Пример 1. В качестве примера для построения дискриминантной функции можно воспользоваться данными А. Н. Бугайца [9] по содержаниям Li_2O, Rb_2O и Cs_2O в микроклинах альбитовых и альбит-сподуменовых пегматитов одного из пегматитовых полей. Характерно, что альбитовые пегматиты обладают повышенными содержаниями Ta, Be, Nb, т. е. являются перспективными на эти элементы. В связи с этим вполне естественно возникновение задачи о построении решающего правила, позволяющего различать микроклины из альбитовых пегматитов от микроклинов из пегматитов альбит-сподуменовых. В таблицах 22 и 23 соответственно приведены значения содержаний Li_2O, Rb_2O и Cs_2O в микроклинах из упомянутых выше типов пегматитов.

Таблица 22

Содержания Li_2O, Rb_2O и Cs_2O в микроклинах из альбитовых пегматитов с высокими содержаниями ценных компонентов *

№ п.п	Li_2O	Rb_2O	Cs_2O	№ п.п	Li_2O	Rb_2O	Cs_2O
1	0,5	22,0	2,2	20	0,2	19,2	1,8
2	0,5	18,0	2,0	21	0,2	15,6	1,7
3	0,5	18,0	1,2	22	1,0	12,4	1,4
4	0,5	22,0	4,2	23	0,2	12,4	1,9
5	0,5	22,0	4,0	24	2,0	23,2	3,4
6	0,5	53,0	14,5	25	0,2	18,0	2,8
7	0,5	38,0	8,3	26	0,2	24,6	2,0
8	0,5	15,0	1,2	27	1,0	39,2	9,0
9	1,5	84,0	16,0	28	0,2	68,0	10,0
10	0,7	7,0	1,2	29	0,2	47,2	5,5
11	3,5	90,0	17,0	30	0,2	46,0	5,6
12	1,2	80,0	15,0	31	1,5	13,0	2,6
13	2,5	80,0	13,0	32	1,0	13,6	2,5
14	2,5	88,0	13,0	33	0,2	52,0	15,0
15	0,5	25,0	7,5	34	0,2	19,4	3,7
16	0,5	40,0	4,2	35	0,2	21,6	4,0
17	0,5	40,0	3,8	36	0,2	18,4	3,0
18	0,8	40,0	5,0	38	0,2	16,6	3,4
19	0,2	27,2	2,5	39	1,0	16,4	3,2

* Содержания всех трех окислов (в %) умножены на 100.

Таблица 23

Содержания Li_2O, Rb_2O и Cs_2O в микроклинах из альбит-сподуменовых пегматитов *

№ п/п	Li_2O	Rb_2O	Cs_2O	№ п/п	Li_2O	Rb_2O	Cs_2O
1	0,5	43,0	3,2	15	1,2	51,2	3,6
2	0,8	40,0	3,5	16	2,4	51,2	3,4
3	1,3	46,0	3,9	17	1,2	43,0	3,5
4	2,5	42,0	4,2	18	0,5	51,0	8,0
5	1,5	42,7	3,9	19	0,5	43,0	4,5
6	0,6	38,0	4,0	20	0,6	48,0	4,4
7	0,5	40,0	3,5	21	2,2	56,0	6,0
8	1,1	46,0	3,7	22	1,2	50,0	4,7
9	0,5	39,0	2,9	23	0,5	43,0	4,0
10	1,0	50,0	3,3	24	0,5	40,0	4,5
11	0,7	50,0	3,1	25	0,5	37,0	3,7
12	1,0	48,6	2,1	26	0,5	37,0	17,0
13	0,2	50,0	3,8	27	0,2	40,0	7,8
14	1,4	46,0	3,8				

* Содержания всех трех окислов (в %) умножены на 100.

По приведенным в табл. 22 и 23 данным были вычислены оценки векторов средних и ковариационных матриц. Так, для альбитовых пегматитов

$$\overline{X} = \{0{,}74 \quad 33{,}93 \quad 5{,}68\} \cdot 10^{-2},$$

$$S_1 = \begin{pmatrix} 0{,}58 & 10{,}76 & 1{,}97 \\ 10{,}76 & 567{,}61 & 104{,}49 \\ 1{,}97 & 104{,}49 & 23{,}21 \end{pmatrix} \cdot 10^{-4},$$

а для альбит-сподуменовых

$$\overline{Y} = \{0{,}95 \quad 44{,}80 \quad 4{,}59\} \cdot 10^{-2},$$

$$S_2 = \begin{pmatrix} 0{,}39 & 1{,}38 & -0{,}30 \\ 1{,}38 & 26{,}99 & -3{,}09 \\ -0{,}30 & -3{,}09 & 7{,}83 \end{pmatrix} \cdot 10^{-4}.$$

Предварительная проверка гипотез о равенстве векторов средних и о равенстве ковариационных матриц, которую представляется провести читателю самостоятельно, показала, что обе эти гипотезы уверенно отклоняются, т. е. векторы средних значений и ковариационные матрицы следует считать различными. Этот результат означает, что построение некоторого решающего правила для классификационного отнесения микроклинов неизвестного происхождения к одному из рассматриваемых типов имеет смысл.

В связи с тем, что ковариационные матрицы нельзя признать равными, квадратичное решающее правило является предпочти-

тельнее линейного, которое и построено в соответствии с формулой (8.28), т. е. критическая область R_1, определяющая правило отнесения неизвестных микроклинов к совокупности альбитовых пегматитов, будет охарактеризована следующим неравенством:

$$R_1 : \{z_1 - 0{,}0095z_2 - 0{,}04480z_3 - 0{,}4594\}$$

$$\begin{pmatrix} 31373{,}66 & -1530{,}98 & 608{,}18 \\ -1530{,}98 & 462{,}73 & 123{,}37 \\ 608{,}18 & 123{,}37 & 1350{,}07 \end{pmatrix} \begin{pmatrix} z_1 - 0{,}0095 \\ z_2 - 0{,}4480 \\ z_3 - 0{,}4590 \end{pmatrix} - \{z_1 - 0{,}0074z_2 -$$

$$-0{,}3393z_3 - 0{,}0568\} \begin{pmatrix} 26574{,}87 & -515{,}24 & 62{,}57 \\ -515{,}24 & 112{,}90 & -464{,}53 \\ 62{,}57 & -464{,}53 & 2516{,}95 \end{pmatrix} \times$$

$$\times \begin{pmatrix} z_1 - 0{,}0095 \\ z_2 - 0{,}4480 \\ z_3 - 0{,}4590 \end{pmatrix} \geqslant -20{,}8873 - (-23{,}4676) = 2{,}5803.$$

Необходимо отметить, что матрицы, входящие в квадратичные формы левой части неравенства, представляют собой матрицы и $\mathbf{S}_1^{-1}$ и $\mathbf{S}_2^{-1}$, обратные матрицам $\mathbf{S}_1$ и $\mathbf{S}_2$. Числа — 20,8873 и — 23,4676 являются соответственно $\ln|\mathbf{S}_1|$ и $\ln|\mathbf{S}_2|$.

Таким образом, образец микроклина, для которого $Z = \{z_1, z_2, z_3\}$, где $z_1 = = Li_2O$, $z = Rb_2O$, $z_3 = Cs_2O$, после подстановки в полученное выражение приведет к тому, что вычисленное значение окажется больше или равно 2,5803, следует отнести к альбитовым пегматитам. Если же окажется, что вычисленное значение меньше 2,5803, то данный образец микроклина нужно отнести к совокупности микроклинов из альбит-сподуменовых пегматитов.

Пример 2. Для второго примера построения дискриминантной функции воспользуемся данными об измерении характеристик черепов ореодонтов, приведенными на с. 117, из книги Р. Миллера и Дж. Кана [32]. В данном случае мы воспользуемся только двумя из сети приведенных выборок, а именно результатами измерения характеристик черепов двух видов: Merychoidodon culbertsoni и Prodesmatochoerus meeni:

По приведенным выборкам вычислены следующие статистические характеристики:
векторы средних значений

$$\overline{X} = \{47{,}71 \quad 90{,}78 \quad 17{,}57 \quad 6{,}61\},$$

$$\overline{Y} = \{39{,}33 \quad 83{,}25 \quad 15{,}67 \quad 4{,}42\}.$$

и выборочная ковариационная матрица

$$S=\begin{vmatrix} 8{,}75 & 3{,}01 & 1{,}32 & 0{,}40 \\ 3{,}01 & 8{,}67 & 1{,}45 & -0{,}08 \\ 1{,}32 & 1{,}45 & 2{,}92 & 0{,}19 \\ 0{,}40 & -0{,}08 & 0{,}19 & 0{,}54 \end{vmatrix},$$

где $\overline{X}$ и S_1 относятся к первому изучаемому виду, а $\overline{Y}$ и S_2 — ко второму.

Предварительная проверка гипотезы о равенстве четырехмерных средних показала их существенное различие, что дает основание для построения соответствующего решающего правила. Необходимо отметить, что в главе 5 настоящей книги мы исключили из данных Р. Миллера и Дж. Кана для простоты третий признак — максимальную длину черепной коробки. В данном примере этот признак учтен.

В соответствии с выражением (8.15) линейная дискриминантная функция будет определена выражением

$$ZS^{-1}\{\overline{X}-\overline{Y}\}',$$

а критическое значение, определяющее области R_1 и R_2 для классификационного отнесения изучаемой особи к первому или второму виду, будет представлено формулой

$$\frac{1}{2}\{\overline{X}+\overline{Y}\}\,S^{-1}\{\overline{X}-\overline{Y}\}'.$$

Таким образом, для построения решающего правила нам, кроме имеющейся в нашем распоряжении матрицы S, понадобятся разность $\{\overline{X}-\overline{Y}\}$ и сумма $\{\overline{X}+\overline{Y}\}$, которые соответственно равны

$$\{\overline{X}-\overline{Y}\}=\{8{,}38 \quad 7{,}53 \quad 1{,}90 \quad 2{,}19\},$$

$$\{\overline{X}+\overline{Y}\}=\{87{,}04 \quad 174{,}03 \quad 33{,}24 \quad 11{,}03\}.$$

В результате, располагая этими данными, можно вычислить набор коэффициентов линейной дискриминантной функции

$$S^{-1}\{\overline{X}-\overline{Y}\}'=\begin{vmatrix} 8{,}75 & 3{,}01 & 1{,}32 & 0{,}40 \\ 3{,}01 & 8{,}67 & 1{,}45 & -0{,}08 \\ 1{,}32 & 1{,}45 & 2{,}92 & 0{,}19 \\ 0{,}40 & -0{,}08 & 0{,}19 & 0{,}54 \end{vmatrix}^{-1}\begin{vmatrix} 8{,}38 \\ 7{,}53 \\ 1{,}90 \\ 2{,}19 \end{vmatrix}=$$

$$=\begin{vmatrix} 0{,}139 & -0{,}043 & -0{,}035 & -0{,}097 \\ -0{,}043 & 0{,}140 & -0{,}055 & 0{,}072 \\ -0{,}035 & -0{,}055 & 0{,}393 & -0{,}121 \\ -0{,}097 & 0{,}072 & -0{,}121 & 1{,}977 \end{vmatrix}\begin{vmatrix} 8{,}38 \\ 7{,}53 \\ 1{,}90 \\ 2{,}19 \end{vmatrix}=\begin{vmatrix} 0{,}5592 \\ 0{,}7469 \\ 0{,}2222 \\ 3{,}8301 \end{vmatrix}.$$

Таким образом, дискриминантная функция определяется выражением

$$Z\mathbf{S}^{-1}\{\overline{X}-\overline{Y}\}' = 0{,}5592z_1 + 0{,}7469z_2 - 0{,}2222z_3 + 3{,}8301z_4.$$

Теперь, используя имеющиеся у нас данные, вычислим критическое значение

$$\frac{1}{2}\{\overline{X}+\overline{Y}\}\,\mathbf{S}^{-1}\{\overline{X}-\overline{Y}\}' = \frac{1}{2}\{87{,}04 \quad 174{,}03 \quad 33{,}24 \quad 11{,}03\} \times$$

$$\times \begin{vmatrix} 8{,}75 & 3{,}01 & 1{,}32 & 0{,}40 \\ 3{,}01 & 8{,}67 & 1{,}45 & -0{,}08 \\ 1{,}32 & 1{,}45 & 2{,}92 & 0{,}19 \\ 0{,}40 & -0{,}08 & 0{,}19 & 0{,}54 \end{vmatrix}^{-1} \begin{vmatrix} 8{,}38 \\ 7{,}53 \\ 1{,}90 \\ 2{,}19 \end{vmatrix} = \frac{123{,}5126}{2} = 106{,}7563.$$

В результате критическая область R_1, которая является областью классификационного отнесения к первому виду, будет определена неравенством

$$R_1 : 0{,}5592\ z_1 + 0{,}7469\ z_2 - 0{,}2222\ z_3 + 3{,}8301\ z_4 \geqslant 106{,}7563.$$

Область R_2 определяется строгим неравенством с обратным знаком.

Допустим, что измерения, проведенные на черепе изучаемого нами представителя ореодонтов, дали следующие результаты:

$$z_1 = 34{,}0;\ z_2 = 77{,}0;\ z_3 = 16{,}0;\ z_4 = 4{,}8.$$

Подставив эти значения в нашу линейную дискриминантную функцию, получим

$$0{,}5592 \cdot 34{,}0 + 0{,}7469 \cdot 77{,}0 - 0{,}2222 \cdot 16{,}0 +$$

$$+ 3{,}8301 \cdot 4{,}8 = 91{,}3534.$$

Так как вычисленное значение 91,3534 значительно меньше, 106,7563, то изучаемого представителя ореодонтов можно уверенно отнести к Prodesmatochoerus meeni.

ЗАДАЧИ РАЗГРАНИЧЕНИЯ

В предыдущей главе мы рассматривали ситуации, когда группы геологических объектов заранее заданы и к ним следует отнести объекты, принадлежность которых априори неизвестна. Однако заранее заданное разграничение на группы в геологических исследованиях не всегда удается построить заранее, а чаще всего приходится сталкиваться с ситуацией, когда априори неизвестны ни сами группы, ни их число, а также неизвестно, можно ли какие-либо группы выделять.

Тем не менее задачи нахождения границ между геологическими объектами или наборами наблюдений по комплексу признаков очень часто возникают при геологических исследованиях. Примером могут служить задачи расчленения немых толщ, построения классификации пород, ископаемых организмов и т. п. Ясно, что границы должны проводиться там, где рассматриваемые признаки испытывают наибольшие изменения, и что нельзя проводить рубеж там, где признаки ведут себя достаточно стабильно. Следует отметить, что очень часто нельзя ответить на вопрос о стабильном или нестабильном поведении того или иного признака без привлечения соответствующих объективных методов. Так, например, получению ответа на этот вопрос была посвящена работа А. Б. Вистелиуса [14], и им же в последующей работе [15] для проведения одной границы на плоскости по измеренным значениям одного признака использовались отношения правдоподобия.

Однако определение положения границы по одному признаку нередко приводит к противоречивым результатам, когда для разграничения используются разные признаки независимо один от другого. Естественно, эти противоречия не возникнут, если для разграничения будет использован совместно весь комплекс изучаемых признаков.

Прежде чем перейти к формальной постановке задач разграничения необходимо рассмотреть используемую в дальнейшем математическую модель геологического объекта.

9.1. МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ГЕОЛОГИЧЕСКОГО ОБЪЕКТА В ЗАДАЧАХ РАЗГРАНИЧЕНИЯ

9.1.1. Общие положения

Математика, будучи абстрактной наукой, требует формализации реальных понятий, т. е. построения их абстрактных математических моделей. Например, точка, рассматриваемая в математике, ничего не имеет общего с реальной точкой, поставленной каранда-

шом на бумаге. Точно так же скважина, изображаемая на чертеже в виде линии, не имеет ничего общего с прямой линией, которая используется в математике. Однако абстрактное понятие — множество точек прямой линии — можно использовать как математическую модель реальной скважины, канавы или профиля, по которому проводилось опробование изучаемого геологического объекта.

При исследовании плоского сечения какого-либо тела, например среза гранитного массива, математической моделью реального сечения может служить множество точек плоскости. Если же геологический объект изучается в трех направлениях, то в качестве его модели можно использовать множество точек в трехмерном объеме.

Необходимо отметить, что так как наблюдения, которые производятся при геологических исследованиях, всегда охватывают конечное число участков или точек, нет необходимости использовать в качестве моделей непрерывные множества, а можно ограничиться дискретными множествами, содержащими конечное число элементов. Таким образом, в качестве математической модели объекта геологического исследования, без учета каких-либо его свойств, мы будем рассматривать множество T точек t. Множество T мы будем рассматривать как фиксированное и в соответствии с терминологией теории множеств будем называть T пространством.

До сих пор мы не касались вопроса учета в нашей модели того или иного числа геологических характеристик. Как уже было отмечено в главе 1, в качестве модели геологического наблюдения, охарактеризованного комплексом признаков, удобно рассматривать m-мерную случайную величину $\Xi = \{\xi_1, \ldots, \xi_m\}$. В результате каждой точке $t \in T$ можно поставить в соответствие случайную m-мерную величину $\Xi_t = \{\xi_{t1}, \xi_{t2}, \ldots, \xi_{tm}\}$, что позволяет в качестве математической модели изучаемого геологического объекта (Ξ^T) рассматривать дискретное множество T точек t, на котором задана m-мерная случайная функция Ξ_t.

Теперь нам необходимо рассмотреть некоторые формальные понятия, связанные с пространством T, которые нам потребуются в дальнейшем изложении.

Пусть пространство T содержит n точек t. Обозначим через A произвольное множество, образованное элементами t пространства T. В нашей интерпретации множество A соответствует произвольному участку изучаемого геологического объекта. Обозначим через A произвольный класс множеств A, порожденных пространством T, и пусть L — множество индексов l.

Назовем разбиением порядка k пространства T такой класс A множеств A_l, для которого выполнены соотношения

$$\bigcup_{l=1}^{k} A_l = T, \tag{9.1}$$

для любых l_1 и l_2

$$A_{l_1} \cap A_{l_2} = \varnothing \qquad l_1,\ l_2 \in L, \tag{9.2}$$

где $\bigcup$ — знак объединения множеств, $\bigcap$ — знак пересечения множеств, $\varnothing$ — пустое множество. Таким образом, все множества разбиения в сумме составляют T, два любых множества разбиения не имеют общих точек.

Обозначив через R^k полное множество разбиений порядка k, можем записать

$$R^k = \left\{ \mathscr{A} : \bigcup_{l=1}^{k} A_l = T, \quad A_{l_1} \cap A_{l_2} = \varnothing, \quad l_1 \neq l_2 \right\}. \tag{9.3}$$

Элемент множества R^k обозначим r^k. В реальной действительности любому элементу $r^k \in R^k$ соответствует набор k участков изучаемого геологического объекта, которые в сумме образуют весь объект и попарно не пересекаются. Естественно, что при $k = 1$ множество R^1 состоит только из одного элемента — пространства T. Однако уже при $k = 2$ число элементов, образующих R^2, резко возрастает и находится в пределах от $2^{n-1}\left(1 - \frac{1}{\sqrt{n+1}}\right)$ до 2^{n-1}, т. е. растет по показательному закону вместе с n. Заметим, что максимум числа элементов в R^k соответствует $k = \frac{n}{2}$ при четном n и $k = \frac{n+1}{2}$ и $k = \frac{n-1}{2}$ при нечетном n, а при $k = n$ содержит только один элемент, как и в случае R^1.

