МИНИСТЕРСТВО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

В.Г.Ворошилов

Математическое моделирование в геологии

Учебное пособие

Томск 2001

УДК 550.8:519.2

Ворошилов В.Г. Математическое моделирование в геологии: Учебное пособие. Томск: Изд. ТПУ, 2001. - 124 с.

В учебном пособии изложены вероятностно-статистические методы обработки геологической информации и методы математического моделирования свойств геологических объектов и явлений, используемые в геологической практике.

Пособие подготовлено на кафедре геологии и разведки месторождений полезных ископаемых ТПУ и предназначено для студентов специальности 080200 «Геология и разведка месторождений полезных ископаемых» направления 553200 «Геология и разведка полезных ископаемых».

Печатается по постановлению Редакционно-издательского Совета Томского политехнического университета

Рецензенты:

Парначев В.П., д-р.г.-м.н., профессор, зав.кафедрой динамической геологии Томского государственного университета.

Летувнинкас А.И., к.г.-м.н., профессор кафедры минералогии и кристаллографии Томского государственного университета

ВВЕДЕНИЕ

Курс «Математическое моделирование в геологии» введен в вузах сравнительно недавно, поэтому слабо обеспечен специальной учебной литературой. Имеющиеся справочники и монографии по применению статистики и математического моделирования в геологии, как правило, трудны для начального изучения предмета.

Настоящее пособие ставит своей целью в сжатой, но доступной форме ознакомить читателя с основами применяемых в геологии методов математического моделирования. Приводимые приемы статистической обработки данных сопровождаются примерами из конкретной геологической практики. Поскольку круг затронутых вопросов весьма широк, не все они освещены одинаково подробно. Более детально обсуждаются понятия, имеющие первостепенное значение для дальнейшего восприятия материала, а также методы, наиболее часто используемые в практике.

Материал излагается в порядке возрастания сложности, поэтому его нужно осваивать последовательно. В целях большей доступности работы для малоподготовленного читателя, все математические формулы и выкладки приводятся без доказательств, лишь с пояснениями, необходимыми для понимания их смысла. С этой же целью очень кратко даются основные понятия теории вероятностей и матричной алгебры.

В пособии не рассматриваются приемы и методы компьютерной обработки данных, ввиду их разнообразия. Все широко известные программные продукты (типа Statistica for Windows) сопровождаются подробными описаниями, цитировать которые не имеет смысла. Особенности работы с авторскими компьютерными программами изложены в специальном методическом пособии.

1. КРАТКИЕ ИСТОРИЧЕСКИЕ СВЕДЕНИЯ О ПРИМЕНЕНИИ МАТЕМАТИЧЕСКИХ МЕТОДОВ В ГЕОЛОГИИ

Первые попытки использования методов математической

статистики для обработки геологических наблюдений относятся к XVIII-XIX векам. В этот период их применяли, в основном, для группировки данных в минералогии, палеонтологии и других областях геологии. Систематический характер такие исследования приобретают с конца XIX века. Так, в 1899 году сибиряк Н. Псарев, исходя из нормального закона распределения золота в россыпях, вычислял ошибку оценки среднего содержания золота в россыпях и определял количество проб, необходимых для оценки среднего содержания с заданной точностью. Применение вероятностно-статистических методов в минералогии и петрографии на рубеже XIX-XX в.в. связано с именами Г. Ниггли и Ф.Ю. Левинсона-Лессинга. Именно этими методами они выделили главные семейства горных пород.

В XX-ом веке можно условно выделить три периода использования математических методов в геологии. Первый охватывает отрезок времени до 30-х годов и характеризуется единичными работами ученых по применению математической статистики при опробовании месторождений (Н.Н. Курек, С.Ю.Деборжинский, В.В.Котульский, К.Л.Пожарицкий, Л.И.Шаманский), группировке анализов горных пород и минералов (П.Е.Чирвинский, Ф.Ю.Левинсон-Лессинг), для характеристики изменчивости свойств ископаемых организмов (Д.В. Наливкин).

Во второй период, с 1930 по 1965 годы простейшие статистические методы стали широко применять для оценки изменчивости свойств месторождений, анализа распределения химических элементов в породах и рудах, для обоснования плотности разведочной сети. Серьезные статистические исследования по этим проблемам проводились В.Г.Соловьевым, Н.В.Барышевым, Н.К.Разумовским, И.П.Шараповым, Д.А.Зенковым, П.Л.Каллистовым, Криге Д.Г., Дж.С.Девисом, Л.И.Шаманским, Д.А.Казаковским, В.В.Богацким и другими.

Третий период начался с середины 60-х годов. Широкое внедрение ЭВМ в практику геологических исследований резко расширило круг решаемых задач и способствовало проникновению математики во все области геологии. Компьютерная революция, докатившаяся до нашей страны к началу 90-х годов, практически сняла технические ограничения,

препятствовавшие ранее применению наиболее трудоемких в вычислительном отношении методов. Современное состояние математических методов в геологии отражено в десятках монографий и сотнях публикаций, из которых следует особо отметить работы Д.А. Родионова, В.Н. Бондаренко, А.Б. Каждана, Н.Н. Боровко, Р.И. Дубова. Среди зарубежных авторов назовем прежде всего тех, чьи работы переведены на русский язык: Крамбейн Ч., Лоули Д., Максвелл А., Матерон Ж., Миллер Р.Л., Криге Д.Г. , Дж.С.Девис и ряд других. В настоящее время в мире опубликовано несколько тысяч книг, посвященных математическому моделированию в геологии. Особенно продуктивно эти методы используются в США, Франции, Японии, ФРГ, Великобритании, что в немалой степени обусловлено высоким уровнем развития компьютерных технологий в этих странах.

2. ПОНЯТИЕ О ГЕОЛОГО-МАТЕМАТИЧЕСКОМ МОДЕЛИРОВАНИИ ОБЪЕКТОВ И ЯВЛЕНИЙ

Необходимость применения моделей при описании природных объектов связана с тем, что геологические системы управляются одновременно многими факторами различной физической природы и не поддаются строгому количественному описанию. В отличие от закона, имеющего характер абсолютной истины, модель дает лишь приближенное представление об объекте, точнее, о тех его свойствах, для изучения которых осуществлялось моделирование. Создание геолого-математической модели осуществляется в следующей последовательности:

1)Получение исходных данных об объекте или явлении путем измерения и определения его свойств.

2)Создание геологической модели объекта и формулировка геологической задачи.

3)Выражение поставленной задачи в математической форме. Создание математической модели. При этом может возникнуть необходимость в получении дополнительных данных или в уточнении геологических представлений об объекте.

4)Математические расчеты в соответствии с принятой моделью.

5)Проверка соответствия полученных результатов фактическим данным. Если геологических моделей было несколько (это обычный случай), можно оценить, какая из них лучше соответствует действительности.

Поскольку полученная модель учитывает лишь отдельные свойства объекта, ее можно последовательно усложнять и детализировать. Чем сложнее модель, тем более достоверно она отражает изучаемый объект и позволяет более надежно прогнозировать его свойства. Однако в реальных условиях существует оптимальная степень сложности математических моделей, которая определяется с учетом требований к точности решения поставленной задачи. Степень сложности модели может также ограничиваться возможностями аналитических решений и электронно-вычислительной техники.

Таким образом, в геологии моделируются не сами объекты, а изменчивость их свойств, наблюдаемая на данном уровне изучения объекта. Характер этой наблюдаемой изменчивости зависит не только от природы явления, но и от детальности геологических исследований и методики их проведения. В связи с этим необходимо рассмотреть понятие геологической совокупности.

Под *геологической совокупностью* понимают множество геологических объектов, объединенных каким-либо признаком. Например, совокупность образцов гранитов Тигертышского комплекса, совокупность галек русла реки Томи. В первом случае объединяющим признаком является принадлежность всех образцов к гранитам Тигертышского комплекса, во втором случае - принадлежность всех галек к руслу реки Томи. Такую геологическую совокупность мы будем называть *изучаемой*. Понятно, что далеко не вся изучаемая совокупность доступна нам для наблюдения. Геологу чаще всего приходится довольствоваться лишь отдельными обнажениями, характеризующими часть изучаемого объекта. Отсюда ясно, что необходимо различать изучаемую и *опробуемую* совокупность и всегда отдавать себе отчет в том, насколько вторая представительна по отношению к первой. В том случае, когда обнаженность объекта позволяет произвольно формировать опробуемую совокупность, объем ее и степень

представительности определяется, исходя из имеющихся данных и личного опыта геолога. Однако источник возможных ошибок не ограничивается несовпадением изучаемой и опробуемой совокупности. Последняя также не может быть исследована в полном объеме. Геолог обычно ограничивается определенным количеством образцов, проб, замеров и т.д. Множество всех произведенных над опробуемой совокупностью наблюдений образует *выборочную* совокупность, или просто *выборку*. Очевидно, что выборочная совокупность во много раз меньше опробуемой. В то же время именно по результатам выборочных наблюдений делаются выводы не только по опробуемой, но и по всей изучаемой совокупности. Это обстоятельство всегда надо иметь в виду, делая какие-либо выводы, иначе самые точные вычисления не спасут от ошибки.

К выборочным данным предъявляются следующие требования:

1) выборка должна состоять из наблюдений, полученных в одинаковых условиях;

2) наблюдения должны быть независимы друг от друга.

Возможность распространения выводов, полученных по выборочным данным, на всю изучаемую совокупность обеспечивается применением методов математической статистики.

Статистика - это наука, изучающая закономерности, которым подчинены массовые случайные явления. Из этого определения следует, что использование математической статистики для моделирования свойств геологических объектов возможно лишь в том случае, если геологические наблюдения удовлетворяют условию массовости (то есть, их можно многократно повторять при одних и тех же условиях), могут быть представлены в виде схемы случайных событий и выражены случайной величиной. Проведение геологических исследований обычно заключается в замерах значений изучаемого свойства в произвольных точках пространства. Эти замеры можно поэтому рассматривать как серию случайных событий, а получаемые результаты - числовые значения - как случайные величины, поскольку их невозможно предсказать заранее. Замеры эти можно повторять многократно. Следовательно, явления,

изучаемые в процессе геологических исследований, могут рассматриваться как случайные и массовые и для них правомерно использование статистических методов.

Теоретической базой математической статистики является теория вероятностей, отдельные положения которой мы рассмотрим ниже.

3. ОСНОВЫ ТЕОРИИ ВЕРОЯТНОСТЕЙ

3.1. Основные определения и понятия

В статистическом моделировании одним из главных является понятие о вероятности случайного события.

Под *событием* понимается любой факт, который может реализоваться в результате опыта или испытания. Под *опытом* или *испытанием*, в свою очередь, понимается осуществление определенного комплекса условий, причем, не обязательно с участием человека. Классическим примером испытания является подбрасывание вверх монеты, а выпадение герба или цифры, соответственно, является событием. Примером испытания, проходящего без участия человека, можно считать извержение вулкана. К событиям, возникшим в результате данного испытания, относятся средняя мощность лавового потока, процент пирокластики, химический состав лавы и т.д. Событие может заключаться в появлении или непоявлении какого-либо признака в одном из многих испытаний. Например, присутствие золота в одной из многих шлиховых проб, наличие концентрации урана более 0,1% в одной из многих штуфных проб и т.д.

Все события подразделяются на достоверные, невозможные и случайные. *Достоверным* называется событие, которое неизбежно произойдет при каждом испытании данного вида. *Невозможное* событие никогда не реализуется при данном виде испытаний. События третьего типа характеризуются тем, что они могут произойти в данном испытании, а могут и не произойти. Если испытание повторяется многократно, то в одних случаях эти события произойдут, а в других нет. В каких именно случаях события реализуются, мы заранее предсказать не можем, поэтому такие события и называются *случайными*.

Рассмотрим пример случайного события. По одному из

рудных тел медного месторождения отобрано по равномерной сети 1000 проб, содержание меди в которых колеблется от 0,1% до 5%. Кондиционным является содержание в 2%. Наличие меди в любой наугад взятой пробе будет событием достоверным, а вот содержание в ней меди свыше 2% -событие случайное. Если мы разделим количество проб с кондиционным содержанием на общее количество проб, то получим величину коэффициента рудоносности для данного рудного тела. Эта величина будет меняться от одного рудного тела к другому, причем заранее нельзя предсказать, какое значение она примет в каждом конкретном случае, то есть, это величина случайная.

Итак, *случайной* называется величина, принимающая в результате испытания то или иное, заранее неизвестное, значение.

Случайные величины бывают *дискретными* (прерывистыми) и *непрерывными*. При этом значения, которые они принимают, могут ограничиваться какими-либо пределами, а могут и не ограничиваться.

Дискретная величина может принимать только какие-то фиксированные значения и, если задан интервал, то число этих значений конечно. Например, дискретной величиной является число знаков золота в шлиховой пробе, число буровых скважин на участке и т.д. Непрерывная случайная величина может принимать бесконечное множество значений в любом заданном интервале. В рассмотренном выше примере содержание меди в пробах колеблется от 0,1 до 5,0%. Внутри этого интервала величина содержания меди теоретически может принимать бесконечное множество значений, поэтому является величиной непрерывной.

Случайная величина характеризуется тем, что может принимать множество различных значений, однако все эти значения имеют разную возможность проявления. Допустим, в ящике лежит 100 образцов, 90 из которых содержит пирит, 9 - халькопирит и 1 - галенит. Если мы возьмем наугад один образец, то, скорее всего, он будет с пиритом. Возможность вынуть образец с халькопиритом будет значительно ниже, а с галенитом - и вовсе ничтожна. В качестве количественной меры возможности появления случайного события используется

величина, называемая вероятностью.

Вероятность события А - это число, которое характеризует степень объективной возможности появления этого события. Оно обозначается Р(А) или просто р, т.е. р=Р(А).

Существует несколько определений вероятности, из которых мы рассмотрим два: классическое и статистическое. Классическое определение гласит, что вероятность события А равна отношению числа случаев, благоприятствующих событию А, к общему числу случаев. В рассмотренном выше примере вероятность того, что первый, наугад вынутый образец, окажется с халькопиритом равна:

$$P(A) = \frac{9}{100} = 0{,}09$$

На практике классическое определение зачастую неприменимо, так как общее число случаев обычно неизвестно или бесконечно. Кроме того, далеко не всегда можно представить исходы опыта в виде равновозможных и несовместимых событий.

Между тем, давно было замечено, что частота появления событий при многократном повторении опыта имеет тенденцию стабилизации около какой-то постоянной величины. Это свидетельствует о том, что данные события тоже обладают определенной степенью возможности появления в опыте, меру которой можно представить в виде относительной частоты или частости.

Частость (относительная частота) - это отношение числа опытов, благоприятствовавших событию А, к общему числу произведенных испытаний. Швейцарским ученым Яковом Бернулли доказано, что при большом числе испытаний частость стремится воспроизвести вероятность и в пределе совпадает с ней. Следовательно, вероятность Р(А) -это относительная частота появления события В в n произведенных испытаниях (статистическое определение вероятности):

$$P(A) = \frac{m}{n}.$$

Чем больше число *n*, тем вероятность, определенная по этой

формуле, ближе к ее истинному значению. Поэтому на практике всегда надо стремиться к тому, чтобы выборка была достаточно представительной.

Как видно из определений, P(A) изменяется в пределах от 0 до 1. Вероятность достоверного события равна 1, невозможного - 0.

3.2. Закон распределения случайной величины

Законом распределения случайной величины называется зависимость между всеми возможными значениями случайной величины и соответствующими им вероятностями. Закон распределения может быть задан в виде таблицы, графика или функции распределения. Табличный способ наиболее простой, выглядит он следующим образом.

Таблица 1

Задание закона распределения

X	X_1	X_2	X_3	X_4	X_5	X_6	. . . . . X_n
P	P_1	P_2	P_3	P_4	P_5	P_6	. . . . . P_n

Понятно, что табличное задание закона распределения возможно только для дискретной случайной величины с конечным числом значений. На практике непрерывную случайную величину обычно разбивают на ряд интервалов и затем оперируют с центрами интервалов как с дискретной случайной величиной. Графическое изображение такого *ряда распределения* выглядит так.

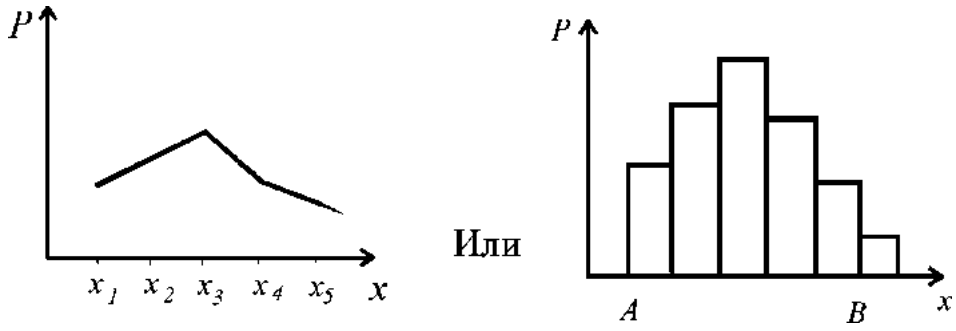


Рис. 1. Графическое изображение рядов распределения

Если истинное значение вероятностей неизвестно, по оси ординат откладывают относительную частоту появления каждого из значений.

Наиболее общей формой задания закона распределения является *функция распределения*. Она определяет вероятность того, что случайная величина ξ примет значение, *меньшее* какого-то фиксированного значения X. Эта вероятность зависит от X и, следовательно, является функцией от X, т.е.

$$F(x)= P(\xi < x) \qquad (1)$$

Если мы построим графическое выражение F(x) по табличным данным, то получим график.

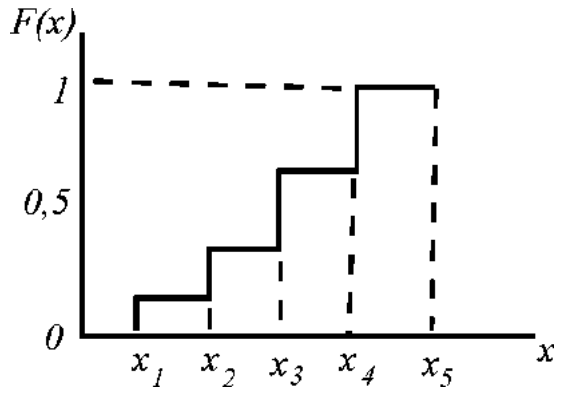


Рис. 2. График интегральной функции распределения дискретной случайной величины

Непрерывная случайная величина имеет график функции распределения в виде плавной кривой.

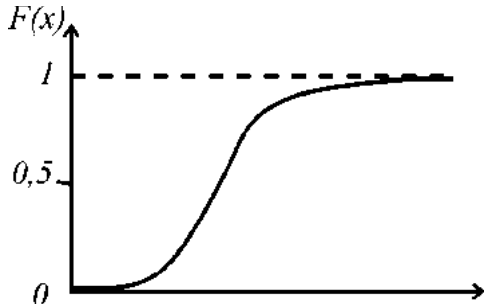


Рис. 3. График интегральной функции распределения непрерывной случайной величины

Описанная функция носит название *интегральной* функции распределения. Отметим ее основные свойства:

1)F(x), как и всякая вероятность, изменяется в пределах от 0 до 1. $P(-\infty) = 0$, $P(+\infty) = 1$;

2) вероятность попадания случайной величины в интервал от A до B равна разности ординат в точках B и A, т.е.

$$P(A \le \xi < B) = F(B) - F(A). \qquad (2)$$

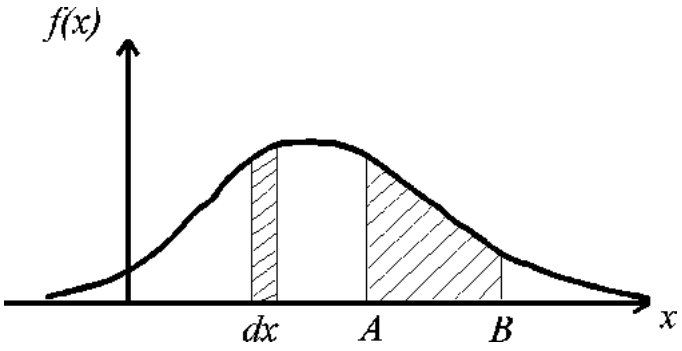


Рис. 4. Графическое изображение функции плотности вероятности (дифференциальная функции распределения)

Непрерывная случайная величина может быть задана не только интегральной, но и *дифференциальной* функцией (или функцией *плотности распределения вероятности*). Она представляет собой первую производную от интегральной функции:

$$f(x) = \frac{dF(x)}{dx} \ . \qquad (3)$$

Выделим на оси x элементарный участок dx **.** Вероятность попадания случайной величины на этот участок, исходя из формулы (3), равна

$$dF(X) = f(X) \cdot dX.$$

То есть, это площадь элементарного прямоугольника со сторонами dx и $f(x)$ (рис.4). Отсюда вытекает вывод о том, что вероятность попадания случайной величины в интервал от A до B численно равна площади криволинейной трапеции, ограниченной графиком f (x), осью A и перпендикулярами в точках A и B. Из курса высшей математики мы знаем, что эта площадь равна интегралу функции $f(x)$ в пределах от A до B.

Итак, отметим основные свойства дифференциальной функции.

1. Поскольку *f (x)* неубывающая функция, то ее первая производная всегда больше или равна нулю. Это означает, что график *f (x)* целиком расположен выше оси *x*.

2. Интегральная функция может быть выражена через дифференциальную по формуле

$$F(x) = \int_{-\infty}^{x} f(x) \cdot dx . \qquad (4)$$

3. Вероятность того, что случайная величина попадет в интервал от A до B равна

$$P(A \le \xi < B) = \int_{A}^{B} f(x) \cdot dx .$$

Помимо вышеизложенных рассуждений, это вытекает также из формул (2) и (4):

$$P(A \le \xi < B) = F(B) - F(A) = \int_{-\infty}^{B} f(x) \cdot dx - \int_{-\infty}^{A} f(x) \cdot dx = \int_{A}^{B} f(x) \cdot dx$$

4. Вся площадь, заключенная под кривой *f (x),* характеризует полную вероятность, поэтому равна 1:

$$\int_{-\infty}^{+\infty} f(x) \cdot dx = 1$$

3.2.1.Основные характеристики положения и рассеяния случайной величины

Закон распределения полностью характеризует случайную величину с вероятностной точки зрения. Однако при решении практических задач обычно нет необходимости знать все возможные значения случайной величины и соответствующие им вероятности. Удобнее пользоваться некоторыми количественными показателями, которые в сжатой форме дают достаточно полную информацию о случайной величине.

Наиболее существенные особенности распределения случайной величины могут быть выражены с помощью числовых характеристик *положения* и *рассеяния*. К важнейшим характеристикам положения относятся математическое ожидание, мода и медиана.

Математическое ожидание характеризует положение

случайной величины на числовой оси, определяя собой некоторое среднее значение, около которого сосредоточены все возможные значения случайной величины. Поэтому математическое ожидание иногда называют просто средним значением случайной величины. Математическое ожидание дискретной случайной величины можно определить как среднее из ее значений, взвешенных по вероятностям их появления:

$$M(x) = \frac{X1P1 + X2P2 + ... + XnPn}{P1 + P2 + ... + Pn} = \frac{\sum XiPi}{\sum Pi}; \qquad (5)$$

$\sum p_i = 1$, поскольку это полная вероятность. Следовательно, $M(x) = \sum x_i p_i$ (5), то есть, математическое ожидание дискретной случайной величины есть сумма произведений всех ее возможных значений на соответствующие им вероятности. Можно доказать, что с увеличением числа испытаний среднее арифметическое ($\bar{x}$)все больше приближается к $M(x)$, а при $n = \infty$ они совпадают.

Математическому ожиданию можно дать механическую интерпретацию. Если вероятности p_i или $f(x)\cdot dx$ принять за веса значений случайной величины, то $M(x)$ есть не что иное, как абсцисса центра тяжести всей системы материальных точек.

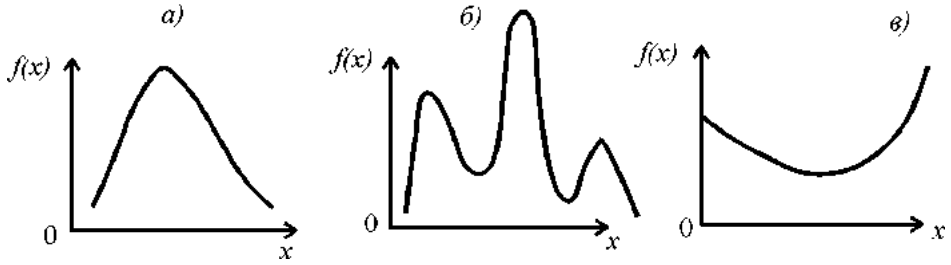


Рис. 5. Одномодальная (а), многомодальная (б) и антимодальная (в) кривые распределения случайной величины.

Модой (Мо) случайной величины называется наиболее вероятное ее значение. Геометрически мода – это абсцисса точки максимума дифференциальной кривой распределения. Кривые распределения могут быть одно- и многомодальными. Есть также кривые, не имеющие максимума, но имеющие минимум. Они называются антимодальными (рис 5).

Медианой (M_e) случайной величины называется такое ее значение, для которого вероятность встречи больших и меньших значений одинакова:

$$F(M_e) = P(\xi < M_e) = P(\xi > M_e) = 0{,}5.$$

С геометрической точки зрения M_e - это абсцисса точки, в которой площадь, ограниченная кривой распределения, делится пополам. Для определения медианы дискретной случайной величины можно расположить все ее значения в порядке возрастания (убывания). В случае четного числа значений, медиана равна полусумме двух средних (по порядку) значений.

Если кривая распределения симметрична относительно среднего значения, то $M(x)$, M_o и M_e равны между собой; в общем случае они не совпадают.

В качестве характеристик рассеяния случайной величины относительно среднего значения обычно используют дисперсию, стандарт и коэффициент вариации.

Дисперсия (δ^2) служит главной характеристикой рассеяния:

$$\delta^2 = \frac{\sum_{i=1}^{n}(Xi - M(x))^2}{n} = \sum_{i=1}^{n} (Xi - M(x))^2 \cdot Pi. \qquad (6)$$

Можно использовать и другую формулу:

$$\delta^2 = \overline{x}^2 - (\overline{x})^2. \qquad (7)$$

Поскольку дисперсия имеет размерность квадрата случайной величины, для оценки разброса значений обычно используют производную от нее характеристику – *стандарт* (среднее квадратическое отклонение):

$$\delta = \sqrt{\delta^2}. \qquad (8)$$

Стандарт выражается в тех же единицах, что и случайная величина и наглядно показывает разброс ее значений. Однако для сравнения степени разброса двух величин, имеющих разную размерность, стандарт применить невозможно. В этом случае используют безразмерный показатель - *коэффициент вариации* (ν):

$$\nu = \frac{\sigma}{M(x)} \cdot 100\%. \qquad (9)$$

Коэффициент вариации с успехом используется для сравнения степени изменчивости различных геологических объектов и явлений.

Кривые распределения случайной величины могут быть симметричными и асимметричными (рис. 6), сжатыми и растянутыми (рис. 7). Эти их свойства отражаются в показателях *асимметрии* (А) и *эксцесса* (Е):

$$\text{A} = \frac{\sum (x_i - M(x))^3 \cdot Pi}{\sigma^3} = \frac{\sum_{i=1}^{n} (x_i - M(x))^3}{n \cdot \sigma^3}, \qquad (10)$$

$$\text{E} = \frac{\sum (x_i - M(x))^4 \cdot Pi}{\sigma^4} - 3 = \frac{\sum_{i=1}^{n} (x_i - M(x))^4}{n \cdot \sigma^4} - 3. \qquad (11)$$

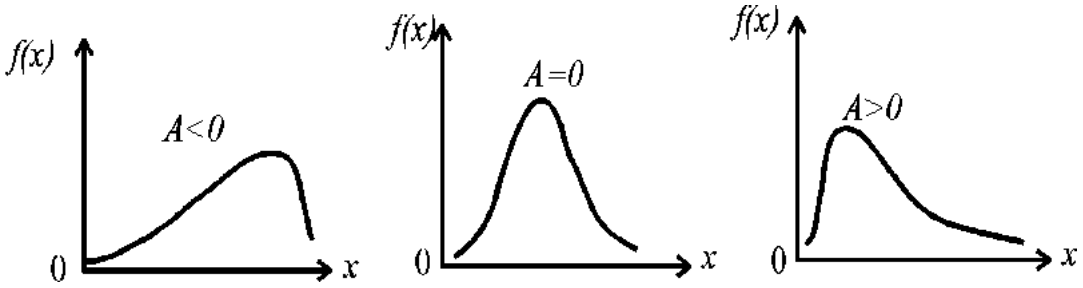


Рис. 6. Симметричность и асимметричность кривых распределения

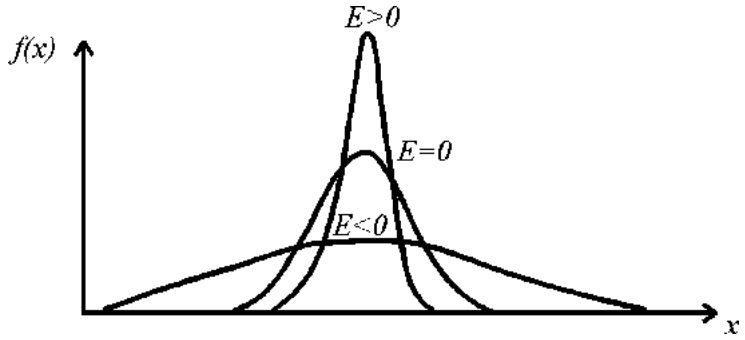


Рис. 7. Сжатость и растянутость кривых распределения

Расчет характеристик положения и рассеяния случайной величины можно осуществить не только по формулам (5) и (11),

но и с помощью моментов случайной величины. В отдельных случаях это оказывается более удобным, особенно при применении вычислительной техники. В учебниках этот способ описан достаточно подробно и здесь, ввиду ограниченности объема пособия, не приводится

3.3. Некоторые теоретические законы распределения случайной величины

Для приближенного описания эмпирически наблюдаемых распределений свойств геологических объектов в практике применяют самые различные теоретические законы распределения случайной величины. При этом часто ограничиваются использованием четырех основных законов: нормального, логнормального, биноминального, Пуассона.