В дальнейшем при решении стратиграфических задач по данным, расположенным в пространстве в определенной последовательности на некоторой линии, нам придется использовать понятие линейно упорядоченного разбиения. Если пространство T линейно упорядочено, т. е. его элементы представляют собой натуральный ряд чисел, то линейно упорядоченным разбиением r_0^k порядка k пространства T будем называть класс $\mathscr{A}$ линейно упорядоченных множеств A_l, для которого выполнены следующие соотношения:

$$\bigcup_{l=1}^{k} A_l = T, \tag{9.4}$$

$$A_l \cap A_{l+1} = \varnothing, \tag{9.5}$$

$$t' < t'' \text{ или } t' > t'' \text{ при } t' \in A_l,\ t'' \in A_{l+1}. \tag{9.6}$$

Тогда множество R_0^k всех линейно упорядоченных разбиений порядка k можно представить следующим образом:

$$R_0^k = \left\{ \mathscr{A} : \bigcup_{l=1}^{k} A_l = T,\ A_l \cap A_{l+1} \neq \varnothing,\ t' < t'' \text{ при } t' \in A_l,\ t'' \in A_{l+1} \right\}. \tag{9.7}$$

Каждому абстрактному элементу r_0^k множества R_0^k соответствует конкретный вариант разбиения изучаемого реального объекта на k участков без нарушения заранее заданного порядка. Так, например, если скважина, по которой сверху вниз отобрано n образцов, разделена на k участков, то каждый из них должен содержать образцы, соответствующие последовательности их номеров в скважине.

Необходимо отметить, что число элементов r_0^k в R_0^k значительно меньше, чем в R^k. Например, при $k=2$ число элементов в R^2 почти равно 2^{n-1}, тогда как число их в R_0^2 равно только $n-1$.

Подводя итог, следует подчеркнуть, что исходя из приведенного выше определения модели геологического объекта, каждому множеству $A \subset T$ можно поставить в соответствие набор случайных величин $\Xi^A = \{\Xi_t,\ t \in A\}$. Каждому классу $\mathscr{A}$ множеств A элементов пространства T соответствует класс множеств случайных величин $\Xi^{\mathscr{A}} = \{\Xi^A,\ A \in \mathscr{A}\}$. Обозначив функцию распределения случайной величины Ξ_t через $F_t(X)$, а ее плотность вероятности (если Ξ_t непрерывна) через $f_t(X)$, тем самым можем поставить в соответствие каждому элементу нашей модели геологического объекта функцию $f_t(X)$, которая обеспечивает этому элементу однозначную характеристику.

9.1.2. Обобщенные модели однородного и неоднородного геологического объекта

Прежде чем ставить вопрос о критериях однородности и методах выявления границ, необходимо дать определения моделей однородного и неоднородного геологического объекта. Таким образом, в рамках изложенной выше модели геологического объекта нужно дать формальное определение его однородности и неоднородности. Пусть, как и раньше, T — пространство точек t, Ξ_t — случайная величина, соответствующая точке t, $F_t(X)$ — функция ее распределения, Ξ^T — множество (пространство) случайных величин Ξ_t, соответствующее пространству T. Через $\mathscr{X}$ мы будем обозначать множество значений X. Заметим также, что две случайные величины Ξ_{t_1} и Ξ_{t_2} называются эквивалентными, если $F_{t_1}(X) = F_{t_2}(X)$ для всех $X \in \mathscr{X}$. Таким образом, если весь набор Ξ^T случайных величин Ξ_t разбит на множества эквивалентных между собой величин, то и T будет представлено в виде

$$T = T_1 \cup T_2 \cup \ldots \cup T_k,\ k \geqslant 1. \tag{9.8}$$

В итоге для любых t_1 и t_2, входящих в одно и то же множество T_l, будет выполнено соотношение

$$F_{t_1}(X) = F_{t_2}(X) \text{ для всех } X \in \mathscr{X}, \tag{9.9}$$

а при t_1 и t_2, входящих в разные T_l, будет иметь место неравенство

$$F_{t_1}(X) \neq F_{t_2}(X), \quad t_1 \in T_l, \quad t_2 \in T_h, \quad l \neq h. \tag{9.10}$$

Если все случайные величины Ξ_t, входящие в Ξ^T, эквивалентны, то $k = 1$, и мы получим модель статистически однородного геологического объекта. Если же $k > 1$, то получим модель статически неоднородного объекта.

Ниже приведены формальные определения, связанные как с однородными, так и неоднородными объектами.

Множество Ξ^A случайных величин Ξ_t называется однородным, если для всех t', $t'' \in A$ будет выполнено равенство

$$F_{t'}(X) - F_{t''}(X) = 0 \text{ для всех } X \in \mathscr{X}, \; t', \; t'' \in A. \tag{9.11}$$

Класс E множеств Ξ^A называется классом однородных множеств, если для любого $A \subset \mathscr{A}$ выполнено равенство

$$F_{t'}(X) - F_{t''}(X) = 0 \text{ для всех } X \in \mathscr{X}, \; t', \; t'' \in A. \tag{9.12}$$

Класс E множеств Ξ^A называется однородным классом, если для любой пары множеств A_1, $A_2 \subset \mathscr{A}$

$$F_{t'}(X) - F_{t''}(X) = 0 \text{ для всех } X \in \mathscr{X}, t' \in A_1, \; t'' \in A_2. \tag{9.13}$$

Естественно, что если множество Ξ^A удовлетворяет условию (9.13) и совпадает с Ξ^T, то все пространство Ξ^T однородно и состоит из элементов Ξ_t, которые эквивалентны друг другу и статистически неразличимы. Кроме того, если пространство Ξ^T однородно, то любое его подмножество Ξ^A также однородно, а любой класс $\mathrm{E}^{\mathscr{A}}$ множеств Ξ^A, порожденных пространством Ξ^T, есть однородный класс.

Подводя итог, необходимо особо подчеркнуть, что статистически однородное пространство Ξ^T является математической моделью таких геологических объектов, в которых проведение каких-либо границ не имеет смысла, так как любое разбиение этого объекта на более дробные участки не приведет к существенным различиям в комплексе признаков.

Теперь перейдем к формальным определениям, связанным с неоднородными объектами.

Множество Ξ^A пространства Ξ^T назовем неоднородным, если существует такая пара t', $t'' \in A$, для которой имеет место неравенство

$$|F_{t'}(X) - F_{t''}(X)| > 0, \text{ хотя бы для одного } X \in \mathscr{X}, \; t', \; t'' \in A. \tag{9.14}$$

Если класс однородных множеств содержит хотя бы одну пару множеств Ξ^{A_1} и Ξ^{A_2} таких, что

$$|F_{t'}(X) - F_{t''}(X)| > 0 \text{ хотя бы для одного } X \in \mathscr{X},\ t' \in A_1,\ t'' \in A_2, \tag{9.15}$$

то такой класс называется неоднородным классом однородных множеств. Если условие (9.15) выполнено для всех пар множеств Ξ^{A_l}, Ξ^{A_h}, входящих в неоднородный класс однородных множеств $\Xi^{\mathscr{A}}$, то такой класс называется полностью неоднородным.

Таким образом, неоднородное множество Ξ^A пространства Ξ^T представляет собой математическую модель произвольного участка геологического объекта, который состоит из более дробных участков, среди которых имеются отличающиеся по комплексу признаков пары. Ясно, что если изучаемый объект и его формальный аналог (пространство Ξ^T) содержат хотя бы один неоднородный участок (множество Ξ^A), то такой объект неоднороден.

Естественно, что полное множество разбиений R неоднородного пространства Ξ^T содержит только один элемент r_1^k, соответствующий полностью неоднородному классу.

9.1.3. Нормальная модель однородного и неоднородного геологических объектов

В предыдущем разделе модели однородного и неоднородного геологического объекта были даны в весьма общем виде, а для того, чтобы ими можно было пользоваться в реальных условиях, необходимо четко определить вид функций распределения $F_t(X)$ случайных величин Ξ_t. Из всего множества возможных видов функций распределения наиболее целесообразным представляется выбрать в качестве модели распределения для Ξ_t m-мерное нормальное распределение. Такое предпочтение нормальному распределению объясняется следующими обстоятельствами.

Дело в том, что точка t абстрактного пространства соответствует некоторому небольшому участку изучаемого геологического объекта. Такой участок А. Б. Вистелиус называет «локальным». Им же в работе [57] было показано, что в качестве модели одномерных распределений геологических характеристик в пределах локальных участков наиболее подходящим является нормальный закон. Кроме того, даже в случаях распределений, существенно отличающихся от нормального, обычно случайные величины можно преобразовать так, что новые случайные величины будут обладать нормальным распределением.

Необходимо также отметить, что нормальное распределение, будучи сравнительно простым и хорошо изученным, является весьма удобным в качестве модели, так как позволяет довольно

легко строить статистические критерии для проверки различных гипотез по результатам наблюдений.

Предположение о нормальности локальных распределений ни в коей мере не означает, что обобщенные распределения геологических характеристик на всей совокупности изучаемых локальных участков будут обязательно нормальными. Получаемая в результате взвешенная смесь различных нормальных распределений может иметь самый разнообразный вид, охватывающий многие случаи наблюдаемых в геологической практике распределений.

Таким образом, в условиях сделанных нами ограничений, сводящихся к m-мерному нормальному распределению комплекса геологических характеристик в пределах локального участка, математической моделью геологического объекта будет фиксированное множество Ξ^T m-мерных случайных величин Ξ_t, которые распределены нормально с плотностями $f(X;\ \mu_t,\ \Sigma_t)$, где $\mu_t = \{\mu_{t1}, \mu_{t2}, \ldots, \mu_{tj}, \ldots, \mu_{tm}\}$ — вектор-строка, составленный из математических ожиданий случайных величин ξ_{tj}, образующих Ξ_t, Σ_t — ковариационная матрица. Как следует из раздела,

$$f(X;\ \mu_t,\ \Sigma_t) = (2\ )^{-m} |\Sigma_t|^{\frac{1}{2}} \exp \times$$

$$\times \left[-\frac{1}{2} \{X-\mu_t\}\, \Sigma_t^{-1}\, \{X-\mu_t\}' \right]. \qquad (9.16)$$

Как и в предыдущем изложении положим, что T конечно и содержит n элементов t. Тогда фиксированное множество Ξ^T будет статистически однородным, если любой паре t', $t'' \in T$ соответствует равенство

$$f(X;\ \mu_{t'},\ \Sigma_{t'}) - f(X;\ \mu_{t''},\ \Sigma_{t''}) = 0 \text{ для всех } X \in \mathscr{X}. \qquad (9.17)$$

Этому выражению равносильно совместное выполнение равенств:

$$\left.\begin{array}{l} \mu_{t'} - \mu_{t''} = \{0, \ldots, 0\} \\ \Sigma_{t'} = \Sigma_{t''} \end{array}\right| \text{ для всех } t',\ t'' \in T. \qquad (9.18)$$

Естественно, что если пространство Ξ^T однородно, то для любой пары множеств Ξ^{A_l}, Ξ^{A_h}, содержащих n_l и n_h элементов, будет справедливо равенство

$$\frac{1}{n_l} \sum_{t \in A_l} \mu_t - \frac{1}{n_h} \sum_{t \in A_h} \mu_t = \{0, \ldots, 0\} \text{ для всех } A_l,\ A_h \subset T. \qquad (9.19)$$

Непосредственно из этого равенства следует, что моделью неоднородного объекта в условиях заданных ограничений будет являться такое фиксированное множество Ξ^T, в котором хотя бы для

одной пары t', $t'' \in T$ будет иметь место неравенство

$$\frac{1}{n_l} \sum_{t' \in A_l} \mu_t - \frac{1}{n_h} \sum_{t'' \in A_h} \mu_t \neq \{0, 0, \ldots, 0\}$$ хотя бы для одной пары

$$t', t'' \in T.$$

Нетрудно также видеть, что если нормальное пространство Ξ^T неоднородно, то множество разбиений второго порядка R^2 содержит хотя бы один элемент r^2, для которого

$$\frac{1}{n_1} \sum_{t \in A_1} \mu_t - \frac{1}{n_2} \sum_{t \in A_2} \mu_t \neq \{0, 0, \ldots, 0\}, \qquad (9.21)$$

где A_1 и A_2 — непересекающиеся множества точек t, в сумме составляющие T, n_1 и n_2 — число элементов в A_1 и A_2 соответственно.

9.1.4. Некоторые дополнительные ограничения

Во многих геологических задачах, связанных с поиском разграничений общей совокупности на однородные множества объектов, помимо предположения о нормальности распределений, можно использовать следующие дополнительные ограничения.

1. Компоненты $\xi_{t1}, \xi_{t2}, \ldots, \xi_{tj}, \ldots, \xi_{tm}$ вектора-строки Ξ_t, соответствующего «локальной» совокупности, являются независимыми. Это ограничение может показаться очень сильным и не соответствующим реальной действительности, так как богатый опыт вычисления оценок коэффициентов корреляции при изучении геологических объектов показывает наличие сильных корреляционных зависимостей. Дело в том, что наше предположение о независимости геологических характеристик касается только ограниченных локальных участков, где оно, как правило, выполняется, и не распространяется на более крупные, обобщенные совокупности, в которых и отмечается обычно наличие корреляции при геологических исследованиях.

Следствием условия независимости компонент вектора Ξ_t является то, что все недиагональные элементы соответствующей ему ковариационной матрицы Σ_t равны нулю и матрица Σ_t становится диагональной матрицей, что в свою очередь значительно упрощает вычисления.

2. Диагональные элементы матриц Σ_t равны для всех $t \in T$, т. е. одномерные случайные величины ξ_{tj}, соответствующие признаку с номером j, имеют одинаковые дисперсии для всех $t \in T$.

Второе допущение, как и первое, практически всегда выполняется, если «локальные» совокупности достаточно малы и приблизительно равновелики. Это предположение также значительно упрощает построение статистических критериев и сильно сокращает объем вычислительных операций при их применении.

Таким образом, в условиях всех сделанных предположений относительно модели геологического объекта показателем однородности пространства Ξ^T может служить функция

$$\delta(r^2) = \frac{1}{n_1} \sum_{t \in A_1} \mu_t - \frac{1}{n_2} \cdot \sum_{t \in A_2} \mu_t, \qquad (9.22)$$

заданная на множестве разбиений R^2 второго порядка пространства T на непересекающим множестве A_1 и A_2. Если эта функция равна нулевому вектору на всех элементах $r^2 \in R^2$, то Ξ^T однородно, а если существует такое непустое подмножество в R^2, на элементах которого $\delta(r^2)$ отличается от нулевого вектора, то Ξ^T неоднородно.

9.2. ГИПОТЕЗА ОБ ОДНОРОДНОСТИ ГЕОЛОГИЧЕСКОГО ОБЪЕКТА И ЕЕ ПРОВЕРКА

9.2.1. Проверяемая гипотеза и альтернативы

В предыдущем разделе было показано, что функцию $\delta(r^2)$, определенную выражением (9.22), можно рассматривать как показатель однородности набора Ξ^T m-мерных случайных величин Ξ.

Следовательно, предположению об однородности Ξ^T равносилен набор H_0 гипотез $\delta(r^2) = \{0, 0, \ldots, 0\}$ для всех $r^2 \in R^2$, т. е.

$$H_0 : \delta(r^2) = \{0, 0, \ldots, 0\} \text{ для всех } r^2 \in R^2, \qquad (9.23)$$

тогда как множество альтернатив H_1 будет представлять собой неравенство

$$H_1 : \delta(r^2) = \{0, 0, \ldots, 0\} \text{ хотя бы для одного } r^2 \in R^2. \qquad (9.24)$$

Для проверки гипотезы H_0, естественно, потребуются эмпирические данные, которые в подавляющем большинстве геологических ситуаций представляют собой следующее. По изучаемому объекту производится n наблюдений (отбор образцов или проб) с последующим их анализом, связанным с определением в них значений отдельных признаков.

Если число образцов n, а число признаков m, то каждое наблюдение с номером t будет представлять собой m-мерный вектор-строку

$$X_t = \{x_{t1}, x_{t2}, \ldots, x_{tj}, \ldots, x_{tm}\}. \qquad (9.25)$$

Таким образом, в условиях нашей модели геологического объекта каждой m-мерной случайной величине Ξ_t, входящей в Σ^T, будет поставлено в соответствие одно выборочное наблюдение X_t. Весь

набор n m-мерных наблюдений будет представлять собой матрицу порядка $n \times m$

$$\begin{vmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ \cdot \\ X_t \\ \cdot \\ \cdot \\ \cdot \\ X_n \end{vmatrix} = \begin{vmatrix} x_{11}x_{12} \ldots x_{1j} \ldots x_{1m} \\ x_{21}x_{22} \ldots x_{2j} \ldots x_{2m} \\ \cdot \\ \cdot \\ \cdot \\ x_{t1}x_{t2} \ldots x_{tj} \ldots x_{tm} \\ \cdot \\ \cdot \\ \cdot \\ x_{n1}x_{n2} \ldots x_{nj} \ldots x_{nm} \end{vmatrix}. \tag{9.26}$$

По этим данным требуется проверить сформулированную выше нулевую гипотезу.

9.2.2. Критерий для проверки гипотезы об однородности

Вопрос построения статистического критерия для проверки гипотезы об однородности изучаемого геологического объекта подробно рассмотрен автором ранее [38], и здесь мы воспользуемся только конечным результатом. В качестве такого критерия рассматривается функция $v(r^2)$, заданная на множестве разбиений R^2 пространства T на две части,

$$v(r^2) = \sum_{j=1}^{m} \frac{n_1 n_2 \left[\bar{x}_j^{(1)} - \bar{x}_j^{(2)}\right]^2}{(n_1 + n_2)\, s_j^2}, \tag{9.27}$$

где $\bar{x}_j^{(1)}$ и $\bar{x}_j^{(2)}$ — средние арифметические признаки с номером j, вычисленные для каждой из двух совокупностей, на которые разделен набор n наблюдений, n_1 и n_2 — число наблюдений в этих совокупностях, s_j^2 — оценка дисперсии признака с номером j, вычисленная в предположении дисперсии обеих групп, на которые разделена совокупность, одинаковая. Эту оценку удобно вычислять по формуле

$$s_j^2 = \frac{1}{n_1 + n_2 + 1}\left[\sum_{t \in A_1} x_{tj}^2 + \sum_{t \in A_2} x_{tj}^2 - \right.$$

$$\left. - \frac{1}{n_1 + n_2}\left(\sum_{t \in A_1} x_{tj} + \sum_{t \in A_2} x_{tj}\right)^2\right], \tag{9.28}$$

где A_1 и A_2 — множества, на которые разделено пространство T,

причем $A_1 \cup A_2 = T$. Функцию $v(r^2)$ удобнее подсчитывать по формуле

$$v(r^2) = \frac{n_1 + n_2 - 1}{(n_1 + n_2) n_1 n_2} \sum_{j=1}^{m} \frac{\left(n_2 \sum\limits_{t \in A_1} x_{tj} - n_1 \sum\limits_{t \in A_2} x_{tj} \right)^2}{\sum\limits_{t \in T} x_{tj}^2 - \frac{1}{n_1 + n_2} \left(\sum\limits_{t \in T} x_{tj} \right)^2}. \quad (9.29)$$

Если проверяемая гипотеза верна, то $v(r^2)$ будет представлять собой значение случайной величины, распределенной как χ^2 с m степенями свободы. Таким образом, гипотеза об однородности принимается, если

$$\max_{r^2 \in R^2} v(r^2) \leqslant \chi^2_{q,\, m}, \quad (9.30)$$

и отклоняется, если

$$\max_{r^2 \in R^2} v(r^2) > \chi^2_{q,\, m}. \quad (9.31)$$

Необходимо отметить, что при достаточно больших значениях в качестве критерия можно воспользоваться величиной

$$\tau = \frac{\max\limits_{r^2 \in R^2} v(r^2) - m}{\sqrt{2m}}, \quad (9.32)$$

которая в условиях нулевой гипотезы, как уже отмечалось ранее, будет распределена приблизительно нормально со средним значением, равным 0, и дисперсией, равной 1. Для критических приложений вполне достаточно выбрать критическое значение, равное 3, и сравнивать с ним τ. Если $\tau > 3$, то гипотеза об однородности уверенно отклоняется.