Нормальным называется закон, для которого интегральная функция распределения имеет вид:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \int_{-\infty}^{x} e^{-\frac{(x-M(x))^2}{2\sigma^2 \cdot dx}} \quad . \qquad (12)$$

Функция плотности вероятности, соответственно, описывается выражением:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-M(x))^2}{2\sigma^2}} \quad . \qquad (13)$$

Графическое ее выражение приведено на рис.8.

Функция распределения достигает максимума в точке x= M(x). Допустим, что M(x) = a. Если вместо случайной величины рассмотреть новую случайную величину

$$t = \frac{x-a}{\sigma} \quad , \qquad (14)$$

то новая величина t будет также распределена нормально со средним значением, равным нулю и дисперсией, равной 1:

$$\sigma_t^{\,2} = \overline{(t - M(t))}^{\,2} = \bar{t}^{\,2} = \left(\frac{\overline{x-a}}{\sigma^2}\right)^2 = \frac{\overline{(x-a)^2}}{\sigma^2} = \frac{\sigma^2}{\sigma^2} = 1.$$

Плотность вероятности величины t имеет вид:

$$f(t) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}} \quad . \qquad (15)$$

Это уравнение иногда называют уравнением Гаусса, а соответствующую кривую - кривой Гаусса. Преобразованное таким образом распределение называется *нормированным* или *стандартным нормальным* распределением.

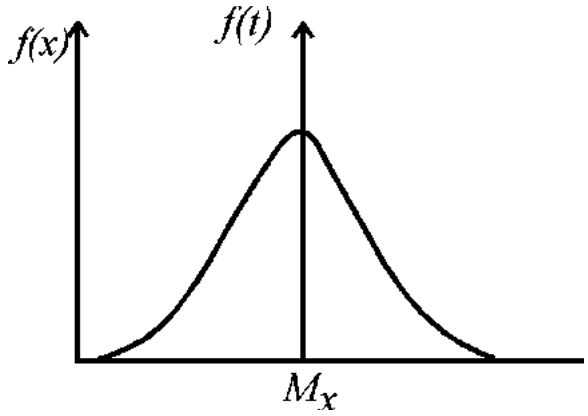


Рис. 8. График функции плотности вероятности нормального распределения

Переход от нормального к стандартному нормальному распределению заключается в переносе центра распределения в начало координат с выражением случайной величины в долях ее стандарта. Необходимость такого преобразования заключается в том, что вычисление вероятностей по формуле (12) представляет собой очень трудоемкую операцию, а составить таблицы для всех возможных значений случайной величины не представляется возможным. Такие таблицы составлены для нормированной (безразмерной) величины t , для которой, как мы увидим ниже, вполне достаточно иметь таблицу значений $F(t)$ в интервале $-3 \le t \le 3$.

Переход от реальных значений случайной величины к нормированным по формуле (14) не представляет никаких трудностей.

В справочниках приводятся таблицы для $f(t)$, $F(t)$ и $\Phi(t)$:

$$F(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-t} e^{-\frac{t^2}{2}} dt \quad , \qquad (16)$$

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-t}^{+t} e^{-\frac{t^2}{2}} dt, \qquad (17)$$

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

Интеграл (16) определяет вероятность попадания случайной величины в интервал от $-\infty$ до t, а интеграл (17) от $-t$ до $+t$. Чтобы определить вероятность попадания случайной величины в интервал от A до B, необходимо вначале нормировать границы интервала:

$t_1 = \frac{A - Mx}{\sigma}$, $t_2 = \frac{B - Mx}{\sigma}$, а затем найти соответствующие значения $F(t)$ в таблице и вычислить искомую вероятность:

$$P = F(t_2) - F(t_1).$$

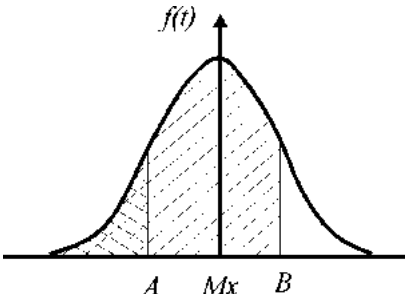


Рис. 9. Вероятность попадания случайной величины в заданный интервал

F(B) = P(x < B);
F(A) = P(x < A);
P(A ≤ x < B)=F(B)-F(A).

Если A и B расположены симметрично относительно M_x , то задача упрощается: находим $t = t_1 = t_2$ и определяем по таблице Ф (t) искомую вероятность: $P = \Phi(t)$.

Итак, перечислим основные свойства функции нормального распределения:

1)Кривая $f(x)$ всегда симметрична относительно ординат в точках $x = Mx$ (или $t = 0$, если распределение нормировано).

2)При $t = \pm\infty$, $f(t)$ стремится к нулю. Собственно, уже при $|t| > 3$, $f(t)$ практически равна 0: Ф(t=1)= 0,6827, Ф(t=2)=0,9545, Ф(t=3)=0,9973). Иначе говоря, практически все значения случайной величины (99,73%) укладываются в интервал $M_x \pm 3\sigma$. На этом свойстве основано широко используемое в геохимии правило "трех сигм", согласно которому концентрации

элементов, превышающие фон более, чем на три стандарта, считаются аномальными.

3)При $t = 0$ плотность вероятности максимальна:

$$f(t=0) = 0{,}3989.$$

Рассмотрим пример использования таблиц нормального распределения (табл. 2).

На одном из золоторудных тел установлено, что среднее содержание золота составляет 7,5 г/т при σ = 3,5. Какова вероятность того, что в наугад взятом образце содержание золота будет колебаться от 11 до 18 г/т.

Ход решения:

$$t_1 = \frac{11 - 7{,}5}{3{,}5} = 1; \quad t_2 = \frac{18 - 7{,}5}{3{,}5} = 3;$$

$$P = F(t_2) - F(t_1) = 0{,}9986 - 0{,}8414 == 0{,}1572.$$

Это означает, что в каждых 15-16 пробах из 100 наугад взятых из данного рудного тела содержание золота составит от 11 до 18 г/т.

Таблица 2

Некоторые значения функций, связанных с нормальным распределением

t	$f(t)$	$F(t)$	$\Phi(t)$	t	$f(t)$	$F(t)$	$\Phi(t)$
-4,0	0,0001	0,00003		0,5	0,352	0,6915	0,3829
-3,5	0,0009	0,0002		1,0	0,242	0,8414	0,6827
-3,0	0,0044	0,0014		1,5	0,129	0,9332	0,8664
-2,5	0,0175	0,0062		2,0	0,054	0,9772	0,9545
-2,0	0,0540	0,0228		2,5	0,017	0,9938	0,9878
-1,5	0,1295	0,0668		3,0	0,004	0,9986	0,9973
-1,0	0,2420	0,1586		3,5	0,001	0,9998	0,9995
-0,5	0,3521	0,3085		4,0	0,000	0,9999	0,9999
0,0	0,3989	0,5000	0,0				

В тесной связи с нормальным находится *логарифмически нормальный* (логнормальный) закон распределения, очень широко применяемый в геологии. Установлено, что этим законом удовлетворительно описывается распределение ряда химических элементов в породах, распределение содержаний

золота в россыпях, распределение диаметра частиц при дроблении и т.д. При логнормальном распределении нормальному закону подчинены не сами значения случайной величины, а их логарифмы. Поэтому вначале находят натуральные (или десятичные) логарифмы всех значений случайной величины, а затем все операции проводят с логарифмами, как с обычными числами: вычисляют их статистические характеристики и по таблицам нормального распределения определяют вероятности. В случае, если в исходной совокупности встречаются нулевые значения, их заменяют минимальными или половиной чувствительности анализа, поскольку логарифмировать нулевые значения нельзя.

Кривая плотности вероятности логнормального распределения, построенная не по логарифмам, а по исходным значениям, является асимметричной и описывается следующим выражением:

$$f(x) = \frac{1}{x\sigma_{\ln}\sqrt{2\pi}} \cdot e^{-\frac{(\ln x - M_{\ln x})^2}{2\sigma_{\ln}^2}}, \qquad (18)$$

где $M_{ln\ x}$ и σ_{ln} математическое ожидание и стандарт логарифмов значений.

Эта функция достигает максимума в точке

$$M_0 = e^{M_{\ln x} - \delta_{\ln}^2}.$$

Медиана (или среднее геометрическое) равна

$$Me = e^{M_{\ln x}}.$$

Математическое ожидание равно

$$Mx = e^{M_{\ln x} + \frac{\sigma_{\ln}}{2}}.$$

Дисперсия определяется соотношением

$$\sigma_x^2 = e^{2M_{\ln x}} \cdot \left(e^{2\sigma_{\ln}} - e^{\sigma_{\ln}}\right).$$

Асимметрия и эксцесс функции положительны.

Таблицы для логнормального распределения отсутствуют, поэтому теоретическую кривую плотности вероятности строят непосредственно по формуле (18).

Биноминальный закон распределения используется в тех случаях, когда в результате одного испытания событие А может либо появиться с вероятностью *p*, либо не появиться с вероятностью *q* = 1-*p*. Подобная схема испытания называется схемой Я.Бернулли. Этим ученым был найден закон биномиального распределения, согласно которому вероятность того, что событие А произойдет в *n* испытаниях ровно *x* раз равна:

$$P_n(x) = C^x_n \cdot P^x \cdot q^{n-x} = \frac{n!}{x!(n-x)!} \cdot p^x \cdot (1-p)^{n-x}. \qquad (19)$$

Здесь *n* и *p* являются параметрами биноминального распределения.

Основные характеристики биноминального распределения определяются следующими выражениями:

$$M_x = np; \qquad \sigma^2 = np(1-p);$$

$$A = \frac{q-p}{\sqrt{npq}}; \qquad E = \frac{1-6pq}{npq}.$$

Биноминальным законом описывается только распределение дискретных величин. Коэффициенты C_n^x при x = 1,2,3... образуют ряд коэффициентов разложения бинома Ньютона, почему распределение и называется биноминальным. Эти коэффициенты можно найти по специальным таблицам (1), или по треугольнику Паскаля (если $x \le 12$).

В тех случаях, когда *n* и *x* очень большие числа, вычисление вероятности по формуле (19) представляет значительные трудности. В этом случае рекомендуется применение приближенной формулы Муавра-Лапласа:

$$p_n(x) \approx \frac{1}{\sqrt{npq}} \cdot f(t), \qquad (20)$$

здесь $f(t)$ - функция плотности вероятности стандартного нормального распределения,

$$t = \frac{x - M_x}{\sigma} = \frac{x-np}{\sqrt{npq}}.$$

Значения *f*(*t)* берутся из табл.2. Рассмотрим пример.

На месторождении было отобрано 10015 проб, из них с кондиционным содержанием металла 5009, то есть частость ≈ 0,5. Необходимо определить вероятность того, что из 10 наугад взятых проб кондиционных будет 0,1,2.... 10 проб. Поскольку проб мало, вычисление ведем по формуле (19).

Таблица 3

Вероятность встречи кондиционных проб из 10 случайных (nx1000)

X	0	1	2	3	4	5	6	7	8	9	10
P	1	10	44	117	205	246	205	117	44	10	10

В частности, для X = 5, $P = 252 \cdot (\frac{1}{2})^5 \cdot (\frac{1}{2})^5 = 0{,}246$.

Допустим теперь, что у нас возникла необходимость определить вероятность того, что из 100 взятых проб кондиционными окажутся 55. Формула (19) в этом случае оказывается малопригодной, поэтому воспользуемся формулой (20):

$$P_n(x) = \frac{1}{\sqrt{100 \cdot 0{,}5 \cdot 05}} \cdot f(t) = 0{,}2 \cdot f(t).$$

По табл.2. $f(t) = 0{,}2420$, следовательно,

$$P_n(x) = 0{,}2 \cdot 0{,}2420 = 0{,}048.$$

При $n \to \infty$ биноминальное распределение стремится к нормальному, но, если при этом p или q стремится к нулю, то случайная величина начинает подчиняться *распределению Пуассона*. Формула Муавра-Лапласа в этом случае становится малопригодной, а при p=0 теряет смысл. Выражение, определяющее вероятность появления маловероятного события в серии из *n.* испытаний, было найдено Пуассоном :

$$P_n(\varepsilon = m) = \frac{\lambda^m \cdot e^{-\lambda}}{m!}, \qquad (21)$$

где $\lambda = n \cdot p$ является единственным параметром распределения. Можно легко убедиться, что

$$M_x = \sigma^2{}_x = \lambda = np, \quad A = \frac{1}{\sqrt{np}} \; ; \quad E = \frac{1}{np} \; .$$

A и E всегда больше нуля.

Функция распределения такой случайной величины представляет собой сумму:

$$P_n(\varepsilon \le m) = \sum_{k=0}^{m} \frac{\lambda \cdot e^{-\lambda}}{k!} \; , \qquad \text{где } k = 0, 1, 2 \; \ldots, m.$$

Если n недостаточно велико, а единичная вероятность p недостаточно мала (> 0,1), то вероятность, вычисляемая по формуле Пуассона, содержит заметную погрешность. Для этих случаев А.Н.Колмогоровым предложена исправленная формула

$$P_m{}' = \frac{\lambda^m \cdot e^{-\lambda}}{m!} - \frac{в}{2} \cdot \frac{\lambda^{m-2} \cdot e^{-\lambda}}{2(m-2)!} \cdot \left(\frac{\lambda^2}{(m-1)m} - \frac{2\lambda}{m-1} + 1 \right)$$

Формула Колмогорова учитывает и возможное изменение единичной вероятности, здесь:

$$\lambda = p_1 + p_2 + p_3 + \ldots . + p_n$$
$$b = p_1{}^2 + p_2{}^2 + p_3{}^2 + \ldots . + p_n{}^2$$

То есть, предложенное Пуассоном значение $\lambda = p \cdot n$ является частным случаем, когда все p равны между собой.

Рассмотрим пример распределения Пуассона.

В бассейне одного из водотоков отобрано 150 шлиховых проб, в отдельных из которых имеются знаки золота (табл.4).

Рассчитаем среднее содержание и дисперсию:

$$\bar{x} = \frac{0{,}32 + 1 \cdot 51 + \ldots + 6 \cdot 1 + 7 \cdot 0}{150} = 1{,}52 \; ,$$

$\bar{x}$

$$S^2 = \frac{(0-1{,}52)^2 \cdot 32 + (1-1{,}52)^2 \cdot 51 + \ldots + (7-1{,}52)^2 \cdot 0}{150} = 1{,}54$$

Как видим, $\bar{x} \approx S^2$, что является одним из признаков распределения Пуассона. Подставив λ=1,52 в формулу (21), рассчитаем P_m. Затем рассчитываем теоретическую частоту, округляя ее до целых чисел.

Почти полное совпадение теоретической и фактической частот свидетельствует о том, что распределение знаков золота в

шлихах данного водотока действительно подчиняется закону Пуассона.

Кроме рассмотренных четырех законов распределения в геологии используются и другие, в частности, распределения, производные от логнормального, распределение Пойа, Лапласа, равномерное и другие. Их описание, при необходимости, можно найти в соответствующих справочниках и учебниках по математической статистике.

Таблица 4

Распределение знаков золота по пробам

Число знаков золота	0	1	2	3	4	5	6	7
Число проб	32	51	36	19	8	2	1	0
Pm	0,22	0,33	0,25	0,12	0,04	0,01	0,01	0,0
N·Pm	33,4	50,1	37,6	18,8	7,1	2,1	1,2	0,1
Теорети-ческая частота	33	50	38	19	7	2	1	0

4. СТАТИСТИКА СЛУЧАЙНЫХ ВЕЛИЧИН

4.1. Статистические оценки неизвестных параметров

Каждая геологическая совокупность может быть разделена на изучаемую, опробуемую и выборочную. Из этого следует, что изучаемая и опробуемая совокупности характеризуются некоторыми неизвестными нам значениями исследуемых свойств, чаще всего средними содержаниями и дисперсиями, о которых мы можем судить на основе выборочных данных. Выборки зачастую бывают ограничены по объему, поэтому вопрос об их использовании для суждения о неизвестных параметрах генеральной совокупности стоит особенно остро.

Полученные по выборочным данным приближенные характеристики каких-либо свойств изучаемой совокупности называются их *оценками*. Например, в качестве оценки неизвестного среднего значения чаще всего используется среднее арифметическое по выборке, хотя возможны и другие варианты

оценок этого параметра: среднее геометрическое, среднее гармоническое и др. В связи с этим всегда возникает вопрос о выборе из набора возможных вариантов оценок параметров тех из них, которые удовлетворяют некоторым требованиям качества.

Статистические оценки могут быть *точечными* и *интервальными*. При точечной оценке неизвестная характеристика оценивается некоторым числом, а при интервальной оценке указывается некоторый интервал значений, в пределах которого с заданной вероятностью должно находиться истинное значение оцениваемой величины.

Точечные оценки должны удовлетворять требованиям *состоятельности, несмещенности* и *эффективности*. Состоятельной называется оценка, сходящаяся по вероятности к оцениваемому параметру с увеличением объема выборки:

$$\lim_{n\to\infty}\{p(\bar{\theta}-\theta)<\varepsilon\}=1.$$

Несмещенной называется оценка, математическое ожидание которой равно оцениваемому параметру при любом объеме выборки (т.е. нет систематической ошибки). Если требование несмещенности не выполняется, это обычно легко устраняется путем введения поправки. Максимально эффективной (или просто эффективной) называется оценка, обладающая минимальной дисперсией из всех возможных оценок. Понятно, что такая оценка (если она не смещена) наиболее предпочтительна, так как обеспечивает максимально тесную группировку результатов около истинного значения неизвестного параметра.

На практике не всегда удается удовлетворить всем трем требованиям. В этом случае выбору оценки всегда должно предшествовать ее критическое рассмотрение со всех точек зрения.

Наиболее важными характеристиками случайной величины являются математическое ожидание и дисперсия. Примем без доказательства, что при нормальном законе распределения состоятельной, несмещенной и эффективной оценкой математического ожидания случайной величины является среднее арифметическое ($\bar{x}$), полученное по выборочным

данным. При логнормальном распределении $\bar{x}$ не является эффективной за счет разброса больших значений, поэтому в практике геохимических работ в этом случае обычно используют среднее геометрическое:

$$\tilde{x} = e^{\overline{\ln x}}. \qquad (22)$$

Выборочная оценка дисперсии (S^2) при нормальном законе распределения определяется по формуле:

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}. \qquad (23)$$

Эта оценка является несмещенной и состоятельной, но не является эффективной. Оценка, удовлетворяющая всем трем требованиям, имеет вид:

$$S_*^2 = \frac{\sum_{i=1}^{n}(Xi - Mx)^2}{n}. \qquad (24)$$

Для ее вычисления необходимо знать математическое ожидание *Mx*, которое, как правило, неизвестно. Поэтому на практике обычно пользуются формулой (23).

Если число наблюдений невелико, то для оценки дисперсии можно использовать размах значений выборки:

$$S^2 = \frac{W_n}{d_n},$$

где $W_n = (Xmax - Xmin)$, d_n- коэффициент, который дается в специальной таблице (1). В геохимической практике чаще применяется коэффициент β:

$$S = \beta \cdot Wn. \qquad (25)$$

или для логнормального распределения

$$S_{ln} = \beta \cdot ln\frac{X\max}{X\min}. \qquad (26)$$

Для β также составлена специальная таблица (16).

Если распределение логнормально, то S^2 оказывается неэффективной даже при выборках большого объема. В этом случае можно воспользоваться максимально правдоподобной

оценкой дисперсии ($в$):

$$в^2 = e^{2\overline{\ln X}} \{ \psi (2 S^2_{\ln}) - \psi (\frac{n-2}{n-1} \cdot S_{\ln}^2) \},$$

где $$\psi(t) = e^t \{ 1 - \frac{t(t+1)}{n} + \frac{t^2(3t^2+22t+21)}{6n^2} \},$$

Выражение это очень громоздкое, поэтому в геохимической практике для оценки стандарта обычно используют стандартный множитель ε:

$$\varepsilon = e^{Sln}$$

Правило "трех сигм" в этом случае имеет вид:

$$Ханом. = \widetilde{x} \cdot \varepsilon^3$$

где *Ханом.*-аномальное значение, $\widetilde{x}$ - среднее геометрическое.

4.2. Точность оценок параметров. Построение доверительных интервалов оценок

Точечная оценка не содержит информации о точности полученного результата. Чем меньше выборка и чем больше изменчивость признака, тем большей может оказаться ошибка определения точечной оценки. Поэтому нам желательно знать тот интервал значений, в который с заданной вероятностью попадает истинное значение изучаемого признака.

Согласно центральной предельной теореме, доверительный интервал, внутри которого с заданной вероятностью будет находиться истинное значение математического ожидания, определяется из соотношения:

$$\lambda = \overline{x} \pm t \cdot \sigma_{\overline{x}} = \overline{x} \pm \frac{t\sigma_x}{\sqrt{n}}$$

или для выборочных данных

$$\lambda = \overline{x} \pm \frac{tS}{\sqrt{n}} \ . \qquad (27)$$

Число t зависит от выбранной доверительной вероятности. Это не что иное, как аргумент табличной функции Ф(t) (табл.2), поэтому его всегда можно найти по таблице. При доверительной вероятности 95% (или уровне значимости 5%, что одно и то же) $t = 1.96 \approx 2$.

Если объем выборки менее 60, то характер распределения величины *t* зависит не только от *Mx* и σ_x, но и от объема выборки. Такое распределение называется распределением Стьюдента. Число *t* в этом случае находится не из таблицы функции Ф(t), а из таблицы распределения Стьюдента, которая имеется в любом учебнике по математической статистике. Допустим, *n* = 15, p = 95%. Находим в таблице распределения Стьюдента число t , соответствующее уровню значимости 5% и числу степеней свободы *k* =*n* -1 = 14. Это будет 2,15.

Таким образом, число t показывает, сколько раз надо отложить $S_{\bar{x}}$ влево и вправо от $\bar{x}$, чтобы накрыть истинное значение M(x) с вероятностью p .

Если случайная величина распределена биноминально, то для нахождения доверительного интервала вводится дополнительная величина

$$\varphi = 2\arcsin\sqrt{p}\ .$$

Стандарт этой величины, выраженный в радианах, приближенно оценивается по формуле

$$S\varphi = \frac{1}{\sqrt{n}}\ .$$

Доверительный интервал для φ равен

$$\lambda_\varphi = \varphi \pm \frac{t}{\sqrt{n}}\ . \qquad (28)$$

Переход от φ к *p* осуществляется с помощью специальных таблиц (1).

Рассмотрим нахождение доверительного интервала для среднего квадратического отклонения *S*.

Ознакомимся предварительно с распределением χ^2 (*хи* - квадрат) или Пирсона. Если в формулу выборочной дисперсии вместо нормально распределенной величины *x* ввести новую случайную величину $\omega = x - M(x)$, то значение S^2 не изменится, а случайная величина ω также будет подчиняться нормальному закону с $M(\omega) = 0$ и дисперсией δ^2. Следовательно:

$$S^2 = \frac{1}{n}\sum(Xi - Mx)^2 = \frac{1}{n}\sum\omega_i^{\,2},$$

откуда $n\,S^2 = \sum\omega_i^{\,2}$.

Разделим обе части на δ^2, тогда

$$\frac{nS^2}{\delta^2} = \sum\left(\frac{\omega_i}{\delta}\right)^2; \qquad \frac{\omega_i}{\delta} = \frac{x_i - Mx}{\delta} = \text{t}.$$

Так как случайная величина ω подчиняется нормальному закону с параметрами (0,δ), то t также имеет нормальный закон распределения с параметрами (0, 1). Значения t_1, t_2, . . ., t_n независимы между собой, следовательно, независимы и их квадраты.

Обозначим $\chi^2 = \frac{nS^2}{\delta^2} = \sum t_i{}^2$

Итак, случайная величина, представляющая собой сумму квадратов независимых случайных величин, каждая из которых подчиняется нормальному закону распределения с параметрами (0,1), называется случайной величиной с χ^2- распределением и $k = n$ степенями свободы.

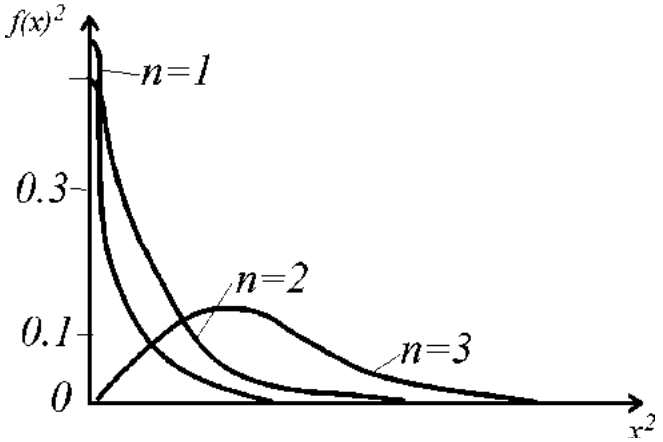


Рис. 10. График плотности вероятности распределения χ^2

Число степеней свободы равно числу независимых переменных минус число связей, накладываемых на эти переменные.

Дифференциальная функция распределения χ^2 имеет вид

$$f(\chi^2) = \text{L}_\text{n} \cdot \chi^{\text{n-2}} \cdot e^{\frac{\chi^2}{2}}. \qquad (29)$$

здесь L_n - коэффициент, зависящий от n. Как видим, распределение χ^2 не зависит от M_x и δ^2 ,а зависит лишь от объема выборки. График функции $f(\chi^2)$ показан на рис.10.

Математическое ожидание распределения χ^2 равно числу степеней свободы $M_x = k$. Можно также доказать, что дисперсия $\delta^2_\chi = 2k$. Для функции распределения χ^2 составлены таблицы, по которым можно вычислить вероятность того, что случайная величина, подчиняющаяся закону χ^2 с известным числом n , не превысит фиксированного значения $\chi^2_{k,\alpha}$.

Построение доверительного интервала дисперсии при заданной доверительной вероятности p = 1- α (α- уровень значимости) осуществляется с помощью выражения:

$$P\left(\frac{nS^2}{\chi^2_2} < \delta^2 \le \frac{nS^2}{\chi^2_1}\right) = 1 - \alpha . \qquad (30)$$

Рассмотрим пример.

Требуется построить доверительный интервал с вероятностью p = 0,96 для дисперсии случайной величины x , распределенной нормально, если $S^2 = 10$, $n = 20$.

По таблице χ^2 -распределения нам необходимо выбрать два таких значения, чтобы площадь, заключенная под кривой $f(\chi^2)$ в интервале χ^2_1 и χ^2_2, равнялась 1-α ; χ^2_1 и χ^2_2 обычно выбирают так, чтобы (рис.11):

$$P(\chi^2 < \chi^2_2) = P(\chi^2 > \chi^2_2) = \frac{\alpha}{2} .$$

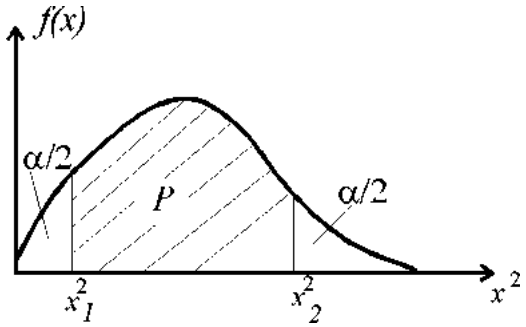


Рис. 11. Выбор точек χ^2_1 и χ^2_2 для нахождения доверительного интервала для δ^2

В нашем примере α = 0,04, $\frac{\alpha}{2}$ = 0,02. Находим по таблицам значения χ^2_1 и χ^2_2 при $p_1 = 0{,}98$, $p_2 = 0{,}02$ и $k=n-1 = 19$.

$$\chi^2_1 = 8{,}6 \; ; \; \chi^2_2 = 33{,}7.$$

Доверительный интервал для δ^2 запишется следующим образом:

$$\frac{20 \cdot 10}{33,7} < \delta^2 \leq \frac{20 \cdot 10}{8,6},$$

или $5,94 < \delta^2 \leq 23,6$.

Для среднего квадратического отклонения:

$$2,43 < \delta \leq 4,82 .$$

5. ПОСТРОЕНИЕ СТАТИСТИЧЕСКИХ РЕШЕНИЙ.

5.1. Статистические гипотезы

Выше мы рассмотрели различные способы получения статистических оценок неизвестных параметров. Для геолога вычисление этих оценок не является самоцелью, а делается для дальнейшего использования при обосновании геологических выводов. Решение многих геологических задач основано на принципе аналогии, когда для объяснения особенностей строения слабо изученного объекта используются закономерности, установленные при исследовании аналогичных объектов. Понятно, что при этом необходимо установить степень сходства объекта - аналога с изучаемым участком. Чаще всего при этом сравниваются средние значения определенных признаков. В результате принимается одно из двух решений: либо разницей между средними можно пренебречь и считать их равными, либо различия между оценками существенные и средние следует признать различными.

Вообще, вопрос о различии или сходстве может возникнуть и при исследовании других статистических параметров: дисперсии, коэффициентов корреляции, асимметрии т.д. Во всех этих случаях для решения вопроса о сходстве или различии геологических объектов используются статистические методы проверки гипотез о равенстве числовых характеристик их свойств.

Под *статистическими гипотезами* подразумеваются такие гипотезы, которые относятся либо к виду, либо к отдельным параметрам распределения случайной величины. Например, статистической является гипотеза о том, что веса проб, отобранные одним человеком по одной методике распределены по нормальному закону.

Поскольку выборочные данные ограничены по объему и

носят случайный характер, при обосновании выводов по статистическим данным вполне возможны ошибочные заключения. При этом ошибки могут быть двух видов:

1) если гипотеза, являющаяся правильной, не принята - это ошибка 1-го рода,

2) если принята ложная гипотеза - ошибка 2-го рода.

Проверяемая гипотеза обычно обозначается H_0 и называется нулевой, конкурирующая или альтернативная гипотеза обозначается H_1. Например: H_0 : $\mu_1 = \mu_2$; H_1 : $\mu_1 \neq \mu_2$. Вероятность p , определяющая область, в пределах которой правильность принятого решения будет событием практически достоверным, называется доверительной вероятностью, а сама область - доверительной областью. Вероятность q=1-p, соответствующая уровню вероятности практически невозможного события, называется уровнем значимости, а ее область - критической (рис.12).

Если эмпирическое значение попадает в область To, то принимаем гипотезу Ho, если в T_1 ,- то гипотезу H_1.

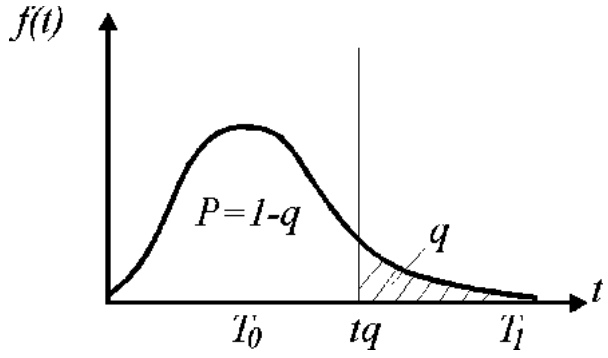


Рис. 12 . Доверительная (To) и критическая (T_1) области принятия гипотезы

Как видно на рис. 12, уровень значимости q определяет вероятность ошибки 1-го рода, и, казалось бы, надо брать q, как можно меньше. Но, к сожалению, это далеко не всегда оправдано. Рассмотрит альтернативу H_1 : $\mu_1 \neq \mu_2$. Очевидно, что событие $t \in T_0$ при условии, что верна H_1, будет способствовать ошибочному решению, т.е. принятию гипотезы H_0 , хотя она не верна. Эта ошибка 2-го рода, и она тем больше, чем меньше q.

Ее вероятность на рис. 13 обозначена β, вероятность ошибки 1-го рода обозначена α.

Таким образом, нужна какая-то золотая середина. Вероятность 1-β называется мощностью критерия

относительно конкурирующей гипотезы. Очевидно, надо стремиться, чтобы 1-β была как можно больше.

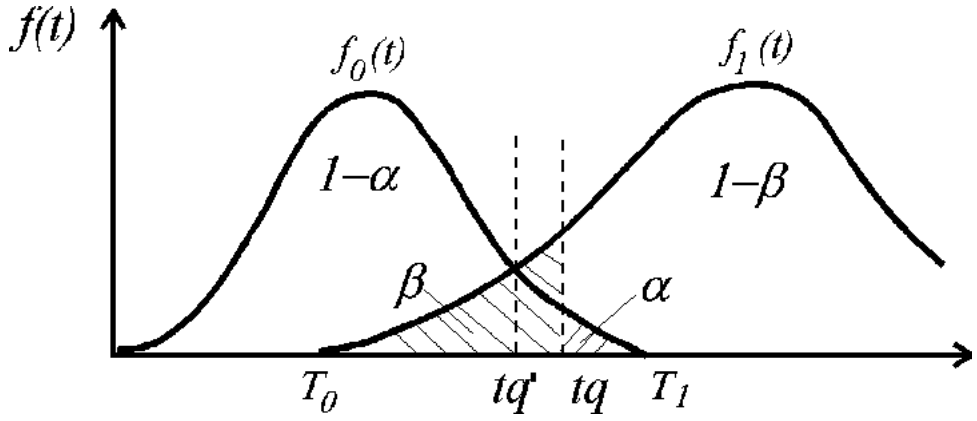


Рис. 13. Соотношение ошибок 1-го и 2-го рода

Рассмотрим пример.

По данным опробования вновь выявленного рудопроявления установлено что $\bar{x}$ = 65 г/т (N = 10 проб). Предполагается, что распределение близко к нормальному с неизвестным математическим ожиданием M_x и δ^2 = 2809 (δ = 53). Существует также два эталонных объекта с μ_1 = 100 г/т и μ_2 = 50 г/т, характеризующих два промышленных типа месторождений. Требуется проверить предположение о том, что изучаемый объект относятся к 1-му промышленному типу, т.е. проверить гипотезу H_0 : μ = 100 г/т при альтернативе H_1 : μ = 50 г/т. Зададимся вначале q = 0,01. По таблицам Ф(t) находим t = -2,33. Затем определяем нижнюю границу доверительного интервала среднего для первого эталонного объекта:

$$B = \mu_1 + \frac{t_q \cdot \delta}{\sqrt{n}} = 100 - \frac{2{,}33 \cdot 53}{10} \approx 61 .$$

Если $\bar{x} \geq$ 61, то гипотеза H_0 принимается.

Таким образом, при q = 0,01 мы бы приняли гипотезу H_0.

Определим теперь мощность критерия относительно альтернативы H_1 :

$$t_{1-\beta} = \frac{(в - \mu_2) \cdot \sqrt{n}}{\delta} = \frac{(61 - 50) \cdot \sqrt{10}}{53} = 0{,}60;$$

$1 - \beta = \mathrm{F}(t_{1-\beta}) = 0{,}73$; т.е. β =0,27.

Вероятность ошибки 2-го рода составляет 27%. Это слишком много, поэтому вычислим новое значение $в$ при q = 0,05:

$$в = 100 - \frac{1{,}65 \cdot 53}{10} = 72.$$

Мощность критерия равна:

$$1 - \beta = F\left(\frac{(72 - 50) \cdot 10}{53}\right) = 0{,}905.$$

Отсюда β = 0,095.

Такая вероятность ошибки 2-го рода уже приемлема, поэтому при проверке гипотезы H_0 лучше использовать уровень значимости 0,05, чем 0,01. Поскольку $\bar{x}$ = 65 < 72, то проверяемая гипотеза отклоняется и принимается альтернатива H_1. Такое решение обеспечивает меньшую вероятность появления ошибки 2-го рода наряду с небольшим значением уровня значимости. На рис. 13 это означает, что мы сместили критическую точку из t_q в t_q', расширив область T_1. В геологии обычно очень трудно оценить вероятность ошибки 2-го рода, поэтому во всех случаях формального выбора доверительная область ограничивается уровнем значимости 5%.

5.2. Статистическая проверка некоторых типовых гипотез

5.2.1. Проверка гипотез о функциях распределения

Для эффективного использования статистических методов в решении геологических задач обычно недостаточно иметь по выборке среднее значение и дисперсию. Необходимо еще знать закон распределения случайной величины. Знание этого закона позволяет сознательно выбирать по возможности эффективные критерии и оценки параметров.

Рассмотрим вначале наиболее общий и строгий способ проверки гипотез о законе распределения, носящий название критерия Пирсона, а затем ознакомимся с менее строгими и несложными методами проверки гипотез о нормальном (логнормальном) распределении.

Допустим, мы имеем выборку объемом 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. Таким образом, задача заключается в проверке гипотезы *Ho* : $F(x)=F_0$(x) при альтернативе H_1:$F(x) \neq F_0(x)$.

Разбиваем область выборочных значений x_1, x_2, x_3, .. x_n на k интервалов, необязательно равных, и подсчитываем частости попадания значений выборки в эти интервалы l=1,2... k .
Если H_0 верна, то число

$$\chi^2 = \sum_{l=1}^{k} \frac{(N_l - n_l)^2}{N_l} \qquad (31)$$

будет распределено как χ^2 с k - m степенями свободы. Здесь $N_l = n \cdot P_l$ - теоретические частоты, $P_l = F_0(a_{l+1}) - F_0(a_l)$, a_l , a_{l+1} - границы интервалов, n_l - частоты попадания значений x. в интервалы.

В случае проверки гипотезы о нормальном распределении теоретические частоты подсчитываются по формуле:

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

где $t_l = \frac{a_l - \bar{x}}{S}$, $\bar{x} = \frac{\sum_{i=1}^{k} n_l \cdot \overline{x_l}}{n}$; $S^2 = \frac{\sum n(\overline{x_l - x})^2}{n-1}$.

Частоты N_l показывают, как бы распределились наши n наблюдений, если бы выборка была взята из нормальной совокупности с $M(x) = \bar{x}$ и $\delta^2 = S^2$. Следовательно, при проверке гипотезы используется три ограничения: $\sum n_l = n$, $M(x) = \bar{x}$, $\delta^2 = S^2$, поэтому число степеней свободы равно $k - 3$.

Если вычисленное значение χ^2 больше, чем $\chi^2_{q\,,k-3}$ взятое из таблицы χ^2-распределения, то гипотеза *Ho* отклоняется, т.е. считаем, что распределение не соответствует нормальному (при заданном уровне значимости q). Если под рукой нет таблицы, можно воспользоваться способом В. И. Романовского: если $\frac{\chi^2 - k}{\sqrt{2\pi}} \geq 3$, то гипотеза H_0 отклоняется. Здесь k - число степеней свободы.

Критерий Пирсона не зависит от вида функции распределения, выбранной для F_0. Неприменим он только при малом n .

В геологической практике, при проверке гипотезы о соответствии эмпирического распределения нормальному (логнормальному) закону, чаще пользуются методом, основанным на рассмотрении оценок асимметрии (A) и эксцесса (E). В условиях нормального распределения случайные величины, значения которых A и E мы наблюдаем, распределены приблизительно нормально со средними значениями, равными 0 и дисперсиями $\approx \sqrt{\frac{6}{n}}$ и $\sqrt{\frac{24}{n}}$ соответственно. Следовательно, числа $t_1 = \frac{A}{\sqrt{\frac{6}{n}}}$ и $t_2 = \frac{E}{\sqrt{\frac{24}{n}}}$ в случае нормального распределения, будут представлять собой значения случайных величин, распределенных приблизительно нормально с параметрами (0,1). Поэтому гипотезу о нормальном распределении следует отклонить, если хоть одно из них, t_1 или t_2 , превысит по абсолютной величине t_q . Обычно принимают $t_q = 3$ (при уровне значимости q = 0,01).

Первое представление о соответствии изучаемого распределения нормальному можно также получить из визуального анализа гистограмм распределения значений и даже таблиц сгруппированных исходных данных.

Проверка гипотезы о логнормальном распределении не представляет особых трудностей и сводится к проверке гипотезы о нормальном распределении логарифмов значений случайной величины.

5.2.2. Проверка гипотез о равенстве средних значений (математических ожиданий)

Необходимость сравнения средних значений возникает при решении самых разнообразных геологических задач, практически во всех разделах геологии. В данном пособии рассматривается три вида подобных гипотез: а) о равенстве неизвестного среднего заданному значению; б) о равенстве двух неизвестных средних и в) о равенстве k неизвестных средних в условиях нормального, логнормального распределения и в случае, если распределение неизвестно. Поскольку в геологической практике точное значение дисперсии обычно неизвестно, речь будет идти о тех

случаях, когда дисперсия оценивается по выборке.

а) Проверка гипотезы о равенстве неизвестного среднего заданному значению.

Критерий для проверки гипотезы имеет вид:

$$t = \frac{(\bar{x} - \mu_0)\sqrt{n}}{S} \qquad (32)$$

где μ_0 - заданное значение, n - объем выборки.

Если гипотеза верна, то t будет представлять собой значение случайной величины, распределенной нормально с параметрами (0,1). Критическое значение t_q берем поэтому из таблиц функции $F(t)$, в зависимости от заданного уровня значимости q.

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

Если n <20, то значения t_q берутся из таблиц распределения Стъюдента.

Если распределение логнормальное, то критерий t имеет вид:

$$t = \frac{\overline{\ln x} + 0.5 S_{\ln}{}^2 - \ln \mu_0}{\sqrt{\frac{S_{\ln}{}^2}{n} + \frac{S_{\ln}{}^4}{2(n-1)}}} \qquad (33)$$

Здесь S^2_{ln} - дисперсия распределения логарифмов значений.

б) Проверка гипотезы о равенстве двух неизвестных средних.

Это наиболее распространенная в геологии задача, так как утверждение о сходстве или различии геологических объектов и явлений основывается на утверждении о равенстве или неравенстве неизвестных средних значений их свойств.

В данном случае наиболее часто применяют параметрический критерий Стъюдента (Вэлча):

$$t = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{\frac{S^2{}_1}{n_1} + \frac{S^2{}_2}{n_2}}} \qquad (34)$$

Если $t > t_{q\ ,\ n1 + n2 - 2}$, взятого из таблицы распределения Стьюдента, то гипотеза о равенстве неизвестных средних отвергается. Указанный критерий применим только в случае, если $\delta^2{}_1 \neq \delta^2{}_2$. Если же выяснится, что $\delta^2{}_1 = \delta^2{}_2$ (проверку этой гипотезы см. ниже), то следует применить следующий критерий:

$$t = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{S\sqrt{\frac{n_1 + n_2}{n_1 \cdot n_2}}} \qquad (35)$$

где $S = \sqrt{\frac{(n_1 - 1)\cdot S^2{}_1 + (n_2 - 1)\cdot S^2}{n_1 + n_2 - 2}}$.

Критическое значение $t_{q\ ,\ n1 + n2 - 2}$ при этом также берется из таблицы распределения Стьюдента.

Если распределение случайной величины логнормальное, то следует использовать критерий Д. А. Родионова (при $\delta^2{}_1 \neq \delta^2{}_2$)

$$t = \frac{\overline{\ln x_1} - \overline{\ln x_2} + 0.5(S^2{}_1 - S^2{}_2)}{\sqrt{\frac{S^2{}_1}{n_1} + \frac{S^2{}_2}{n_2} + \frac{1}{2}\left(\frac{S^4{}_1}{n_1 - 1} + \frac{S^4{}_2}{n_2 - 1}\right)}} \qquad (36)$$

где $S^2{}_1$ и $S^2{}_2$ - дисперсии распределения логарифмов значений. Критическое значение t_q находится по таблице функции нормального распределения $F\ (\ t\ )$. Принятие или отклонение гипотезы H_0, осуществляется так же, как было описано при рассмотрении формулы (32).

Если выясняется, что $\delta^2{}_1 = \delta^2{}_2$, то можно использовать критерий Стьюдента, заменив в формуле (35) $\bar{x}_1$ и $\bar{x}_2$, на $ln\,\bar{x}_1$, и $ln\,\bar{x}_2$.

В случае если закон распределения случайных величин неизвестен, следует воспользоваться непараметрическими критериями Ван-дер-Вардена, Вилкоксона, или Манна-Уитни.

Рассмотрим пример применения критерия Манна-Уитни. Как и во всех критериях подобного типа, вычислительные операции проводится не с самими числами, а с их рангами.

Допустим, мы имеем две выборки X и Y объема n и m и хотим проверить гипотезу о том, что они принадлежат к одной и той же совокупности. Объединим две выборки и расположим все значения в порядке возрастания – от меньшего к большему. Наименьшее значение при этом получит ранг 1, наибольшее – ранг $(n+m)$. Если выборки принадлежат одной совокупности, то естественно ожидать, что ранги одной из выборок будут достаточно равномерно рассеяны в общей последовательности рангов. Критерий Манна-Уитни вычисляется по формуле:

$$T = \sum_{i=1}^{n} R(x_i) - \frac{n(n+1)}{2} \qquad (37)$$

Первый член – это сумма рангов наблюдений первой выборки, n – число наблюдений в первой выборке. Критические значения T для нижнего критического предела приведены в таблице 5 приложения к данному пособию. Предел для верхней критической площади определяется выражением $T_{1-\alpha} = n * m - T_{\alpha}$. Например, если в нашем случае $n = 8$, $m = 10$, вычисленное значение $T = 35$, а уровень значимости 10%,то нижний критический предел будет равен: $T_{0,05} = 21$, верхний предел $T_{0,95} = 8 *10–21=59$. Вычисленное T не выходит за эти пределы, следовательно, с вероятностью 90% можно утверждать, что выборки не различаются, то есть принадлежат одной совокупности.

в) Проверка гипотезы о равенстве k неизвестных средних

Это наиболее общий случай проверки гипотез о равенстве средних. Необходимость в такой проверке возникает довольно часто, при одновременном сравнении нескольких геологических объектов. Иногда эту задачу пытаются решить путем попарных сравнений средних, но такой подход нельзя признать удовлетворительным.

Таким образом, проверяемая гипотеза имеет вид $H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k = \mu_0$, а множество альтернатив можно представить как $H_1 : \mu_i \neq \mu_0$ хотя бы для одного $i = 1, 2, 3... k$.

В условиях нормального распределения, в случае, если $\delta^2_1 = \delta^2_2 = \ldots = \delta^2_k$, эту гипотезу можно проверить с помощью критерия, аналогичного критерию Стьюдента:

$$t_i = \frac{y_i\sqrt{n_i(N-2)}}{\sqrt{N-n_i-n_i y^2{}_i}}\ , \qquad (38)$$

где $N=\sum_{i=1}^{k} n_i$; $y_i = \frac{\bar{x}_i - \bar{x}}{S}$; $\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$;

$$\bar{\bar{x}} = \frac{1}{N}\sum_{i=1}^{k} n_i \cdot \bar{x}_i = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}\ ;\ S^2 = \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)\cdot S^2{}_i\ ;$$

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

Критическое значение $t_{q,N-2}$ берется из таблицы распределения Стьюдента для заданного уровня значимости и числа степеней свободы N - 2. Если хотя бы одно из вычисленных значений t_i превысит табличное, гипотеза H_0 отвергается.

Если дисперсии нельзя признать равными, можно воспользоваться критерием ν :

$$\nu = \sum_{i=1}^{k}\frac{(\bar{x}_i - \bar{\bar{x}})^2 \cdot n_i}{S^2{}_i} \qquad (39)$$

Обозначения те же, что в (38). Если вычисленное значение ν превысит табличное значение $\chi^2_{q,k-1}$ взятое из таблиц χ^2-распределения, то гипотеза о равенстве k средних отвергается. В противном случае гипотеза H_0 принимается как не противоречащая выборочным данным.

Если распределение логнормально, то при равенстве дисперсий логарифмов, гипотезу о равенстве математических ожиданий случайных величин можно свести к гипотезе о равенстве математических ожиданий их логарифмов и использовать вышеприведенные критерии.

Если дисперсии логарифмов не равны, то дальнейшую проверку следует прекратить, так как параметрических критериев для такого случая не существует. В такой ситуации можно воспользоваться более общими непараметрическими (не чувствительными к виду распределения) критериями, например,

критерием Краскла-Уэллиса или Пури-Сена-Тамуры (17). Критерий Краскла-Уэллиса является непараметрическим аналогом однофакторного дисперсионного анализа. Он позволяет проверить гипотезу о том, что все k совокупностей, из которых взяты выборки имеют одинаковое распределение. Вычисление критерия Краскла-Уэллиса сходно с вышеописанной процедурой для критерия Манна-Уитни: все наблюдения из k выборок объединяются и ранжируются от наименьшего к наибольшему. Для каждой выборки вычисляется сумма рангов:

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

где x_{ik} – ранг i-го наблюдения в k-й выборке, n_k – число наблюдений в k-й выборке.

Статистика Краскла-Уэллиса вычисляется по формуле:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R^2{}_k}{n_k} - 3(N+1),$$

где N – общее число наблюдений в k выборках. Критические значения H можно взять из таблицы χ^2 распределения для (k-1) степеней свободы (табл. 4 приложения)

Критерий Пури-Сена-Тамуры используется для проверки гипотез о равенстве многомерных средних в двух объектах. Он опирается на понятия матричной алгебры и требует для своей реализации использования компьютерных технологий.

5.2.3. Проверка гипотез о равенстве дисперсий

а) Проверка гипотез о равенстве двух дисперсий.

Дисперсия является мерой рассеяния результатов наблюдений, поэтому может быть использована для описания изменчивости свойств геологических объектов. Поскольку применение обычного в геологии метода аналогии невозможно без сравнения степени изменчивости рассматриваемых объектов, то ясно, что сравнение дисперсий - задача обычная при геологических исследованиях. Кроме того, как мы видим выше, проверка гипотез о равенстве дисперсий необходима для выбора критерия при проверке гипотезы о равенстве средних.

Итак, нам требуется проверить гипотезу $H_0 : \delta^2{}_1 = \delta^2{}_2$ при аль-

тернативе $H_1 : \delta^2_1 \neq \delta^2_2$. Если распределение не противоречит нормальному, в этом случае обычно пользуются критерием Фишера:

$$F = \frac{S^2_1}{S^2_2}. \qquad (40)$$

В числитель при этом записывается *большая* дисперсия. Критическое значение F берется из таблиц распределения Фишера, которые имеются во всех руководствах по математической статистике. Выбрав таблицу для соответствующей доверительной вероятности 1-q, по горизонтали находим столбец со значением n_1 -1 по вертикали - строку со значением n_2-1. На их пересечении будет искомое критическое значение $F_{1-q\ ,\ n_1 - 1\ ,\ n_2 - 1}$. Здесь n_1 - количество членов в выборке с большей дисперсией.

Если вычисленное значение F превысит табличное, гипотеза о равенстве дисперсий отвергается.

В условиях логнормального распределения критерий Фишера применяется для проверки гипотезы о равенстве дисперсий логарифмов значений.

Если закон распределения не соответствует нормальному (логнормальному), можно воспользоваться ранговым критерием Сиджела-Тьюки (17), который является почти полным аналогом критерия Вилкоксона.

б) Проверка гипотезы о равенстве более, чем двух дисперсий

Критерий для проверки этой гипотезы был предложен в 1937 году Бартлетом и носит его имя. Бартлет показал, что, если $H_0 : \delta^2_1 = \delta^2_2 = \ldots = \delta^2_k = \delta^2_0$ верна, то величина

$$B = \frac{1}{c}\left[\sum_{i=1}^{k}(n_i - 1)\cdot \ln S^2 - \sum_{i=1}^{k}(n_i - 1)\ln S^2_i\right] \qquad (41),$$

будет распределена как χ^2 с (k - 1) степенями свободы. Здесь:

$$c = 1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{\sum_{i=1}^{k}(n_i - 1)}\right]$$

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{k}(n_i - 1)S^2{}_i$$

Если вычисленное значение B окажется больше табличного $\chi^2_{q,k\text{-}1}$, то гипотеза о равенстве дисперсий отвергается. Многомерным аналогом критерия Бартлета является критерий Кулъбака(17).

6. ИССЛЕДОВАНИЕ РАЗЛИЧИЙ МЕЖДУ ГЕОЛОГИЧЕСКИМИ ОБЪЕКТАМИ.

6.1. Дисперсионный анализ

Обычной для геологии является ситуация, когда относительно имеющегося набора наблюдений заранее неизвестно, является ли он однородным или неоднородным и на какое число однородных групп его следует разделить. Поскольку статистическая неоднородность объекта означает его геологическую неоднородность, то ясно, что задача статистического разграничения совокупности наблюдений является типичной при самых различных геохимических, петрографических, палеонтологических и других исследованиях.

Задачи, основанные на проверке гипотезы о статистической однородности геологических объектов разделяются на 3 типа:

1) выделение аномальных значений;

2) разделение неоднородных выборочных совокупностей;

3) оценка степени влияния различных факторов на характер изменчивости свойств объектов (дисперсионный анализ).

1) Задача выявления аномальных значений не имеет универсального статистического решения. В практике геохимических работ обычно используют правило "трех сигм":

$X_{аном.} > \bar{x} + 3S$. Однако этот способ нельзя признать корректным, так как он не гарантирует от ошибок как 1-го, так и 2-го рода, причем, вероятность этих ошибок оценить нельзя.

В тех достаточно редких случаях, когда распределение значений не противоречит нормальному закону, можно использовать критерий аномальности Н.В. Смирнова:

$$t = \frac{x\max - \bar{x}}{S^2_{ст}} \quad , \qquad (42)$$

где $S^2_{см} = S^2 \left(\frac{n-1}{n} \right)$ - смещенная оценка дисперсия.

Критическое значение $t_{1\text{-}q,n}$ берется из таблиц распределения Смирнова (1). Если вычисленное значение t не превышает табличного, следует признать, что выборка не содержит аномальных значений.

Во всех других случаях оптимальным следует признать определение аномальных значений опытным путем, на основе анализа геологических причин изменчивости свойств объекта. Статистические характеристики имеют при этом вспомогательную роль.

2)Разделение неоднородных выборочных совокупностей позволяет решать задачи геологического картирования, выбирать наиболее информативный комплекс геофизических и геохимических методов и т.п. Простейшие методы разделения неоднородных совокупностей основаны на анализе графиков эмпирических кривых распределения. На неоднородность выборки может указывать наличие нескольких максимумов (полимодальность) на кривой распределения. Существуют специальные палетки для подбора эталонных кривых плотности распределения, позволяющие разделять исходную неоднородную выборку на ряд однородных. Виды палеток и правила пользования ими подробно описаны в литературе и здесь не приводятся.

Алгоритмы аналитического решения задачи разграничения подробно рассмотрены в работе Д. А. Родионова (13).

3) Оценка степени влияния различных факторов на характер изменчивости свойств геологических объектов

осуществляется с помощью дисперсионного анализа. Это статистический метод анализа, основанный на разложении общей дисперсии признака на составные части, обусловленные влиянием различных факторов.

Это можно представить как:

$$(x - \bar{x}) = \alpha + \beta + \gamma \ ,$$

где α - отклонение, вызываемое фактором А, β - отклонение,

вызываемое фактором В, γ - отклонение, вызываемое другими неучтенными факторами. Иначе говоря, $\delta^2_x = \delta^2_\alpha + \delta^2_\beta + \delta^2_\gamma$. Сравнивая δ^2_α или δ^2_β с δ^2_γ можно установить степень влияния факторов А и В на величину x по сравнению с неучтенными факторами. Сравнивая δ^2_α и δ^2_β между собой, можно установить сравнительное влияние факторов А и В на x . Существенность влияния какого-либо фактора на исследуемую величину определяется по критерию Фишера:

$$F_A = \frac{S^2_\alpha}{S^2_\gamma}; \qquad F_B = \frac{S^2_\beta}{S^2_\gamma}$$

Если вычисленное значение превышает табличное, то влияние фактора признается значимым.

Рассмотрим пример .

Требуется выяснять, как влияют состав вмещающих пород (А) и гипсометрическое положение рудных тел (В) на среднее содержание в них золота. Данные по средним содержаниям приведены в таблице 5.

Таблица 5.

Содержание золота в рудных телах

А/В	Горизонт +800 м	Горизонт +500 м	Горизонт +200 м	$\overline{x_i}$
Песчаники	1,0	2,0	3,0	2,0
Граниты	5,0	5,0	10,0	7,0
$\overline{x_j}$	3,0	4,0	6,5	4,5

$$S^2_\alpha = \frac{[\sum(\bar{x}_i - \bar{x})^2 \cdot n_1}{n_2 - 1} = \frac{[(2-4.5)^2 + (7-4.5)^2] \cdot 3}{1} = 37.5$$

$$S^2_\beta = \frac{[\sum(\bar{x}_j - \bar{\bar{x}})^2] n_2}{n_1 - 1} = \frac{[(3-4.5)^2 + (4-4.5) + (6.5-4.5)^2] \cdot 2}{2} = 6.5$$

$$S^2_\gamma = \frac{\sum\sum(x_{ij} - \bar{x}_i - \bar{x}_j + \bar{x})^2}{(n_2 - 1)(n_1 - 1)} = [(1\text{-}2\text{-}3\text{+}4.5)^2 + (2\text{-}2\text{-}4\text{+}4.5)^2 + (3\text{-}2\text{-}6.5\text{+}4.5)^{2+} \ldots + (10\text{-}7\text{-}6.5\text{+}4.5)^2]/2 = 1.5$$

здесь n_1 - число столбцов, n_2 - число строк

$$F_A = \frac{37,5}{1,5} = 25 \; ; \qquad F_B = \frac{6,5}{1,5} = 4.3$$

Табличные значения для уровня значимости 0,05 и $\kappa_2 = 2$, $\kappa_1 = 1$ равны: $F_A = 18,5$, $F_B = 19,2$. Таким образом, приходим к выводу, что влияние фактора А (состав вмещающих пород) на содержание золота в руде значимо, а влияние фактора В (гипсометрический уровень) - незначимо.

По количеству исследуемых факторов дисперсионный анализ может быть одно-, двух- и многофакторным. При многофакторном анализе общая идея разложения дисперсии остается той же самой, но сложность вычислений резко возрастает. Например, для 4-факторного комплекса общая дисперсия разлагается уже на 14 составных частей.