Если гипотеза об однородности принята, то это означает, что по имеющимся данным никаких разграничений изучаемого объекта проводить нельзя, так как любое из них не будет обоснованным.

Если же принята альтернатива, то из этого следует, что изучаемый геологический объект по имеющимся данным может быть разделен по крайней мере на две совокупности, причем в качестве приемлемого варианта разграничения выбирается тот, которому соответствует $\max v(r^2)$.

Полученные в результате две совокупности требуют соответствующей проверки гипотезы об однородности, в зависимости от результатов которой они также делятся на две части или же остаются неизменными. Соответствующее теоретическое обоснование оптимальности процедуры разделения совокупности по максимальному значению критерия в условиях отклонения проверяемой гипотезы об однородности было уже рассмотрено автором [38]. Естественно, что такая процедура дихотомического деления должна продолжаться до тех пор пока все выделенные подмножества пространства Ξ^T окажутся однородными.

Основная трудность, связанная с решением задач разграничения, заключается в том, что при проверке гипотезы об однородности требуется провести перебор приблизительно 2^{n-1} вариантов разбиения совокупности n наблюдений на две части с целью поиска максимального значения критерия $v(r^2)$. Более того, если гипотеза об однородности отклоняется, то процедура поиска максимума повторяется в двух выделенных совокупностях, что также порождает перебор очень большого числа вариантов. Учитывая что величина 2^{n-1} растет очень быстро с увеличением n, становится ясным, что перебор такого числа вариантов является затруднительным даже для электронных вычислительных машин при весьма умеренных значениях n. В геологических же исследованиях число наблюдений в выборках бывает весьма значительным и поэтому для обработки этого материала требуется такая процедура поиска максимального значения критерия $v(r^2)$, которая бы не требовала рассмотрения всех элементов множества разбиений R^2 совокупности наблюдений на две части, а ограничивалась их сравнительно небольшой частью.

Автором были предложены [38] два способа сокращения вычислительных процедур при проверке гипотезы об однородности набора геологических наблюдений, но ни один из них нельзя признать полностью удовлетворительным в связи с тем, что оба эти способа не обеспечивали однозначности решения. Первый из них основан на предварительном разбиении всей совокупности наблюдений на небольшие группы с последующим их объединением по минимуму критерия. Второй способ заключается в представлении всей совокупности наблюдений в виде линейно упорядоченных последовательностей и их разграничении, что требует значительно меньшего объема вычислений, с последующим объединением полученных групп по минимуму критерия. Следует отметить, что нередко геологические данные носят такой характер, что требуют применения именно этого способа. Однако, если последовательности выделялись искусственно, то конечный результат заграничения в той или иной степени зависит от того, как выбраны эти последовательности. То же самое можно сказать и о первом способе, конечный результат которого зависит от варианта задания исходных групп.

Ниже предлагается новый способ решения данной задачи, который свободен от перечисленных недостатков, т. е. однозначен и не требует очень большого объема вычислений.

9.3. АЛГОРИТМЫ ЗАДАЧ РАЗГРАНИЧЕНИЯ

В этом разделе мы рассмотрим вычислительные процедуры, связанные с тремя задачами разграничения. Первая из них, являющаяся наиболее общей и наиболее сложной, представляет собой задачу разграничения m-мерных наблюдений, расположенных на плоскости или в трехмерном объеме на статистически однородные группы. К этой же задаче сводится и построение классификаций

пород, минералов, ископаемых организмов и т. п. по комплексу соответствующих характеристик.

Вторая задача является упрощенным вариантом первой и представляет собой разграничение набора линейно упорядоченных наблюдений на статистически однородные группы. Этот алгоритм особенно полезен при расчленении стратиграфических разрезов по комплексу признаков (особенно по микрофаунистическим и палинологическим данным), результаты опробования которых представляют собой упорядоченные последовательности многомерных наблюдений.

Третья задача — сопоставление двух стратиграфических разрезов, включает результаты второй задачи и некоторую часть первого алгоритма, предназначенную для поиска стратиграфических аналогов в двух расчлененных разрезах.

Необходимо отметить, что алгоритмы разграничения состоят, как правило, из трех частей. Первая из них — это проверка гипотезы об однородности изучаемого набора наблюдений, вторая — поиск границ и, наконец, третья — устранение ложных границ. Последняя часть алгоритма связана с тем, что в процессе поиска границ и последовательного дихотомического дробления изучаемой совокупности наблюдений, могут возникнуть две или более группы, которые не обладают существенными различиями друг от друга. Естественно, что такие группы наблюдений следует объединить в одну совокупность. Для выявления и объединения подобных групп наблюдений и предназначена третья часть алгоритмов разграничения. Дальнейшее изложение алгоритмов будет подчинено именно такой структуре.

9.3.1. Обобщенный алгоритм разграничения набора m-мерных наблюдений на плоскости и в объеме

Часть I. Проверка гипотезы об однородности

1. Дана выборка, состоящая из n m-мерных наблюдений

$$\begin{vmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ \cdot \\ X_t \\ \cdot \\ \cdot \\ \cdot \\ X_n \end{vmatrix} = \begin{vmatrix} x_{11}x_{12} \ldots x_{1j} \ldots x_{1m} \\ x_{21}x_{22} \ldots x_{2j} \ldots x_{2m} \\ \cdot \\ \cdot \\ \cdot \\ x_{t1}x_{t2} \ldots x_{tj} \ldots x_{tm} \\ \cdot \\ \cdot \\ \cdot \\ x_{n1}x_{n2} \ldots x_{nj} \ldots x_{nm} \end{vmatrix}, \tag{9.33}$$

где $X_t = \{x_{t1}, x_{t2}, \ldots, x_{tj}, \ldots, x_{tm}\}$ — m-мерный вектор-строка, $t = 1, 2, \ldots, n$. Вся выборка представляет собой матрицу порядка $n \times m$. Множество значений t будем обозначать, как и раньше, через T.

2. Рассматривается n вариантов разбиения совокупности n наблюдений на две части, таких что одна из них содержит только одно наблюдение X_t, а другая — оставшиеся $n-1$ наблюдений. Для каждого из n вариантов такого разбиения на множества A^1 и A^{n-1}, вычисляется значение критерия (9.29), который в данном случае имеет вид

$$v(A^1,\ A^{n-1}) = \frac{1}{n}\sum_{j=1}^{m} \frac{\left[(n-1)\,x_{tj} - \sum\limits_{t\in A^{n-1}} x_{tj}\right]^2}{\sum\limits_{t\in T} x_{tj}^2 - \frac{1}{n}\left(\sum\limits_{t\in T} x_{tj}\right)^2}. \qquad (9.34)$$

Из всех n значений критерия $v(A^1,\ A^{n-1})$ выбирается максимальное, чем определяется соответствующее этому максимуму наблюдение X_{t_1}.

3. Рассматриваются все $n-1$ пар, образованные X_{t_1} и оставшимися $n-1$ наблюдениями, и соответствующие им $n-1$ вариантов разбиения пространства T на два подмножества A^2 и A^{n-2}. Для каждого такого разбиения вычисляется значение критерия (9.29), т. е.

$$v(A^2,\ A^{n-2}) = \frac{n-1}{2(n-2)n}\sum_{=}^{m} \frac{\left[(n-2)\sum\limits_{t\in A^2} x_{tj} - 2\sum\limits_{t\in A^{n-2}} x_{tj}\right]^2}{\sum\limits_{t\in T} x_{tj}^2 - \frac{1}{n}\left(\sum\limits_{t\in T} x_{tj}\right)^2}, \qquad (9.35)$$

и определяется тот вариант из $n-1$ вариантов, которому соответствует $\max v(A^2,\ A^{n-2})$. Таким образом, устанавливается пара наблюдений X_{t_1}, X_{t_2}, включающая X_{t_1}, выявленное на предыдущем этапе.

4. Эта процедура продолжается до тех пор, пока будет достигнуто разбиение на $n/2$ и $n/2$ наблюдений в случае четного n, и на $\frac{n-1}{2}$ и $\frac{n+1}{2}$ при нечетном n. Таким образом, для любого $k \leqslant \frac{n}{2}$ при четном n и $k \leqslant \frac{n-1}{2}$ при нечетном n вычисляется значение критерия

$$v(A^k,\ A^{n-k}) = \frac{n-1}{k(n-k)n}\sum_{j=1}^{m} \frac{\left[(n-k)\sum\limits_{t\in A^k} x_{tj} - k\sum\limits_{t\in A^{n-k}} x_{tj}\right]^2}{\sum\limits_{t\in T} x_{tj}^2 - \frac{1}{n}\left(\sum\limits_{t\in T} x_{tj}\right)^2}. \qquad (9.36)$$

При этом множество A^k включает $n-1$ наблюдений, обеспечивающих максимальное значение критерия на $k-1$ предыдущей стадии вычислений.

5. В результате будет получена последовательность $n/2$ или $(n-1)/2$ максимальных значений критерия, полученных на $n/2$ или $(n-1)/2$ стадиях вычислений. Из всех этих значений выбирается максимальное, которому соответствует разбиение T на A^{k^*} и A^{n-k^*}, т. е. отыскивается значение

$$\max_{k} \max_{A^k \in \mathcal{A}^k} v(A^k,\ A^{a-k}) = v(k^*), \tag{9.37}$$

где $\mathcal{A}^k$ — класс всех множеств A^k, включающих выбранную на предыдущей стадии комбинацию $k-1$ наблюдений.

6. Если

$$v(k^*) \leqslant \chi^2_{q,\,m}, \tag{9.38}$$

где $\chi^2_{q,\,m}$ — заданное значение χ^2, соответствующее уровню значимости q и m степеням свободы, то дальнейшие вычисления прекращаются, так как для данного набора наблюдений гипотеза об однородности не отклоняется, из чего следует, что любые разграничения этой совокупности наблюдений не имеют смысла.

Если же

$$v(k^*) > \chi^2_{q,\,m}, \tag{9.39}$$

то гипотеза об однородности набора наблюдений отклоняется, из чего следует, что изучаемую совокупность наблюдений нужно разделить не менее чем на две части. При этом выбирается тот вариант разбиения на две части, который соответствует $v(k^*)$.

Необходимо отметить, что в практической работе значительно удобнее пользоваться отношением

$$\tau = [v(k^*) - m]/\sqrt{2m}. \tag{9.40}$$

Вычислительные процедуры прекращаются и совокупность рассматривается как однородная, если $\tau \leqslant 3$, и гипотеза об однородности отклоняется, если $\tau > 3$.

Часть II. Поиск границ

7. Если гипотеза об однородности изучаемой совокупности наблюдений отклонена, то эта совокупность делится на две части в соответствии с $v(k^*)$.

8. Каждая из двух новых совокупностей анализируется отдельно по алгоритму, описанному в первой части, в результате чего принимается решение об однородности или неоднородности каждой из совокупностей. Если для какой-либо из этих совокупностей гипотеза об однородности принимается, то дальнейшие вычисления для нее прекращаются. Если же принимается альтернатива, то данная совокупность снова делится на две части, в соответствии с правилом, изложенным в пункте 7, и анализ вновь полученных совокупностей продолжается.

9. Процедура такого дихотомического деления изучаемой совокупности продолжается до тех пор, пока во всех выделенных более дробных совокупностях будет принята гипотеза об однородности. Однако некоторые из полученных разграничений могут оказываться ложными и поэтому нужно перейти к третьей части алгоритма — устранению ложных границ.

Часть III. Устранение ложных границ

10. В результате проведенных вычислений изучаемая выборка, объем которой n, будет разделена на h групп наблюдений. Обозначим через $T_1, T_2, \ldots, T_l, \ldots, T_h$ — непересекающиеся подмножества в T, которые соответствуют выделенным группам наблюдений.

11. Из упомянутых h групп наблюдений можно образовать $h(h-1)/2$ пар и для каждой из них вычислить значение критерия

$$v(T_l, T_s) = \frac{n_l + n_s - 1}{n_l n_s (n_l + n_s)} \sum_{j=1}^{m} \times$$

$$\times \frac{\left(n_s \sum\limits_{t \in T_l} x_{tj} - n_l \sum\limits_{t \in T_s} x_{tj}\right)^2}{\sum\limits_{t \in T_l \cup T_s} x_{tj}^2 - \frac{1}{n_l + n_s}\left(\sum\limits_{t \in T_l \cup T_s} x_{tj}\right)^2}. \qquad (9.41)$$

В результате будет получена треугольная матрица, содержащая $h(h-1)/2$ значений критерия

$$\begin{bmatrix} v(T_1, T_2) & v(T_1, T_3) \ldots vT(_1, T_h) \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ & v(T_l, T_s) \ldots v(T_l, T_h) \\ & v(T_{h-1}, T_h) \end{bmatrix}. \qquad (9.42)$$

12. Из всех этих значений выбирается минимальное, которое сравнивается с допустимым $\chi^2_{q,m}$ при заданном уровне значимости q и m степенях свободы.

Если $\min\limits_{l,s} v(T_l, T_s) > \chi^2_{q,m}$, то дальнейшие вычисления прекращаются и все выделенные группы наблюдений рассматриваются как существенно отличающиеся одна от другой. Если же $\min\limits_{l,s} v(T_l, T_s) \leqslant \chi^2_{q,m}$, то та пара групп T_l, T_s, на которой достигнуто это минимальное значение, объединяется в одну группу T_{l_1}.

13. В результате число групп будет $h-1$, и процедура проверки продолжается для данного уменьшенного набора групп. Для этого достаточно вычислить значения критерия для всех возможных пар, которые образует T_{l_1} с остальными $h-2$ группами. Значения критерия для тех пар, в которые не входит T_{l_1}, можно взять из мат-

рицы, определенной в пункте 11. Из всех этих значений критерия опять выбирается минимальное, которое сравнивается с критическим $\chi^2_{q,\,m}$.

14. Такая последовательная процедура проверки, использующая парные объединения, продолжается до тех пор, пока минимальное значение превысит допустимое $\chi^2_{q,\,m}$. Необходимо отметить, что на практике бывает удобно в качестве критерия использовать отношение

$$\tau = \frac{\min\limits_{l,\,s} v(T_l,\ T_s) - m}{\sqrt{2m}}. \tag{9.43}$$

Процедура объединения прекращается как только будет достигнуто неравенство $\tau > 3$.

15. Полученные в результате группы наблюдений следует рассматривать как статистические однородные, отличающиеся одна от другой совокупности.

9.3.2. Алгоритм разграничения совокупности линейно упорядоченных многомерных наблюдений

Часть I. Проверка гипотезы об однородности

1. Исходные данные в этой задаче, как и в предыдущем случае, представляют собой матрицу порядка $n \times m$, состоящую из n строк (наблюдения) и m столбцов (признаки), т. е.

$$\begin{vmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ \cdot \\ X_t \\ \cdot \\ \cdot \\ \cdot \\ X_n \end{vmatrix} = \begin{vmatrix} x_{11}x_{12} \ldots x_{1j} \ldots x_{1m} \\ x_{21}x_{22} \ldots x_{2j} \ldots x_{2m} \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ x_{t1}x_{t2} \ldots x_{tj} \ldots x_{tm} \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ x_{n1}x_{n2} \ldots x_{nj} \ldots x_{nm} \end{vmatrix}. \tag{9.44}$$

В отличие от предыдущей задачи, где положение наблюдений X_t может быть произвольным, и конечный результат не зависит от расположения X_t, в данной задаче положение наблюдений строго зафиксировано. Так, например, положение точек наблюдения в скважине соответствует именно такой схеме.

2. Для проверки гипотезы об однородности изучаемого набора наблюдений рассматривается n—1 возможных вариантов разбиения всей совокупности наблюдений на две части без изменения

расположения X_t. Каждому разбиению соответствует значение критерия.

$$v_k = \frac{n-1}{kn(n-k)} \sum_{j=1}^{m} \frac{\left[(n-k)\sum_{t=1}^{k} x_{tj} - k \sum_{t=k+1}^{n} x_{tj}\right]^2}{\sum_{t=1}^{n} x_{tj}^2 - \frac{1}{n}\left(\sum_{t=1}^{n} x_{tj}\right)^2}, \quad (9.45)$$

где k — номер наблюдения, после которого проводится граница, делящая совокупность наблюдений на две части. Таким образом, $k = 1, 2, \ldots, n-1$.

3. Из всех $n-1$ вычисленных значений критерия выбирается максимальное и сравнивается с критическим значением $\chi^2_{q,m}$. Если

$$\max_k v_k \leqslant \chi^2_{q,m}, \quad (9.46)$$

то процедура прекращается и весь изучаемый набор линейно упорядоченных наблюдений рассматривается как однородный, т. е. не имеющий никаких границ. Естественно, что при изучении стратиграфических данных подобный результат свидетельствует о том, что никаких стратиграфических границ в исследуемом разрезе по имеющимся данным проводить нельзя, так как они будут необоснованы.

Если же

$$\max_k v_k > \chi^2_{q,m}, \quad (9.47)$$

то из этого следует, что рассматриваемый набор линейно упорядоченных наблюдений содержит по крайней мере одну границу.

Заметим, что в качестве критерия, как и в предыдущем разделе, можно воспользоваться отношением

$$\tau = \frac{\max_k v_k - m}{\sqrt{2m}}. \quad (9.48)$$

Решения принимаются аналогично описанным при $\tau \leqslant 3$ и $\tau > 3$ соответственно.

Часть II. Поиск границ

4. Если гипотеза об однородности изучаемого набора наблюдений отклонена, то вся линейно упорядоченная совокупность наблюдений разделяется на две части по максимальному значению критерия и каждая из этих частей анализируется отдельно изложенным выше способом.

5. Процедура последовательного дихотомического дробления совокупности наблюдений продолжается до тех пор, пока для каждой из них будет иметь место неравенство

$$\max_k v_k \leqslant \chi^2_{q,m} \quad (9.49)$$

или

$$\tau \leqslant 3. \tag{9.50}$$

В результате набор линейно упорядоченных наблюдений будет разделен на статистически однородные совокупности. Однако среди выделенных между этими совокупностями границ могут оказаться ложные и поэтому заключительная процедура алгоритма сводится к устранению ложных разграничений.

Часть III. Устранение ложных границ

6. Допустим, что в результате предыдущих вычислений, набор изучаемых нами линейно упорядоченных наблюдений разделен на h групп, в соответствии с чем фиксированное множество индексов T разделено на подмножества $T_1, T_2, \ldots, T_l, \ldots, T_h$. В рассматриваемой ситуации ложные границы возможны только для смежных подмножеств T_l, T_{l+1} и, следовательно, проверка границ должна проводиться только для таких смежных подмножеств.

7. Таким образом, начиная с T_1 и T_2 вычисляется значение критерия

$$v(T_1, T_2) = \frac{n_1 + n_2 - 1}{n_1 n_2 (n_1 + n_2)} \sum_{j=1}^{m} \frac{\left[n_2 \sum\limits_{t \in T_1} x_{tj} - n_1 \sum\limits_{t \in T_2} x_{tj} \right]^2}{\sum\limits_{t \in T_1 \cup T_2} x_{tj}^2 - \frac{1}{n_1 + n_2} \left(\sum\limits_{t \in T_1 \cup T_2} x_{tj} \right)^2}. \tag{9.51}$$

Если $v(T_1, T_2) \leqslant \chi^2_{q,m}$, то граница признается ложной, и совокупности наблюдений, соответствующие T_1 и T_2, объединяются в одну группу, которая сравнивается описанным способом с T_3. Если же $v(T_1, T_2) \geqslant \chi^2_{q,m}$, то граница между T_1 и T_2 принимается как обоснованная и проверка продолжается для T_2 и T_3.

8. Такая последовательная проверка продолжается до тех пор, пока будет проверена последняя граница, предшествующая совокупности T_h. В результате или все выделенные границы сохранятся, или же некоторые из них будут устранены как ложные, и тогда число выделенных однородных совокупностей станет несколько меньше. Как это уже делалось выше, в качестве критерия можно использовать отношение

$$\tau = \frac{v(T_l, T_{l+1}) - m}{\sqrt{2m}}. \tag{9.52}$$

9.3.3. Алгоритм сопоставления двух стратиграфических разрезов

Теоретические основы процедуры статистического сопоставления двух стратиграфических разрезов были ранее изложены автором [38], и поэтому здесь мы ограничимся только описанием алгоритма.

1. Исходные данные представляют собой два набора m-мерных линейно упорядоченных наблюдений $(X_t', \ t = 1, 2, \ldots, n_1)$ и $(X_t'', \ t = 1, 2, \ldots, n_2)$. В результате применения алгоритма расчленения линейно упорядоченных последовательностей наблюдений оба сравниваемых разреза расчленяются на k_1 и k_2 стратиграфических подразделений, которым соответствуют множества T_1', $T_2', \ldots, T_{k_1}'$, и T_1'', $T_2'', \ldots, T_{k_2}''$ значений индекса t.

2. В результате попарного сопоставления наборов наблюдений соответствующих T_l' и T_s'' с помощью критерия

$$v(T_l', T_s'') = \frac{n_l' + n_s'' - 1}{n_l' n_s'' (n_l' + n_s'')} \sum_{j=1}^{m} \times \times \frac{\left(n_s'' \sum\limits_{t \in T_l'} x_{tj}' - n_l' \sum\limits_{t \in T_s''} x_{tj} \right)^2}{\sum\limits_{t \in T_l' \cup T_s''} x_{tj}^2 - \frac{1}{n_l' + n_s''} \left(\sum\limits_{t \in T_l' \cup T_s''} x_{tj} \right)^2} \tag{9.53}$$

будет получена прямоугольная матрица порядка $k_1 \times k_2$

$$\begin{vmatrix} v(T_1', T_1'') & v(T_1', T_2'') \ldots v(T_1', T_{k_2}'') \\ v(T_2', T_1'') & v(T_2', T_2'') \ldots v(T_2', T_{k_2}'') \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ v(T_{k_1}', T_1'') & v(T_{k_1}', T_2'') \ldots v(T_{k_1}', T_{k_2}'') \end{vmatrix} \tag{9.54}$$

3. Из всех $k_1 \times k_2$ значений критерия отыскивается минимальное $\min\limits_{l, s} v(T_l', T_s'')$, которое сравнивается с допустимым значением $\chi^2_{q, m}$, соответствующим уровню значимости q и m степеням свободы. Если окажется, что

$$\min_{l, s} v(T_l', T_s'') > \chi^2_{q, m}, \tag{9.55}$$

то разрезы следует рассматривать как не имеющие никаких общих элементов, т. е. как несопоставимые.

Если же окажется, что

$$\min_{l, s} v(T_l', T_s'') \leqslant \chi^2_{q, m}, \tag{9.56}$$

то разрезы рассматриваются как имеющие хотя бы один общий стратиграфический элемент, и тогда пара подразделений T_l' и T_s'', которой соответствует минимум критерия, объединяется.

4. Обозначим те значения l и s, которым соответствует минимальное значение критерия, через l_0 и s_0. В результате вычеркивания строки с номером l_0 и столбца с номером s_0 матрица (9.54) раз-

бивается на четыре подматрицы. В силу специфики стратиграфической задачи, запрещающей перекрестное сопоставление выделенных совокупностей после объединения групп с номерами l_0 и s_0, из дальнейшего рассмотрения исключаются две подматрицы — правая верхняя и левая нижняя. Каждая из оставшихся двух подматриц, левая верхняя и правая нижняя, анализируются отдельно. В этих подматрицах также отыскиваются минимальные значения критерия, которые сравниваются с допустимым значением $\chi^2_{q,\,m}$. Если окажется, что найденное минимальное значение критерия в данной подматрице окажется меньше допустимого $\chi^2_{q,\,m}$, то процедура ее разделения на четыре подматрицы с последующим устранением из рассмотрения правой верхней и левой нижней подматриц повторяется. Если же для данной подматрицы минимальное значение критерия превысит допустимое $\chi^2_{q,\,m}$, то соответствующие этой подматрицы участки сравниваемых стратиграфических разрезов рассматриваются как несопоставимые.

5. Процедура продолжается до тех пор, пока минимальные значения критерия во всех последовательных выделенных подматрицах окажутся больше, чем допустимое значение $\chi^2_{q,\,m}$. В итоге те стратиграфические подразделения сравниваемых разрезов, которые близки по комплексу признаков, будут объединены, и одновременно будут выделены выклинивающиеся слои, т. е. не имеющие аналогов в одном из сравниваемых разрезов.

9.4. ОБЩИЕ ПРИНЦИПЫ ПОСТРОЕНИЯ СТАТИСТИЧЕСКИХ МЕТОДОВ РАЗГРАНИЧЕНИЯ ГЕОЛОГИЧЕСКИХ ОБЪЕКТОВ

Как уже отмечалось, одной из самых распространенных геологических задач является нахождение границ в геологических объектах по комплексу признаков. С этой задачей приходится сталкиваться при расчленении стратиграфических разрезов, картирования магматических комплексов, при построении классификаций в палеонтологии, петрографии и др. Можно уверенно утверждать, что практически во всех отраслях геологии задача разграничения является типичной, с которой приходится иметь дело геологам в их повседневной деятельности.

В общей постановке эта задача выглядит следующим образом. Некоторый геологический объект или группа объектов, рассматриваемые как генеральная совокупность, опробуются с последующим определением в каждой из проб значений заданного набора изучаемых характеристик. По имеющимся результатам наблюдений требуется определить, можно ли рассматривать данный объект как однородный, и если нет, то определить группировку наблюдений, обеспечивающую однородность внутри групп при различиях между ними. Такая группировка определит положение границ в изучаемом объекте.

Еще раз напомним формальную постановку задачи, введя следующие обозначения. Пусть T дискретное множество точек $t = 1, 2, \ldots, n$ и A — произвольное подмножество в T. Каждой точке t поставим в соответствие m-мерную случайную величину Ξ_t с функцией распределения $F_t(X, \Theta)$, где X — m-мерное значение аргумента, Θ — набор параметров. Обозначим через X_t выборочное значение Ξ в точке t, а через Θ_t — соответствующее значение Θ. Набор случайных величин Ξ_t, соответствующих множеству A, обозначим через Ξ^A. Пусть Θ_0 — заданный набор значений множества параметров Θ. Тогда множество Ξ^A случайных величин Ξ_t будем считать однородным, если

$$\Theta_t = \Theta_0 \text{ для всех } t \in A. \tag{9.57}$$

Таким образом, нетрудно видеть, что условие однородности набора случайных величин Ξ_t зависит от того, какие параметры распределения рассматриваются в наборе Θ. В подавляющем большинстве практических приложений при статистическом изучении m-мерных наблюдений рассматриваются m-мерное среднее μ и ковариационная матрица Σ, т. е. $\Theta = (\mu, \Sigma)$.

В этих условиях возможны три варианта определения однородного набора Ξ^A случайных величин Ξ_t, который рассматривается как математическая модель однородного геологического объекта.

1. Множество Ξ^A случайных величин Ξ_t рассматривается как однородное, если для всех $t \in A$ $\mu_t = \mu_0$.

2. Множество Ξ^A случайных величин Ξ_t рассматривается как однородное, если для всех $t \in A$ $\Sigma_t = \Sigma_0$.

3. Множество Ξ^A случайных величин Σ_t рассматривается как однородное, если для всех $t \in A$ $\mu_t = \mu_0$, $\Sigma_t = \Sigma_0$.

Первый вариант соответствует случаю, когда однородность понимается как равенство средних значений комплекса признаков во всех участках геологического объекта, на которые распространяется влияние наблюдения в точке t. Обычно рассматривается именно этот вариант.

Во втором случае средние значения из рассмотрения исключаются, и однородность объекта определяется относительно ковариационных матриц, соответствующих участкам геологического объекта, на которые распространяется влияние одного наблюдения.

Третий вариант является обобщением двух первых, причем это определение однородности является более сильным, чем два предыдущих, так как требует одновременного выполнения равенства средних и равенства ковариационных матриц.

Таким образом, первый шаг построения статистического метода разграничения заключается в выборе того или иного варианта определения однородности, который зависит от характера геологического исследования и от априорной информации об информативности тех или иных параметров. Естественно, что тремя перечисленными вариантами определения однородности не исчерпывается

их множество, и возможны другие определения, которые зависят от особенностей решаемых геологических задач.

Каждый из трех перечисленных вариантов определения однородности можно рассматривать как статистическую гипотезу, которую требуется проверить по результатам наблюдения X_t над случайными величинами Ξ_t. Таким образом, вторым важным шагом построения метода разграничения является выбор статистического критерия для проверки гипотезы об однородности. Необходимо отметить, что для проверки одной и той же гипотезы может быть в наличии несколько статистических критериев и естественно желание выбрать такой из них, который бы обеспечивал решение поставленной задачи при минимальном числе наблюдений и был бы по возможности менее трудоемким. Так, например, для проверки гипотезы о равенстве m-мерных средних можно воспользоваться критерием Уилкса, критерием Готелинга, информационным критерием Кульбака или же упрощенным критерием, предложенным автором и др.

Однако первые три критерия, учитывающие зависимости между рассматриваемыми характеристиками, не позволяют сравнивать выборки меньшего объема, чем число m рассматриваемых признаков. В связи с этим применение упомянутых критериев не обеспечивает достаточную детальность разграничения и уничтожает эффект, получаемый за счет привлечения большого числа признаков. Упрощенный критерий (9.27) свободен от этих недостатков и наряду с простотой вычислений обеспечивает достаточную детальность разграничения. Аналогичная задача выбора статистического критерия возникает и в условиях других определений однородности, рассматриваемых как нулевая гипотеза.

Третий шаг построения метода разграничения заключается в выборе логической процедуры последовательности рассмотрения вариантов разбиения совокупности наблюдений на две части. Дело в том, что проверка любой гипотезы об однородности набора наблюдений представляет собой множество проверок гипотез о равенстве заданных параметров по двум выборкам, соответствующим каждому варианту разбиения изучаемой совокупности на две части. Характер процедуры рассмотрения вариантов разбиения в значительной степени определяется геологической задачей. Так, например, если задача заключается в расчленении статистического разреза по комплексу характеристик, то положение точек наблюдения строго зафиксировано на одной линии, и для проверки гипотезы об однородности достаточно рассмотреть n—1 вариантов разбиения совокупности наблюдений на две части. Если же требуется проверить гипотезу об однородности набора наблюдений, расположенного на плоскости, то в данном случае существует несколько подходов к построению процедуры перебора вариантов разбиения. Так, можно рассмотреть все 2^{n-1} вариантов разбиения, поставив каждому из них в соответствие значение критерия. Однако число 2^{n-1} весьма велико даже для умеренных значений n,

и поэтому такая процедура трудно осуществима даже для быстродействующих электронных вычислительных машин. Можно также, если наблюдения на плоскости представляют совокупность профилей, рассмотреть каждый профиль отдельно, применяя процедуру расчленения разреза, а полученные в результате однородные участки подвергнуть попарному сопоставлению с последующим объединением несущественно различающихся пар. Однако результаты, получаемые таким способом, в некоторой степени зависят от ориентации системы профилей и не всегда могут быть однозначными. Если набор наблюдений, исходя из геологических предпосылок, можно разделить на некоторое число (пусть большое) групп, содержащих больше одного наблюдения, но которые априори рассматриваются как однородные, то процедуру можно свести ко второй части описанного выше способа, заключающегося в попарном сравнении выделенных групп. Этими примерами возможный перечень процедур далеко не исчерпывается и можно привести еще ряд других вариантов, определяемых как объемом выборки, так и геологическими условиями решаемой задачи.

Подводя итоги, необходимо отметить следующее.

1. Построение статистических методов разграничения геологических объектов по комплексу признаков складывается из трех последовательно решаемых задач:

а) построения математической модели однородного объекта, т. е. определения требований к однородности, рассматриваемых как статистические гипотезы;

б) выбора или построения статистического критерия для проверки гипотезы об однородности;

в) выбора процедуры последовательного рассмотрения вариантов разбиения совокупности наблюдений на две части.

2. Так как каждая из трех задач имеет несколько вариантов решения, это приводит к множеству методов разграничения, конкретный выбор которых определяется главным образом характером геологической задачи и имеющегося в наличии фактического материала.

3. Большинство возможных статистических методов разграничения еще не выявлено, поэтому их изучение является одной из основных задач дальнейших исследований.

9.5. ОГРУБЛЕНИЕ РЕЗУЛЬТАТОВ РАЗГРАНИЧЕНИЯ

Описанный выше критерий, применяемый в задачах разграничения, весьма чувствителен, что позволяет установить такие границы, которым соответствуют весьма незначительные применения в комплексе признаков, что обычно выражается в очень небольшом превышении данным критерием критического значения. Методы разделения главных и второстепенных границ изложены в главе 6, и ими можно воспользоваться для устранения второстепенных разграничений.

Однако возможны такие ситуации, когда за счет очень плавных изменений значений изучаемых признаков и чувствительности критерия совокупность наблюдений будет разделена на большое число небольших групп, незначительно, хотя и статистически существенно отличающихся одна от другой. В наших случаях весьма желательно сделать разграничение более грубым, т. е. менее детальным, чтобы уловить основные тенденции, устранив локальные различия. Это можно сделать за счет заведомого увеличения критического значения, т. е. сознательно допуская пренебрежение некоторыми различиями между сопоставляемыми наблюдениями.

Более того, последовательно увеличивая критическое значение, можно получить несколько вариантов разграничения, каждому из которых соответствует определенная степень детальности, а выбрать тот, детальность которого соответствует поставленной геологической задаче.

Проще всего такое огрубление разграничения провести следующим образом. Если в качестве критерия для проверки гипотезы об однородности использовать случайную величину $\tau = \frac{V - m}{\sqrt{2m}}$, которая в условиях нулевой гипотезы распределена приблизительно нормально со средним значением, равным нулю, и дисперсией, равной единице, то в качестве критического значения для отклонения нулевой гипотезы целесообразно выбрать значение 3, как соответствующее уровню значимости 0,001. Если в результате процедуры разграничения при этом критическом значении будет получено большое число мелких групп наблюдений, трудно поддающихся интерпретации, то это означает, что изучаемая совокупность опробована с излишней детальностью и чувствительность метода позволяет улавливать локальные флуктуации. Для того, чтобы сделать разграничение более грубым, следует задать большее критическое значение, например 4, 5, 6 и т. д., или 4, 6, 8 и т. д. и провести расчет серии вариантов разграничения с последующим выбором того из них, детальность которого окажется приемлемой.

Пример такого выбора нужного варианта будет приведен в следующем разделе на данных опробования вод придонной части залива Кара-Богаз-Гол.

9.6. ПРИМЕРЫ

Пример 1. Чтобы наглядно показать весь процесс вычислений, связанных с задачами разграничения геологических объектов по комплексу признаков, мы воспользуемся очень простым примером, включающим только три признака и не очень большое число наблюдений, так что все вычисления при желании можно провести на настольной вычислительной машине.

В данном примере мы рассмотрим задачу изучения соотношений песчаных и известковистых (бентосных и планктонных) форамини-

фер в стратиграфическом разрезе. Известно, что одним из показателей характера палеогеографических условий морского бассейна является соотношение песчаных и известковистых форм фораминифер, живших в нем. Изменения в этих соотношениях обычно бывают связаны с изменениями фациальной обстановки, тогда как при стабильных условиях эти соотношения бывают приблизительно постоянными. В данном примере известковые формы разделены на бентосные и планктонные.

Если в любой точке разреза количество песчаных, бентосных и планктонных фораминифер выражено числовыми характеристиками, то участки разреза, однородные в смысле этих характеристик, можно рассматривать как обладавшие стабильными фациальными условиями, а границы между однородными, отличающимися один от другого участками,— как соответствующие моментам смены фациальной обстановки.

В распоряжении автора имелись данные об изменении числовых характеристик содержания песчаных, бентосных и планктонных форм фораминифер в породах палеогенового возраста по разрезу одной из скважин в Западной Туркмении.

Образцы, в которых проводились подсчеты по методике, описанной ранее [39], были отобраны через интервал 5—10 м. В качестве числовых характеристик количества особей каждого вида была использована шестибалльная шкала (0, 1, 2, 3, 4, 5), где 0 означает полное отсутствие особей данного вида, 1 — наличие единичных форм, 2 — число особей данного вида не превышает 20, 3 — число особей находится в интервале от 20 до 100, 4 — число особей порядка нескольких сотен, 5 — фауна образца практически полностью представлена особями данного вида. Такие полуколичественные характеристики очень похожи на результаты полуколичественного спектрального анализа. Полученные таким образом числовые характеристики затем суммировались для каждого образца по всем видам одной из групп форм песчаной, бентосной или планктонной. Результаты этих подсчетов приведены в табл. 24, причем вес образцов, в которых проводились подсчеты, составлял 100—150 г.

Прежде чем говорить о каких-либо тенденциях в изменении числовых характеристик рассматриваемых групп по разрезу, необходимо ответить на вопрос, можно ли рассматривать имеющийся набор трехмерных данных как однородный или же, наоборот, как неоднородный, из чего будет следовать обоснованность задачи о поиске границ.

Таким образом, нам необходимо проверить утверждение, заключающееся в том, что каких-либо существенных изменений в соотношениях песчаных, бентосных и планктонных фораминифер при образовании пород рассматриваемого разреза не происходило. Этому предположению равносильна статистическая гипотеза, заключающаяся в том, что набор, состоящий из 28 линейно упорядоченных результатов наблюдений над тремя признаками, является

однородным. Геологическое альтернативное утверждение будет заключаться в том, что соотношения рассматриваемых трех групп фораминифер существенно изменялись в процессе формирования пород разреза. Статистическим аналогом этого утверждения будет множество гипотез, заключающихся в условии неоднородности рассматриваемого набора данных.

Таблица 24

Числовая характеристика содержания песчаных, бентосных и планктонных форм фораминифер в палеогеновых породах Западной Туркмении

Номер образца	1	2	3	4	5	6	7	8
	Значения признаков			v_k	v_k	v_k	v_k	v_k
1	0	1	0				2,69	
2	0	2	0					
3	0	4	0					
4	0	4	0			14,44		14,47
5	3	7	0					
6	4	9	1				2,75	
7	2	5	0					
8	4	10	1					
9	3	6	0		32,11			17,68
10	5	13	0					
11	6	15	0				4,59	
12	6	17	2					
13	7	13	2		11,27			11,27
14	3	17	1					
15	3	20	2					
16	4	17	2					
17	4	16	3					
18	6	26	3				7,04	
19	5	37	3					
20	5	31	4					
21	6	27	1					
22	3	18	5					
23	4	13	2					
24	4	20	9	30,44				20,03
25	0	0	5					
26	0	10	7					
27	1	6	0		3,23			
28	2	1	2					

Следует отметить, что даже беглое рассмотрение табл. 24 приводит к выводу о существенном изменении комплекса трех признаков в интервале между 24-м и 25-м образцами. Изменения здесь на-

столько явны, что вывод будет достаточно надежен и без применения статистических методов.

Однако вынести однозначное решение относительно однородности или неоднородности данных в интервале от образца с номером 5 до 24-го образца включительно при простом рассмотрении таблицы не представляется возможным. Сравнение верхней и нижней групп образцов этого участка наводит на мысль о возможных различиях, но при этом вопрос о точке проведения границы остается открытым.

В столбце 4 табл. 24 приведено максимальное значение критерия v_k, определенного формулой (9.45). Так как оно сильно превышает допустимое значение $\chi^2_{0,05;\,3} = 7{,}815$, соответствующее уровню значимости 0,05 и трем степеням свободы, гипотезу об однородности следует отклонить и принять альтернативу, равносильную утверждению, что соотношения между песчаными, бентосными и планктонными фораминиферами меняются по разрезу. Максимальное значение критерия в этом столбце равно 30,44 и соответствует интервалу между 24-м и 25-м образцами. Разделив всю таблицу по этому интервалу на две части, проанализируем их отдельно одну за другой.