7. КОРРЕЛЯЦИОННАЯ ЗАВИСИМОСТЬ МЕЖДУ СВОЙСТВАМИ ГЕОЛОГИЧЕСКИХ ОБЪЕКТОВ

До сих пор мы вели речь о характере распределения одной, изолированной случайной величины. Но в геологической практике обычно приходится иметь дело не с одной, а одновременно с несколькими случайными величинами. Например, при опробовании золоторудной жилы одновременно производят замеры ее мощности, определяют содержание сульфидов, цвет кварца, степень брекчирования, элементы залегания и т.п. Эти изучаемые свойства могут быть независимы, но могут быть и определенным образом взаимосвязаны. Задача исследователя состоит в том, чтобы установить, есть ли эта связь, и, если есть - рассчитать уравнение зависимости.

Отметим прежде всего, что связь между величинами может быть функциональной и стохастической. *Функциональной* называется такая связь, когда одному значению x соответствует одно, строго определенное значение y . Примерами такой связи являются, к примеру, формулы физики. *Стохастическая* - это такая связь, когда одна случайная величина реагирует на изменение значений другой величины изменением своего закона распределения. В геологии обычно используется частный случай стохастических связей - *статистическая (корреляционная)* зависимость (когда среднее значение

одной величины является функцией от значения, принимаемого другой величиной). Форма и теснота корреляционной связи могут быть выражены аналитически, но обычно исследование начинают не с расчетов, а с графического анализа зависимости в двухмерном пространстве. По оси абсцисс откладывают значения одного свойства, по оси ординат - другого. Совокупность наблюдений образует облако точек (рис. 14).

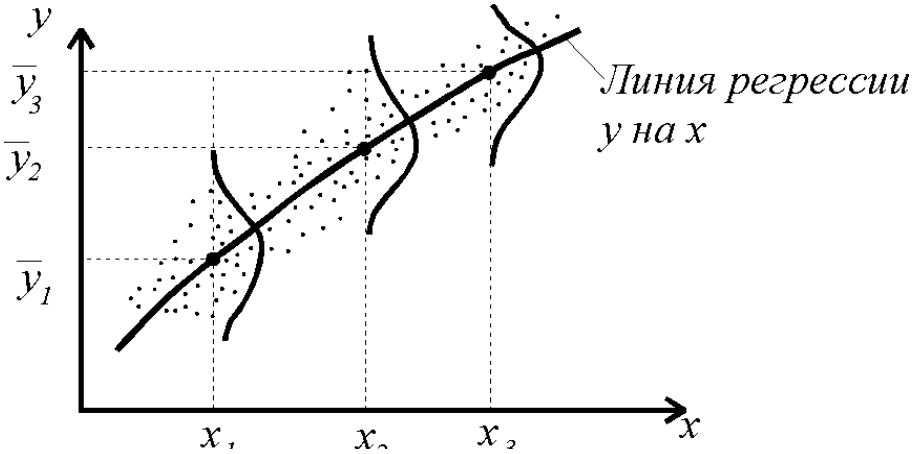


Рис.14. Облако точек, условные центры распределения и линия регрессии y на x.

Графический анализ заключается в изучении формы и ориентировки облака точек. Если все точки расположены вдоль линии, то связь функциональная, если облако точек изометричное - связь отсутствует. Чаще облако точек вытянуто в виде эллипса в каком-то направлении, характеризуя нестрогую корреляционную зависимость между свойствами.

Если мы возьмем на оси x произвольные точки x_1, x_2 , x_3, то каждой из них будут соответствовать наборы значений y со своими средними значениями $\bar{y}_1, \bar{y}_2, \bar{y}_3$ (рис.14). Эти средние называются условными центрами распределения (среднее значение y равно $\bar{y}_i$ *при условии*, что $x = x_i$).

Соединив между собой множество условных центров распределения, мы получаем *линию регрессии*, которая является графическим выражением формы связи между x и y. Уравнение этой линии называется функцией или *уравнением регрессии*. Системе из 2-х величин всегда будет соответствовать две линии регрессии: $y_x = f(x)$ и $x_y = f(y)$. Регрессия может быть линейной (когда линии регрессии - прямые линии) и нели-

нейной. Для линейной регрессии уравнения будут иметь следующий вид;

$y = a_1 + в_1 X$ (регрессия y на x);

$\boldsymbol{x} = a_2 + в_2 y$ (регрессия x на y)

Линии регрессии пересекаются в точке, имеющей координаты $\bar{x}$ и $\bar{y}$.

Коэффициенты уравнения регрессии могут быть получены непосредственно с графика (рис. 15). Угол γ характеризует при этом *т е с н о т у* связи между x и y **.** Чем меньше γ , тем ближе связь к функциональной.

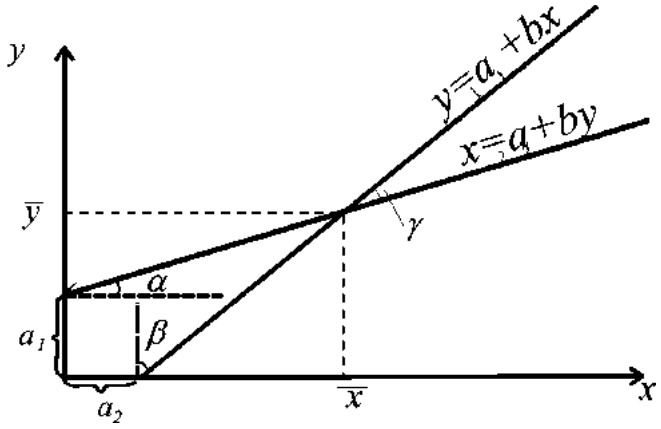


Рис. 15. Графики уравнений линейной регрессии

$в_1 = tg\ \alpha$; $в_2 = tg\ \beta$

Более точный, аналитический способ нахождения коэффициентов уравнения линейной регрессии из результатов опыта предложен Лежандром и Гауссом:

$$a_1 = \frac{\sum y \sum x^2 - \sum x \sum yx}{n\sum x^2 - (\sum x)^2} \quad , \qquad (43)$$

$$в_1 = \frac{n\sum yx - \sum x \sum y}{n\sum x^2 - (\sum x)^2} \quad , \qquad (44)$$

При нелинейной регрессии коэффициенты уравнения регрессии подбирают таким образом, чтобы сумма квадратов отклонений всех точек от линии зависимости была минимальной (метод наименьших квадратов):

$$\sum \delta^2{}_i \rightarrow \min$$

Вид аппроскимирующей функция задается либо, исходя из теоретических соображений, либо путем эмпирического подбора. Это могут быть уравнения параболы, синусоиды, показательной функции и т.д. В каждом из этих уравнений присутствуют коэффициенты *а , в , с,* которые определяют расположение кривой на графике. Следовательно, сумма квадратов отклонений также зависит от значений коэффициентов, т.е. является их функцией:

$$\sum \delta^2_{\text{i}} = f(a, в, c)$$

Чтобы найти минимум этой функции, надо приравнять нулю частные производные по неизвестным коэффициентам:

$$\frac{\partial f}{\partial a} = 0 ; \qquad \frac{\partial f}{\partial в} = 0 ; \qquad \frac{\partial f}{\partial c} = 0$$

В результате будет получена система уравнений, решая которую, мы найдем коэффициенты *а, в, с.*

Уравнения регрессии характеризуют форму связи между величинами, однако ничего не говорят о тесноте этой связи, то есть, близости ее к функциональной.

Теснота связи характеризуется такими показателями, как *ковариация, коэффициент корреляции, корреляционное отношение*. Ковариация - это математическое ожидание произведения отклонений двух случайных величин от их математического ожидания. Для выборочных данных формула расчета ковариации имеет вид:

$$Cov\ (x,y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \qquad (45)$$

Ковариация обладает размерностью, поэтому в практике обычно пользуются коэффициентом корреляции, который представляет собой ковариацию, нормированную по стандартам *х* и у :

$$\rho = \frac{\text{cov}(x, y)}{\delta_x \cdot \delta_y}$$

При определении коэффициента корреляции по выборочным данным можно использовать несколько модификаций расчетной формулы:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\cdot S_x \cdot S_y}\ ; \qquad (46)$$

$$r = \left(\frac{\sum xy}{n} - \bar{x}\cdot\bar{y}\right) \Big/ (S_x \cdot S_y); \qquad (47)$$

$$r = \frac{\sum x_i y_i - \frac{1}{n}\sum x_i y_i}{\sqrt{\left[\sum x^2{}_i - \frac{1}{n}(\sum x_i)^2\right]\left[\sum y_i{}^2 - \frac{1}{n}(\sum y_i)^2\right]}}. \qquad (48)$$

Если нет необходимости вычислять $\bar{x}$, $\bar{y}$, S_x , S_y для каких-либо других целей, то наиболее удобна для расчетов формула (48).

Если известны коэффициенты уравнения линейной регрессии, то для вычисления r можно воспользоваться еще двумя модификациями:

$$r = \frac{a_1 S_x}{S_y}\ ; \qquad (49)$$

$$r = \sqrt{в_1 \cdot в_2}\ . \qquad (50)$$

Приближенная оценка коэффициента корреляции может быть получена также графическим путем. Для этого облако точек делится на 4 квадранта линиями, проведенными в точках, соответствующих медианам x и y (рис. 16).

Коэффициент корреляция подсчитывается по формуле:

$$r = \frac{n_{i,3} - n_{2,4}}{N}$$

где N - общее количество точек, $n_{1,3}$ - количество точек в квадрантах 1 и Ш, $n_{2,4}$ - то же, в квадрантах П и IV.

Коэффициент корреляции определяет тесноту линейной связи между двумя величинами. Его значения изменяются от -1 до +1. При r = 0 связь между величинами отсутствует. При \r\ = 1 связь функциональная. Знак ± показывает, прямой, или обратной пропорциональной является взаимосвязь.

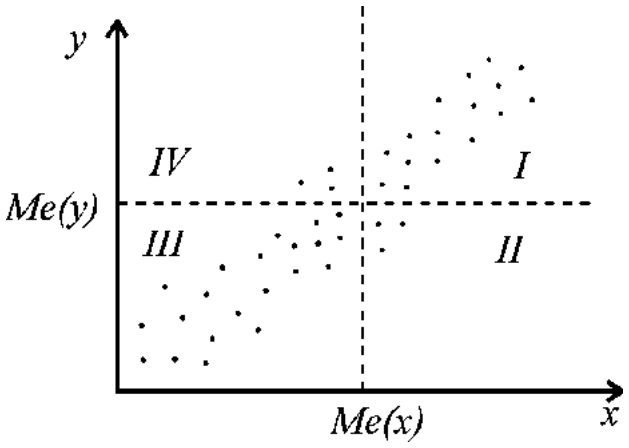


Рис. 16. Определение коэффициента корреляции графическим путем

Следовательно, проверка гипотезы о наличии корреляционной связи заключается в оценке значимости отличия от нуля вычисленных по выборке значений r :

$$H_0 : \rho = 0 ; \qquad H_1 : \rho \neq 0.$$

Критерий для оценки значимости отличия r от 0 предложен Фишером:

$$t = \frac{r\sqrt{n \cdot 2}}{\sqrt{1 - r^2}}.$$

Если вычисленное значение t больше, чем $t_{1-\frac{q}{2}}$,$n\text{-}2$ взятое из таблицы распределения Стьюдента, то отличие r от 0 признается значимым. В геологической практике иногда пользуются упрощенным критерием:

$$\left| r_{крит.} \right| = 2\sqrt{\frac{1}{N}} \qquad \text{(для уровня значимости 0,05).}$$

Во многих руководствах по математической статистике и задачниках по геохимии есть специальные таблицы критических значений коэффициента корреляции в зависимости от числа наблюдений N (16).

К сожалению, коэффициент корреляции очень чувствителен к виду функции распределения величин, входящих в двумерную систему. Поэтому, если эти распределения отличаются от нормальных и не поддаются нормализации, для проверки гипотезы о наличии корреляционной связи следует использовать

ранговый коэффициент корреляции Спирмена. При этом каждому значению *x* и *y* присваивается ранг в порядке возрастания их значений. Если значения повторяются, им присваивается средний между повторяющимися значениями ранг. Выражение для *r* имеет вид:

$$r = 1 - \frac{6\sum_{i=1}^{n} d^2{}_i}{n(n^2-1)}, \qquad (51)$$

где *n* - количество пар значений в выборке, *d* - разность рангов сопряженных значений *x* и *y* **.**

Для оценки значимости отличия рангового коэффициента корреляции от 0 существует специальная таблица критических значений (1, 16). Можно также воспользоваться выражением:

$$r_{крит.} = \frac{\varphi(p)}{\sqrt{n-1}}$$

где $\varphi(p)$ - значение обратной функции нормального распределения при доверительной вероятности *p* (берется из таблицы).

Если вычисленное значение *r* окажется больше $r_{крит.}$, отличие его от нуля считается значимым. В противном случае считаем, что линейная связь между величинами не установлена.

Рассмотрим пример вычисления рангового коэффициента корреляции. Требуется определить наличие корреляционной связи между мощностью кварцевой жилы и содержанием в ней золота по данным опробования (таблица 6).

Вычисленное значение *r* равно:

$$r = 1 - \frac{6 \cdot 153{,}5}{8(64-1)} = 1 - 1.827 = -0.827$$

Критическое значение *r*, взятое из таблицы для уровня значимости 0,05 и числа наблюдений *n*=8, равно 0,738. Следовательно, считаем, что между мощностью жилы и содержанием в ней золота существует значимая отрицательная корреляционная связь.

Таблица 6.

Расчет рангового коэффициента корреляции

№№ проб	Содержание		Мощность		d_i	d^2_i
	г/т	ранги	м	ранги		
1	2,5	1	2,6	8	7	49
2	3,8	3,5	2,5	7	3,5	12,25
3	12,1	6	1,4	3	3	9
4	3,4	2	2,1	5	3	9
5	3,8	3,5	2,1	5	1,5	2,25
6	13,2	7	1,1	1	6	36
7	6,4	5	2,1	5	0	0
8	14,1	8	1,2	2	6	36
						153,5

В случае, если корреляционная связь имеет нелинейный характер, она может существовать и при r=0. В этой ситуации необходимо вычисление *корреляционного отношения* (η). Корреляционное отношение показывает, какую долю от общей дисперсии составляет дисперсия, учтенная уравнением регрессии (закономерная составляющая дисперсии):

$$\eta_{y/x} = \frac{S_{\bar{y}_i}}{S_y}; \quad (52) \qquad \eta_{x/y} = \frac{S_{\bar{x}_i}}{S_x} \quad . \quad (53)$$

Незакономерная, случайная составляющая дисперсии характеризует разброс значений вокруг линии регрессии. Таким образом:

$$S^2_y = S^2_{\bar{y}_i} + S^2_{\text{случ.}}$$

Отсюда ясно, что, чем меньше случайная составляющая, т. е., чем меньше разброс значений от линии регрессии, тем выше значение корреляционного отношения. Закономерная составляющая дисперсии рассчитывается по формулам:

$$S_{\bar{y}_i} = \sqrt{\frac{1}{N}\sum_{i=1}^{m}(\bar{y}_i - \bar{y})^2 \cdot n_i} \quad ; \quad (54)$$

$$S_{\bar{x}_i} = \sqrt{\frac{1}{N}\sum_{i=1}^{m}(\bar{x}_i - \bar{x})^2 \cdot n_i} \quad . \qquad (55)$$

Выборочные данные при этом разбиваются на группы, для каждой из которых подсчитываются $\bar{x}_i$ (или $\bar{y}_i$). В формуле (54) N - общее число наблюдений, m - число групп, n_i - число наблюдений в i группе. Для расчета S_x и S_y следует воспользоваться формулой (23).

Значения η изменяются от 0 (связь отсутствует) до 1 (связь функциональная). В случае линейного характера взаимосвязи:

$$\eta_{x/y} = \eta_{y/x} = |r| . \qquad (56)$$

Таким образом, коэффициент корреляции можно рассматривать как частный случай корреляционного отношения. Закономерная составляющая дисперсии в этом случае связана с коэффициентом корреляции соотношением:

$$S^2_{\bar{y}_i} = S^2_{y} \cdot r^2$$

Значимость отличия η от нуля проверяется по критерию:

$$\Theta_y = \frac{\eta^2_{y/x}(N-m-2)}{(1-\eta^2_{y/x})(m-2)} \cdot \sqrt{\frac{(m-2)(N-m-4)}{2(N-4)}} \quad . \qquad (57)$$

Условные обозначения те же, что в формуле (54). Если $\eta = 0$, то Θ_y распределена по нормальному закону с параметрами (0,1). Следовательно, если вычисленное значение Θ_y превысит 3 (при доверительной вероятности 0,99) или 2 (при доверительной вероятности 0,95), считаем, что корреляционная связь существует.

Поскольку уравнения линейной регрессии являются наиболее простыми, на практике всегда необходимо выяснять причины нелинейности взаимосвязи величин и, по возможности, устранять их. К примеру, нелинейность связи может быть обусловлена неоднородностью выборочных данных, специфическими условиями эксперимента (оконтуривание рудных тел по заданной минимальной мощности) и т.д. В любом случае, вначале необходимо выяснить, линейный или нелинейный характер имеет установленная взаимосвязь.

Критерий для этой цели предложен Фишером. Он основан на сравнении значений η и r , которые в случае линейного характера связи должны быть равны.

$$F = \frac{\eta^{2}{}_{y/x} - r^2}{1 - \eta^{2}{}_{y/x}} \cdot \frac{N - m}{m - 2} \cdot \qquad (58)$$

Обозначения те же, что в формуле (54). Если вычисленное значение превысит $F_{q\ ,\ m-2\ ,\ N\ -\ m,}$ взятое из таблиц распределения Фишера, то связь признается нелинейной.

Рассмотрим пример вычисления корреляционного отношения.

Необходимо выяснить, существует ли зависимость между содержаниями золота и свинца на одном из месторождений, по данным опробования

Таблица 7

Содержания золота и свинца по данным опробования

Au, г/т	2,5	1,2	3,6	1,1	4,8	12,1	8,2	4,9
Pb, $n \cdot 10^{-3}$%	8	8	8	10	5	15	5	3
Au, г/т	6,8	13,2	15,1	7,8	8,8	9,1	2,6	5,5
Pb, $n \cdot 10^{-3}$%	4	8	10	5	8	6	6	3

Вычисленное значение коэффициента корреляции оказалось незначимым, но форма облака точек, позволяет предполагать наличие нелинейной зависимости между величинами.

Вычисление корреляционного отношения начинается с упорядочения выборки – значения зависимой переменной должны быть расположены в порядке возрастания, чтобы можно было объединить наблюдения в группы. Вычисления удобно проводить в виде таблицы (для удобства *Au* обозначим *x*, а *Pb* – *y*).

Корреляционное отношение равно:

$$\eta^{2}_{y/x} = \frac{94,17}{142} = 0,66$$

Оценим значимость отличия η от нуля:

$$\theta_y = \frac{0,66(16-5-2)}{(1-0,66)(5-2)} \cdot \sqrt{\frac{(5-2)(16-5-4)}{2(16-4)}} = 5,8 \cdot \sqrt{0,87} = 5,4 > 3$$

Таблица 8

Вычисление корреляционного отношения

x	y	$(y_i - \bar{y})^2$	$\overline{y_i}$	$(\overline{y_i} - \bar{y})^2 n_i$
1,1 1,2 2,5	10 8 8	9 1 1	8,7	$(8,7 - 7)^2 \cdot 3 = 8,67$
2,6 3,6 4,8 4,9	6 8 5 3	1 1 4 16	5,5	$(5,5 - 7)^2 \cdot 4 = 9,0$
5,5 6,8 7,8	3 4 5	16 9 4	4,0	$(4,0 - 7)^2 \cdot 3 = 27$
8,2 8,8 5,1	5 8 6	4 1 1	6,3	$(6,3 - 7)^2 \cdot 3 = 1,5$
12,1 13,2 15,1	15 8 10	64 1 9	11,0	$(11 - 7)^2 \cdot 3 = 48$
	$\bar{y} = 7$	$\sum$ = 142		$\sum = 94,17$

Значение критерия превышает допустимый разброс около 0, следовательно корреляционная зависимость существует.

Коэффициенты корреляции и уравнения регрессия чрезвычайно широко используются в геологической практике. Уравнение регрессии чаще всего применяются для предсказания значений одной случайной величины по значениям другой.

Например, в ряде случаев можно предсказать содержания попутных компонентов по содержаниям основных: кадмия по цинку, гафния по цирконию и т. д. Еще более широкое применение уравнения множественной регрессии находят при решении задач распознавания образов и классифицирования объектов. С этими методами, как и с использованием корреляционных матриц для решения прогнозных и классификационных задач мы познакомимся ниже, при рассмотрении многомерных

моделей.

8. МНОГОМЕРНЫЕ СТАТИСТИЧЕСКИЕ МОДЕЛИ.

8.1. Элементы матричной алгебры

Выше мы рассмотрели способы выявления и исследования взаимосвязей между двумя какими-либо свойствами геологического объекта или явления. Однако всякое геологическое явление характеризуется множеством признаков, которые можно наблюдать и измерять. Решение большинства геологических задач, как правило, требует совместного рассмотрения комплекса характеристик изучаемого объекта, каждый из которых представляет собой многомерную случайную величину, или случайный вектор. Совокупность этих векторов образуют матрицы наблюдений. В связи с тем, что во всех задачах многомерной статистики приходится выполнять операции с матрицами, рассмотрим некоторые понятия матричной алгебры.

Матрицей порядка $n \times m$ называется прямоугольная таблица, состоящая из n строк и m столбцов. Матрицы обычно обозначаются жирными заглавными буквами, а их элементы - маленькими буквами с нижними индексами:

$$\boldsymbol{A} = \begin{bmatrix} 4 & 3 & 2 \\ 2 & 1 & 6 \end{bmatrix} ; a_{12} = 3 \quad ; \qquad a_{21} = 2.$$

В геологии типичным примером матрицы является таблица химических анализов проб и вообще, любая таблица наблюдений, при условии, что она не имеет пустых клеток.

Если матрица имеет порядок $m \times m$, она называется *квадратной*.

Квадратная матрица, для всех элементов которой $x_{ty} = x_{yt}$ называется *симметричной*. Элементы этой матрицы симметричны относительно главной диагонали:

$$\boldsymbol{A} = \begin{bmatrix} 2 & 1 & 5 \\ 1 & 3 & 4 \\ 5 & 4 & 7 \end{bmatrix}$$

Если все элементы квадратной матрицы кроме лежащих на главной диагонали, равны 0, матрица называется *диагональной*:

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 5 \end{bmatrix}$$

Диагональная матрица, элементы которой равны 1, называется *единичной*. Она обозначается I и играет в матричной алгебре ту же роль, что и цифра 1 в операциях с обычными числами:

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Нулевая матрица, все элементы которой равны 0, играет в матричной алгебре ту же роль, что 0 в обычной алгебре.

Матрица порядка $1 \times N$ или $N \times 1$ называется, соответственно, *вектором-строкой* или *вектором-столбцом*:

$$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \qquad \mathbf{Y} = [\, y_1 \;\; y_2 \;\; y_3 \,]$$

Матрица порядка 1×1 называется *скаляр* , т. е. это просто число.

Если строки матрицы порядка $n \times m$ преобразовать в столбцы, то мы получим матрицу порядка $m \times n$, которая называется транспозицией первой матрицы. *Транспонированная* матрица обозначается A' (или A^T):

$$\boldsymbol{A} = \begin{bmatrix} 3 & 7 \\ 2 & 4 \\ 2 & 6 \end{bmatrix} ; \qquad \boldsymbol{A^T} = \begin{bmatrix} 3 & 2 & 2 \\ 7 & 4 & 6 \end{bmatrix}$$

Две матрицы равны, если порядок их одинаков и все элементы одной матрицы равны соответствующим элементам другой:

$$\begin{bmatrix} 2 & 1 & 3 \\ 1 & 4 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 3 \\ 6 & 4 & 1 \end{bmatrix} \neq \begin{bmatrix} 2 & 1 & 3 & 0 \\ 6 & 4 & 1 & 0 \end{bmatrix}$$

Операция *сложения* или *вычитания* определена только для матриц, имеющих один и тот же порядок. При этом складываются (вычитаются) одноименные элементы матриц:

$$A + B = D; \qquad d_{ij} = a_{ij} + b_{ij} .$$

Отметим, что:

$$A + B = B + A, \qquad (A \pm B)^T = A^T \pm B^T$$

Умножение матриц возможно только тогда, когда число

столбцов матрицы A равно числу строк матрицы B (матрица A слева). Если A (3 x 2), а B (2 x 4), то AB (3x4). Отметим, что A x $B \neq B$ x A.

Умножение вектора-строки на вектор-столбец есть скаляр:

$$[6\ \ 5\ \ 3] \text{ x } \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix} = 6*3 + 5*2 + 3*1 = 31$$

При перемножения матриц каждая строка левой матрицы умножается на каждый столбец правой матрицы. Каждое из полученных произведений является элементом новой матрицы:

$$\text{P=AB} = \begin{bmatrix} 3 & -1 \\ 0 & 4 \\ 2 & 6 \end{bmatrix} \cdot \begin{bmatrix} 4 & 0 & 2 & 3 \\ 0 & 2 & 6 & 10 \end{bmatrix} = \begin{bmatrix} 12 & -2 & 0 & -1 \\ 0 & 8 & 24 & 40 \\ 8 & 12 & 40 & 66 \end{bmatrix}$$

Например, элемент p_{23} вычислен так:

$$P_{23} = \begin{bmatrix} 0 & 4 \end{bmatrix} \times \begin{bmatrix} 2 \\ 6 \end{bmatrix} = 0*2 + \ 4*6 = 24.$$

Любая матрица может быть умножена на скаляр. При этом на скаляр умножается каждый элемент матрицы.

При перемножении нескольких матриц вначале перемножаются две левых матрицы, затем полученное произведение умножается на следующую матрицу и т.д.:

$$ABCD \ \rightarrow (AB)\cdot C \rightarrow (ABC)\cdot D$$

Отметим также, что:

$$(AB)^T = B^T A^T$$
$$(CDE)^T = E^T D^T C^T \quad ,$$

т. е. транспозиция располагается в обратном порядке.

Если вектор-строку умножить на ее транспозицию, мы получим сумму квадратов значений:

$$X \cdot X^T = \begin{bmatrix} 6 & 5 & 3 \end{bmatrix} \cdot \begin{bmatrix} 6 \\ 5 \\ 3 \end{bmatrix} = 6^2 + 5^2 + 3^2 = \sum x^2_1.$$

Эта операция широко используется в компьютерных программах.

Если транспозицию матрицы перемножить на саму матрицу, то получим новую матрицу квадратов значений и смешанных произведений:

$$X^T \cdot X = \begin{bmatrix} \sum x_1^2 & \sum x_i x_2 & . & \sum x_1 x_m \\ . & . & . & . \\ \sum x_m x_1 & \sum x_m x_2 & . & \sum x_m^2 \end{bmatrix}$$

Если эту матрицу умножить на скаляр $(\frac{1}{N-1})$, то в полученной *ковариационной* матрице диагональные элементы будут являться .дисперсиями случайных величин $x_1, x_2, \ldots x_м$ а недиагональные - ковариациями. Если же вместо x взять стандартизованные переменные, то получится *корреляционная* матрица, диагональные элементы которой равны единицам, а недиагональные - обычным коэффициентам корреляции:

$$R = \begin{bmatrix} 1 & r_{x_1 x_2} & . & r_{x_1 x_n} \\ . & . & . & . \\ . & . & . & . \\ r_{x_n x_1} & r_{x_n x_2} & . & 1 \end{bmatrix}$$

Эти операции также чрезвычайно широко используются в компьютерных программах.

Ковариационная (или корреляционная) матрица определяет статистические свойства исходной матрицы. Ниже мы познакомимся с методами многомерной статистики, позволяющими исследовать структуру и свойства корреляционных матриц, а предварительно рассмотрим еще несколько понятий.

Если AB = I (матрица квадратная), то $B = A^{-1}$ называется *обращенной* матрицей (то есть, обратной по отношению к A)

$$A^{-1} \cdot A = A \cdot A^{-1}$$

Чтобы показать возможности применения обращенных матриц, рассмотрим пример. Допустим, мы имеем систему линейных уравнений:

$$4\beta_0 + 8\beta_1 = 16$$
$$8\beta_0 + 20\beta_1 = 36$$

Эту систему можно записать в матричной форме:

$S \cdot \beta = q$, где $S = \begin{bmatrix} 4 & 8 \\ 8 & 20 \end{bmatrix}$; $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$; $q = \begin{bmatrix} 16 \\ 36 \end{bmatrix}$

Умножим обе части на $\boldsymbol{S}^{-1}$:

$S^{-1} \cdot S \cdot \beta = S^{-1} \cdot q$, т. е. $I\beta = S^{-1} \cdot q$

Следовательно, если мы найдем матрицу S^{-1} то, умножив ее на q , мы найдем и вектор β. Для нашего примера:

$S^{-1} = \begin{bmatrix} 1{,}25 & -050 \\ -0{,}50 & 0{,}25 \end{bmatrix}$; $\beta = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$; $\beta_0 = 2$; $\beta_1 = 1$.

Обращение матриц - задача довольно сложная, существуют различные способы ее реализации, например, метод Дулиттла (9). В настоящее время обращение матриц реализовано во всех программных продуктах, предназначенных для работы с матрицами.