В столбце 5 табл. 24 приведены максимальные значения критерия v_k, полученные при анализе выделенных двух частей таблицы. Нетрудно видеть, что верхнюю часть, охватывающую образцы с 1 по 24, нельзя рассматривать как однородную, так как в соответствующем ей наборе значений v_k имеются значения, превышающие 7,815. Максимальное же значение v_k в нижней части таблицы оказалось меньше критического 7,815, что позволяет рассматривать группу образцов с 25-го по 28-й как однородную и дальнейший ее анализ прекратить.

Вернемся к верхней части таблицы. Максимальное из соответствующих ей значений критерия равно 32,11 и приурочено к интервалу между 9-м и 10-м образцами. Разделив верхнюю часть таблицы по этому значению на две части и проанализировав их независимо одну от другой, получим значения критерия v_k, приведенные в столбце 6 табл. 24.

Так как в нижней части таблицы (столбец 6) максимальное значение критерия, равное 11,27, превышает допустимое 7,815, то интервал от 10 до 24 образца следует разделить на две части между 13 и 14 образцами. Максимальное значение критерия v_k в верхней части таблицы, равное 14,44, значительно превышает 7,815 и в связи с этим верхнюю часть матрицы следует рассматривать как неоднородную. После того как в верхней части таблицы была применена уже описанная процедура разделения по максимальному значению критерия v_k с последующим анализом полученных двух групп, оказалось, что обе группы можно рассматривать как однородные. Соответствующие значения критерия v_k приведены в столбце 7 табл. 24.

В результате изучаемый набор 28 образцов оказался разделенным на 5 однородных групп. В первую из них входят образцы, начиная с 1-го по 4-й номер включительно, во вторую — с 5-го по 9-й, в третью — с 10-го по 13-й, в четвертую — с 14-го по 24-й, а в пятую — с 25-го по 28-й. Попарное сравнение смежных групп показало, что среди четырех найденных границ не содержится ложных, так как соответствующие им значения критерия v_k оказались больше, чем 7,815. Эти значения приведены в столбце 8 табл. 24.

Таким образом, в результате статистического изучения разрез скважины расчленен на пять участков, однородных в смысле количественных соотношений между песчаными, бентосными и планктонными фораминиферами. Каждый из этих участков можно рассматривать как соответствующий стабильному периоду фациальных условий бассейна. Кроме того, в данном разрезе выделяются четыре отчетливых момента смены соотношений между песчаными, бентосными и планктонными формами, которые можно рассматривать как моменты смены фациальных условий.

Пример 2. В качестве второго примера разграничения совокупности геологических наблюдений на статистически однородные группы в случае, когда совокупность наблюдений линейно упорядочена, рассмотрим данные по анализу проб керна одной из скважин Кольского полуострова, которые были любезно представлены автору С. Ф. Соболевым. В данном случае рассматривается 47 проб, отобранных из керна в интервале глубины от 1814,0 до 2186,4 м из пород, представленных сланцами, диабазами, алевролитами, филлитами, туфами, габбро-диабазами, туффитами и песчаниками. В каждой пробе проводились определения содержаний следующих 13 компонент: Co, Mn, Zn, Cu, Ti, Li, V, Sr, W, Pb, Ag, K, Na. Результаты этих анализов приведены в табл. 25.

Из табл. 26 видно, что полученные результаты границы статистически однородных участков в четырех случаях из шести совпадают с границами между различными геологическими образованиями. Кроме того, две границы не связаны с изменениями петрографического состава пород, что свидетельствует об изменении изучаемого комплекса химических элементов внутри габбро-диабазов. Среди трех выделенных разновидностей габбро-диабазов наблюдаются, с одной стороны, слабые различия (между участками 4 и 5), так как вычисленное значение $v_k = 22{,}38$ незначительно превосходит критическое 22,36, а с другой стороны, сильные, о чем свидетельствует вычисленное значение $v_k = 42{,}5$ для пятого и шестого участков, намного превышающее критическое.

Необходимо отметить, что на основании проведенного анализа удалось установить, что комплекс изучаемых элементов практически не меняется во втором интервале от второй до девятнадцатой пробы, несмотря на частую смену петрографических разновидностей пород. Это свидетельствует о том, что в данном интервале петрографический фактор не влияет на поведение изучаемого комплекса элементов, что также относится и к седьмому участку.

Таблица 25

Содержания элементов в керне скважины

Номер п.п	Co	Mn	Zn	Cu	Ti	Li	V	Sr	W	Pb	Ag	K	Na
1	260	170	48	260	19	12	190	6	3	94	5	0,12	0,055
2	110	50	50	240	4	6	190	5	5	60	25	0,15	0,56
3	100	57	60	330	4	2	160	5	9	48	5	0,12	1,69
4	65	57	48	820	5	2	170	3	3	13	5	0,13	3,05
5	230	140	63	330	5	10	400	4	3	33	5	0,27	1,69
6	150	88	37	260	25	20	190	5	3	20	5	0,37	1,65
7	270	56	50	153	26	19	190	6	5	36	50	0,48	3,55
8	41	58	36	210	4	1	200	4	3	57	10	0,35	1,32
9	70	50	43	410	5	2	130	5	20	11	5	0,28	1,62
10	60	270	130	250	9	24	160	10	3	9,6	5	0,20	0,48
11	280	80	20	520	12	35	290	3	37	42	90	2,44	1,31
12	5	10	62	820	3	3	210	3	9	17	5	0,13	1,34
13	240	42	50	150	3	11	60	7	36	34	100	2,07	1,85
14	8	70	40	160	3	8	70	6	12	17	5	0,23	2,87
15	120	130	90	200	5	15	90	8	10	8,3	5	0,14	1,39
16	190	80	28	300	8	9	60	4	25	38	70	1,58	2,22
17	87	76	27	350	7	7	80	4	5	24	20	0,13	2,10
18	110	60	62	450	3	2	140	4	3	32	5	0,27	2,04
19	40	30	70	56	3	2	200	4	3	10	45	0,21	1,73
20	7	17	43	540	3	3	210	4	11	20	5	0,09	2,39
21	170	60	20	280	4	21	30	3	3	3,5	5	0,26	2,63
22	330	100	24	270	5	16	30	3	3	5,9	5	0,08	2,04
23	140	40	50	280	3	3	30	3	5	10	30	0,15	3,47
24	100	50	9	130	3	16	30	3	3	0,25	5	0,01	0,01
25	65	60	52	850	3	3	130	3	3	21	5	0,05	2,96
26	95	90	64	1000	3	3	210	3	3	16	5	0,08	1,80
27	80	70	70	1000	6	3	240	3	7	23	5	0,10	2,15
28	5	5	25	120	3	3	130	3	10	9,7	5	0,05	2,96
29	40	35	50	410	3	3	120	3	5	15	30	0,05	2,22
30	40	30	50	280	3	3	170	3	5	16	40	0,13	1,47
31	5	510	35	600	3	1	190	5	3	13	5	0,12	2,22
32	120	100	55	840	3	2	180	4	3	14	5	0,08	1,94
33	150	100	47	750	3	3	180	3	17	5,1	5	0,40	1,24
34	180	60	50	330	3	3	170	3	5	15	40	0,16	2,56
35	60	60	54	840	3	2	160	3	3	0,25	5	0,01	0,01
36	65	40	48	540	3	2	110	3	7	21	5	0,08	2,27
37	100	90	64	260	33	10	180	8	60	16	5	0,06	2,64
38	143	53	50	360	25	6	160	7	5	20	30	0,05	2,32
39	220	60	45	210	17	6	150	4	54	38	50	0,86	0,79
40	104	48	50	380	16	7	200	4	5	20	70	0,05	1,30
41	65	70	43	650	7	4	270	4	3	32	5	0,09	1,62
42	130	65	25	240	8	7	200	4	76	29	80	2,23	0,5
43	140	65	24	210	8	6	220	4	35	35	60	1,72	1,48
44	180	58	25	160	5	4	170	3	40	40	50	1,53	1,06
45	160	60	37	240	8	5	380	3	52	0,25	5	0,01	0,01
46	190	64	25	160	49	2	360	9	29	38	60	1,65	0,19
47	140	65	21	140	31	2	200	7	39	36	110	2,84	0,20

Таблица 26

Результаты разграничения совокупности проб из керна скважины по данным анализа на 13 компонент

Номер п/п	Средняя глубина	Породы	Номер однород-ного участка	v_k	$\chi^2_{q,\,m}$
1	1814,0	Сланцы по туффитам	1	23,25	22,36
2					
3	1823,5	Диабазы			
4					
5		Переслаивание алевролитов			
	1833,0	и филлитов			
6					
7					
8		Диабазы			
	1866,0				
9					
10	1878	Туффиты	2		
11		Переслаивание алевролитов			
	1915,8	и филлитов			
12					
13					
14					
	1920,5	Туфы			
15					
16					
	1941,5	Диабазы			
17					
18					
19	1975,2	Диабазы			
20				29,67	22,36
21					
22	1985,5	Туфы алевролитовые	3		
23					
24				28,04	22,36
25					
26		Габбро-диабазы	4		
27				22,88	22,36
28					
29	2020,0	Габбро-диабазы			
30					
31					
32			5		
33					
34					
35					
36				42,50	22,36
37					
38					
39	2125,5	Габбро-диабазы	6		
40					
41					
				35,74	22,36

Номер п/п	Средняя глубина	Породы	Номер однородного участка	v_k	$\chi^2_{q,\,m}$
42	2155,0	Песчаники с прослоями филлитов			
43					
44		Песчаники и алевролиты с прослоями филлитов	7		
45	2178,1				
46					
47					

Характерно также, что алевролитовые туфы очень отчетливо отделяются по данному комплексу элементов от вышележащих диабазов и расположенных ниже габбро-диоритов.

Ниже в табл. 27 приведены статистические характеристики выделенных однородных участков. В этой таблице $\bar{x}$ — среднее арифметическое, s — оценка стандартного отклонения.

В результате применения метода, описанного в разделе 9.3.2, изучаемый интервал разреза был разделен на 7 статистически однородных, относительно данного комплекса признаков, участков, которые представлены в табл. 26.

Пример 3. В качестве примера разграничения совокупности многомерных наблюдений, расположенных на плоскости, воспользуемся результатами анализа проб воды из придонной части залива Кара-Богаз-Гол на 12 компонент — плотность (d), сухой остаток (p), CO_3, HCO_3, SO_4, Cl, P_2O_5, Br, Ca, Mg, K, Na. Всего было взято 68 проб, схема расположения которых приведена на рис. 12. Все эти данные, которые приведены в табл. 28, любезно представлены автору В. П. Фединым.

Задача заключалась в том, чтобы по приведенным в таблице данным выделить зоны стабильного поведения комплекса перечисленных выше компонент, установив между ними границы.

В результате применения метода, описанного в разделе 9.3.1, изучаемая совокупность была разделена на 14 групп, самая большая из которых содержит восемь наблюдений, а несколько минимальных по три наблюдения. Результаты этого разграничения представлены в табл. 29, где приведены значения критерия τ, соответствующие всем парам выделенных групп. Нетрудно видеть, что все они превышают критическое значение, равное 3, что позволяет признать все границы обоснованными.

Полученные результаты очень трудно поддаются интерпретации в связи с тем, что за счет чувствительности критерия и высокой детальности опробования улавливаются мелкие локальные изменения изучаемого комплекса компонент. Для того чтобы результаты стали более приемлемыми для интерпретации, необходимо

или сделать критерий более грубым, как это описано в разделе 9.5, что дает возможность избежать влияния локальной изменчивости и выявить главные изменения, или же разредить сеть опробования, что также даст возможность избежать излишней детальности.

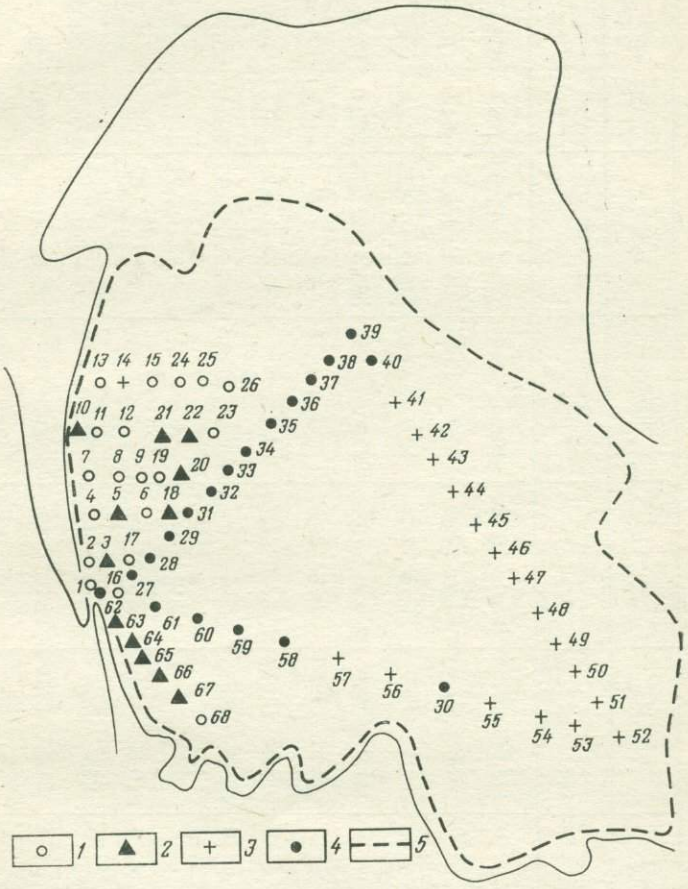

Рис. 12. Результаты разграничения совокупности проб воды из придонной части залива Кара-Богаз-Гол

Пробы: *1* — 1-й группы; *2* — 2-й группы; *3* — 3-й группы; *4* — 4-й группы; *5* — границы акватории

В табл. 30 приведены статистические характеристики и перечень проб каждой группы.

С этой целью были проведены различные варианты разграничения, соответствующие разным критическим значениям τ. Ниже приведены критические значения τ и соответствующие им числа полу-

Таблица 27

Статистические характеристики однородных участков

Номер п/п	Элементы	$n_1 = 1$		$n_2 = 19$		$n_3 = 4$		$n_4 = 3$		$n_5 = 9$		$n_6 = 5$		$n_7 = 9$	
		$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s
1	Co	260,0	0	114,9	88,8	185,0	100,0	80,0	15,0	73,0	62,5	126,4	59,17	195,5	70,2
2	Mn	170,0	0	74,8	57,2	62,5	26,3	73,3	15,3	104,4	155,3	64,2	16,62	67,4	14,56
3	Zn	48,0	0	53,1	25,0	25,75	17,4	62,0	9,16	46,0	9,75	50,4	8,20	25,3	5,63
4	Cu	260,0	0	371,2	203,6	240,0	73,5	949,9	86,6	523,3	257,4	372,0	170,5	195,5	48,5
5	Ti	19,0	0	7,21	6,88	3,75	0,96	4,0	1,73	3,0	0	19,6	9,84	15,4	15,2
6	Li	12,0	0	9,53	9,29	14,0	7,7	3,0	0	2,44	0,73	6,6	2,19	7,8	12,2
7	V	190,0	0	168,4	83,3	30,0	0	193,3	56,86	156,6	29,15	192,0	47,6	215,5	104,8
8	Sr	6,0	0	4,95	1,81	3,0	0	3,0	0	3,33	0,71	5,40	1,94	4,8	2,04
9	W	3,0	0	10,8	10,9	3,5	1,0	4,33	2,31	6,44	4,56	25,4	28,9	44,2	19,56
10	Pb	94,0	0	27,9	16,1	5,51	4,0	20,0	3,60	12,82	6,23	25,2	9,34	26,8	16,0
11	Ag	5,0	0	24,5	31,2	11,2	12,5	5,0	0	15,55	16,0	32,0	28,42	53,3	33,07
12	K	0,12	0	0,51	0,7	0,12	0,10	0,07	0,02	0,12	0,11	0,22	0,35	1,51	0,96
13	Na	0,055	0	1,9	0,70	2,04	1,47	2,30	0,59	1,188	0,87	1,73	0,75	0,44	0,55

Таблица 28

Значение изучаемых компонентов в пробах воды со дна залива Кара-Богаз-Гол

Номер п/п	d	p	CO_3	HCO_3	SO_4	Cl	P_2O_5	Br	Ca	Mg	K	Na	Примечания
1	1,196	0,052	0,061	5,493	8,851	0,008	2,738	0,377	2,961	20,810	0,045	0,035	Для точности определений пробы анализировались с контролем путем дублирования анализов
2	1,169	0,044	0,050	4,525	8,841	0,008	2,362	0,365	3,279	19,800	0,035	0,031	
3	1,101	0,028	0,044	2,962	5,953	0,018	1,512	0,200	2,311	13,200	0,022	0,022	
4	1,242	0,058	0,063	6,185	12,008	0,008	3,151	0,473	4,586	23,820	0,048	0,039	
5	1,098	0,026	0,044	2,729	5,547	0,018	1,343	0,197	2,255	12,280	0,020	0,020	
6	1,240	0,064	0,060	6,291	12,132	0,000	3,278	0,484	4,481	27,270	0,048	0,041	
7	1,254	0,062	0,066	6,874	12,531	0,000	3,362	0,489	4,843	28,600	0,039	0,039	
8	1,217	0,066	0,076	6,110	11,231	0,008	2,921	0,435	4,499	25,720	0,042	0,036	
9	1,167	0,048	0,052	4,819	9,033	0,008	2,620	0,234	3,120	20,220	0,037	0,029	
10	1,111	0,028	0,044	3,383	6,449	0,009	1,543	0,239	2,772	14,610	0,022	0,022	
11	1,187	0,048	0,058	5,279	9,791	0,008	2,552	0,446	3,840	22,230	0,041	0,033	
12	1,163	0,042	0,059	4,649	8,830	0,009	2,279	0,322	3,498	19,830	0,035	0,030	
13	1,193	0,050	0,057	5,381	10,402	0,008	2,579	0,449	4,228	23,340	0,041	0,033	
14	1,263	0,062	0,073	7,090	13,106	0,000	3,531	0,519	4,986	29,500	0,052	0,044	
15	1,217	0,056	0,060	5,953	11,453	0,000	2,971	0,437	4,468	25,710	0,046	0,038	
16	1,230	0,054	0,069	6,247	11,736	0,008	3,419	0,461	3,973	23,200	0,048	0,036	
17	1,252	0,058	0,074	6,704	12,640	0,008	3,246	0,490	5,044	23,420	0,050	0,041	
18	1,127	0,012	0,065	3,359	6,400	0,008	1,735	0,279	2,318	14,314	0,029	0,024	
19	1,191	0,028	0,076	4,933	8,700	0,008	25,500	0,372	3,019	19,380	0,040	0,033	
20	1,147	0,027	0,060	4,000	7,367	0,008	1,995	0,380	2,730	16,883	0,033	0,027	
21	1,131	0,021	0,056	3,634	5,858	0,018	1,733	0,293	2,065	14,000	0,038	0,027	
22	1,144	0,023	0,060	3,950	6,076	0,008	1,935	0,297	2,612	16,170	0,028	0,025	
23	1,960	0,036	0,067	5,005	9,670	0,000	2,636	0,386	3,527	21,713	0,043	0,033	
24	1,225	0,031	0,083	5,822	11,059	0,000	3,019	0,440	4,049	24,730	0,046	0,039	
25	1,208	0,040	0,071	5,404	10,303	0,000	2,726	0,500	3,876	23,134	0,043	0,035	
26	1,185	0,036	0,064	4,828	9,495	0,000	2,424	0,486	3,652	21,210	0,037	0,033	
27	1,230	0,081	0,081	5,818	11,968	0,004	3,713	0,553	3,286	25,634	0,057	0,048	
28	1,227	0,081	0,082	5,615	11,699	0,004	3,645	0,553	3,142	25,030	0,055	0,045	
29	1,231	0,087	0,075	6,012	11,978	0,004	3,692	0,553	3,428	26,104	0,057	0,047	