Ознакомимся еще с одним понятием - *скалярного произведения* (*c*), которое представляет собой произведение вектора-строки X ($1\times m$) на матрицу ($m\times m$) и затем на вектор-столбец Y (m x 1). В итоге получается скаляр.

$$C = x \cdot S \cdot Y = (x_1 \; x_2 \ldots x_m) \cdot \begin{bmatrix} x_{11} & x_{12} & . & x_{1m} \\ . & . & . & . \\ . & . & . & . \\ x_{m1} & x_{m2} & . & x_{mm} \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \\ . \\ y_m \end{bmatrix}$$

Например:

$$C = \begin{bmatrix} 2 & 4 \end{bmatrix} \cdot \begin{bmatrix} 2 & 3 \\ 4 & 2 \end{bmatrix} \cdot \begin{bmatrix} 3 \\ 3 \end{bmatrix} = \begin{bmatrix} 20 & 14 \end{bmatrix} \cdot \begin{bmatrix} 3 \\ 3 \end{bmatrix} = 60 + 42 = 102.$$

Эта операция используется в тех случаях, когда результат необходим в виде одного числа.

В дальнейшем мы столкнемся с понятием детерминанта или *определителя* квадратной матрицы. Определителем матрицы называется многочлен вида

$$|A| = \sum \pm a_{1\alpha} \cdot a_{2\beta} \; a_{m\gamma} ,$$

где α, β. . γ - произвольная перестановка чисел от 1 до m. Суммирование ведется по произвольным перестановкам, поэтому определитель содержит $m!$ членов, половина из которых - четные (со знаком +), а половина - нечетные (со знаком -). Простейший

пример определителя:

Если $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$, то $|A| = a_{11}\, a_{22} - a_{12}\, a_{21}$

8.2. Статистические методы классифицирования геологических объектов

Задача классифицирования является одной из главнейших при любом геологическом исследовании. С необходимостью классифицировать различные природные объекты геолог постоянно сталкивается при решении как прогнозных, так и поисковых задач. Классификация при этом понимается как распределение объектов по классам по принципу их сходства.

Методы статистической классификации геологических объектов с использованием корреляционных и ковариационных матриц можно подразделить на следующие группы:

1)Методы анализа матриц с позиций теории графов.

2)Метод корреляционных профилей.

3)Методы, опирающиеся на понятие компактности.

4) Иерархическое группирование (кластер-анализ).

5) Каноническая корреляция.

6) Регрессионный анализ.

7) Дискриминантный анализ.

8) Факторный анализ.

Все методы основаны на группировании, то есть, разделении исходного массива данных на классы (кластеры, таксоны и т.д.), максимально однородные внутри и максимально разобщенные между собой.

8.2.1. Методы анализа корреляционных матриц с позиций теории графов

Этот метод позволяет представить результаты анализа в графическом виде, поэтому является наиболее простым и наглядным. При этом способе изучаемые элементы располагаются по окружности и те из них, которые связаны значимой корреляционной связью, соединяются прямыми линиями. Значимые отрицательные связи можно показывать другим цветом или пунктиром. В итоге получается наглядная

картина обособления отдельных групп элементов (рис. 17).

Если число линий велико, можно использовать не парные, а частные коэффициенты корреляции, которые меньше подвержены влиянию общих для всех элементов факторов, снижающих контрастность выделенных групп (рис. 17, б). Частные коэффициенты корреляции вычисляются по формуле :

$$r_{123} = \frac{r_{12} - r_{23} \cdot r_{13}}{\sqrt{(1 - r^2_{23})(1 - r^2_{13})}}. \qquad (59)$$

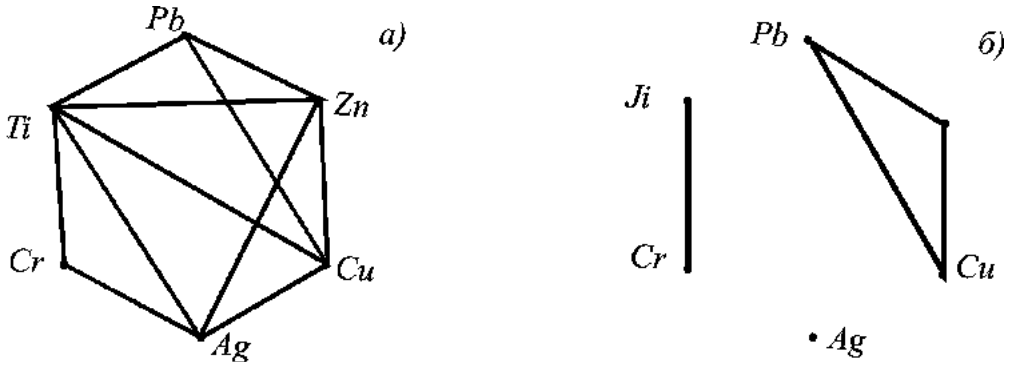


Рис .17. Графы корреляционных связей, построенных по парным (а) и частным (б) коэффициентам корреляции

Это коэффициент корреляции между первым и вторым элементом, очищенный от влияния связи их с третьим элементом

Еще более наглядным является построение взвешенных графов. При этом способе длина линий, соединяющих элементы, обратно пропорциональна величине парного коэффициента корреляции. В итоге получается картина, подобная изображенной на рис. 18.

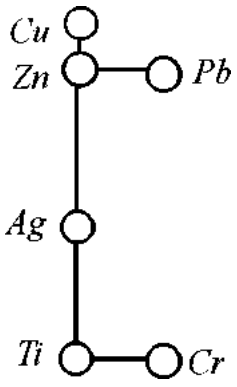


Рис. 18. Взвешенный граф корреляционных связей элементов

8.2.2. Метод корреляционных профилей

Сущность метода заключается в том, что на оси абсцисс

наносятся элементы или их символы, а ординаты точек соответствуют трансформированным коэффициентам корреляции Z (рис. 19). Z находится из соотношения:

$$Z = 0.5 \ln \frac{1+r}{1-r}.$$

Для Z составлена специальная таблица (15).

Если уверенное выделение групп по графику осуществить нельзя, то дальнейшее уточнение состава групп осуществляется о помощью критерия:

$$\max \left| Z_{kh} - Z_{lh} \right| < 2.77S . \qquad (60)$$

Здесь $S = \frac{1}{\sqrt{N-3}}$ (стандартное отклонение Z).

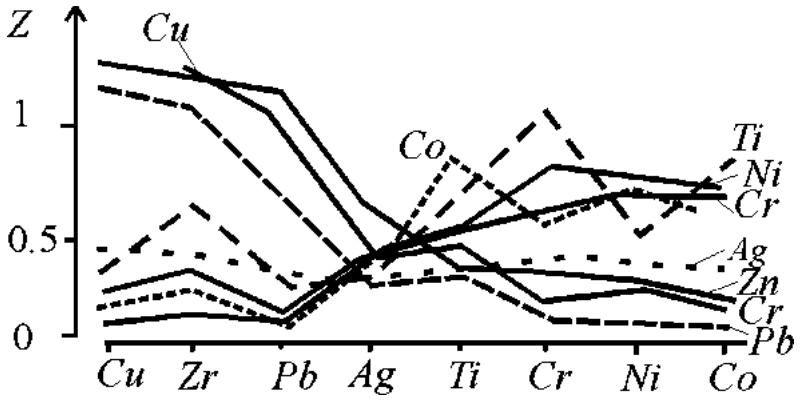


Рис .19. Корреляционные профили

Если неравенство (60) выполняется, то коэффициенты Z_{kh} и Z_{lh} можно считать равными, а элементы *k* и *l* можно объединять в одну группу.

Рассмотрим пример: (табл. 9, 10)*:*

Число проб в выборке равно 39.

$$2{,}77\,S = \frac{2{,}77}{\sqrt{36}} = 0{,}461.$$

Максимальное расхождение между столбцами *Zn* и *Pb* равно 0,11 <. 0,461, между *Ni* и *Cr* - 0,10 < 0,461, остальные расхождения больше 0,461.

Таблица 9

Значение коэффициента Z.

	Pb	Zn	Co	Ni	Cr
Pb	1	1.10	0.42	0.10	0.20
Zn	1.10	1	0.31	0.20	0.20
Co	0.42	0.31	1	0.30	0.30
Ni	0.10	0.20	0.30	1	0.87
Cr	0.20	0.20	0.30	0.87	1

Таблица 10

Значения *max (Z kh - Z lh)*

	Pb	Zh	Co	Ni	Cr
Pb		0.11	0.79	0.90	0.90
Zn	H_0		0.68	1.0	0.90
Co	H_1	H_1		0.57	0.57
Ni	H_1	H_1	H_1		0.10
Cr	H_1	H_1	H_1	H_0	

Таким образом, можно выделить три обособленные группы элементов:

(*Pb , Zn*) - *Co* - (*Ni, Cr*).

8.2.3. Методы, опирающиеся на понятие компактности

В качестве оценки компактности выделенных групп в данном случае используется соотношение $k = \frac{\overline{R}jj(m)}{\overline{R}ij(m)}$ **,** где *Rjj (m)* - средняя внутригрупповая связь при *m* выделенных группах, *Rij(m)* - средняя межгрупповая связь между *m* группами. Оптимальным признается вариант группирования, при котором *k* максимально.

Рассмотрим пример (таблица 11):

Рассчитаем *k* для двух вариантов группирования:

а) *(Pb* , Zn) - (*Co*) - (*Ni, Cr*);

б) (*Pb, Zn, Co)* - (*Ni* , *Cr*).

а) средняя внутригрупповая связь рассчитывается как средневзвешенное на количество элементов в каждой группе. В таблице эти значения закрашены.

$$\overline{Rjj}(3) = \frac{(3.6/4)\cdot 2 + (1/1)\cdot 1 + (3.4/4)\cdot 2}{5} = 0{,}9.$$

Средняя межгрупповая связь рассчитывается как среднее арифметическое средних значений межгрупповых связей. В таблице эти связи обведены пунктиром.

$$\overline{Rjj}(3) = \frac{0.35 + 0.175 + 0.3}{3} = 0{,}275$$

отсюда: $k_1 = \frac{0{,}9}{0{,}275} = 3{,}27.$

Таблица 11

Корреляционная матрица R

	Pb	Zn	Co	Ni	Cr
Pb	1.0	0.8	0.4	0.1	0.2
Zn	0.8	0.3	0.3	0.2	0.2
Co	0.4	1.0	1.0	0.3	0.3
Ni	0.1	0.3	0.3	1.0	0.7
Cr	0.2	0.3	0.3	0.7	1.0

б) коэффициенты рассчитываем по той же методике:

$$\overline{Rjj}(2) = \frac{0.67\cdot 3 + 0.85\cdot 2}{5} = 0{,}74,$$

$$\overline{Rij}(2) = \frac{0.1 + 0.2 + 0.2 + 0.2 + 0.3 + 0.3}{6} = 0{,}217$$

$k_2 = 3{,}41.$

Как видим, более оптимальным является второй вариант группирования.

8.2.4. Иерархическое группирование (кластер-анализ)

При этом методе объединение элементов в группы производится на основе определения "расстояния" между ними (меры близости). Объединение в группы представляет собой пошаговую процедуру и вычисление внутригрупповых и межгрупповых "расстояний" производится на каждом шаге. Как только вычисленное "расстояние" превысит заданное критическое значение, дальнейшее объединение элементов в группы прекращается.

В качестве мер для определения "расстояний" используются специальные метрики, обладающие свойствами рефлексивности, симметричности и транзитивности. Не вдаваясь в рассмотрение этих понятий, отметим, что коэффициент корреляции свойством транзитивности не обладает. Поэтому в качестве мер близости используется не сам коэффициент корреляции (г), а производные от него метрики: евклидово расстояние, дистанционный коэффициент и т.д. (15,17).

В качестве примера рассмотрим использование дистанционного коэффициента (d_T):

$$d_T = arccos\ r$$

Переход от d_T к r и обратно легко выполнить, воспользовавшись специальной таблицей (15).

Итак, в исходную корреляционную матрицу вместо значений r подставляем d_T .

Допустим, n = 37. Определим по таблицам критические значения r для доверительной вероятности 0,95 и 0,99. Они равны 0,325 и 0,418. Критические значения $d_T = arccos\ r_{крит.}$ равны, соответственно 1,24 и 1,14. Дальнейшие вычисления оформляются в виде таблицы, где d_T –внутригрупповая связь выделяемой группы, h_{min}- минимальная межгрупповая связь.

Группирование начинается с элементов, имеющих минимальное "расстояние" между собой и заканчивается после того, как $h._{min}$ или d_T превысит $arccos\ r_{крит.}$ В рассмотренном примере, при доверительной вероятности 99 %, объединение следует прекратить после 2-го шага, а при р = 95 % - после 3-го шага.

Таблица 12.

Матрица дистанционных коэффициентов (d_T)

	Pb	Zn	Co	Ni	Cr
Pb	0	0.64	1.16	1.47	1.37
Zn		0	1.27	1.37	1.37
Co			0	1.27	1.27
Ni				0	0.80
Cr					0

Таблица 13.

Расчет вариантов группирования

№ шага	Число групп	Элементы групп	d_T	h_{min}	arccos $r_{крит}$
1	4	Pb, Zn	0,64	0,80	1,24 для 95 %
2	3	Ni, Cr	0,80	1,22	1,14 для 99 %
3	2	Pb, Zn, Co	1,02	1,35	
4	1	Pb, Zn, Co, Ni, Cr	1,20	-	

Для большей наглядности результаты группирования можно представить в виде дендрографа (рис.20). Он строится таким образом: по оси абсцисс располагаются исследуемые элементы и откладываются межгрупповые расстояния, по оси ординат откладываются значения d_T

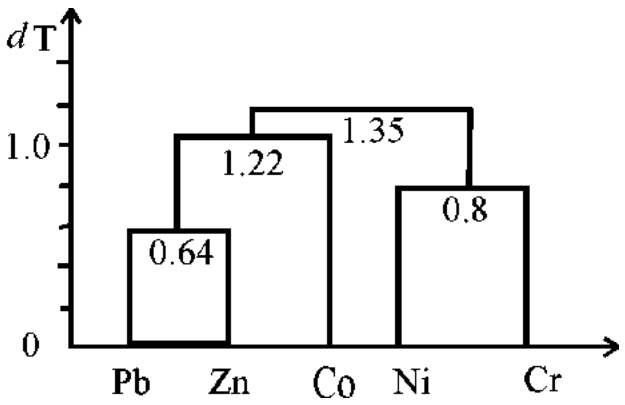


Рис.20. Дендрограф, построенный по данным табл. 11.

Для других метрик критические значения выбираются иначе, составлены специальные алгоритмы и программы. Однако в любом случае, предварительно необходим геологический анализ целесообразности деления совокупности на определенное число групп. Только в том случае, если с геологических позиций нельзя сказать ничего определенного по этому поводу, порог группирования задается формальными методами.

8.2.5. Каноническая корреляция

Суть метода заключается в вычислении коэффициента канонической корреляции между намеченными группами и сравнении полученного коэффициента ν с табличным значением χ^2 для заданного уровня значимости q и числа степеней свободы $n_1 \times n_2$ с помощью критерия J . Коэффициент ν находится из уравнения: $R_{12}R^{-1}{}_{22}R_{21} - \nu^2 R_{11} = 0$. (61)

Здесь RJJ - подматрицы коэффициентов корреляции внутри выделенных групп, Rij - подматрицы коэффициентов корреляции между элементами разных групп.

Значимость корреляции между группами определяется с помощью критерия J :

$$J = (N - n_2 - 1)\sum_{i=1}^{n_1} \frac{\nu^2{}_i}{1-\nu^2{}_i}, \qquad (62)$$

где $n_1 < n_2$, $N = n_1 + n_2$

Если вычисленное значение J больше табличного $\chi^2{}_{q\ ,\ n1\ x\ n2,}$ то связь значимая, и рассматриваемые группы можно объединять в одну. В противном случае группы объединять нельзя. Метод канонической корреляции употребляется обычно для уточнения состава заметно перекрывающихся групп.

Таким образом, общим моментом для рассмотренных выше методов является объединение в группы элементов, максимально связанных между собой, при отсутствии значимой связи между выделенными группами элементов.

8.2.6 Регрессионный анализ

Уравнение множественной регрессии можно записать в следующем виде

$$y = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n = a_0 + \sum_{i=1}^{n} a_i x_i \qquad (63)$$

Найденное уравнение позволяет наилучшим образом оценить совместное влияние многих параметров на переменную *y*. По значениям коэффициентов a_i можно также судить, каково влияние на *y* каждого отдельного параметра. Например, если уравнение имеет вид:

$$C_{Au} = 4{,}5 + 11{,}1c_{\mathrm{Pb}} + 1{,}3c_{Cu} + 0{,}02c_{Zn} - 6c_{Mo} - 0{,}003c_{Cb},$$

где c_i - концентрация элемента i , то можно сделать следующие выводы:

а)Содержание золота прямо пропорционально концентрации в жиле, свинца, меди и цинка и обратно пропорционально концентрации молибдена и висмута. Помимо предсказания содержаний золота, этот факт можно также использовать при выводе коэффициента геохимической зональности.

б)Наиболее информативными элементами являются свинец и молибден, в меньшей мере - медь. Цинк и висмут для простоты расчетов можно из уравнения безболезненно исключить.

в) Полученное уравнение может быть с успехом использовано для решения геологических вопросов (связь с определенной стадией минерализации и т.д.), а, следовательно, и для прогнозных целей.

Таким образом, задача регрессионного анализа сводится к нахождению коэффициентов уравнения множественной регрессии. Они определяются из соотношения:

$$a_i = \frac{Sy}{Si} \sum r_{ij} (r_{ij})^{-1}, \qquad (64)$$

где S_y - стандартное отклонение зависимой переменной;

S_i - стандартное отклонение i -й независимой переменной (значения S_i находятся по диагонали ковариационной матрицы);

r_{ij} - парная корреляция между y и i-.й независимой переменной; $(r_{ij})^{-1}$- обратная величина парной корреляции между независимыми переменными (r_{ij} берется из корреляционной матрицы).

Свободный член вычисляется по формуле:

$$a_0 = \bar{y} - \sum a_i \bar{x}_i . \qquad (65)$$

Поскольку уравнение регрессии есть смысл отыскивать лишь в том случае, если корреляция между *y* и набором переменных x_i

существует, то предварительно следует вычислить коэффициент множественной корреляции:

$$R = \sqrt{1 - \frac{|L|}{a_{11}|L'|}} \qquad (66)$$

где \$L\$ - определитель ковариационной матрицы;

\$L'\$ - определитель ковариационной матрицы без первого столбца и первой строки.

8.2.7. Дискриминантный анализ

Пусть мы имеем две матрицы наблюдений U и V из двух эталонных совокупностей. Суть дискриминантного анализа заключается в нахождении такого решающего правила (дискриминантной функции), которое позволило бы отнести новую оцениваемую выборку к одной из двух эталонных совокупностей (при условии, что выборка относится к одной из них). Дискриминантная функция строится следующим образом.

1) Вычисляем по данным выборок U и V выборочную ковариационную матрицу B :

$$B = \frac{1}{n_1 + n_2 - 2}(S_u + S_v),$$

где S_u и S_v - матрицы сумм центрированных квадратов и смешанных произведений (т.е. $S_u = u^T \cdot u$).

2)Обозначим B^{-1} через $C = (C_{ij})$ Тогда коэффициенты можно вычислить по формуле:

$$a_i = \sum c_{ij}(\bar{u}_j - \bar{v}_j) \qquad (67)$$

Дискриминантная функция равна:

$$D(x) = \sum a_i x_i \qquad (68)$$

3) Теперь необходимо найти критическое значение D_0 , такое, что если $D(x) > D_0$, то выборка относится к первой совокупности, а, если $D(x) < D_0$ то ко второй совокупности. D_0 находится из соотношения:

$$D_0 = \frac{1}{2}\sum a_i(\bar{u}_j + \bar{v}_i) \qquad (69)$$

Подставляя в формулу (68) выборочные значения x, получаем

одно из названных неравенств, что позволяетотнести изучаемый объект к одной из эталонных совокупностей.

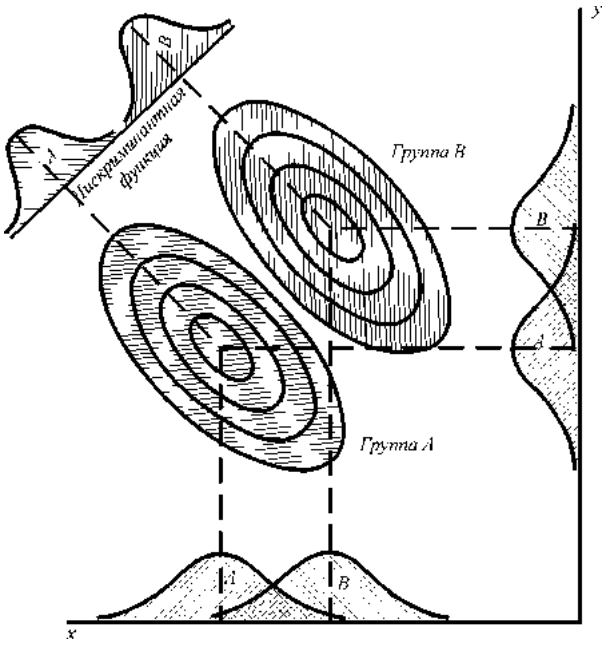


Рис. 21. Графическое изображение проекции двух двумерных случайных величин на линейную дискриминантную функцию.

Разумеется, построение дискриминантной функции имеет смысл только при условии, что $\mu_1 \neq \mu_2$, где μ_1 и μ_2 многомерные средние объектов U и V **,** имеющих k -мерное нормальное распределение.

Критерий для проверки гипотезы о равенстве многомерных средних можно найти из соотношения:

$$q = \frac{n_1 n_2 (n_1 + n_2 - k - 1)}{k(n_1 + n_2)(n_1 + n_2 - 2)} \sum_{i=1}^{k} \sum_{j=1}^{k} c_{ij} (\bar{u}_i - \bar{v}_i)(\bar{u}_j - \bar{v}_j)$$

Если вычисленное значение q больше $F_{\alpha, k, (n1\text{ -}n2\text{ -}k\text{-}1)}$ взятого из таблиц распределения Фишера, то расхождение между средними считается значимым. Мерой надежности принимаемых решений может служить "обобщенное расстояние" или критерий Махаланобиса (D^2)*:*

$$D^2 = \frac{1}{|R|} \sum_{i=1}^{n} \sum_{j=1}^{m} R_{ij} \frac{\bar{x}_{1i} - \bar{x}_{2i}}{\delta_i} \cdot \frac{\bar{x}_{1j} - \bar{x}_{2j}}{\delta_j}, \qquad (70)$$

где R - корреляционная матрица, R_{ij} - алгебраическое дополнение элемента, стоящего на пересечении i строки и j столбца.

Чем больше D^2, тем надежнее дискриминация. Численное выражение ошибки классификации можно найти из соотношения

$$P = 1 - \Phi(D^2),$$

где Ф(. . .) - функция нормального распределения.

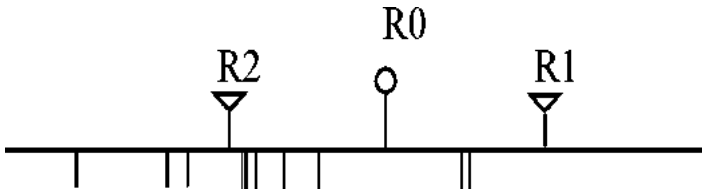


Рис.22. Графическое изображение результатов дискриминантного анализа.

R_1 и R_2 – многомерные средние двух эталонных объектов. R_0 – граничное значение для разделения двух совокупностей. Черточками показаны пробы испытуемой совокупности.

На рисунке 21 показана графическая интерпретация понятия дискриминантной функции для двумерного случая. Рисунок 22 иллюстрирует пример использования дискриминантного анализа для отнесения вновь выявленного рудопроявления к одному из эталонных типов месторождений.

8.2.8. Факторный анализ

С возрастанием количества анализируемых признаков быстро растет трудность изучения и классификации характеризуемых ими объектов. Между тем, любые сложнопостроенные системы, как правило, управляются сравнительно небольшим набором факторов. Выявлению и анализу этих факторов посвящен широкий круг вычислительных процедур, обычно объединяемых названием «факторный анализ». Следует однако, помнить, что в названной области выделяется несколько самостоятельных процедур: метод главных компонент (МГК), R–метод факторного анализа, Q–метод факторного анализа, анализ главных координат, анализ соответствия (5). Все эти методы основаны на выделении собственных значений и собственных векторов ковариационной или корреляционной матрицы, поскольку заранее предполагается, что в наборе многомерных наблюдений скрыта простая структура, выражающаяся через дисперсии и ковариации переменных.

Метод главных компонент позволяет выявить группы элементов, наиболее тесно связанных с тем или иным мощным фактором. Элементы, однонаправлено изменяющие свое состояние под действием общего фактора, могут быть объединены в комбинации, называемые главными компонентами. Число последних намного меньше исходного числа параметров, в то же время они несут практически всю полезную информацию об изменчивости свойств, заключенную в исходной совокупности.

Главные компоненты вычисляются по формулам:

$$1\text{ГК} = \sum \omega_{i1} x_i = \omega_{11} \cdot x_1 + \omega_{21} x_2 + \ldots + \omega_{n1} x_n \quad ; \qquad (71)$$

$$2\text{ГК} = \sum \omega_{i2} x_i \quad ;$$

$$3\text{ГК} = \sum \omega_{i3} x_i \quad \text{и т.д..}$$

Здесь x_i - значения параметров, ω_{ij} - факторные нагрузки (это влияние j -го фактора на i -й элемент, т.е. своего рода коэффициент корреляции между ними).

Таким образом, для нахождения главных компонент нам необходимо вычислить матрицу факторных нагрузок W. Она определяется из соотношения:

$$W = u\Lambda^{\frac{1}{2}} \qquad (72)$$

где u - матрица собственных векторов, а Λ - матрица собственных чисел корреляционной матрицы R . Элементы матрицы Λ определяются как корни характеристического уравнения:

$$\left|R - \lambda I\right| = 0 \quad , \qquad \text{где I - единичная матрица.}$$

Вычислив этот определитель, получаем уравнение, степень которого и число полученных корней равны числу строк в корреляционной матрице R . При этом $\lambda_1 > \lambda_2 > \lambda_3 \ldots > \lambda_n$, а $\sum \lambda_i = n$. Матрица u, находится из выражения:

$$(R - \lambda 1)\, u = 0$$

Подставляя в это уравнение найденные значения λ_i , получаем для каждого λ_i вектор значений u_i.

Допустим, в результате вычислений нами найден вектор значений λ для корреляционной матрицы размерностью (5 х 5):

$\lambda_1 = 2{,}41$; $\lambda_2 = 1{,}40$; $\lambda_3 = 0{,}71$; $\lambda_4 = 0{,}32$; $\lambda_5 = 0{,}17$.

Поскольку $\sum \lambda_i = n$, то вклад каждого фактора в общую изменчивость можно определить по формуле:

$$v_k = \frac{\lambda_k}{n} \cdot 100\%. \qquad (73)$$

Отсюда:

$v_1 = \frac{2,41}{5} \cdot 100\% = 48.2\%$; $v_2 = 28\%$; $v_3 = 14.1\%$; $v_4 = 6.3\%$; $v_5 = 3.4\%$.

Таким образом, вклад первых трех факторов в общую изменчивость составляет более 90%, поэтому при анализе матрицы *W* можно ограничиться рассмотрением первых трех главных компонент.

Таблица 14.

Матрица факторных нагрузок *W*

Фактор	*Fe*	*Mg*	*Ca*	*Na*	*K*
F_1	0,75	0,75	0,65	0,63	0,67
F_2	-0,57	-0,52	-0,05	0,66	0,61
F_3	-0,10	-0,28	0,76	-0,15	-0,17

Как видим, 1-й фактор значимо влияет на все элементы. Такой фактор обычно называют генеральным. Генеральный фактор отрицательно сказывается на контрастности корреляционной матрицы, обуславливая перекрытие выделяемых групп. "Очистка" связей осуществляется с помощью анализа факторных нагрузок 2-го и 3-го факторов. Дать главным факторам геологическую интерпретацию не всегда возможно, но когда это удается, информативность метода резко возрастает. В частности, в рассмотренном примере со 2-м фактором, очевидно, связано проявление щелочного метасоматоза. Понятно, что геометризация на площади участка значений 2-й главной компоненты позволит в этом случае оконтурить зоны щелочного метасоматоза. Дать интерпретацию 1-му фактору сложнее. Возможно, это влияние расстояния от контакта с гранитоидной интрузией. С 3-м фактором, видимо, связан процесс карбонатизации пород.

Помимо группирования, метод главных компонент можно использовать и для распознавания образов. Для этого в координатах двух ГК выносятся значения для эталонных

объектов и локализуются области, отвечающие этим объектам (рис.23.)

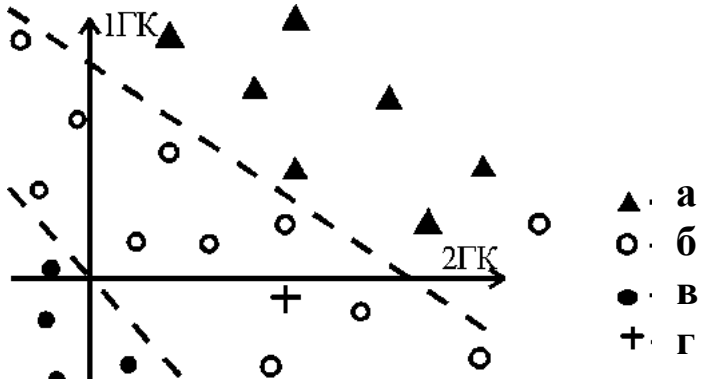


Рис. 23. Определение промышленного типа месторождения по методу главных компонент.

а - 1-й промтип, б - 2-й промтип, в – непромышленнные объекты,
г - изучаемое рудопроявление.