Продолжение табл. 28

Номер п/п	d	p	CO_3	HCO_3	SO_4	Cl	P_2O_5	Br	Ca	Mg	K	Na	Примечания
30	1,235	0,085	0,077	6,044	12,065	0,004	3,796	0,550	3,302	26,110	0,056	0,047	
31	1,243	0,086	0,073	6,168	12,300	0,004	3,721	0,550	3,685	26,980	0,058	0,048	
32	1,245	0,082	0,076	6,469	12,171	0,004	3,761	0,548	3,640	23,876	0,060	0,050	
33	1,245	0,069	0,086	6,529	12,161	0,004	3,734	0,548	3,707	27,121	0,060	0,050	
34	1,246	0,078	0,082	6,559	12,147	0,004	3,700	0,554	3,765	27,357	0,060	0,051	
35	1,245	0,086	0,073	6,537	12,493	0,004	3,706	0,554	3,990	27,808	0,061	0,051	
36	1,248	0,086	0,076	6,607	11,928	0,004	3,776	0,544	3,523	25,973	0,058	0,051	
37	1,246	0,080	0,076	6,595	12,154	0,004	3,683	0,548	3,838	27,328	0,058	0,050	
38	1,250	0,082	0,072	6,782	12,384	0,004	3,788	0,546	3,876	27,934	0,060	0,050	
39	1,251	0,087	0,074	6,662	12,373	0,004	3,809	0,545	3,779	27,761	0,061	0,050	
40	1,255	0,079	0,085	6,561	12,723	0,000	3,775	0,605	3,989	27,952	0,058	0,050	
41	1,252	0,067	0,085	6,716	12,176	0,000	3,709	0,520	3,874	27,482	0,053	0,036	
42	1,255	0,065	0,085	6,656	12,382	0,000	3,643	0,529	4,097	27,90	0,055	0,042	
43	1,261	0,071	0,075	6,931	12,946	0,000	3,733	0,54	4,42	29,124	0,058	0,050	
44	1,26	0,074	0,075	7,015	12,789	0,000	3,703	0,534	4,418	29,854	0,057	0,044	
45	1,256	0,072	0,077	6,851	12,46	0,000	3,732	0,529	4,073	28,090	0,056	0,043	
46	1,263	0,074	0,075	7,119	13,055	0,000	3,788	0,538	4,480	29,494	0,058	0,048	
47	1,256	0,069	0,077	6,846	12,535	0,000	3,721	0,536	4,154	29,062	0,057	0,044	
48	1,258	0,071	0,072	6,910	12,530	0,000	3,700	0,525	4,228	28,323	0,058	0,046	
49	1,250	0,069	0,076	6,952	12,648	0,000	3,737	0,530	4,213	25,576	0,058	0,046	
50	1,260	0,072	0,073	6,934	12,770	0,000	3,732	0,528	4,318	28,939	0,058	0,048	
51	1,260	0,070	0,075	6,980	12,910	0,000	3,777	0,528	4,339	28,832	0,058	0,048	
52	1,260	0,070	0,076	6,786	12,383	0,000	3,619	0,529	4,205	28,959	0,058	0,048	
53	1,255	0,074	0,077	6,882	12,158	0,000	3,610	0,524	4,135	27,910	0,058	0,042	
54	1,256	0,070	0,076	6,820	12,372	0,000	3,580	0,529	4,287	28,040	0,057	0,048	
55	1,255	0,076	0,079	6,730	12,604	0,000	3,705	0,532	4,163	27,986	0,056	0,048	
56	1,257	0,070	0,074	6,898	12,373	0,000	3,581	0,527	4,325	28,234	0,056	0,046	
57	1,254	0,076	0,072	6,740	12,251	0,000	3,718	0,530	3,912	27,703	0,054	0,046	
58	1,250	0,068	0,096	6,705	12,501	0,000	3,751	0,601	3,956	27,838	0,054	0,050	
59	1,245	0,073	0,088	6,548	12,115	0,000	3,715	0,536	3,736	27,339	0,058	0,050	

Продолжение табл. 28

Номер п/п	d	p	CO_3	HCO_3	SO_4	Cl	P_2O_5	Br	Ca	Mg	K	Na	Примечания
60	1,235	0,073	0,089	6,039	11,975	0,000	3,711	0,542	3,408	23,214	0,057	0,049	
61	1,235	0,075	0,087	5,921	11,592	0,000	3,702	0,545	3,119	25,507	0,057	0,049	
62	1,230	0,073	0,097	5,886	11,819	0,000	3,754	0,542	3,142	25,900	0,057	0,049	
63	1,124	0,023	0,056	3,516	6,789	0,018	1,749	0,242	2,653	15,208	0,028	0,024	
64	1,154	0,035	0,055	4,234	8,340	0,018	2,113	0,290	1,470	16,800	0,025	0,027	
65	1,118	0,027	0,052	3,333	6,483	0,018	1,628	0,224	1,609	14,430	0,025	0,020	
66	1,128	0,021	0,056	3,555	6,904	0,009	1,646	0,272	2,933	15,417	0,026	0,023	
67	1,131	0,021	0,056	3,567	7,161	0,009	1,756	0,264	2,904	15,820	0,029	0,023	
68	1,170	0,029	0,067	4,428	8,902	0,008	2,363	0,379	3,234	19,770	0,038	0,030	

Таблица 29

$\tau > 3$

Номер п/п	2	3	4	5	6	7	8	9	10	11	12	13	14
1	6,22	10,00	15,72	15,84	10,74	10,99	16,38	12,16	17,68	10,84	10,28	16,21	9,86
2		6,17	10,74	11,60	16,03	11,26	18,63	13,48	20,83	12,55	10,55	20,15	11,39
3			4,97	6,60	15,52	9,10	16,29	11,13	18,52	10,98	8,87	17,87	8,92
4				7,72	20,66	13,76	21,10	15,94	23,39	15,55	13,41	22,80	13,49
5					20,04	11,81	19,14	14,00	21,47	13,91	11,56	21,12	11,71
6						8,96	15,24	9,16	15,56	10,38	9,49	13,40	8,27
7							3,20	4,57	6,29	7,13	4,81	10,56	6,06
8								7,70	6,70	13,35	11,19	16,26	11,59
9									4,41	7,47	4,65	8,83	3,19
10										13,27	11,21	16,90	10,19
11											3,64	8,86	6,60
12												8,75	7,84
13													4,45

Таблица 30

Результаты разграничения при $\tau > 3$

Номер группы	Состав группы	Компоненты	Среднее арифметическое	Оценки дисперсий
1	11, 19, 26, 1, 25, 24 13, 23 $n_1 = 8$	d	1,293	0,07277870
		p	0,040	0,00000069
		CO_3	0,067	0,00008269
		HCO_3	5,268	0,10913100
		SO_4	9,787	0,63034100
		Cl	0,003	0,00001828
		P_2O_3	5,514	64,6574300
		Br	0,432	0,00240423
		Ca	3,644	0,20845900
		Mg	22,129	2,44197000
		K	0,042	0,00000828
		Na	0,034	0,00000450
2	9, 2, 68, 12 $n_2 = 4$	d	1,167	0,00000957
		p	0,040	0,00006758
		CO_3	0,057	0,00005933
		HCO_3	4,605	0,02848650
		SO_4	8,902	0,008886535
		Cl	0,008	0,00000024
		P_2O_3	2,406	0,021900330
		Br	0,325	0,00426566
		Ca	3,254	0,02491390
		Mg	19,905	0,04470310
		K	0,036	0,00000224
		Na	0,030	0,00000066
3	64, 32, 20 $n_3 = 3$	d	1,148	0,00002633
		p	0,028	0,00003733
		CO_3	0,058	0,00000833
		HCO_3	4,061	0,02298520
		SO_4	7,560	0,49335100
		Cl	0,011	0,00003333
		P_2O_3	2,014	0,00820130
		Br	0,322	0,00250633
		Ca	2,270	0,048428100
		Mg	16,617	0,15200300
		K	0,023	0,00001633
		Na	0,026	0,00000133
4	21, 18, 63, 66, 67 $n_4 = 5$	d	1,128	0,00000418
		p	0,019	0,00001879
		CO_3	0,057	0,00001620
		HCO_3	3,525	0,01053460
		SO_4	6,622	0,25811000
		Cl	0,012	0,00002629
		P_2O_3	1,727	0,00215570
		Br	0,271	0,00042100
		Ca	2,574	0,14215900
		Mg	14,951	0,58713100
		K	0,029	0,00002150
		Na	0,024	0,00000269

Номер группы	Состав группы	Компоненты	Среднее арифметическое	Оценки дисперсий
5	3, 5, 10, 65 $n_5 = 4$	d	1,107	0,00000466
		p	0,027	0,00000091
		CO_3	0,046	0,00001600
		HCO_3	3,101	0,09701690
		SO_4	6,108	0,10844700
		Cl	0,015	0,00000025
		P_2O_3	1,507	0,01374690
		Br	0,215	0,00040200
		Ca	2,239	0,22923800
		Mg	13,630	1,20259800
		K	0,022	0,00000425
		Na	0,021	0,00000133
6	4, 6, 8, 7, 16, 15, 17 $n_6 = 7$	d	1,236	0,00023166
		p	0,059	0,00001923
		CO_3	0,066	0,00004147
		HCO_3	6,337	0,10922600
		SO_4	11,974	0,27453900
		Cl	0,004	0,00001828
		P_2O_3	3,191	0,03511590
		Br	0,466	0,00054966
		Ca	4,556	0,11298600
		Mg	26,962	1,43575500
		K	0,045	0,00001547
		Na	0,038	0,00000428
7	14, 43, 46 $n_7 = 3$	d	1,262	0,00000132
		p	0,069	0,00003900
		CO_3	0,074	0,00000133
		HSO_4	7,046	0,01024430
		SO_4	13,035	0,00667858
		Cl	0,000	0,000000000
		P_2O_3	3,684	0,01831300
		Br	0,532	0,00012699
		Ca	4,628	0,09702270
		Mg	29,372	0,04638670
		K	0,055	0,00001200
		Na	0,047	0,00000933
8	44, 50, 51 49, 48, 52 $n_8 = 6$	d	1,259	0,00000069
		p	0,071	0,00000319
		CO_3	0,074	0,00000269
		HCO_3	6,929	0,00625991
		SO_4	12,671	0,03570260
		Cl	0,000	0,00000000
		P_2O_3	3,710	0,00280857
		Br	0,529	0,00000879
		Ca	4,286	0,00730123
		Mg	28,723	0,05599970
		K	0,057	0,00000026
		Na	0,046	0,00000226

Номер группы	Состав группы	Компоненты	Среднее арифметическое	Оценки дисперсий
9	41, 42, 58, 40	d	1,252	0,00000502
		p	0,069	0,00003958
	$n_9 = 4$	CO_3	0,087	0,00003024
		HCO_3	6,659	0,00499216
		SO_4	12,445	0,05224990
		Cl	0,000	0,00000000
		P_2O_3	3,719	0,00334493
		Br	0,563	0,00207024
		Ca	3,979	0,00852584
		Mg	27,793	0,04515070
		K	0,055	0,00000466
		Na	0,044	0,00004533
10	57, 55, 45, 47, 56, 54	d	1,255	0,00000096
		p	0,072	0,00000861
	$n_{10} = 7$	CO_3	0,076	0,00000533
		HCO_3	6,823	0,00432713
		SO_4	12,399	0,02499350
		Cl	0,000	0,00000000
		P_2O_3	3,663	0,00488837
		Br	0,529	0,00001428
		Ca	4,149	0,01872920
		Mg	28,007	0,02988170
		K	0,056	0,00000157
		Na	0,045	0,00000557
11	28, 27, 30, 29	d	1,230	0,00001090
		p	0,083	0,00000900
	$n_{11} = 4$	CO_3	0,078	0,00001091
		HCO_3	5,872	0,03938290
		SO_4	11,927	0,02010190
		Cl	0,003	0,00000000
		P_2O_3	3,711	0,00393159
		Br	0,552	0,00000224
		Ca	3,289	0,01370230
		Mg	25,719	0,26101100
		K	0,056	0,00000091
		Na	0,046	0,00000158
12	61, 60, 62	d	1,233	0,00000832
		p	0,073	0,00000133
	$n_{12} = 3$	CO_3	0,091	0,00002800
		HCO_3	5,948	0,00642609
		SO_4	11,795	0,03709030
		Cl	0,000	0,00000000
		P_2O_3	3,723	0,00083696
		Br	0,543	0,00000299
		Ca	3,225	0,02525030
		Mg	25,873	0,12548400
		K	0,057	0,00000000
		Na	0,049	0,00000000

Номер группы	Состав группы	Компоненты	Среднее арифметическое	Оценки дисперсий
13	31, 32, 33, 36, 34, 37, 59 $n_{13} = 7$	d	1,245	0,00000228
		p	0,079	0,00004080
		CO_3	0,079	0,00003258
		HCO_3	6,495	0,02303250
		SO_4	12,145	0,01450470
		Cl	0,003	0,00000225
		P_2O_3	3,727	0,00109100
		Br	0,546	0,00003130
		Ca	3,702	0,01045640
		Mg	27,139	0,04092400
		K	0,058	0,00000157
		Na	0,049	0,00000114
14	35, 38, 39 $n_{14} = 3$	d	1,248	0,00001032
		p	0,085	0,00000699
		CO_3	0,073	0,00000099
		HCO_3	6,660	0,01500840
		SO_4	12,416	0,00440025
		Cl	0,003	0,00000000
		P_2O_3	3,767	0,00296233
		Br	0,548	0,00002433
		Ca	3,881	0,01115440
		Mg	27,834	0,00799550
		K	0,060	0,00000033
		Na	0,050	0,00000000

Таблица 31

$\tau > 10$

Номер п/п	2	3	4	5	6	7
1	23,90	15,30	28,67	28,18	21,42	32,49
2		37,60	42,80	42,51	36,31	48,86
3			19,70	17,74	15,55	23,69
4				10,67	18,74	22,41
5					18,95	17,68
6						15,46

ченных групп наблюдений:

критическое значение τ	3	4	5	6	7	8	9	10	15	20
число групп	14	11	9	9	8	8	8	7	6	4

Как видно, число стабильных групп наблюдений, отражающее главные тенденции изменения комплекса изучаемых характеристик,

достигается только при значениях τ, превышающих 15. Здесь нет возможности привести результаты всех вариантов счета и поэтому мы ограничимся лишь двумя, соответствующими критическим значениям, равным 10 и 20. В табл. 31 приведены вычисленные значения критерия τ, превышающие критическое значение, равное 10.

Конкретный состав полученных семи групп приведен в табл. 32.

Результаты вычислений при критическом значении, равном 20, приведены в таблицах 33 и 34.

Таблица 32

Результаты разграничения при $\tau > 10$

Номер группы	Состав группы	Компоненты	Среднее арифметическое	Оценки дисперсий
1	19, 1, 11, 26, 68, 2, 9, 12 $n_1 = 8$	d	1,178	0,00015985
		p	0,040	0,00008098
		CO_3	0,060	0,00006926
		HCO_3	4,869	0,13242400
		SO_4	9,055	0,14629200
		Cl	0,007	0,00000841
		P_2O_3	5,347	65,7421100
		Br	0,372	0,00573339
		Ca	3,326	0,09648500
		Mg	20,046	0,78758800
		K	0,038	0,00001142
		Na	0,031	0,00000421
2	64, 20, 22, 21, 67, 18, 66, 63, 10, 65, 3, 5 $n_2 = 12$	d	1,126	0,00030006
		p	0,024	0,00003151
		CO_3	0,054	0,00004672
		HCO_3	3,518	0,17802100
		SO_4	6,685	0,56971700
		Cl	0,013	0,00002474
		P_2O_3	1,726	0,04602020
		Br	0,265	0,00254033
		Ca	2,386	0,22992500
		Mg	14,927	1,96066900
		K	0,027	0,00002499
		Na	0,023	0,00000624
3	23, 13, 25, 24, 8, 15, 16, 4, 6, 7,17 $n_3 = 11$	d	1,294	0,04906500
		p	0,052	0,00013761
		CO_3	0,067	0,00006136
		HCO_3	5,997	0,32137500
		SO_4	11,389	0,91615100
		Cl	0,003	0,00001745
		P_2O_3	3,027	0,08429750
		Br	0,458	0,00112147
		Ca	4,324	0,19769100
		Mg	23,605	4,86702500
		K	0,044	0,00001229
		Na	0,037	0,00000821

Номер группы	Состав группы	Компоненты	Среднее арифметическое	Оценки дисперсий
4	14, 46, 43, 44, 51, 50, 49, 52, 43 $n_4 = 9$	d	1,260	0,00000278
		p	0,070	0,00001274
		CO_3	0,074	0,00000202
		HCO_3	6,768	0,00992512
		SO_4	12,792	0,05715840
		Cl	0,000	0,00000000
		P_2O_3	3,701	0,00651552
		Br	0,530	0,00003949
		Ca	4,400	0,05791770
		Mg	23,943	0,15027900
		K	0,057	0,00000386
		Na	0,046	0,00000399
5	41, 42, 57, 53, 45, 55, 47, 56, 54 $n_5 = 9$	d	1,255	0,00000210
		p	0,071	0,00001474
		CO_3	0,078	0,00001974
		HCO_3	6,793	0,00715587
		SO_4	12,372	0,02419750
		Cl	0,000	0,00000000
		P_2O_3	3,666	0,00395722
		Br	0,528	0,00002077
		Ca	4,113	0,02240770
		Mg	27,937	0,05286020
		K	0,055	0,00000244
		Na	0,043	0,00001411
6	28, 27, 29, 30, 60, 62, 61 $n_6 = 7$	d	1,231	0,00001012
		p	0,079	0,00003257
		CO_3	0,083	0,00006766
		HCO_3	5,905	0,02350200
		SO_4	11,870	0,02990660
		Cl	0,002	0,00000467
		P_2O_3	3,716	0,00230741
		Br	0,548	0,00002657
		Ca	3,261	0,01644420
		Mg	25,785	0,17912800
		K	0,056	0,00000061
		Na	0,047	0,00000223
7	58, 40, 59, 38, 39, 35, 37, 34, 33, 36, 32, 31 $n_7 = 12$	d	1,247	0,00001186
		p	0,079	0,00004387
		CO_3	0,079	0,00005729
		HCO_3	6,560	0,02252090
		SO_4	12,290	0,04753040
		Cl	0,002	0,00000327
		P_2O_3	3,743	0,00155532
		Br	0,556	0,00049298
		Ca	3,792	0,02105210
		Mg	27,433	0,16202000
		K	0,058	0,00000402
		Na	0,049	0,00000062

Как видно из приведенных таблиц, наиболее приемлем для интерпретации последний результат, когда с помощью метода разграничения были выделены четыре группы проб; в этом случае нам удалось освободиться от влияния мелких локальных флуктуаций и выявить наиболее существенные изменения состава воды на площади дна залива Кара-Богаз-Гол. Полученные результаты представлены также на рис. 12.

На основании полученных данных можно сделать вывод, что территория залива опробована избыточным числом проб, необходимых для выявления наиболее существенных закономерностей. Возникает вопрос об определении минимума числа проб, который бы позволял выявить основные закономерности и не приводил к переопробованию. Эта задача особенно важна в связи с тем, что после перекрытия пролива Кара-Богаз-Гол ожидается значительное снижение уровня воды в заливе, и опробование из-за заиленности берегов будет возможно производить только с вертолета, что, естественно, приведет к существенному удорожанию этих работ.

Таблица 33

$\tau > 20$

	2	3	4
1	37,07	37,46	35,04
2		63,80	63,95
3			33,82

Для решения такой задачи был проведен следующий эксперимент. Имеющаяся в нашем распоряжении исходная выборка, состоящая из 68 проб, была разделена на две, причем в первую попали только четные номера, а во вторую нечетные. Каждая из двух полученных выборок анализировалась отдельно при критическом значении τ, равном 3,00. В результате одна из выборок была разделена на 9 групп, а вторая — на 7. Это свидетельствует о том, что при выборках объема 34 проб по-прежнему сохраняется излишняя детальность разграничения.

После этого исходная выборка тем же способом была разделена на четыре малых выборки, и в результате разграничения при критическом значении, равном 3,00, каждая из них оказалась разделенной на четыре части. Таким образом, при выборках объема 17 проб достигается способность метода улавливать наиболее существенные закономерности, но этот объем можно будет признать приемлемым, если окажется, что состав полученных групп во всех малых выборках хорошо согласуется с составом групп в исходной большой выборке. Такое сравнение проведено в табл. 35, из которого видно, что результаты разграничения по малым выборкам хорошо согласуются с результатами разграничения по большой выборке. Действительно, во всех четырех вариантах разграничения, насчитывающего в общей сложности 16 групп, выявилось только три несовпадения, т. е. 9, 12, 68, попавшие во вторую группу первого и четвертого вариантов, не содержатся во второй группе,

Таблица 34

Результаты разграничения при $\tau > 20$

Номер групп	Состав групп	Компоненты	Среднее арифметическое	Оценка дисперсий
1	19, 1, 11, 26, 68, 2 9, 12, 23, 13, 25, 24, 8, 15, 16, 4, 6, 7, 17 $n_1 = 19$	d	1,245	0,03078540
		p	0,047	0,00014137
		CO_3	0,064	0,00007343
		HCO_3	5,522	0,55776600
		SO_4	10,406	1,96732700
		Cl	0,005	0,00001609
		P_2O_3	4,004	26,9981200
		Br	0,422	0,00475235
		Ca	3,904	0,40366710
		Mg	23,441	9,80208300
		K	0,042	0,00002184
		Na	0,034	0,00001405
2	64, 20, 22, 21, 67, 18, 66, 63, 10, 65, 3,5 $n_2 = 12$	d	1,126	0,00030008
		p	0,024	0,00003151
		CO_3	0,054	0,00004672
		HCO_3	3,518	0,17802100
		SO_4	6,685	0,56971700
		Cl	0,013	0,00002474
		P_2O_3	1,726	0,04602020
		Br	0,265	0,00254033
		Ca	2,386	0,22992600
		Mg	14,927	1,96066900
		K	0,027	0,00002499
		Na	0,023	0,00000624
3	14, 46, 32, 51, 44, 50, 49, 48, 52, 56, 54, 47, 55, 45, 57, 53, 42, 41 $n_3 = 18$	d	1,257	0,00000932
		p	0,070	0,00001305
		CO_3	0,076	0,00001359
		HCO_3	6,880	0,01515920
		SO_4	12,582	0,08512870
		Cl	0,000	0,00000000
		P_2O_3	3,684	0,00523538
		Br	0,529	0,00002900
		Ca	4,256	0,05963740
		Mg	28,440	0,36342100
		K	0,056	0,00000343
		Na	0,045	0,00001056
4	28, 61, 62, 27, 60, 29, 30, 59, 31, 36 32, 37, 33, 34, 35, 39, 38, 40, 58 $n_4 = 19$	d	1,241	0,00007009
		p	0,079	0,00003770
		CO_3	0,081	0,00003867
		HCO_3	6,318	0,12702400
		SO_4	12,133	0,08233470
		Cl	0,002	0,00000334
		P_2O_3	3,733	0,00189759
		Br	0,553	0,00032705
		Ca	3,596	0,087339500
		Mg	23,829	0,83012200
		K	0,057	0,00000383
		Na	0,049	0,00000232

полученной в результате разграничения исходной выборки. Для всех остальных групп отмечается полное совпадение.

Таблица 35

Сравнение результатов разграничения одной выборки ($n = 68$) при критическом значении, равном 20, и четырех выборок по 17 наблюдений при $\tau_{кр} = 3$

Номер групп	Состав групп большой выборки	Состав групп в четырех малых выборках				
		I	II	III	IV	Число несовпадений
1	19, 1, 11, 26, 68, 2 9, 12, 23, 13, 25 24, 8, 15, 16, 4, 6, 7, 17	1, 13, 25 17	19, 7, 11, 23, 15	26, 2 6	24, 8 4,16	Нет
2	64, 20, 22, 21, 67, 18, 66, 63, 10, 65 3, 5	V 9, 21, 5, 65	3, 67, 63	66, 18, 22, 10	V V 12, 68, 20, 64	3 *
3	14, 46, 43, 51, 44, 50, 49, 48, 52, 56 54, 47, 55, 45, 57, 53, 42, 41	41, 57 49, 53, 45	55, 47, 43, 51	50, 54, 42, 46, 14	44, 56 52, 48	
4	28, 61, 62, 27, 60 29, 30, 59, 31, 36, 32, 27, 33, 34, 35, 39, 38, 40, 58	37, 33, 29, 61,	31, 27, 59, 39, 35	34, 30, 58, 38, 62	28, 60, 36, 40, 32	Нет

* Галочками отмечены несовпадения.

Полученный результат свидетельствует о том, что при опробовании залива Кара-Богаз-Гол можно ограничиться значительно более редкой сетью, чем использовалась до сих пор, и судя по полученным данным более редкая сеть должна насчитывать 17—20 проб.

Для контроля полученного вывода было проведено дальнейшее дробление исходной выборки. На этот раз она была разделена на 8 частей по 8 проб в каждой и 4 оставшиеся пробы отброшены. Каждая часть была подвергнута процедуре разграничения при критическом значении, равном 3,00.

В результате три из восьми выборок оказались разделенными на три группы, четыре — на две и одна выборка вообще не разделилась. Таким образом, было установлено, что выборки, объем которых близок к 8, не обеспечивают детальности разграничения, достаточной для выявления наиболее существенных закономерностей.

СПИСОК ЛИТЕРАТУРЫ

1. *Айвазян С. А.* Статистическое исследование зависимостей. М., Металлургия, 1968.

2. *Андерсон Т.* Введение в многомерный статистический анализ. М., Физматгиз, 1963.

3. *Барышев Н. В.* Контроль опробования.— В кн.: Мат. по метод. разв. и подсчета запасов, 1948, вып. 2, с. 88.

4. *Блэкуел Д.* и *Гиршик М.*— Теория игр и статистических решений. М., Изд-во иностр. лит., 1958.

5. *Большев Л. Н.* Асимптотические пирсоновские преобразования.— Теория вероятностей и ее применение, т. 8, вып. 2, 1963, с. 129—154.

6. *Большев Л. Н., Смирнов Н. В.* Таблицы математической статистики. М., Наука, 1965.

7. *Бондаренко В. Н.* Статистические решения некоторых задач геологии. М., Недра, 1970.

8. *Бондаренко В. Н.* Сравнительный анализ геологических объектов с закономерной изменчивостью свойств. М., Недра, 1978.

9. *Бугаец А. Н.* Статистические методы при поисках и оценке пегматитов по геохимическим данным. М., Недра, 1970.

10. *Бугаец А. Н., Л. Н. Дуденко.* Математические методы при прогнозировании месторождений полезных ископаемых. М., Недра, 1976.

11. *Ван дер Варден Б. Л.* Математическая статистика. М., Изд-во иностр. лит., 1960.

12. *Вершковская О. В., Краснова В. С., Родионов Д. А.* Распределение содержаний галлия в сфалеритах флюоритово-сульфидных месторождений.— В кн.: Тр. ИМГРЭ АН СССР, № 6. 1961, с. 38.

13. *Вистелиус А. Б.* Мера связи между членами парагенезиса и методы ее изучения.— Зап. Всес. минер. об-ва, 1948, ч. 77, № 2, с. 147—158.

14. *Вистелиус А. Б.* Расчленение немых толщ по количественным минералогическим, петрографическим и химическим признакам.— Зап. Всес. минер. об-ва, 1957, ч. 86, № 1, с. 99—115.

15. *Вистелиус А. Б., Яновская Т. Б.* Программирование задач геологии и геохимии при использовании универсальных электронных вычислительных машин.— Геология рудных месторождений 1963, № 3, с. 34—48.

16. *Воронин Ю. А., Еганов Э. А., Черемисина Е. Н.* Использование моделирования на ЭВМ для эффективного направления детальных поисковых работ — В кн.: Основы научн. прогноза месторождений рудн. и нерудн. полезн. ископаемых. Л., 1971, с. 178—179.

17. *Гнеденко Б. В.* Курс теории вероятностей. М., Наука, 1965.

18. *Дорофеюк А. А.* Обучение машины распознаванию образов без поощрения.— В кн.: Вопросы технической кибернетики, М., 1966, с. 87—92.

19. *Жуков Н. Н.* Вероятностно-статистические методы анализа геолого-геофизической информации. Киев, Вища школа, 1975.

20. *Зенков Д. А.* Эволюция статистических характеристик содержания компонентов в связи с методами исследований при разведке.— В кн.: Методика опробования рудных месторождений при разведке и эксплуатации. Свердловск, 1960, с. 8—10.

21. *Иванов В. В., Родионов Д. А., Тархов Ю. А.* О характере распределения и среднем содержании индия в некоторых минералах из месторождений различных генетических типов.— Труды ИМГРЭ (Геохимия), 1963, № 11, с. 1016—1026.

22. *Использование* алгоритма распознавания образов для решения задач промысловой геофизики/Ш. А. Губерман, М. Л. Извекова, А. И. Холин, Я. И. Хургин.— Докл. АН СССР, т. 154, вып. 5, 1964, с. 1082—1083.

23. *Каллистов П. Л.* Методы экспериментального определения рациональных схем обработки проб.— Советская геология, 1938, т. 8, № 10, с. 82—98.

24. *Камалов М. К.* Распределение квадратичных форм. Ташкент, Изд-во АН СССР, 1968.
25. *Коган Р. И.* Интервальные оценки запасов полезных ископаемых. М., Недра, 1972.
26. *Колмогоров А. Н.* Основные понятия теории вероятностей. М.—Л., ОНТИ, 1936.
27. *Крамбейн У. и Грейбилл Ф.* Статистические модели геологии. М., Мир, 1969.
28. *Кульбак С.* Теория информации и статистика. М., Наука, 1967.
29. *Лбов Г. С.* Выбор эффективной системы зависимых признаков. Вычислительные системы.— В кн.: Сб. трудов Ин-та математики СО АН СССР, вып. 19, 1965, с. 21—34.
30. *Левинсон-Лессинг Ф. Ю.* Разделение базальтов и андезитов.— Изв. геол. ком. 44 (4), 1925, с. 411—422.
31. *Левинсон-Лессинг Ф.* Ю. Разделение дацитов и липаритов.— Докл. АН СССР, сер. А, № 8, 1930, с. 179—184.
32. *Миллер Р., Кан Дж. С.* Статистический анализ в геологических науках. М., Мир, 1965.
33. *Ониси Х., Санделл Е. Б.* Геохимия мышьяка.— В кн.: Геохимия редких элементов. М., Изд-во иностр. лит., 1959, с. 435—493.
34. *Прохоров Ю. В., Розанов Ю. А.* Теория вероятностей М. Наука, 1967.
35. *Разумовский Н. К.* Характер распределения содержаний металлов в рудных месторождениях.— Докл. АН СССР, 1940, т. 28, № 9, с. 815—817.
36. *Разумовский Н. К.* Логнормальный закон распределения и его особенности.— Зап. Ленингр. горного ин-та, вып. 20, 1948, с. 105—121.
37. *Родионов Д. А.* Функции распределения содержаний элементов и минералов в изверженных горных породах.— М., Наука 1964.
38. *Родионов Д. А.* Статистические методы разграничения геологических объектов по комплексу признаков, М., Недра,, 1968.
39. *Родионов Д. А., М. К. Родионова, Т. М. Забелина.* Полуколичественный анализ в биостратиграфии и палеоэкологии. М., Недра, 1973.
40. *Смирнов Б. В.* Интервальные оценки запасов месторождений природного газа. Математические методы исследований в геологии.— Экспресс-информация ВИЭМС, сер. 2, 1974.
41. *Струве Е. А.* Сборник анализов изверженных и метаморфических пород СССР. М., Изд-во АН СССР, 1940.
42. *Турекьян К. К., Калп Дж. Л.* Геохимия стронция.— В кн.: Геохимия редких элементов. М., Изд-во иностр. лит., 1959, с. 69—156.
43. *Уокер Ф., Польдерварт А.* Долериты Карру Южно-Африканского Союза.— В кн.: Геология и петрография трапповых формаций. М., Изд-во иностр. лит., 1950, с. 8—182.
44. *Феллер В.* Введение в теорию вероятностей и ее приложения. М., Мир, 1964.
45. *Ahrens L. H.* A fundamental low of geochemistry.— Nature, 172 (4390), 1953, p. 1148.
46. *Ahrens L. H.* The lognormal distribution of elements (I—II).— Geoch. et Cosmoch. Acta 5 (2), 1954, pp. 49—73; 6 (2—3), 1954, pp. 121—131.
47. *Ahrens L. H.* Lognormal type distribution (III).— Geoch. et Cosmoch. Acta, 11 (4), 1957, pp. 205—212.
48. *Aitchison J., Brown J. A. C.* The lognormal distribution. Cambrige University Press, 1957.
49. *Aubrey K. V.* Frequency distribution of elements in igneous rocks.— Geoch. et Cosmoch. Acta, 9 (1—2), 1956, pp. 83—89.
50. *Berry A. C.* The accuracy of Gaussian approximation to the sum of independent variates.— Trans. Am. Math. Soc. 1941, 41 (1), pp. 122—136.
51. *Durovič S.* The lognormal distribution of elements.—Geol. Sbornic (Bratislava), vol. 8, 1957.
52. *Fisher R. A.* Statistical methods for research workers. Edinburgh, 1958.
53. *James G. S.* Tests of linear hypotheses in univariate analysis when

the ratios of population variances are unknown.— Biometrika, vol. 41, 1954, pp. 19—43.

54. *Miller R. L., Goldberg E. D.* The lognormal distribution of geochemistry.— Geoch. et Cosmoch. Acta, 8 (1—2), 1955, pp. 53—62.

55. *Rao C. R.* Advanced Statistical Methods in Biometrical Research. New York, 1952.

56. *Rao C R.* Maximum likelihood estimation for the multinomial distribution, Sankhya, 1957, vol. 18, pp. 139—148.

57. *Vistelius A. B.* Skew frequency distribution and fundamental law of geochemical processes.— J. Geol. 68 (1), 1960, pp. 1—22.

ПРИЛОЖЕНИЕ 1. ФУНКЦИЯ (0, 1)-НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt$

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,500	0,496	0,492	0,488	0,484	0,480	0,476	0,472	0,468	0,464
—0,1	460	456	452	448	444	440	436	433	429	425
—0,2	421	417	413	409	405	401	397	394	390	386
—0,3	382	378	374	371	367	363	359	356	352	348
—0,4	345	341	337	334	330	326	323	319	316	312
—0,5	309	305	302	298	295	291	288	284	281	278
—0,6	274	271	268	264	261	258	255	251	248	245
—0,7	242	239	236	233	230	227	224	221	218	215
—0,8	212	209	206	203	200	198	195	192	189	187
—0,9	184	181	179	176	174	171	169	166	164	161
—1,0	159	156	154	152	149	147	145	142	140	138
—1,1	136	134	131	129	127	125	123	121	119	117
—1,2	115	113	111	109	107	106	104	102	100	099
—1,3	097	095	093	092	090	089	087	085	084	082
—1,4	081	079	078	076	075	074	072	071	069	068
—1,6	055	054	053	052	051	050	048	047	046	046
—1,7	045	044	043	042	041	040	039	038	038	037
—1,8	036	035	034	034	033	032	031	031	030	029
—1,9	029	028	027	027	026	026	025	024	024	023
—2,0	023	022	022	021	021	020	020	019	019	018
—2,1	018	017	017	017	016	016	015	015	015	014
—2,2	014	014	013	013	013	012	012	012	011	011
—2,3	011	010	010	010	010	009	009	009	009	008
—2,4	008	008	008	008	007	007	007	007	007	006
—2,5	006	006	006	006	006	005	005	005	005	005
—2,6	005	005	004	004	004	004	004	004	004	004
—2,7	003	003	003	003	003	003	003	003	003	003
—2,8	003	002	002	002	002	002	002	002	002	002
—2,9	002	002	002	002	002	002	002	001	001	001
—3,0	001	001	001	001	001	001	001	001	001	001
—3,1	001	001	001	001	001	001	001	001	001	001
—3,2	001	001	001	001	001	001	001	001	001	001
—3,3	000	001	000	000	000	000	000	000	000	000
0,0	0,500	0,504	0,508	0,512	0,516	0,520	0,524	0,528	0,532	0,536
0,1	540	544	548	552	556	560	564	567	571	575
0,2	579	583	587	591	595	599	603	606	610	614
0,3	618	622	626	629	633	637	641	644	648	652
0,4	655	659	663	666	670	674	677	681	684	688
0,5	691	695	698	702	705	709	712	716	719	722
0,6	726	729	732	736	739	742	745	749	752	755
0,7	758	761	764	767	770	773	776	779	782	785
0,8	788	791	794	797	800	802	805	808	811	813
0,9	816	819	821	824	826	829	831	834	836	839
1,0	841	844	846	849	851	853	855	858	860	862
1,1	864	867	869	871	873	875	877	879	881	883
1,2	885	887	889	891	893	894	896	898	900	901
1,3	903	905	907	908	910	912	913	915	916	918
1,4	919	921	922	924	925	926	928	929	931	932

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
1,5	933	934	936	937	938	939	941	942	943	944
1,6	945	946	947	948	949	951	952	953	954	954
1,7	955	956	957	958	959	960	961	962	962	963
1,8	964	965	966	966	967	968	969	969	970	971
1,9	971	972	973	973	974	974	975	976	976	977
2,0	977	978	978	979	979	980	980	981	981	982
2,1	982	983	983	983	984	984	985	985	985	986
2,2	986	986	987	987	987	988	988	988	989	989
2,3	989	990	990	990	990	991	991	991	991	992
2,4	992	992	992	992	993	993	993	993	993	994
2,5	994	994	994	994	994	995	995	995	995	995
2,6	995	995	996	996	996	996	996	996	996	996
2,7	997	997	997	997	997	997	997	997	997	997
2,8	997	998	998	998	998	998	998	998	998	998
2,9	998	998	998	998	998	998	998	999	999	999
3,0	999	999	999	999	999	999	999	999	999	999
3,1	999	999	999	999	999	999	999	999	999	999
3,2	999	999	999	999	999	999	999	999	999	999
3,2	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000

Примечание. Квантиль U_q порядка q определяется как значение z, для которого $\Phi(z) = q$.