Таким образом МГК сводится к линейному преобразованию M исходных переменных в m новых переменных, каждая из которых является линейной комбинацией исходных переменных. При этом МГК не является статистическим методом и мы практически не имеем формальных критериев для отбрасывания некоторых переменных или компонент, дающих очень малый вклад в суммарную дисперсию. О правильности своих действий мы можем судить только после проведения анализа МГК.

В отличие от МГК, факторный анализ считается статистическим методом, поскольку в его основе лежат некоторые предположения о природе изучаемой совокупности. Предполагается, что связь между m переменными является отражением корреляционной зависимости каждой из переменных с p взаимно некоррелированными факторами, причем $p<m$ (если $p = m$, модель эквивалентна МГК). Поэтому дисперсию для m переменных можно вычислить с помощью дисперсии p – факторов плюс вклад, происхождение которого одинаково для всех переменных.

Модель R – метода выражается в следующем виде:

$$Xj = \sum_{i=1}^{m}(a_{ik} \cdot f_k) + \varepsilon_j$$

где f_k – $к$-й общий фактор, ε – случайная компонента, присущая исходной переменной a_{ik} – факторная нагрузка i-го элемента на k-й фактор.

В Q-методе факторного анализа, в отличие от R-метода, анализируются взаимосвязи между наблюдениями, а не переменными.

Одно из главных препятствий в применении геологами различных модификаций факторного анализа заключено в абстрактности понятий собственных векторов и собственных значений корреляционных матриц. Между тем, эти категории имеют вполне определенный содержательный и геометрический смысл. Рассмотрим для примера корреляционную матрицу 2×2, позволяющую представлять рассматриваемые понятия в виде двумерных графиков. Как известно, любую матрицу можно представить геометрически в многомерном пространстве как множество векторов. Будем считать, что каждая строка матрицы дает координаты концевых точек вектора, представляющую эту строку. Допустим, исходная корреляционная матрица имеет вид:

$$\begin{bmatrix} 1{,}00 & 0{,}86 \\ 0{,}86 & 100 \end{bmatrix}$$

Ее собственный вектор $\mathrm{I} = \begin{bmatrix} 0{,}707 \\ 0{,}707 \end{bmatrix}$, собственное значение 1,86 (или 93%).

Собственный вектор $\mathrm{II} = \begin{bmatrix} -0{,}707 \\ 0{,}707 \end{bmatrix}$, собственное значение 0,14 (или 7%).

На рис. 24 видно, что строки корреляционной матрицы можно представить как произвольные оси двумерного эллипсоида, тогда собственные вектора, дают направление главных осей эллипсоида, а корень из величины собственного значения – длину главных полуосей. Поскольку собственные значения включают в себя дисперсии переменных, очевидно, что и факторы отражают дисперсии (точнее, стандартные отклонения). При этом наклон и длина главных осей эллипсоида наглядно свидетельствуют о влиянии фактора на значения конкретной переменной.

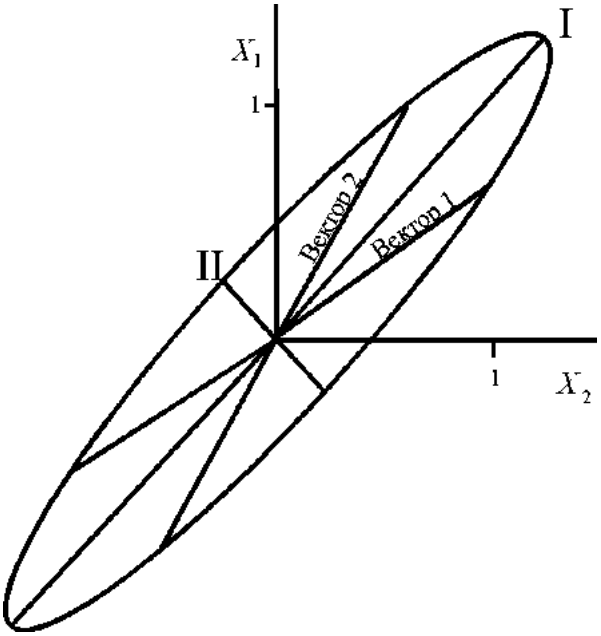


Рис. 24. Графическое изображение собственных векторов корреляционной матрицы.

В нашем случае матрица факторных нагрузок имеет вид:

$$\begin{array}{c} \\ \textit{переменные} \end{array} \begin{array}{c} \textit{факторы} \\ \begin{array}{cc} \text{I} & \text{II} \end{array} \\ \begin{array}{c} 1 \\ 2 \end{array} \begin{bmatrix} 0{,}964 & -0{,}264 \\ 0{,}964 & 0{,}264 \end{bmatrix} \end{array}$$

Как видим, факторы одинаково влияют на $1^{\underline{ю}}$ и $2^{\underline{ю}}$ переменную, поэтому оси эллипсоида расположены под углом 45º к осям координат (это неибежное следствие работы с матрицей 2×2, для матриц высоких порядков такое соотношение нарушается). Относительный вклад каждого фактора в дисперсию переменных различен и это отражается на длинах главных осей эллипсоида. Мы можем вместо двух исходных переменных оперировать значениями $1^{\underline{го}}$ фактора, учитывающего по 93% дисперсии каждой из переменных. Сокращение размерности пространства в этом случае обернется для нас потерей 7% информации.

Очевидно, что наши рассуждения применимы к корреляционным матрицам любой размерности, хотя геометрическое представление собственных векторов для матриц высоких порядков затруднительно.

Поскольку одна из главных задач факторного анализа - сокращение размерности исходного пространства признаков, важнейшим вопросом является выбор количества сохраненных факторов. Формального ответа на этот вопрос не существует, поэтому в большинстве случаев рекомендуется сохранять столько факторов, сколько имеется собственных чисел, больших 1, то есть сохраняются факторы, вклад которых в дисперсию больше, чем у каждой из исходных переменных. Эта рекомендация полезна в тех случаях, когда исходные данные хорошо скоррелированы и первые 2-3 фактора дают основной вклад в общую дисперсию. Если же переменные скоррелированы слабо, то половина и даже больше факторов может иметь собственные числа большие единицы. Число факторов получается слишком большим, причем вклад каждого из них в дисперсию невелик, а содержательная интерпретация затруднительна. В таких случаях применение факторной модели следует признать нецелесообразным.

В ряде случаев бывает затруднительно дать интерпретацию факторов даже если переменные хорошо скоррелированы. Перекрытие групп переменных зачастую обусловлено тем, что положение p ортогональных факторных осей в m-мерном пространстве определяется положением $m–p$ ненужных ортогональных осей в выборочном пространстве. Исключив из рассмотрения ненужные оси, мы можем произвести вращение оставшихся факторных осей таким образом, чтобы выделенные группы наилучшим образом расположились в новых координатах. В наиболее часто используемом методе (метод варимакс Кайзера) вращение осуществляется до тех пор, пока проекции каждой переменной на факторные оси не окажутся близкими либо к нулю, либо к ±1. Чаще всего такое вращение приводит к тому, что для каждого фактора мы получаем несколько больших значений нагрузок и много близких к нулю. Это существенно облегчает содержательную интерпретацию факторов. Если же вращение факторных осей лишь ухудшает первоначальный результат, это свидетельствует либо о взаимной коррелированности факторов, либо о неприменимости выбранной факторной модели.

Графическое представление процедуры вращения факторных осей для двумерного случая дано на рис. 25.

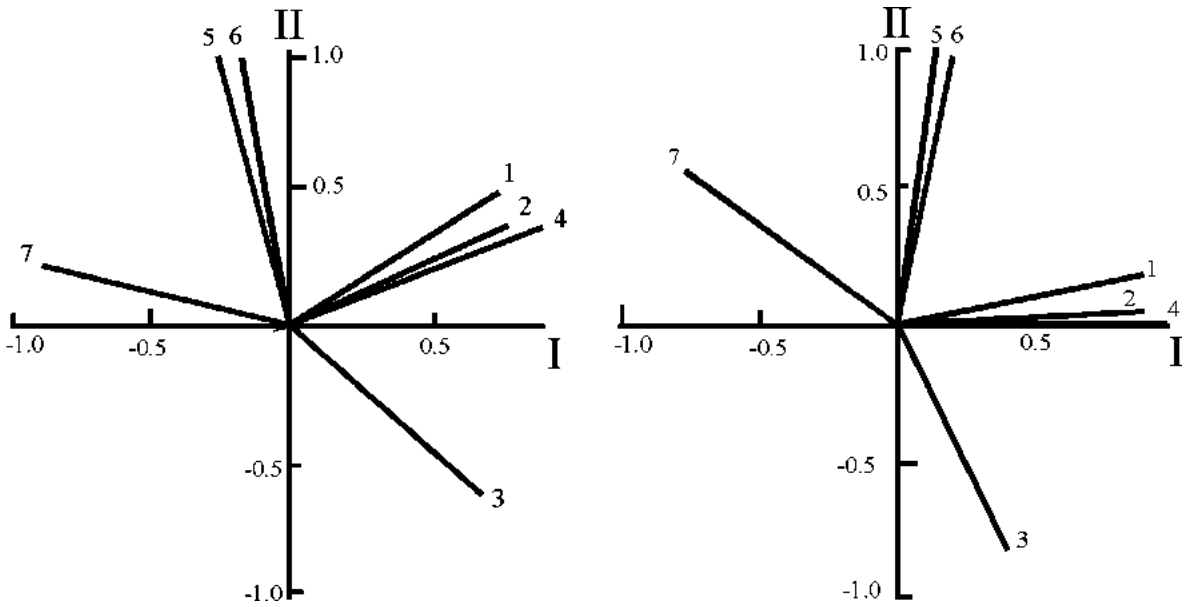


Рис. 25. Вращение факторных осей для двумерного случая.

Проекции векторов переменных на факторные оси соответствуют их факторным нагрузкам. Видно, что после вращения разделение элементов на группы значительно улучшилось. При этом длина векторов и их относительное положение не изменились.

Таким образом, факторный анализ сочетает в себе преимущества и возможности как методов группирования, так и распознавания образов. В частности, он может быть использован как вариант множественной регрессии для вычисления востановленных значений переменной:

$$Хвосст. = S \cdot \omega_j \cdot Z'_j + \bar{x} \cdot \varepsilon',$$

где S – диагональная матрица m x m оценок стандартов m переменных;

ω_j – факторная нагрузка j фактора;

Z'_j – вектор-строка значений фактора j;

$\bar{x}$ - среднее значение параметра по выборочным данным;

ε' -вектор-строка размером N (число наблюдений) вида {1, 1, 1, ... 1}.

Таким способом можно оценить влияние каждого

выделенного фактора (процесса) на распределение конкретного элемента и геометризовать в пространстве интенсивность этого влияния. Эта задача обычна при создании генетических моделей и прогнозо-поисковых комплексов.

9. МОДЕЛИРОВАНИЕ ПРОСТРАНСТВЕННОЙ ИЗМЕНЧИВОСТИ СВОЙСТВ ГЕОЛОГИЧЕСКИХ ОБЪЕКТОВ. ТРЕНД-АНАЛИЗ.

Рассмотренные нами выше статистические модели достаточно полно характеризуют интенсивность изменчивости и ее средний уровень, но не учитывают информации о пространственном размещении точек наблюдений. Между тем, важность исследования пространственных закономерностей изменчивости трудно переоценить. Выявление закономерностей изменения в пространстве параметров рудных тел обоснованно позволяет выбирать плотность и геометрию разведочной сети, а также более целенаправленно организовать проведение поисковых и разведочных работ на фланги и глубину месторождения. Той же цели поисковых и разведочных работ служат методы выявления зональности месторождений (т.е. особенностей пространственного изменения концентраций различных элементов и минералов или их свойств) и закономерностей пространственного размещения месторождений относительно интрузий, разломов и других элементов (моделирование дискретных полей).

Для исследования характера изменчивости признаков используются горно-геометрические и аналитические (аппроскимация полиномами) методы моделирования.

9.1. Горно-геометрическое моделирование

Основоположником методов горно-геометрического моделирования является русский ученый П. К. Соболевский. Им установлены основные принципы моделирования, главный из которых гласит, что значение параметра в каждой точке геологического тела есть функция координат пространства:

$$a = f(x, y, z).$$

Отсюда следует, что если мы знаем математическое выражение этой функции, то можем определить значение

параметра в любой заданной точке объекта.

Геометрическое моделирование графически выполняется в виде изолиний значений параметра. Примерами таких моделей могут служить карты изолиний концентраций какого-либо элемента, карты стратоизогипс угольного пласта, различные геофизические и геохимические карты, наконец, обычная топографическая карта. Поскольку аналитическое выражение таких поверхностей практически невозможно, П.К.Соболевским разработан специальный математический аппарат, позволяющий производить с топоповерхностями любые арифметические и алгебраические действия.

Горно-геометрические модели наглядно отображают плавные, закономерные изменения, но, к сожалению, не дают информации о многочисленных случайных отклонениях от топоповерхности в отдельных точках наблюдения. Поэтому П. Л. Каллистовым в 1956 году было предложено разделять общую изменчивость признака на две составляющих: закономерную и случайную. При случайной изменчивости значения признака в разных точках не зависят друг от друга и от расстояния между точками. При закономерной изменчивости значения признака в различных точках функционально связаны между собой, то есть, являются функцией пространственных координат.

Конечно, в реальной обстановке всегда присутствуют оба вида изменчивости. Чтобы разделить их, П. Л. Каллистовым предложено сглаживать эмпирические данные методом "скользящего окна". Пример сглаживания наблюдений окном из трех проб показан на рис. 26. Суть сглаживания заключается в следующем. Вычисляем среднее по первым трем пробам и присваиваем это значение второй точке. Затем смещаемся на один интервал и вновь вычисляем среднее по трем пробам (второй, третьей и четвертой). Это значение присваивается третьей точке. И так далее, до конца профиля, или выработки.

Осредняющая кривая будет характеризовать закономерную изменчивость, а для характеристики случайной изменчивости используется .дисперсия, рассчитанная через отклонения каждого частного значения от осредняющей кривой.

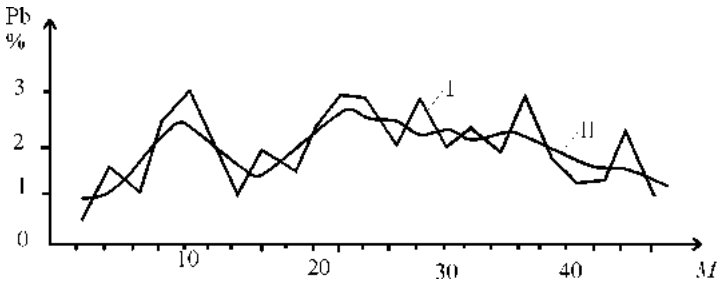


Рис.26. Распределение свинца по простиранию рудного тела

I - эмпирические данные; II - осреднящая кривая

Аналогичным образом можно построить и осредняющую поверхность. Средние значения признака рассчитываются при этом для центра каждой разведочной ячейки, а затем проводятся изолинии. Дисперсия случайной составляющей определяется через отклонение частных значений от построенной топоповерхности.

Рассмотренный метод сглаживания не является строго математическим и обладает рядом недостатков: при сглаживании происходит сдвиг наблюденных значений максимумов и минимумов, при объединении проб в одном окне мы искусственно делаем их взаимосвязанными, хотя они могут быть и независимыми, наконец, соотношение случайной и закономерной составляющих изменчивости зависит не только от свойств самого ряда, но и от способа сглаживания. Все это ограничивает возможности метода Каллистова.

9.2. Аналитические методы моделирования пространственной изменчивости

Дальнейшим развитием идей Соболевского и Каллистова явилась разработка аналитических методов исследования пространственной изменчивости, получивших название тренд-анализа. Тренд-анализ включает в себя и различные методы сглаживания данных, но обычно тренд-поверхности строятся путем аппроксимации исходных данных полиномами различных степеней (аппроксимация - это замена одних математических объектов .другими, более простыми). Каждое свойство объекта описывается при этом непрерывным полем, а значение параметра

в любой точке пространства определяется как значение аппроксимирующей функции в этой точке, плюс случайная переменная:

$$a = f(x, y) + \varepsilon$$

С помощью тренд-анализа решается три типа задач: 1) проверка гипотезы о *наличии* какой-либо закономерности в пространственной изменчивости параметра;

2)выделение и описание наиболее общих закономерностей (*поверхностей* тренда);

3) выявление отклонений от поверхностей тренда (так называемых "аномальных" участков, или *«остатков»* тренда).

Проверка гипотезы о наличия тренда заключается в сравнения полученной эмпирической последовательности значений признака с такой теоретической последовательностью, в которой закономерная составляющая заведомо отсутствует.

Обычно наличие тренда проверяется двумя простыми способами: а) *смены знака* и б) *количества скачков.*

Точкой смены знака называется такой элемент последовательности, в котором знак приращения изменяется на противоположный. В случайной последовательности математическое ожидание числа точек смены знака определяется выражением:

$$M(t) = \frac{2N-4}{3}, \qquad (74)$$

где N - общее количество наблюдений. Значимость отличия эмпирически подсчитанного числа точек смены знака (t) от теоретического ($M(t)$) определяется по критерию:

$$Z = \frac{t - M(t)}{\sqrt{\delta^2_t}}, \qquad (75)$$

где $\delta^2_t = \frac{16N-29}{90}$.

Если Z < -1,65 (для 5% уровня значимости), то тренд существует. В противном случае считаем, что закономерная составляющая пространственной изменчивости отсутствует.

Рассмотрим пример. Требуется установить, существует ли закономерность в распределении содержаний золота по оси штрека (рис. 27.)

В приведенном примере N = 18, t = 8, $M(t) = \frac{36-4}{3}$ = 11. Определим значимость отличия t от $M(t)$:

$$Z = \frac{8-11}{\sqrt{\frac{16 \cdot 18 - 29}{90}}} = -\frac{3}{1,7} = -1,76$$

Если Z< -1,65, следовательно, принимаем гипотезу о наличии тренда.

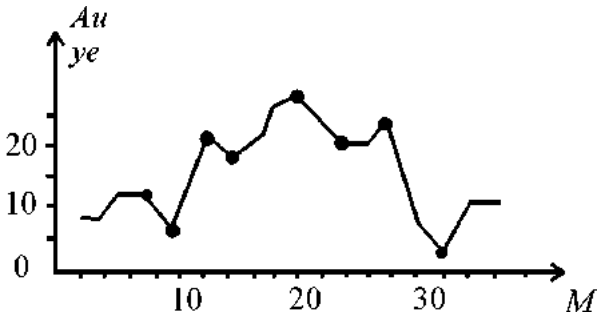


Рис. 27. Распределение содержаний Au по оси штрека.

Метод скачков заключается в том, что вся последовательность делится на "скачки", представляющие собой группы соседних элементов последовательности, имеющих один знак - "+" (если все значения больше медианы) или "-" (если значения признака меньше медианы).

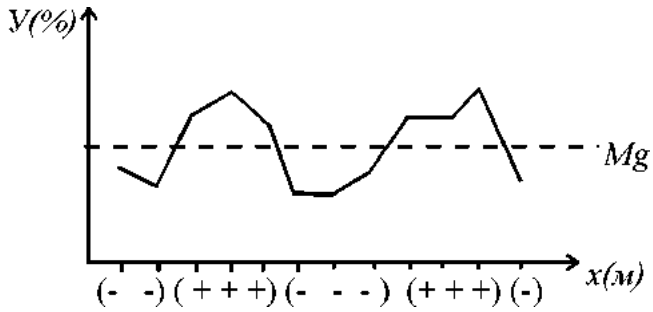


Рис. 28. Разбивка последовательности значений признака на скачки

Теоретическое число скачков при случайной последовательности равно:

$$M(u) = \frac{2n_1n_2}{n_1 + n_2} + 1, \qquad (76)$$

где n_1 - число значений со знаком "+, n_2 - со знаком "-".

Значимость отличия эмпирического числа скачков от теоретического определяется по критерию:

$$Z = \frac{u - M(u)}{\sqrt{\delta^2_u}} \quad , \qquad (77)$$

где $\delta^2_u = \frac{2n_1 \cdot n_2(2n_1 \cdot n_2 - n_1 - n_2)}{(n_1 + n_2)(n_1 + n_2 - 1)}$.

Если вычисленное значение Z превышает (-1,65), считаем, что тренд отсутствует (при уровне значимости 0,05).

Способ смены знаков обычно употребляется для выявления локальных закономерностей, способом скачков устанавливаются региональные закономерности. Для принятия гипотезы о наличии тренда достаточно, чтобы она подтвердилась хотя бы одним способом.

Выделение общих, региональных закономерностей изменчивости осуществляется путем построения поверхностей тренда различных порядков на основе аппроскимации исходных данных полиномами различных степеней. При этом полином первой степени описывает общую для всего участка тенденцию к возрастанию или убыванию значений признака по определенному направлению. Полином более высоких степеней отражает закономерности более высоких порядков. Обычно достаточно вычисление поверхностей тренда не выше 3-4 порядков. Для описания периодических закономерностей используют тригонометрические полиномы.

Поверхность тренда 1-го порядка - это плоскость, уравнение которой выглядит следующим образом:

$$x_{ij} = a_{00} + a_{10}M_i + a_{01}N_j \quad ,$$

где M и N - координаты пространства. Для поверхности 2-го порядка число коэффициентов увеличивается до 6, поверхности более высоких порядков описываются еще более сложными выражениями. Вручную возможно вычисление только поверхностей 1-го порядка.

Коэффициенты ортогональных полиномов находятся методом наименьших квадратов из уравнения:

$$S \cdot \alpha = g.$$

Для поверхности 1-го порядка это выражение можно представить в виде следующих матриц:

$$\boldsymbol{S} = \begin{bmatrix} n & \sum M & \sum N \\ \sum M & \sum M^2 & \sum MN \\ \sum N & \sum MN & \sum N^2 \end{bmatrix}; \quad \boldsymbol{\alpha} = \begin{bmatrix} a_{00} \\ a_{10} \\ a_{01} \end{bmatrix};$$

$$\boldsymbol{g} = \begin{bmatrix} 4.06 \\ 103.1 \\ 160.8 \end{bmatrix}$$

Вычислив значения матриц $\boldsymbol{S}$ и $\boldsymbol{g}$, находим обращенную матрицу S^{-1} и вычисляем матрицу коэффициентов $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha} = \boldsymbol{S}^{-1} \cdot \boldsymbol{g}.$$

Допустим, нами получены следующие значения названных матриц:

$$\boldsymbol{S} = \begin{bmatrix} 12 & 240 & 450 \\ 240 & 6300 & 9000 \\ 450 & 9000 & 21875 \end{bmatrix}, \quad \boldsymbol{\alpha} = \begin{bmatrix} a_{00} \\ a_{10} \\ a_{01} \end{bmatrix}, \quad \boldsymbol{g} = \begin{bmatrix} 4{,}06 \\ 103{,}1 \\ 160{,}8 \end{bmatrix}$$

тогда**:**

$$\boldsymbol{\alpha} = \boldsymbol{S}^{1}\boldsymbol{g} = \begin{bmatrix} 0{,}63 & -0{,}01 & -0{,}0077 \\ -0{,}01 & 0{,}0007 & 0{,}000 \\ -0{,}0077 & 0{,}000 & 0{,}0002 \end{bmatrix} \cdot \begin{bmatrix} 4{,}06 \\ 103{,}1 \\ 160{,}8 \end{bmatrix} = \begin{bmatrix} -0{,}018 \\ 0{,}015 \\ 0{,}002 \end{bmatrix}$$

Искомый полином примет вид:

$$x = -0{,}018 + 0{,}015M + 0{,}002N.$$

Подставляя вместо M и N координаты пространства, мы можем построить поверхность тренда 1-го порядка. Затем в каждой точке считаем отклонение реального значения от плоскости тренда и строим карту остатков от тренда. Ту же операцию можно проделать для поверхностей более высоких порядков и также построить карты отклонений от тренда. Чем выше порядок поверхности, тем большую часть общей изменчивости (дисперсии) она учитывает. В пределе, очевидно, возможен подбор аппроксимирующей поверхности бесконечно высокого порядка, которая учитывала бы 100% всей изменчивости признака, но на практике обычно бывает достаточно построение поверхности, описывающей 90-95% всей дисперсии. Оценка "силы" тренда осуществляется следующим образом. Допустим, что общая сумма квадратов отклонений x от

$\bar{x}$ для всей площади равна 0,60. Это 100% всей изменчивости признака. Подсчитав сумму квадратов отклонений от поверхности тренда 1-го порядка (допустим 0,26), находим, что остаток этот составляет 44% от 0,60. Следовательно, поверхность тренда 1-го порядка учитывает 56% всей изменчивости. Допустим далее, что сумма квадратов отклонений от поверхности тренда 2-го порядка составила 0,05, т.е. 7,5% от 0,60. Следовательно, поверхность 2-го порядка учитывает 92,5% всей изменчивости признака (в том числе и 56% учтенных поверхностей 1-го порядка). Плоскостью 3-го порядка можно было бы учесть еще большую часть изменчивости, но в данном случае, вычисление кубической поверхности уже излишне.

Анализ поверхностей тренда и, особенно, остатков от них может дать богатую информацию для различных геологических выводов, поэтому данная процедура является одной из важнейших в тренд-анализе.

На рисунке 29 показан пример использования поверхностей тренда для выявления зональности золоторудного поля. В рудном поле известно два промышленных месторождения, три рудопроявления и ряд точек минерализации. С целью выявления характерных геохимических признаков промышленной минерализации на площади проведена литогеохимическая съемка по вторичным ореолам рассеяния. В результате выявлены многочисленные аномалии золота и его элементов-спутников, в размещении которых визуально никакой закономерности не устанавливается. Выявленные факторным анализом ассоциации элементов размещаются более упорядоченно, тем не менее картина геохимического поля по-прежнему сложная и трудно интерпретируемая (рис.29а).

Последовательная аппроксимация значений факторов поверхностями тренда позволила выявить в строении рудного поля 3 зоны – ядерную, для которой характерна ассоциация Cu-Co-Ni, промежуточную, с пониженными концентрациями всех элементов, и фронтальную, где накапливаются элементы ассоциации Au-As-Ag. Наиболее отчетливо зональность рудного поля отражается в поверхностях тренда 3-го порядка. При этом промышленные месторождения локализуются во фронтальной

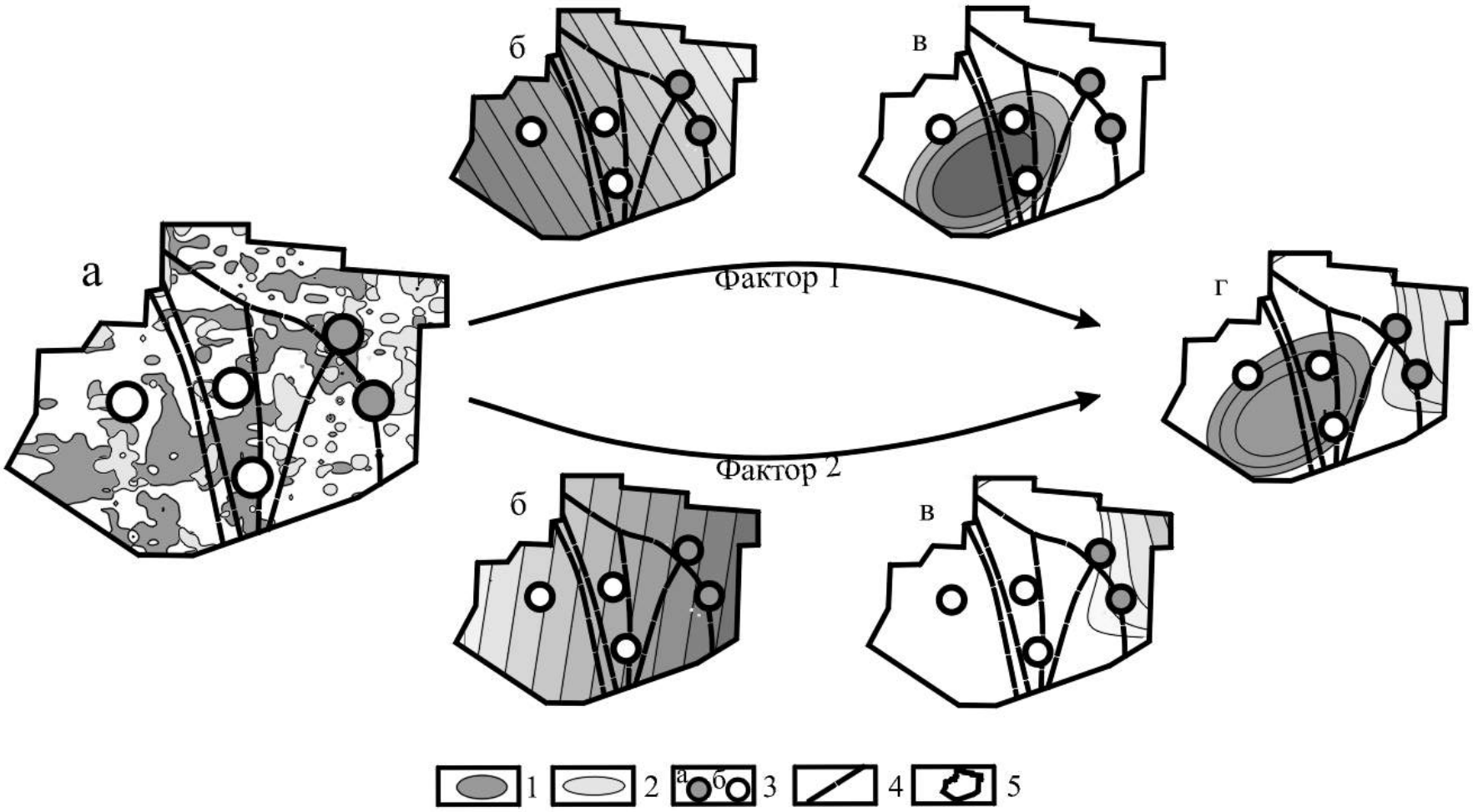


Рис. 29. Использование тренд-поверхностей для выявления зональности рудного поля

1,2 - участки повышенных значений факторов №1 (Cu,Co,Ni) и №2(Au,As,Ag); 3 - промышленные месторождения (а) и рудопроявления (б); 4 - разрывные нарушения; 5 - контур участка геохимической съемки;
а - исходная карта повышенных значений факторов, б - поверхности тренда 1-го порядка, в- поверхности тренда 3-го порядка, г - сводная карта зональности по двум факторам.