ПРИЛОЖЕНИЕ 2. КВАНТИЛИ $\chi^2_q(n)$ РАСПРЕДЕЛЕНИЯ χ^2

n	0,01	0,025	0,05	0,10	0,90	0,95	0,975	0,99
1	0,00	0,00	0,00	0,02	2,71	3,84	5,02	6,64
2	0,02	0,05	0,10	0,21	4,61	5,99	7,37	9,21
3	0,12	0,22	0,35	0,58	6,25	7,82	9,35	11,35
4	0,30	0,48	0,71	1,06	7,78	9,49	11,14	13,28
5	0,55	0,83	1,15	1,61	9,24	11,07	12,83	15,09
6	0,87	1,24	1,64	2,20	10,65	12,59	14,45	16,81
7	1,24	1,69	2,17	2,83	12,02	14,07	16,01	18,48
8	1,65	2,18	2,73	3,49	13,36	15,51	17,54	20,09
9	2,09	2,70	3,33	4,17	14,68	16,92	19,02	21,67
10	2,56	3,25	3,94	4,87	15,99	18,31	20,48	23,21
11	3,05	3,82	4,58	5,58	17,28	19,68	21,92	24,73
12	3,57	4,40	5,23	6,30	18,55	21,03	23,34	26,22
13	4,11	5,01	5,89	7,04	19,81	12,36	23,74	27,69
14	4,66	5,63	6,57	7,79	21,06	23,69	26,12	29,14
15	5,23	6,26	7,26	8,55	22,31	25,00	27,49	30,58
16	5,81	6,91	7,96	9,31	23,54	26,30	28,85	32,00
17	6,41	7,56	8,67	10,09	24,77	27,59	30,19	33,41
18	7,02	8,23	9,39	10,87	25,99	28,87	31,53	34,81
19	7,63	8,91	10,12	11,65	27,20	30,14	12,85	36,19
20	8,26	9,59	10,85	12,44	28,41	31,41	34,17	37,57
21	8,90	10,28	11,59	13,24	29,62	32,67	35,48	38,93
22	9,54	10,98	12,34	14,04	30,81	33,92	36,78	40,29

n	0,01	0,025	0,05	0,10	0,90	0,95	0,975	0,99
23	10,20	11,69	13,09	14,85	32,01	35,17	38,08	41,64
24	10,86	12,40	13,65	15,66	33,20	36,42	39,36	42,98
25	11,52	13,12	14,61	16,47	34,38	37,65	40,65	44,31
26	12,20	13,84	15,38	17,29	35,56	38,89	41,92	45,64
27	12,88	14,57	16,15	18,11	36,74	40,11	43,19	46,96
28	13,57	15,31	16,93	18,94	37,92	41,34	44,46	48,28
29	14,26	16,05	17,71	19,77	39,09	42,56	45,72	49,59
30	14,95	16,79	18,49	20,60	40,26	43,77	46,98	50,89
31	15,66	17,54	19,28	21,43	41,42	44,99	48,23	52,19
32	16,36	18,29	20,07	22,27	42,59	46,19	49,48	53,49
33	17,07	19,05	20,87	23,11	43,75	47,70	50,73	54,78
34	17,79	19,81	21,67	23,95	44,90	48,60	51,97	56,06
35	18,51	20,57	22,47	24,80	46,06	49,80	53,20	57,34
36	19,23	21,34	23,27	25,64	47,21	51,00	54,44	58,62
37	19,96	22,11	24,08	26,49	48,36	52,19	55,67	59,89
38	20,69	22,88	24,88	27,34	49,51	53,38	56,90	61,16
39	21,43	23,65	25,70	28,20	60,66	54,57	58,12	62,43
40	22,16	24,43	25,51	29,05	51,81	55,76	59,34	63,69
41	22,91	25,22	27,33	29,91	52,95	56,94	60,56	64,95
42	23,65	26,00	28,14	30,77	54,09	58,12	61,78	66,21
43	24,40	26,79	28,97	31,63	55,23	59,30	62,99	67,41
44	25,15	27,58	29,79	32,49	56,37	60,48	64,20	68,71
45	25,90	28,37	30,61	33,35	57,51	61,66	65,41	69,96
46	26,66	29,16	31,44	34,22	58,64	62,83	66,62	71,20
47	27,42	29,96	32,27	35,08	59,77	64,00	67,82	72,44
48	28,18	30,76	33,10	35,95	60,91	65,17	69,02	73,68
49	28,94	31,56	33,93	36,82	62,04	66,34	70,22	74,92
50	29,71	32,36	34,76	37,69	63,17	67,51	71,42	76,15
51	30,48	33,16	35,60	38,56	64,30	68,67	72,62	77,39
52	31,25	33,97	36,44	39,43	65,42	69,83	73,81	78,62
53	32,02	34,78	37,28	40,31	66,55	70,99	75,00	79,84
54	32,79	35,59	38,12	41,18	67,67	72,15	76,19	81,07
55	33,57	36,40	38,96	42,06	68,80	73,31	77,38	82,29
56	34,35	37,21	39,80	42,94	69,92	74,47	78,57	83,51
57	35,13	38,03	40,65	43,82	71,04	75,62	79,75	84,73
58	35,91	38,84	41,49	44,70	72,16	76,78	80,94	85,95
59	36,70	39,66	42,34	45,58	73,28	77,93	82,12	87,17
60	37,49	40,48	43,19	46,46	74,40	79,08	83,30	88,38
61	38,27	41,30	44,04	47,34	75,51	80,23	84,48	89,59
62	39,06	42,13	44,89	48,23	76,63	81,38	85,65	90,80
63	39,86	42,95	45,74	49,11	77,75	82,53	86,83	92,01
64	40,65	43,78	46,60	50,00	78,86	83,68	88,00	93,22
65	41,44	44,60	47,45	50,88	79,97	84,82	89,18	94,42
66	42,24	45,43	48,31	51,77	81,09	85,97	90,35	95,63
67	43,04	46,26	49,16	52,66	82,20	87,11	91,52	96,83
68	43,84	47,09	50,02	53,55	83,31	88,25	92,69	98,03
69	44,64	47,92	50,88	54,44	104,42	89,39	93,86	99,23
70	45,44	48,76	51,74	55,33	85,53	90,53	95,02	80,43
71	46,25	49,59	52,60	56,22	86,64	91,67	96,19	101,62
72	47,05	50,43	53,46	57,11	87,74	92,81	97,35	102,82
73	47,86	51,27	54,33	58,01	88,85	93,95	98,52	104,01
74	48,87	52,10	55,19	58,90	89,96	95,08	99,68	105,20
75	49,48	52,94	56,05	59,80	91,06	96,22	100,84	106,39

n	0,01	0,025	0,05	0,10	0,90	0,95	0,975	0,99
76	50,29	53,78	56,92	60,69	92,17	97,35	102,00	107,58
77	51,10	54,62	57,79	61,59	93,27	98,48	103,16	108,77
78	51,91	55,47	58,65	62,48	94,37	99,62	104,32	109,96
79	52,73	56,31	59,52	63,38	95,48	100,75	105,75	111,14
80	53,54	57,15	60,39	64,28	96,58	101,88	106,63	112,33
81	54,36	58,00	61,26	65,18	97,68	103,01	107,78	113,51
82	55,17	58,85	62,13	66,08	98,78	104,14	108,94	114,70
83	55,99	59,69	63,00	66,98	99,88	105,27	110,09	115,88
84	56,81	60,54	63,88	67,88	100,98	106,40	111,24	117,06
85	57,63	61,39	64,75	68,78	102,08	107,52	112,39	118,24
86	58,46	62,24	65,62	69,68	103,18	108,65	113,54	119,41
87	59,28	63,09	66,50	70,58	104,28	109,77	114,69	120,59
88	60,10	63,94	67,37	71,48	105,37	110,90	115,84	121,77
89	60,93	64,79	68,25	72,39	106,47	112,02	116,99	122,94
90	61,75	65,65	69,13	73,29	105,57	113,15	118,14	124,12
91	62,58	66,50	70,00	74,20	108,66	114,27	119,28	125,29
92	63,41	67,36	70,88	75,10	109,76	115,39	120,43	126,46
93	62,24	68,21	71,76	76,01	110,85	116,51	121,57	127,63
94	65,07	69,07	72,64	76,91	111,94	117,63	122,72	128,80
95	65,90	69,93	73,52	77,82	113,04	118,75	123,86	129,97
96	66,73	70,78	74,40	78,73	114,13	119,87	125,00	131,14
97	67,56	71,64	75,28	79,63	115,22	120,99	126,14	132,31
98	68,40	72,50	76,16	80,54	116,32	122,11	127,28	133,48
99	69,23	73,36	77,05	81,45	117,41	123,23	128,42	134,64
100	70,07	74,22	77,93	82,36	118,50	124,34	129,56	135,81

Примечание. q — порядок квантиля, n — число степеней свободы.

ПРИЛОЖЕНИЕ 3. ЗНАЧЕНИЕ МАТЕМАТИЧЕСКОГО ОЖИДАНИЯ d_n РАЗМАХА (0,1)-НОРМАЛЬНОЙ СОВОКУПНОСТИ И КОЭФФИЦИЕНТЫ $\overline{d}_n$ ДЛЯ РАСЧЕТА ОЦЕНКИ СРЕДНЕГО КВАДРАТИЧЕСКОГО ОТКЛОНЕНИЯ ПО РАЗМАХУ

n	d_n	$\overline{d}_n$	n	d_n	$\overline{d}_n$	n	d_n	$\overline{d}_n$	n	d_n	$\overline{d}_n$
2	1,13	0,886	7	2,70	0,370	12	3,26	0,307	17	3,59	0,279
3	1,69	0,591	8	2,85	0,351	13	3,34	0,300	18	3,64	0,275
4	2,06	0,486	9	2,97	0,337	14	3,41	0,294	19	3,69	0,271
5	2,33	0,430	10	3,08	0,325	15	3,47	0,288	20	3,73	0,268
6	2,53	0,395	11	3,37	0,315	16	3,53	0,283			

Примечание. В приложениях 3 и 4 n — число наблюдений выборки.

ПРИЛОЖЕНИЕ 4. КВАНТИЛИ $d_n(q)$ РАЗМАХА (0, 1)-НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ

n	0,005	0,025	0,05	0,95	0,975	0,995	n	0,005	0,025	0,05	0,95	0,975	0,995
2	0,01	0,04	0,09	2,77	3,17	3,97	12	1,55	1,88	2,07	4,62	4,92	5,54
3	0,13	0,30	0,43	3,31	3,68	4,42	13	1,64	1,97	2,16	4,68	4,99	5,60
4	0,34	0,59	0,76	3,63	3,98	4,69	14	1,72	2,06	2,24	4,74	5,04	5,65
5	0,55	0,85	1,03	3,86	4,20	4,89	15	1,80	2,14	2,32	4,80	5,09	5,70
6	0,75	1,06	1,25	4,03	4,36	5,03	16	1,88	2,21	2,39	4,85	5,14	5,74
7	0,92	1,25	1,44	4,17	4,49	5,15	17	1,94	2,27	2,45	4,89	5,18	5,78
8	1,08	1,41	1,60	4,29	4,61	5,26	18	2,01	2,34	2,51	4,93	5,22	5,82
9	1,21	1,55	1,74	4,39	4,70	5,34	19	2,07	2,39	2,57	4,97	5,26	5,95
10	1,33	1,67	1,86	4,47	4,79	5,42	20	2,12	2,45	2,62	5,01	5,30	5,89
11	1,45	1,78	1,97	4,55	4,86	5,49							

ПРИЛОЖЕНИЕ 5. ПРЕДЕЛЫ r_q ДОПУСТИМЫХ ЗНАЧЕНИЙ ДЛЯ $|\check{r}|$ — МОДУЛЯ ВЫБОРОЧНОГО КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ НЕЗАВИСИМЫХ НОРМАЛЬНО РАСПРЕДЕЛЕННЫХ СЛУЧАЙНЫХ ВЕЛИЧИН ($P\{|\check{r}|<r_q\}=q$)

n	q			n	q		
	0,90	0,95	0,99		0,90	0,95	0,99
1	0,988	0,997	1,000	16	0,400	0,468	0,590
2	900	950	0,990	17	389	456	575
3	805	878	959	18	378	444	561
4	729	811	917	19	369	433	549
5	669	754	875	20	360	423	537
6	0,621	0,707	0,834	25	0,323	0,381	0,487
7	582	666	798	30	296	349	449
8	549	632	765	35	275	325	418
9	521	602	735	40	257	304	393
10	497	576	708	45	243	288	372
11	0,476	0,553	0,684	50	0,231	0,273	0,354
12	457	532	661	60	211	250	325
13	441	514	641	70	195	232	302
14	426	497	623	80	183	217	283
15	412	482	606	90	173	205	267
				100	0,164	0,195	0,254

Число степеней свободы ν находится по формуле $\nu=n-2$, где n — число пар наблюдений, по которым вычисляется r. Пределы для выборочного частного коэффициента корреляции определяются по значениям $\nu=n-k-2$, где k — число фиксируемых компонент.

ПРИЛОЖЕНИЕ 6. КВАНТИЛИ t_q-РАСПРЕДЕЛЕНИЯ СТЬЮДЕНТА

n	0,90	0,95	0,975	0,99	0,995	n	0,90	0,95	0,975	0,99	0,995
1	3,078	6,314	12,706	31,821	63,657	3	1,638	2,353	3,182	4,541	5,841
2	1,886	2,920	4,303	6,965	9,925	4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032	30	1,310	1,697	2,042	2,457	2,750
6	1,440	1,943	2,447	3,143	3,707	32	1,309	1,694	2,037	2,449	2,739
7	1,415	1,895	2,365	2,998	3,500	34	1,307	1,691	2,032	2,441	2,728
8	1,397	1,860	2,306	2,897	3,355	36	1,306	1,688	2,028	2,435	2,720
9	1,383	1,833	2,262	2,821	3,250	38	1,304	1,686	2,024	2,429	2,712
10	1,372	1,813	2,228	2,764	3,169	40	1,303	1,684	2,021	2,423	2,705
11	1,363	1,796	2,201	2,718	3,016	42	1,302	1,682	2,018	2,419	2,698
12	1,356	1,782	2,179	2,681	3,055	44	1,301	1,680	2,015	2,414	2,692
13	1,350	1,771	2,160	2,650	3,012	46	1,300	1,679	2,013	2,410	2,687
14	1,345	1,761	2,145	2,625	2,977	58	1,299	1,677	2,011	2,407	2,682
15	1,341	1,753	2,131	2,603	2,947	50	1,299	1,673	2,009	2,403	2,678
16	1,337	1,746	2,120	2,584	2,921	55	1,297	1,667	2,004	2,396	2,668
17	1,333	1,740	2,110	2,567	2,898	60	1,296	1,671	2,000	2,390	2,660
18	1,330	1,734	2,101	2,552	2,878	65	1,295	1,669	1,997	2,385	2,654
19	1,328	1,729	2,093	2,540	2,861	70	1,294	1,667	1,994	2,381	2,648
20	1,325	1,725	2,086	2,528	2,845	80	1,292	1,664	1,990	2,374	2,639
21	1,323	1,721	2,080	2,518	2,831	90	1,291	1,662	1,987	2,369	2,632
22	1,321	1,717	2,074	2,508	2,819	100	1,290	1,660	1,984	2,364	2,626
23	1,320	1,714	2,069	2,500	2,807	120	1,289	1,658	1,980	2,358	2,617
24	1,318	1,711	2,064	2,492	2,797	150	1,287	1,655	1,976	2,352	2,609
25	1,316	1,708	2,060	2,485	2,787	200	1,286	1,653	1,972	2,345	2,601
26	1,315	1,706	2,056	2,479	2,779	250	1,285	1,651	1,970	2,341	2,596
27	1,314	1,703	2,052	2,473	2,771	300	1,284	1,650	1,968	2,339	2,592
28	1,313	1,701	2,048	2,467	2,763	400	1,284	1,649	1,966	2,336	2,588
29	1,311	1,699	2,045	2,462	2,756	500	1,283	1,648	1,965	2,334	2,586

Примечание. n — число степеней свободы, q — порядок квантиля.

ПРИЛОЖЕНИЕ 7. 95%-ные квантили $F_{0,95}(\nu_1, \nu_2)$ распределения Фишера с ν_1 и ν_2 степенями свободы

ν_2	ν_1									
	1	2	3	4	5	6	7	8	9	10
1	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,39	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,30
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,37
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,22
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,02	1,94	1,88	1,83

ПРОДОЛЖЕНИЕ ПРИЛОЖ. 7

ν_2	ν_1								
	12	15	20	24	30	40	60	120	∞
1	243,91	245,95	248,01	249,05	250,09	251,14	252,20	253,25	254,32
2	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,37
6	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,29
8	3,28	3,22	3,10	3,12	3,08	3,04	3,01	2,97	2,93
9	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	2,91	2,87	2,77	2,74	3,70	2,66	2,62	2,58	2,54
11	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	2,31	2,23	2,16	2,11	2,07	2,03	2,98	1,93	1,88
20	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

ПРИЛОЖЕНИЕ 8. 5 %-НЫЕ ЗНАЧЕНИЯ *B*-РАСПРЕДЕЛЕНИЯ ФИШЕРА

	1	2	3	4	5	6	7
0	3,8416	5,9912	7,8148	9,4876	11,0703	12,5919	14,0670
0,04	3,9940	6,1108	7,9186	9,5821	11,1589	12,6750	14,1474
0,16	4,4394	6,4613	8,2254	9,8627	11,4217	12,9247	14,3868
0,36	5,1320	7,0209	8,7220	10,3202	11,8515	13,3349	14,7802
0,64	6,0050	7,7590	9,3881	10,9402	12,4383	13,8965	15,3225
1,00	7,0018	8,6424	10,2023	11,7073	13,1704	14,6000	16,0040
1,44	8,0946	9,6466	11,1462	12,6061	14,0340	15,4363	16,8166
1,96	9,2714	10,7558	12,2045	13,6242	15,0203	16,3952	17,7527
2,56	10,5294	11,9605	13,3671	14,7517	16,1186	17,4691	18,8035
3,24	11,8673	13,2569	14,6276	15,9824	17,3222	18,6486	19,9639
4,00	13,2853	14,6406	15,9808	17,3089	18,6261	19,9318	21,2281
4,84	14,7833	16,1098	17,4248	18,7299	20,0256	21,3130	22,5920
5,76	16,3612	17,6627	18,9564	20,2410	21,5185	22,7892	24,0522
6,76	18,0912	19,3002	20,5744	21,8416	23,1024	24,3572	25,6066
7,84	19,7571	21,0195	22,2775	23,5283	24,7745	26,0161	27,2526
9,00	21,5751	22,8216	24,0639	25,3019	26,5349	27,7634	28,9875
10,24	23,4731	24,7059	25,9346	27,1597	28,3801	29,5980	30,8114
11,56	25,4510	26,6710	27,8879	29,1017	30,3116	31,5192	32,7230
12,96	27,5090	28,7178	29,9242	31,1275	32,3272	33,5253	34,7204
14,44	29,6469	30,8458	32,0424	33,2352	34,4276	35,6158	36,8024
16,00	31,8649	33,0545	34,2412	35,4275	36,6098	37,7918	38,9701
17,64	34,1629	35,3442	36,5227	37,7008	38,8765	40,0499	41,2215
19,36	36,5408	37,7143	38,8864	40,0562	41,2241	42,3918	43,5574
21,16	38,9988	40,1652	41,3295	42,4934	43,6551	44,8163	45,9752
23,04	41,5367	42,6958	43,8549	45,0120	46,1679	47,3234	48,4764
25,00	44,1547	45,3077	46,4606	47,6128	48,7637	49,9128	51,0610

ОГЛАВЛЕНИЕ

Дмитрий Алексеевич Родионов

СТАТИСТИЧЕСКИЕ РЕШЕНИЯ В ГЕОЛОГИИ

Редактор издательства *Л. М. Старикова*
Переплет художника *И. М. Пучкова*
Художественный редактор *Е. Л. Юрковская*
Технический редактор *Н. С. Гришанова*
Корректор *М. П. Курылева*

ИБ № 4183

Сдано в набор 02.12.80. Подписано в печать 08.05.81. Т-09140. Формат $60\times90^{1}/_{16}$. Бумага типографская № 2. Гарнитура литературная. Печать высокая. Усл. печ. л. 14,5. Усл. кр.-отт. 14,5. Уч.-изд. л. 14,0. Тираж 4000 экз. Заказ 459/8244—14. Цена 1 р. 10 к.

Издательство «Недра», 103633, Москва, К-12, Третьяковский проезд 1/19

Ленинградская типография № 4 ордена Трудового Красного Знамени Ленинградского объединения «Техническая книга» им. Евгении Соколовой Союзполиграфпрома при Государственном комитете СССР по делам издательств, полиграфии и книжной торговли. 191126, Ленинград, Социалистическая ул., 14.