зоне, непромышленные рудопроявления тяготеют к ядерной зоне, а в промежуточной зоне известны лишь локальные точки минерализации (рис. 29г).

Вообще говоря, строгого статистического метода разделения тренда и остатка нет. Всякий остаток включает в себя поверхности более высоких порядков, и на каком порядке остановиться - каждый исследователь решает самостоятельно. Чаще всего ограничиваются двумя-тремя порядками.

Анализ остатков тренда всегда производят, исходя из геологических соображений. Процедура эта в большей степени интуитивная, нежели формальная. Нередко "аномальные" отклонения от поверхности тренда могут быть непосредственно использованы для поисков тел полезных ископаемых.

В каждом конкретном случае геолог самостоятельно выбирает различные эмпирические способы анализа остатков тренда, исходя из конкретной геологической ситуации.

9.3.Особенности применения тренд-анализа

Для того, чтобы продуктивно использовать данные тренд-анализа, необходимо учитывать ряд моментов, игнорирование которых может существенно исказить результаты анализа.

1)число точек наблюдения должно превосходить число коэффициентов в полиномиальном уравнении регрессии. Поскольку величина критических значений коэффициентов корреляции и вероятность ошибки второго рода резко возрастают при малом объеме наблюдений, всегда необходимо стремиться к максимальной представительности выборки, особенно при работе с поверхностями высокого порядка.

2)наклоны, существующие на краях карты, без всяких ограничений экстраполируются за ее границы. Это явление, получившее название «краевого эффекта», может приводить к появлению интенсивных аномалий вблизи границ карты, особенно при построении поверхностей высокого порядка, поэтому желательно иметь вокруг карты некую «буферную зону», с немногочисленными контрольными точками. Ширина этой зоны пропорциональна расстоянию между точками на основной карте. Кстати, явление краевого эффекта свойственно

не только тренд-анализу, но и другим методам построения изолиний с помощью аппроксимирующих поверхностей.

3)расположение точек в пределах карты также влияет на форму тренд-поверхности. Размещение наблюдений в виде узкой полосы неизбежно приводит к вытянутости в этом направлении поверхностей высокого порядка. В этом случае добавление даже нескольких точек на удалении от этой полосы значительно улучшает результат. Идеальным случаем является равномерное размещение точек по всему участку. Отметим, что распределение точек в виде отдельных групп (сгущений) не столь сильно влияет на искажение поверхностей тренда, как можно было бы ожидать. Такое размещение обычно при анализе региональных закономерностей на основе опробования конкретных рудных полей и месторождений. Однако влияние локальных отклонений в этом случае очень значительно, что необходимо учитывать при проведении анализа.

9.4. Моделирование дискретных полей

Дискретные поля используются для анализа особенностей пространственного размещения геологических объектов (месторождений). Данную проблему можно подразделять на две задачи: а) проверка гипотезы о случайном расположении объектов (общая задача), б) выделение областей относительного сгущения или разрежения объектов (локальная задача),

Для решения общей задачи вся площадь карты разбивается на квадратные ячейки одного размера. При этом часть ячеек (p) будет содержать хотя бы один объект, а другая часть (1- p) окажется пустой. Затем разбиваем площадь на новые квадраты; каждый из которых содержит N первоначальных ячеек. При случайном расположении объектов вероятность того, что наугад взятый новый квадрат окажется пустым, равна:

$$P_{N(T)} = (1 - p) . \qquad (78)$$

Если эмпирическое значение P_N окажется больше теоретического, это свидетельствует о сгущении объектов в отдельных участках (поскольку число пустых квадратов повышенное). Если $P_N \le P_{N(T)}$, значит, объекты расположены на площади случайно, незакономерно.

Для решения локальной задачи обычно пользуются специальными палетками в виде концентрических окружностей или квадратов. Центр палетки последовательно помещается в различные точки изучаемой площади и для каждой точки подсчитывается избыточная плотность расположения объектов:

$$\nu = \frac{m}{p} - n . \qquad (79)$$

Здесь m - число объектов в пределах меньшей фигуры, n - число объектов в пределах большей фигуры, p - отношение площади меньшей фигуры к большей. При случайном расположении точек v должно быть равно нулю. При сгущении точек $v>o$, при разрежении - $v<0$. Значимость отличия избыточной плотности от нуля при $v>o$ определяют путем вычисления вероятности случайного попадания *не менее*, чем m объектов из общего числа n. в область с относительными размерами p :

$$p_1 = \sum_{i=m}^{n} c_n^{\,i} p^i (1-p)^{n-i} . \qquad (80)$$

Это формула биноминального распределения (см.гл.2). При $v<0$, вероятность случайного попадания *не более*, чем m объектов при тех же условиях находится из выражения:

$$p_2 = \sum_{i=0}^{m} c_n^{\,i} p^i (1-p)^{n-i} \qquad (81)$$

Вычислив вероятность p_1 (или p_2) для полученных значений m , n и p , мы можем сравнивать ее с заданной доверительной вероятностью и оконтуривать таким образом области относительного сгущения (или разрежения) геологических объектов.

Более подробно анализ пространственной изменчивости признаков и моделирование дискретных полей рассмотрены в работах (3,6,9).

10. СЛУЧАЙНЫЕ ФУНКЦИИ

10.1. Основные характеристики случайных функций

Выше мы рассматривали примеры изменения значений

какого-либо параметра в пространстве и выяснили, что эти изменения могут быть описаны аналитически в виде функции и графически в виде профиля или поверхности.

Подобным образом можно выразить, например, распределение содержаний элемента вдоль горной выработки. Если мы опробуем эту выработку несколько раз, то графики каждый раз будут отличаться друг от друга из-за различных причин: ошибок в анализах, несовпадения точек опробования и т.д. В итоге график будет представлять собой набор кривых (рис.30).

В результате каждого испытания функция пространственного распределения принимает определенный вид, причем заранее неизвестно, какой именно. Следовательно, распределение содержаний элемента вдоль выработки представляет из себя *случайную* функцию. Результат каждого опыта называется при этом *реализацией* случайной функции, т.е. реализация представляет из себя обычную, неслучайную функцию.

Если зафиксировать значение аргумента в какой-либо точке l_1 ,то мы получим набор точек x_{1j} , который называется *сечением* случайной функции при $l = l_1$. Это сечение представляет из себя обычную случайную величину, для которой можно определить $\bar{x}$ и S^2_1.

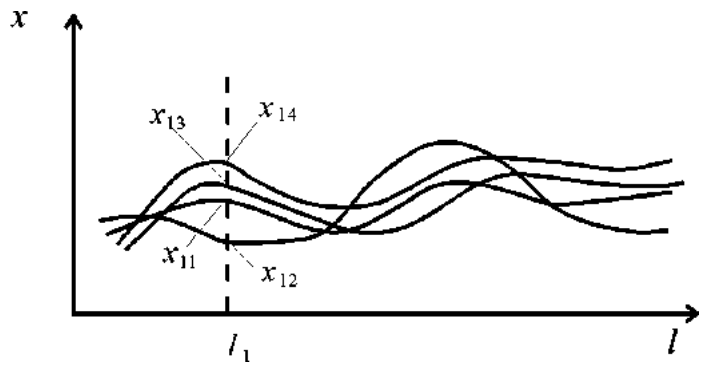


Рис. 30. Распределение содержания элемента вдоль горной выработки по ряду реализаций

Если мы проведем n сечений случайной функции и для каждого определим $\bar{x}_i$, то, соединив затем все $\bar{x}_i$ между собой, получим график, усредненной функции, которая носит название *математического ожидания* случайной функции.

Дисперсией случайной функции называется обычная функция, которая в каждой точке l равна дисперсии соответствующего сечения случайной функции. Дисперсия характеризует ширину полосы разброса отдельных реализаций случайной функции относительно ее математического ожидания.

Поскольку математическое ожидание и дисперсия не дают информации о внутренних связях между отдельными реализациями случайной функции, для описания свойств случайной функции используется еще одна характеристика - *корреляционная (автокорреляционная)* функция. Она представляет из себя неслучайную функцию, которая при каждой паре значений аргумента равна ковариации соответствующих сечений случайной величины:

$$K_x(i, j) = \frac{\sum(x_{in} - \bar{x}_i)(x_{jn} - \bar{x}_j)}{n-1}, \qquad (82)$$

где i, j - сечения случайной функции, n - количество реализаций.

Корреляционную функцию обычно нормируют, деля на произведение S_iSj. В итоге получают *коэффициент корреляции* (или *автокорреляции*) между сечениями случайной функции:

$$\rho(i, j) = \frac{Kx(i, j)}{SiSj}, \qquad (83)$$

Очевидно, что при $i = j$, K_x представляет из себя дисперсию, а $\rho(i, j) = 1$.

Чем больше количество реализаций случайной функции, тем с большей точностью могут быть вычислены ее характеристики. Между тем, на практике мы чаще всего имеем дело лишь с одной реализацией. Возникает вопрос, можно ли по одной реализации судить о характеристиках случайной функции? Оказывается, можно, если эта функция обладает свойствами *стационарности* и *эргодичности*. Стационарной называется такая случайная функция, для которой перечисленные выше характеристики не изменяются при любом сдвиге аргументов по оси l. Эргодичной является такая стационарная функция, которая обладает одинаковыми значениями Mx, Dx , Kx для всех реализаций.

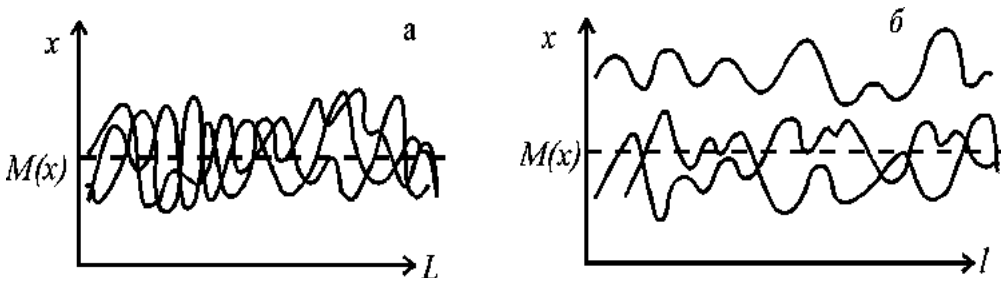


Рис. 31. Эргодичная (а) и неэргодичная (б) стационарные функции

Для стационарной функции значение Kx, (l_1 , l_2) зависят не от самих значений l_1 и l_2, а лишь от расстояния между ними:

$$Kx(\, l_1\, l+r) = Kx(\, r)$$

Следовательно, корреляционная функция в этом случае представляет из себя функцию не двух, а одного аргумента. Это существенно упрощает операции над случайной функцией.

Коэффициент автокорреляция для стационарной случайной функция можно вычислять по формуле:

$$\rho_x(r) = \frac{Kx(r)}{Dx} = \frac{\left[\sum_{i=1}^{n-r}(x_i - \bar{x})(x_{i+r} - \bar{x})\right](n-1)}{(n-1-r)\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (84)$$

Графическим выражением функции ρ_x (r) является кривая экспоненциального типа (рис. 32).

Расстояние $к$ называется *пределом корреляции*. Оно показывает максимальное расстояние, на котором еще проявляется корреляция между замеренными значениями параметра. При дальнейшем увеличении расстояния наблюдения становятся независимыми.

Функция автокорреляция используется при выборе расстояния между разведочными выработками. Если значение какого-либо параметра оруденения необходимо геометризовать (например, провести изолинии мощностей рудного тела), то расстояние между подсечениями рудного тела не должны превышать вычисленного предела корреляции (радиуса автокорреляции). Корреляционная функция используется также: 1) для проверки гипотезы о наличии тренда (оценивается значимость отличия от нуля первых *2-3* значений коэффициента

автокорреляции), 2) для классификации геологических объектов по характеру пространственной изменчивости, 3) для разделения общей изменчивости на случайную и закономерную составляющие:

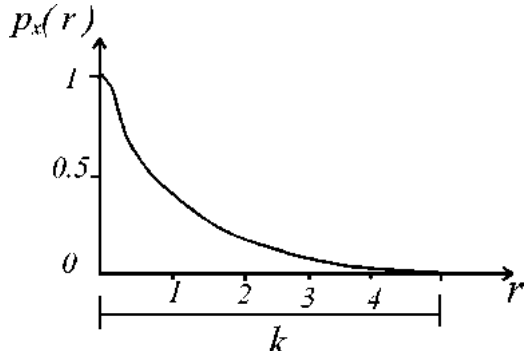


Рис. 32. График нормированной автокорреляционной функции

$$\delta^2_{случ.} = \delta^2 (1 - \rho^2_x(r)), \qquad (85)$$

$$\delta^2_{законом.} = \delta^2 - \delta^2_{случ.} . \qquad (86)$$

Соотношение случайной и закономерной составляющих дисперсии можно использовать для оценки степени разведанности месторождения: чем выше доля закономерной составляющей, тем выше степень разведанности. Значения $\delta_{случ.}$ используются также для вычисления достоверности оценки средних значений параметров при подсчете запасов. При этом в формулу интервальной оценки среднего вместо общей дисперсии подставляют случайную составляющую, подсчитанную по формуле (85):

$$\lambda = \bar{x} \pm \frac{t \cdot \delta_{случ.}}{\sqrt{n}} .$$

Эта процедура носит название "введение поправки за связь". Очевидно, что, чем меньше расстояние между выработками, тем меньше значение $\delta^2_{случ.}$, следовательно, выше точность оценки среднего. При расстоянии между выработками большем радиуса автокорреляции, точность оценки зависит только от количества подсечений (наблюдений).

Необходимо иметь в виду, что реальные геологические поля не являются строго стационарными, поэтому эмпирические графики корреляционных функций характеризуются тем, что не

приближаются асимптотически к нулю, а совершают периодические колебания около этого значения. В ряде случаев эти колебания свидетельствуют о наличии периодической составляющей пространственной изменчивости, анализ которой осуществляется с помощью модели полигармонической случайной функции.

10.2. Полигармонический анализ случайных функций

Геологические объекты очень часто обладают периодическим характером изменчивости свойств. Периодичность нередко отмечается в размещении тектонических нарушений, в пространственном распределении содержаний различных элементов, физических свойств пород, в составе осадочных и метаморфических толщ и т.д. Естественно, периодические колебания при этом осложняются и затушевываются случайными, нерегулярными, флуктуациями.

Для выделения и описания периодической закономерной составляющей изменчивости обычно применяется модель полигармонической случайной функции. Математическое ожидание этой функции выражается тригонометрическим полиномом вида:

$$Mx(l) = A_0 + \sum_{k=1}^{v} Ak \cdot \cos(\omega_k l + \varphi_k), \qquad (87)$$

где A_0 - константа, V - количество гармоник, $A_к$, $\omega_к$, $\varphi_к$ - соответственно, амплитуда, частота и фаза каждой гармоники.

С помощью этой модели любой ряд значений признака, при равном расстоянии между точками (r), можно описать функцией:

$$x(r) = Mx(r) + h_x(r), \qquad (88)$$

где $h_x(r)$ - случайная составляющая, осложняющая периодические колебания.

Модель полигармонической случайной функции наиболее универсальна из всех рассмотренных нами ранее моделей. При отсутствии периодической составляющей она превращается в модель стационарной случайной функции, а при отсутствии автокорреляции - в обычную статистическую модель.

Таким образом, чтобы выявить закономерную

составляющую периодического явления, необходимо правильно подобрать амплитуды, частоты и фазы соответствующих гармоник, отражающих периодичность разных порядков. Подобная операция широко применяется в радиотехнике (частотная модуляция) и других областях техники. В настоящее время разработано множество методов и специальных приборов для выявления скрытых периодичностей (14).

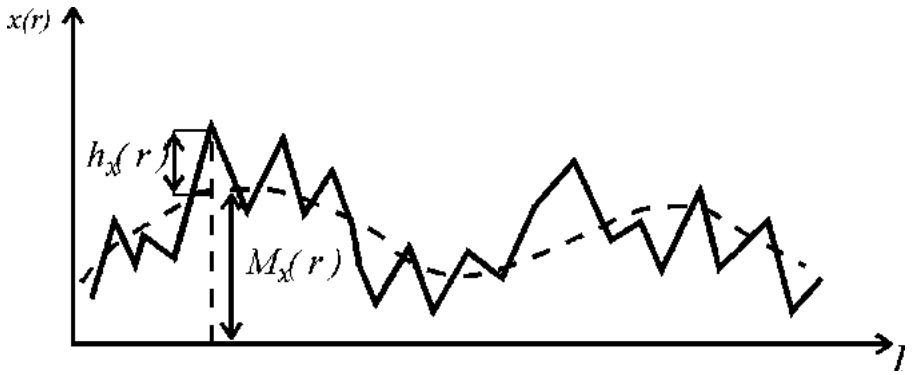


Рис. 33. Сглаживание эмпирических наблюдений тригонометрической функцией
(показана только одна гармоника)

В геологии обычно используют метод, основанный на оценке спектральной плотности дисперсии $S_x(\omega)$, получаемой в результате разложения в ряд Фурье корреляционной функции:

$$S_x(\omega) = \frac{2}{\pi}\int_0^{\infty} Kx(r)\cdot\cos\omega r\cdot dr \qquad (89)$$

Для дискретного ряда наблюдений спектральная плотность заменяется линейчатым спектром амплитуд. Каждое значение линейчатого спектра вычисляется по формуле:

$$d_к = \frac{1}{2}A_k^{\,2}, \qquad (90)$$

где $A_к$ - амплитуда *к*-й гармоники. Сумма амплитуд всех *к* гармоник равна при этом 1. Рассмотрим для наглядности графическое изображение названных характеристик (рис.34).

На левом графике мы видим, как общая дисперсия признака распределяется по частотам колебаний (общая площадь равна полной дисперсии). Правый график отражает распределение общей дисперсии между отдельными гармониками. Здесь ω - число периодов, приходящихся на длину профиля.

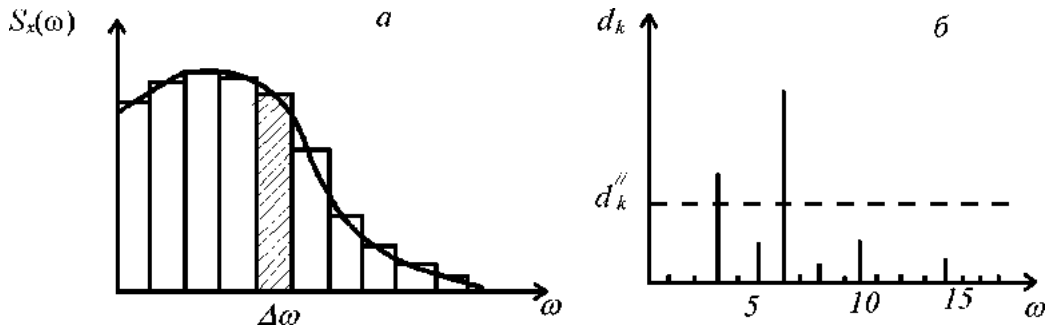


Рис. 34. Графики спектральной плотности дисперсии (а) и частотного спектра амплитуд (б)

Значение $d''_к$, определяющее уровень случайных флуктуаций, определяется по формуле (для 5% уровня значимости):

$$d''_к = \frac{1}{N} + 2\sqrt{\frac{N-1}{N^2(N-1)}}, \qquad (91)$$

где N - число значений спектра амплитуд. Отсюда можно с наперед заданной доверительной вероятностью проверить гипотезу о принадлежности тех или иных пиков спектра к случайным колебаниям.

Доля закономерной составляющей в общей изменчивости вычисляется по формуле:

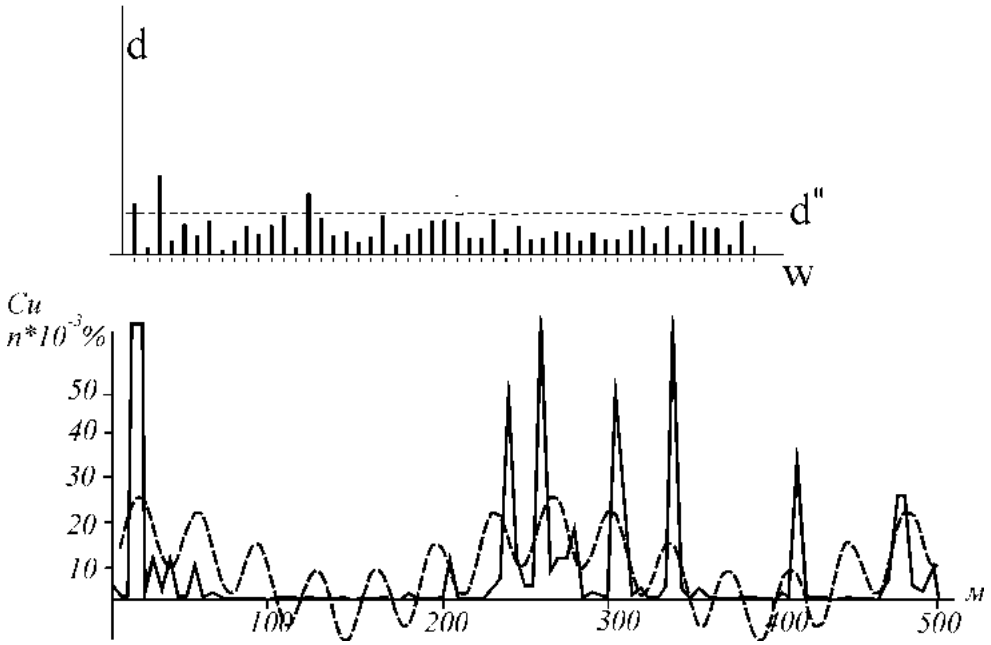
$$S' = \sum d_k(аном.) - \frac{1-\sum d_k(аном.)}{N-m}, \qquad (92)$$

где $\sum d_к$(аном.) - сумма значений спектра, N - число значений спектра, m - число аномальных значений спектра.

Вычисленные значения спектра амплитуд подставляют в тригонометрический полином и получают соответствующее уравнение регрессии, для которого аргументом является координата пространства (или время).

На рис. 35 показан пример применения полигармонической модели при анализе размещения концентраций меди вдоль оси субмеридионального штрека на Синюхинской золоторудном месторождении в Горном Алтае. Здесь выявлены два отчетливых пика, превышающих уровень случайных флуктуаций, соответствующие 270 м и 35 м. Точно такие же закономерности

установлены в размещении других элементов-примесей, а также в изменении интенсивности трещиноватости. Следовательно, на данном месторождении можно говорить о “шаге оруденения”



двух иерархических уровней в субмеридиональном направлении.

Рис. 35. Моделирование пространственных закономерностей в распределении Cu на Синюхинском золоторудном месторождении с помощью преобразований Фурье.

Вверху график частотного спектра амплитуд, внизу график исходных и предсказанных по уравнению Фурье содержаний Cu.

Рассмотренная модель является несколько упрощенной, поскольку так называемая случайная составляющая в общем случае может быть и закономерной, но непериодической. Для того, чтобы оценить функцию непериодической составляющей, необходимо из эмпирической корреляционной функции последовательно вычесть выявленные гармонические составляющие. Если полученная в итоге корреляционная функция колеблется около 0, то непериодической закономерной составляющей нет, есть только случайная. Если же и после вычитания корреляционная функция имеет вид, характерный для стационарной случайной функции (см. рис. 32), значит, есть и непериодическая закономерная составляющая.

Таким образом, гармонический анализ позволяет разделить общую изменчивость любого свойства на три составляющих: а) координированную, которая отражает закономерности,

присущие участку в целом, и описывается тригонометрическим полиномом; б) коррелированную, описывающую закономерные изменения в локальной области с помощью корреляционной функции; в) случайную составляющую, описывающую незакономерные, случайные колебания свойства («белый шум»).

Координированная составляющая используется для выявления неоднородности в строении объектов и предсказаний параметров в любой точке участка. Знание длин периодов и амплитуд закономерных колебаний в значениях важнейших геологоразведочных параметров позволяет судить не только об оптимальных формах, но и об оптимальных размерах разведочной сети. В частности, по эмпирическим корреляционным функциям можно определить «запрещенные»размеры шага наблюдений, совпадающие с длинами отчетливо выраженных периодов. Гармонический анализ позволяет судить о размерах геологических неоднородностей различного масштаба и представляется наиболее подходящим методом для выделения условно однородных участков и подсчетных блоков.

Коррелированная составляющая используется для распространения данных отдельных наблюдений на соответствующий объем недр (в пределах радиуса автокорреляции).

Случайная составляющая изменчивости отражает неполноту знаний о полезном ископаемом. Ее выделение необходимо для вычисления средних значений параметров и вероятных погрешностей их подсчетов в зависимости от принятой геометрии проб.

Более подробно использование гармонического анализа при геологоразведочных работах рассмотрено в работе (6). Математический аппарат для выявления скрытых периодичностей детально анализируется в работе (14).

10.3. Полувариограммы и крайгинг.

Аппарат математического моделирования случайных функций широко используется в специальной области прикладной статистики, получившей название геостатистики (Матерон, 1968). Ключевым понятием геостатистики является понятие регионализированной переменной, которая имеет

свойства, промежуточные между свойствами полностью случайных величин и полностью детерминированных переменных. В отличие от случайных, регионализированные переменные непрерывны от точки к точке, но изменение их настолько сложны, что не могут быть описаны какой-либо детерминированной функцией. Сведения о регионализированной переменной мы получаем только в точках опробования. Для того чтобы оценивать значение переменной в любой точке и характер ее изменения в любом из трех измерений, в геостатистике разработан оригинальный математический аппарат, одной из важнейших характеристик в котором является понятие полудисперсии.

Полудисперсия используется для выражения скорости изменения регионализированных переменных вдоль заданного направления. Если пробы в этом направлении берутся через одинаковые интервалы, то выражение для полудисперсии выглядит следующим образом:

$$\gamma_h = \frac{1}{2n}\sum_{i=1}^{n-h}(x_i - x_{i+h})^2$$

Здесь *Xi* значение переменной в точке *i*; *Xi+h* – значение переменной, взятое через *h* интервалов; *n* – число точек.

Если вычислить значения *γ* для различных *h*, можно получить график, называемый *полувариограммой*. Если переменная *X* стандатизирована, полувариограмма представляет собой зеркальное отражение автокорреляционной функции (рис.36).

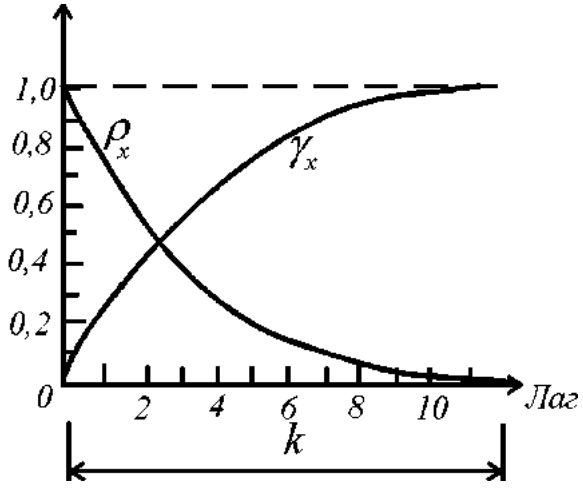


Рис. 36. Соотношение авто-корреляционной функции ρ(x) и полувариограммы (γ_x).

k – ранг регионализированной переменной

Как видим, по мере увеличения лага полудисперсия возрастает и на определенном расстоянии, называемом р а н г о м регионализированной переменной, становится постоянной, равной по величине дисперсии переменной. В пределах окрестности точки, определяемой величиной ранга, любая другая точка связана с центральной и, следовательно, может быть использована для оценки значений переменной в центральной точке. Для вычисления этих оценок в геостатистике создан специальный математический аппарат, называемый в честь южноафриканского геолога Д.Г.Криге, первым применившего статистические методы при подсчете запасов, *к р и г и н г о м* (в старой транскрипции крайгингом).

В отличие от рассмотренных выше методов сглаживания, кригинг позволяет рассчитать значение признака не только в точках наблюдения, но и в любой другой точке, где еще сохраняется пространственная непрерывность моделируемого геологического поля. Степень этой непрерывности выражается вариограммой, следовательно, может быть оценена в любой другой точке. Кроме того, в сравнении с другими методами оценивания, оценки, полученные процедурой кригинга, имеют наименьшую возможную ошибку и обеспечивают определение величины этой ошибки.

Процедуру точечного кригинга можно рассмотреть на примере простейшего случая вычисления значения параметра Y в точке p по трем известным наблюдениям Y_1, Y_2, и Y_3 в других точках. Каждое из этих наблюдений имеет в точке p свой вес, значение которого можно вычислить, решив систему уравнений:

$$W_1\,\gamma(h_{11}) + W_2\gamma(h_{12}) + W_3\gamma(h_{13}) = \gamma(h_{1p}),$$
$$W_1\gamma(h_{12}) + W_2\gamma(h_{22}) + W_3\gamma(h_{23}) = \gamma(h_{2p}),$$
$$W_1\gamma(h_{13}) + W_2\gamma(h_{23}) + W_3\gamma(h_{33}) = \gamma(h_{3p}).$$

Здесь W_i – вес наблюдений i в точке p; $\gamma(h_{ij})$ – значение полувариограммы на расстоянии h между точками i и j. Эти значения берутся непосредственно с графика полувариограммы или из математического выражения, описывающего ее вид. После вычисления весов, значение параметра в точке p можно определить из выражения:

$$\overline{Y_p} = W_1Y_1 + W_2Y_2 + W_3Y_3 .$$

Важно подчеркнуть, что точечный кригинг перестает работать при наличии устойчивого тренда. В этом случае линейная оценка не будет несмещенной, а будет сдвигаться вверх или вниз, в зависимости от размещения контрольных точек.

Универсальный кригинг рассматривает такую нестационарную регионализированную переменную как состоящую из двух компонент – *дрифта* (закономерная часть, которая может быть описана поверхностью тренда) и *остатка* (разность между наблюдаемыми значениями и дрифтом). Очевидно, что если из наблюденных значений вычесть дрифт, то остатки будут стационарными и к ним можно применить процедуру кригинга. Таким образом, универсальный кригинг состоит из трех операций: а) оценка и устранение дрифта, б) кригинг стационарных остатков, в) комбинирование полученных кригингом значений с дрифтом с целью получения истинной поверхности.

Поскольку дрифт вычисляется для окрестностей каждой точки, алгоритм универсального кригинга построен таким образом, что вычисление дрифта и кригирование осуществляются решением одной системы уравнений и мы сразу получаем веса кригинга, включаюшие эффект от заданного дрифта в локальной окрестности точки. Следует отметить, что дрифт – это произвольная конструкция, которая вводится лишь для удовлетворения требования стационарности переменной. Это требование может быть выполнено при самых различных комбинациях модели дрифта, размера окрестностей и оценки полувариограммы, поэтому, в отличие от тренд–поверхности, дрифт на имеет какого-либо содержательного геологического смысла.

В геологии универсальный кригинг широко используется при построении изолиний значений параметров, при подсчете запасов (оценка средних содержаний по блокам) и в ряде других операций.

ЛИТЕРАТУРА

1.Большев Л.Н., Смирнов Н.В. Таблицы математической статистики. - М.: ВЦ АН СССР, 1968. 474 с.
2.Бондаренко В.Н. Статистические решения некоторых задач геологии. - М.: Недра, 1970. 246 с.
3.Боровко Н.Н. Статистический анализ пространственных геологических закономерностей.- М.: Недра, 1971. 173 с.
4.Вентцель Е.С. Теория вероятностей. - М.: Наука, 1969. 576 с.
5.Девис Дж. С.Статистический анализ данных в геологии. Кн. 1, 2.- М.: Недра. 1990. 319 с., 428 с.
6.Каждан А. Б. Методологические основы разведки полезных ископаемых. - М.: Недра, 1974. 271 с.
7.Каждан А.Б., Гуськов 0.И. Математические методы в геологии.- М.: Недра, 1990. 251 с.
8.Каллистов П.Л. Изменчивость оруденения и плотность наблюдений при разведке и опробовании. Сов. геология, 1956, сб.53. С. 119-151.
9.Крамбейн У., Грейбилл Ф. Статистические модели в геологии.- М.: Мир, 1969. 398 с.
10.Математическая статистика/Под ред. А.М. Длина.- М.: Высшая школа, 1975. 398с.
11.Матерон Ж. Основы прикладной геостатистики. - М.: Мир, 1968. 408 с.
12.Миллер Р.Л., Кан Дж. С. Статистический анализ в геологических науках. - М.: Мир, 1965. 482 с.
13.Родионов Д.А. Статистические решения в геологии. - М.: Недра, 1981.- 231 с.
14.Серебренников М.Г., Первозванский А.А. Выявление скрытых периодичностей. - М.: Наука, 1965. 244 с.
15.Смирнов Б.И. Корреляционные методы при парагенетическом анализе. - М.: Недра, 1981. 197 с.
16.Соловов А.П., Матвеев А. А. Геохимические методы поисков рудных месторождений. - М.: МГУ, 1985. 228 с.
17.Справочник по математическим методам в геологии. - М.: Недра, 1987.
18.Шестаков Ю.Г. Математические методы в геологии.- Красноярск: Изд-во КГУ, 1988. 208 с.

ПРИЛОЖЕНИЯ

Таблица 1

Стандартная нормальная функция распределения

$$F(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-u^2/2}du\,,\ [\,F(-x)=1-F(x)]$$

x	F(x)	x	F(x)	x	F(x)	x	F(x)	x	F(x)
1	2	3	4	5	6	7	8	9	10
0.00	0.5000	0.21	0.5832	0.42	06628	0,63	0,7352	0,84	0,7995
0.01	0.5040	0.22	0.5871	0.43	0.6664	0,64	0,7389	0,85	0,8023
0.02	0.5080	0.23	0.5910	0.44	0.6700	0,65	0,7422	0,86	0,8051
0.03	0.5120	0.24	0.5948	0.45	0.6736	0,66	0,7454	0,87	0,8078
0.04	0.5160	0.25	0.5987	0.46	0.6722	0,67	0,7486	0,88	0,8106
0.05	0.5199	0.26	0.6026	0.47	0.6808	0,68	0,7517	0,89	0,8133
0.06	0.5239	0.27	0.6064	0.48	0.6844	0,69	0,7549	0,90	0,8159
0.07	0.5279	0.28	0.6103	0.49	0.6879	0,70	0,7580	0,91	0,8186
0.08	0.5319	0.29	0.6141	0.50	0.6915	0,71	0,7611	0,92	0,8212
0.09	0.5359	0.30	0.6179	0,51	0,6950	0,72	0,7642	0,93	0,8238
0.10	0.5398	0.31	0.6217	0,52	0,6985	0,73	0,7673	0,94	0,8264
0.11	0.5438	0.32	0.6255	0,53	0,7019	0,74	0,7703	0,95	0,8289
0.12	0.5478	0.33	0.6293	0,54	0,7054	0,75	0,7734	0,96	0,8315
0.13	0.5517	0.34	0.6331	0,55	0,7088	0,76	0,7764	0,97	0,8340
0.14	0.5557	0.35	0.6368	0,56	0,7123	0,77	0,7794	0,98	0,8365
0.15	0.5596	0.36	0.6406	0,57	0,7157	0,78	0,7823	0,99	0,8389
0.16	0.5636	0.37	0.6443	0,58	0,7190	0,79	0,7853	1,00	0,8413
0.17	0.5675	0.38	0.6480	0,59	0,7224	0,80	0,7881	1,01	0,8438
0.18	0.5714	0.39	0.6517	0,60	0,7257	0,81	0,7910	1,02	0,8461
0.19	0.5753	0.40	0.6554	0,61	0,7291	0,82	0,7939	1,03	0,8485
0.20	0.5793	0.41	0.6591	0,62	0,7324	0,83	0,7967	1,04	0,8508

Продолжение таблицы 1

x	F(x)	x	F(x)	x	F(x)	x	F(x)	x	F(x)
1	2	3	4	5	6	7	8	9	10
1,05	0,8531	1,27	0,8980	1,49	0,9319	1,71	0,9564	1,93	0,9732
1,06	0,8554	1,28	0,8997	1,50	0,9332	1,72	0,9572	1,94	0,9738
1,07	0,8577	1,29	0,9015	1,51	0,9345	1,73	0,9582	1,95	0,9744
1,08	0,8599	1,30	0,9032	1,52	0,9367	1,74	0,9591	1,96	0,9750
1,09	0,8621	1,31	0,9049	1,53	0,9370	1,75	0,9599	1,97	0,9756
1,10	0,8643	1,32	0,9066	1,54	0,9382	1,76	0,9608	1,98	0,9761
1,11	0,8665	1,33	0,9082	1,55	0,9394	1,77	0,9616	1,99	0,9767
1,12	0,8686	1,34	0,9099	1,56	0,9406	1,78	0,9625	2,00	0,9772
1,13	0,8707	1,35	0,9115	1,57	0,9418	1,79	0,9633	2,02	0,9783
1,14	0,8729	1,36	0,9131	1,58	0,9429	1,80	0,9641	2,04	0,9793
1,15	0,8749	1,37	0,9147	1,59	0,9441	1,81	0,9649	2,06	0,9803
1,16	0,8770	1,38	0,9162	1,60	0,9452	1,82	0,9656	2,08	0,9812
1,17	0,8790	1,39	0,9177	1,61	0,9463	1,83	0,9664	2,10	0,9821
1,18	0,8810	1,40	0,9192	1,62	0,9474	1,84	0,9671	2,12	0,9830
1,19	0,8830	1,41	0,9207	1,63	0,9484	1,85	0,9678	2,14	0,9838
1,20	0,8849	1,42	0,9222	1,64	0,9495	1,86	0,9686	2,16	0,9846
1,21	0,8869	1,43	0,9236	1,65	0,9505	1,87	0,9693	2,18	0,9854
1,22	0,8888	1,44	0,9251	1,66	0,9515	1,88	0,9699	2,20	0,9861
1,23	0,8907	1,45	0,9265	1,67	0,9525	1,89	0,9706	2,22	0,9868
1,24	0,8925	1,46	0,9279	1,68	0,9535	1,90	0,9713	2,24	0,9875
1,25	0,8944	1,47	0,9292	1,69	0,9545	1,91	0,9719	2,26	0,9881
1,26	0,8962	1,48	0,9306	1,70	0,9554	1,92	0,9726	2,28	0,9887

Окончание таблицы 1

x	F(x)	x	F(x)	x	F(x)	x	F(x)	x	F(x)
1	2	3	4	5	6	7	8	9	10
2,30	0,9893	2,48	0,9934	2,66	0,9961	2,84	0,9977	3,20	0,9993
2,32	0,9898	2,50	0,9938	2,68	0,9963	2,86	0,9979	3,40	0,9996
2,34	0,9904	2,52	0,9941	2,70	0,9965	2,88	0,9980	3,60	0,9998
2,36	0,9909	2,54	0,9945	2,72	0,9967	2,90	0,9981	3,80	0,9999
2,38	0,9913	2,56	0,9948	2,74	0,9969	2,92	0,9982		
2,40	0,9918	2,58	0,9951	2,76	0,9971	2,94	0,9984		
2,42	0,9922	2,60	0,9953	2,78	0,9973	2,96	0,9985		
2,44	0,9927	2,62	0,9956	2,80	0,9974	2,98	0,9986		
2,46	0,9931	2,64	0,9959	2,82	0,9976	3,00	0,9987		

Таблица 2

Таблица распределения Стьюдента ($k = n_1 + n_2 - 2$)

N	α=0.1	α=0.05	α=0.02	α=0.01	α=0.001	N	α=0.1	α=0.05	α=0.02	α=0.01	α=0.001
1	6.31	12.71	31.82	63.66	636.62	18	1.73	2.10	2.55	2.88	3.92
2	2.92	4.30	6.97	9.93	31.60	19	1.73	2.09	2.54	2.86	3.88
3	2.35	3.18	4.54	5.84	12.94	20	1.73	2.09	2.53	2.85	3.85
4	2.13	2.78	3.75	4.60	8.61	21	1.72	2.08	2.52	2.83	3.82
5	2.02	2.57	3.37	4.03	6.86	22	1.72	2.07	2.51	2.82	3.79
6	1.94	2.45	3.14	3.71	5.96	23	1.71	2.07	2.50	2.81	3.77
7	1.90	2.37	3.00	3.50	5.41	24	1.71	2.06	2.49	2.80	3.75
8	1.86	2.31	2.90	3.36	5.04	25	1.71	2.06	2.49	2.79	3.73
9	1.83	2.26	2.82	3.25	4.78	26	1.71	2.06	2.48	2.78	3.71
10	1.81	2.23	2.76	3.17	4.59	27	1.70	2.05	2.47	2.77	3.69
11	1.80	2.20	2.72	3.11	4.44	28	1.70	2.05	2.47	2.76	3.67
12	1.78	2.18	2.68	3.06	4.32	29	1.70	2.05	2.46	2.76	3.66
13	1.77	2.16	2.65	3.01	4.22	30	1.70	2.04	2.46	2.75	3.65
14	1.76	2.15	2.62	2.98	4.14	40	1.68	2.02	2.42	2.70	3.55
15	1.75	2.13	2.60	2.95	4.07	60	1.67	2.00	2.39	2.66	3.46
16	1.75	2.12	2.58	2.92	4.02	120	1.66	1.98	2.36	2.62	3.37
17	1.74	2.11	2.57	2.90	3.97	∞	1.65	1.96	2.33	2.58	3.29

Таблица 3

Таблица распределения Фишера
значения $f_{0,95}$ (верхние 5%-ные точки)

k	1									
	1	2	3	4	5	6	8	12	24	∞
1	2	3	4	5	6	7	8	9	10	11
1	161,40	199,5	215,70	224,60	230,20	234,00	238,90	243,90	249,00	254,30
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07

Продолжение таблицы 3

1	2	3	4	5	6	7	8	9	10	11
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
19	4,33	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,05	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,23	2,03	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,38	2,20	2,00	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,18	1,98	1,73
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,16	1,96	1,71
26	4,22	3,37	2,98	2,74	2,59	2,47	2,32	2,15	1,95	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,30	2,13	1,93	1,67
28	4,20	3,34	2,95	2,71	2,56	2,44	2,29	2,12	1,91	1,65
29	4,18	3,33	2,93	2,70	2,54	2,43	2,28	2,10	1,90	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,79	1,51
60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	1,92	1,70	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,02	183	1,61	1,25
∞	3,84	2,99	2,60	2,37	2,21	2,09	1,94	175	1,52	1,00

Продолжение таблицы 3

значения $f_{0,99}$ (верхние 1%-ные точки)

k	1									
	1	2	3	4	5	6	8	12	24	∞
1	2	3	4	5	6	7	8	9	10	11
1	4052	4999	5403	5625	5764	5859	5981	6106	6234	6366
2	98,49	99,00	99,17	99,25	99,30	99,33	99,36	99,42	99,46	99,50
3	34,12	30,81	29,46	28,71	28,24	27,91	27,49	27,05	26,60	26,12
4	21,20	18,00	16,69	15,98	15,52	15,21	14,80	14,37	13,93	13,,46
5	16326	13,37	12,06	11,39	10,97	10,67	10,27	9,89	9,47	9,02
6	13,74	10,92	9,78	9,15	8,75	8,47	8,10	7,72	7,31	6,88
7	12,25	9,55	8,45	7,85	7,46	7,19	6,84	6,47	6,07	5,65
8	11,26	8,65	7,59	7,01	6,63	6,37	6,03	5,67	5,28	4,86
9	10,55	8,02	6,99	6,42	6,06	5,80	5,47	5,11	4,73	4,31
10	10,04	7,65	6,55	5,99	5,64	5,39	5,06	4,71	4,33	3,91
11	9,65	7,20	6,22	5,67	5,32	5,07	4,74	4,40	4,02	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,50	4,16	3,78	3,36
13	9,07	9,70	5,74	5,20	4,86	4,62	4,30	3,96	3,59	3,16
14	8,86	6,51	5,56	5,03	4,69	4,46	4,14	3,80	3,43	3,00
15	8,68	6,36	5,42	4,89	4,56	4,32	4,00	3,67	3,29	2,87

Окончание таблицы 3

1	2	3	4	5	6	7	8	9	10	11
16	8,53	6,23	5,29	4,77	4,44	4,20	3,89	3,55	3,18	2,75
17	8,40	6,11	5,18	4,67	4,34	4,10	3,79	3,45	3,08	2,65
18	8,28	6,01	5,09	4,58	4,25	4,01	3,71	3,37	3,00	2,57
19	8,18	5,93	5,01	4,50	4,17	3,94	3,63	3,30	2,92	2,49
20	8,10	5,85	4,94	4,43	4,10	3,87	3,56	3,23	2,86	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,51	3,17	2,80	2,36
22	7,94	5,72	4,82	4,31	3,99	3,76	3,45	3,12	2,75	2,31
23	7,88	5,66	4,76	4,26	3,94	3,71	3,41	3,07	2,70	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,36	3,03	2,66	2,21
25	7,77	5,57	4,68	4,18	3,86	3,63	3,32	2,99	2,62	2,17
26	7,72	5,53	4,64	4,14	3,82	3,59	3,29	2,96	2,58	2,13
27	7,68	5,49	4,60	4,11	3,78	3,56	3,26	2,93	2,55	2,10
28	7,64	5,45	4,57	4,07	3,75	3,53	3,23	2,90	2,52	2,06
29	7,60	5,42	4,54	4,04	3,73	3,50	3,20	2,87	2,49	2,03
30	7,56	5,39	4,51	4,02	3,70	3,47	3,17	2,84	2,47	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	2,99	2,66	2,29	1,80
60	7,18	4,98	4,13	3,65	3,34	3,12	2,82	2,50	2,12	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,66	2,34	1,95	1,38
∞	6,64	4,60	3,78	3,32	3,02	2,80	2,51	2,18	1,79	1,00

Таблица 4

Таблица χ^2 – распределения (к = п – 1)

к \ α	0,01	0,02	0,05	0,10	0,20	0,30	0,50	0,70	0,80	0,90	0,95	0,98	0,99
1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0,00	0,00	0,00	0,01	0,06	0,14	0,45	1,64	1,64	2,71	3,84	5,41	6,64
2	0,02	0,04	0,10	0,21	0,41	0,71	1,38	2,41	3,22	4,60	5,99	7,82	9,21
3	0,11	0,18	0,35	0,58	1,00	1,42	2,37	3,66	4,64	6,25	7,82	9,84	11,34
4	0,29	0,42	0,71	1,06	1,64	2,20	3,36	4,88	5,99	7,78	9,49	11,67	13,28
5	0,55	0,75	1,14	1,61	2,34	3,00	4,35	6,06	7,29	9,24	11,07	13,39	15,09
6	0,87	1,13	1,63	2,20	3,07	3,83	5,35	7,23	8,56	10,64	12,59	15,03	16,81
7	1,23	1,56	2,17	2,83	3,82	4,67	6,35	8,38	9,80	12,02	14,07	16,62	18,48
8	1,64	2,03	2,73	3,49	4,59	5,53	7,34	9,52	11,03	13,36	15,51	18,17	20,1
9	2,09	2,53	3,32	4,17	5,38	6,39	8,34	10,66	12,24	14,68	16,92	19,68	21,7
10	2,56	3,06	3,94	4,86	6,18	7,27	9,34	11,78	13,44	15,99	18,31	21,2	23,2
11	3,05	3,61	4,58	5,58	6,99	8,15	10,34	12,90	14,63	17,28	19,68	22,6	24,7
12	3,57	4,18	5,23	6,30	7,81	9,03	11,34	14,01	15,81	18,55	21,0	24,1	26,2
13	4,11	4,76	5,89	7,04	8,63	9,93	12,34	15,12	16,98	19,81	22,4	25,5	27,7
14	4,66	5,37	6,57	7,79	9,47	10,82	13,34	16,22	18,15	21,1	23,7	26,9	29,1
15	5,23	5,98	7,26	8,55	10,31	11,72	14,34	17,32	19,31	22,3	25,0	28,3	30,6

Продолжение таблицы 4

1	2	3	4	5	6	7	8	9	10	11	12	13	14
16	5,81	6,61	7,96	9,31	11,15	12,62	15,34	18,42	20,5	23,5	26,3	29,6	32,0
17	6,41	7,26	8,67	10,08	12,00	13,53	16,34	19,51	21,6	24,8	27,6	31,0	33,4
18	7,02	7,91	9,39	10,86	12,86	14,44	17,34	20,6	22,8	26,0	28,9	32,3	34,8
19	7,63	8,57	10,12	11,65	13,72	15,35	18,34	21,7	23,9	27,2	30,1	33,7	36,2
20	8,26	9,24	10,58	12,44	14,58	16,27	19,34	22,8	25,0	28,4	31,4	35,0	37,6
21	8,90	9,92	11,59	13,24	15,44	17,18	20,3	23,9	26,2	29,6	32,7	36,3	38,9
22	9,54	10,60	12,34	14,04	16,31	18,10	21,3	24,9	27,3	30,8	33,9	37,7	40,3
23	10,20	11,29	13,09	14,85	17,19	19,02	22,3	26,0	28,4	32,0	35,2	39,0	41,6
24	10,86	11,99	13,85	15,66	18,06	19,94	23,3	27,1	29,6	33,2	36,4	40,3	43,0
25	11,52	12,70	14,61	16,47	18,94	20,9	24,3	28,2	30,7	34,4	37,7	41,6	44,3
26	12,20	13,41	15,38	17,29	19,82	21,8	25,3	29,2	31,8	35,6	38,9	42,9	45,6
27	12,88	14,12	16,15	18,11	20,7	22,7	26,3	30,3	32,9	36,7	40,1	44,1	47,0
28	13,56	14,85	16,93	18,94	21,6	23,6	27,3	31,4	34,0	37,9	41,3	45,4	48,3
29	14,26	15,57	17,71	19,77	22,5	24,6	28,3	32,5	35,1	39,1	42,6	46,7	49,6
30	14,95	16,31	18,49	20,6	23,4	25,5	29,3	33,5	36,2	40,3	43,8	48,0	50,9

Таблица 5

Критические значения для критерия Манна — Уитни (T_a);
значения приведены для нижнего критического предела; соответствующий предел для верхней критической площади дается значением
$T_{1-a} = nm - T_a$.

n	α	m=2	3	4	6	8	10	11	12	13	14	15	46	17	18	19	20
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
2	.01	0	0	0	0	0	0	0	0	1	1	1	1	1	1	2	2
	.05	0	0	0	1	2	2	2	3	3	4	4	4	4	5	5	5
	.10	0	1	1	2	3	4	4	5	5	5	6	6	7	7	8	8
3	.01	0	0	0	0	1	2	2	3	3	3	4	4	5	5	5	6
	.05	0	1	1	3	4	5	6	6	7	8	8	9	10	10	11	12
	.10	1	2	2	4	6	7	8	9	10	11	11	12	13	13	15	16
4	.01	0	0	0	2	3	4	5	6	6	7	9	8	9	10	10	11
	.05	0	1	2	4	6	8	9	10	11	12	13	15	16	17	18	19
	.10	1	2	4	6	8	11	12	13	14	16	17	18	19	21	22	23
5	.01	0	0	1	3	5	7	8	9	10	11	12	13	14	15	16	17
	.05	1	2	3	6	9	12	13	14	16	17	19	20	21	23	24	26
	.10	2	3	5	8	11	14	16	18	19	21	23	24	26	28	29	31
6	.01	0	0	2	4	7	9	10	12	13	14	16	17	19	20	21	23
	.05	1	3	4	8	11	15	17	18	20	22	24	26	27	29	31	33
	.10	2	4	6	10	14	18	20	22	24	26	28	30	32	35	37	39
7	.01	0	1	2	5	8	12	13	15	17	18	20	22	24	25	27	29
	.05	1	3	5	9	14	18	20	22	25	27	29	31	34	36	38	40
	.10	2	5	7	12	17	22	24	27	29	32	34	37	39	42	44	47

Продолжение таблицы 5

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
8	.01	0	1	3	7	10	14	16	18	21	23	25	27	29	31	33	35
	.05	2	4	6	11	16	21	24	27	29	32	34	37	40	42	45	48
	.10	3	6	8	14	20	25	28	31	34	37	40	43	46	49	52	55
9	.01	0	2	4	8	12	17	19	22	24	27	29	32	34	37	39	41
	.05	2	5	7	13	19	25	28	31	34	37	40	43	46	49	52	55
	.10	3	6	10	16	23	29	32	36	39	42	46	49	53	56	59	63
10	.01	0	2	4	9	14	20	23	25	28	31	34	37	39	42	45	48
	.05	2	5	8	15	21	28	32	35	38	42	45	49	52	56	59	63
	.10	4	7	11	18	25	33	37	40	44	48	52	55	59	63	67	71
11	.01	0	2	5	10	16	23	26	29	32	35	38	42	45	48	51	54
	.05	2	6	9	17	24	32	35	39	43	47	51	55	58	62	66	70
	.10	4	8	12	20	28	37	41	45	49	53	58	62	66	70	74	79
12	.01	0	3	6	12	18	25	29	32	36	39	43	47	50	54	57	61
	.05	3	6	10	18	27	35	39	43	48	52	56	61	65	69	73	78
	.10	5	9	13	22	31	40	45	50	54	59	64	68	73	78	82	87
13	.01	1	3	6	13	21	28	32	36	40	44	48	52	56	60	64	68
	.05	3	7	11	20	29	38	43	48	52	57	62	66	71	76	81	85
	.10	5	10	14	24	34	44	49	54	59	64	69	75	80	85	90	95

Окончание таблицы 5

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
14	.01	1	3	7	14	23	31	35	39	44	48	52	57	61	66	70	74
	.05	4	8	12	22	32	42	47	52	57	62	67	72	78	83	88	93
	.10	5	11	16	26	37	48	53	59	64	70	75	81	86	92	98	103
15	.01	1	4	8	16	25	34	38	43	48	52	57	62	67	71	76	81
	.05	4	8	13	24	34	45	51	56	62	67	73	78	84	89	95	101
	.10	6	11	17	28	40	52	58	64	69	75	81	87	93	99	105	111
16	.01	1	4	8	17	27	37	42	47	52	57	62	67	72	77	83	88
	.05	4	9	15	26	37	49	55	61	66	72	78	84	90	96	102	108
	.10	6	12	18	30	43	55	62	68	75	81	87	94	100	107	113	120
17	.01	1	5	9	19	29	39	45	50	56	61	67	72	78	83	89	94
	.05	4	10	16	27	40	52	58	65	71	78	84	90	97	103	110	116
	.10	7	13	19	32	46	59	66	73	80	86	93	100	107	114	121	128
18	.01	1	5	10	20	31	42	48	54	60	66	71	77	83	89	95	101
	.05	5	10	17	29	42	56	62	69	76	83	89	96	103	110	117	124
	.10	7	14	21	35	49	63	70	78	85	92	99	107	114	121	129	136
19	.01	2	5	10	21	33	45	51	57	64	70	76	83	89	95	102	108
	.05	5	11	18	31	45	59	66	73	81	88	95	102	110	117	124	131
	.10	8	15	22	37	52	67	74	82	90	98	105	113	121	129	136	144
20	.01	2	6	11	23	35	48	54	61	68	74	81	88	94	101	108	115
	.05	5	12	19	33	48	63	70	78	85	93	101	108	116	124	131	139
	.10	8	16	23	39	55	71	79	87	95	103	111	120	128	136	144	152

СОДЕРЖАНИЕ

стр.

Введение 3
1.Краткие исторические сведения о применении математических методов в геологии..... 3
2. Понятие о геолого-математическом моделировании объектов и явлений..... 5
3. Основы теории вероятностей
3.1. Основные определения и понятия..... 8
3.2. Закон распределения случайной величины..... 11
3.2.1.Основные характеристики положения и рассеяния случайной величины..... 14
3.3. Некоторые теоретические законы распределения случайной величины..... 18
4. Статистика случайных величин
4.1. Статистические оценки неизвестных параметров распределения..... 26
4.2.Точность оценок параметров. Построение доверительных интервалов оценок..... 29
5. Построение статистических решений
5.1. Статистические гипотезы..... 33
5.2. Статистическая проверка некоторых типовых гипотез..... 36
5.2.1. Проверка гипотез о функциях распределения..... 36
5.2.2. Проверка гипотез о равенстве средних значений (математических ожиданий)..... 38
5.2.3. Проверка гипотез о равенстве дисперсий..... 43
6. Исследование различий между геологическими объектами. Дисперсионный анализ..... 45
7. Корреляционная зависимость между свойствами геологических объектов..... 48
8. Многомерные статистические модели. Статистические методы классифицирования геологических объектов
8.1. Элементы матричной алгебры..... 59
8.2. Статистические методы классифицирования геологических объектов..... 64
8.2.1. Анализ корреляционных матриц с позиций теории графов..... 64

8.2.2. Метод корреляционных профилей..... 66
8.2.3. Методы, опирающиеся на понятие компактности 67
8.2.4. Иеархическое группирование (кластер-анализ).... 69
8.2.5. Каноническая корреляция..... 71
8.2.6. Регрессионный анализ..... 71
8.2.7. Дискриминантный анализ..... 73
8.2.8. Факторный анализ..... 75
9.Моделирование пространственной изменчивости
свойств геологических объектов. Тренд-анализ..... 83
9.1. Горно-геометрическое моделирование..... 83
9.2. Аналитические методы моделирования
пространственной изменчивости..... 85
9.3. Особенности применения тренд-анализа..... 92
9.4. Моделирование дискретных полей..... 93
10. Случайные функции
10.1. Основные характеристики случайных функций..... 94
10.2. Полигармонический анализ случайных функций... 99
10.3. Полувариограммы и крайгинг..... 103
Литература..... 107
Приложения 108

Валерий Гаврилович Ворошилов

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ В ГЕОЛОГИИ

Учебное пособие

Научный редактор А.Ф. Коробейников, проф., д.г.-м.н. зав.кафедрой геологии и разведки месторождений полезных ископаемых

Редактор Р.Д. Игнатова

Подписано к печати 05.10.2001
Формат 60x84/16. Бумага ксероксная
Печать плоская. Усл.печ.л. 7.2 Уч.-изд. 6.53
Тираж 200 экз. Заказ № 63. Цена с.д.
ИПФ ТПУ. Лицензия ЛТ №1 от 18.07.94.
Типография ТПУ. 634034, Томск, пр.Ленина, 30.