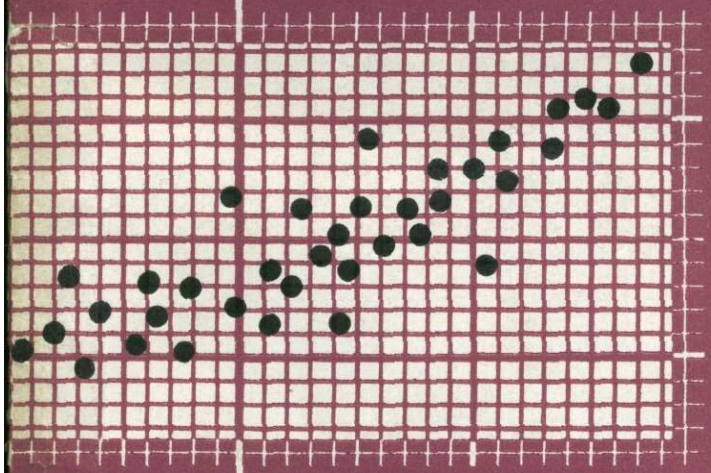
Ю. А. ТКАЧЕВ
Я. Э. ЮДОВИЧ

Статистическая обработка геохимических данных



АКАДЕМИЯ НАУК СССР
КОМИ ФИЛИАЛ
ИНСТИТУТ ГЕОЛОГИИ

Ю. А. ТКАЧЕВ, Я. Э. ЮДОВИЧ

Статистическая обработка геохимических данных

МЕТОДЫ И ПРОБЛЕМЫ

ИЗДАТЕЛЬСТВО «НАУКА»
ЛЕНИНГРАДСКОЕ ОТДЕЛЕНИЕ
ЛЕНИНГРАД, 1975

УДК 519.24 : 550.84+550.8

Статистическая обработка геохимических данных. Методы и проблемы. Ткачев Ю. А., Юдович Я. Э. Изд-во «Наука», Ленингр. отд., Л., 1975, 233 с.

В работе, состоящей из 10 глав, рассмотрены некоторые актуальные вопросы прикладной математической статистики применительно к задачам геохимии и близких к ней областей геологоразведочного дела. Разобраны вопросы точности анализов, методы внутреннего и внешнего контроля, рекомендованы рациональные способы обработки массовых анализов. Приводится критический обзор проблем оценивания природных частотных распределений по выборочным данным (подчеркнуты эвристический и оценочный аспекты проблемы), затронуты вопросы многомерной статистики, в частности проблемы «ложной корреляции», вскрыта сущность «многократной корреляции», показаны возможности и ограничения этого метода. Разработаны способы оценивания параметров объединенных совокупностей по выборкам различной представительности. Критически рассмотрена проблема вычисления кларков, вскрыты свойства кларков как случайных величин. Центральное место отведено проблемам корреляционного и регрессионного анализов. Детально описана процедура использования уравнения регрессии в оценочных и эвристических целях, развито представление о наблюдаемых коэффициентах корреляции как о взвешенных суммах, раскрыты взаимосвязи между регрессионным анализом, функциональной зависимостью и анализом поверхностей отклика. Книга адресована геохимикам, химикам-аналитикам, а также лицам, проводящим опробование и подсчет запасов полезных ископаемых. Библ. — 200, табл. — 16, рис. — 30.

Редакторы:

кандидат геолого-минералогических наук Ю. Н. ПРИХОДЬКО,
кандидат физико-математических наук В. Ф. БУШУЕВ

Т $\frac{20805\text{-}588}{055\ (02)\text{-}75}$ 319-75

ВВЕДЕНИЕ

В последнее десятилетие происходит бурный процесс метаматизации геологии. С одной стороны (работы Ю. А. Воронина и его сотрудников), делаются попытки формализовать основные понятия геологии, чтобы затем дать математическое описание основных геологических объектов и процессов, выполнить анализ существующих классификаций с позиций математической логики и т. д. С другой стороны, продолжается успешное применение в геологии традиционных методов математической статистики. Эти методы весьма эффективны, поскольку переменные величины, с которыми оперирует в своей работе геолог, обычно обладают основными свойствами случайных величин.

Непрерывно нарастает поток литературы, посвященной результатам применения математических методов, а также методике и проблемам такого применения. Если в 1940—1950 гг. в нашей стране публикации по данной тематике принадлежали в сущности только одному человеку — основателю математической геологии А. Б. Вистелиусу, то в наши дни ежегодно публикуется несколько десятков работ. В реферативном журнале «Геология» уже ряд лет существует специальный раздел, посвященный математическим методам. Геологу становится все труднее ориентироваться в этом количестве литературы.

Положение осложняется тем, что последние годы ознаменованы внедрением в практику работы геологов электронных вычислительных машин (ЭВМ), которые позволяют обрабатывать и получать громадное количество информации за весьма короткое время. Применительно к задачам математической статистики можно сказать, что для ЭВМ нет таких задач, решить которые нельзя только вследствие вычислительных трудностей: большинство типовых статистических задач решаются на ЭВМ за считанные минуты или часы. Однако уже выяснилось, что для программирования (по готовому алгоритму) на ЭВМ или получения каких-то численных результатов совсем не обязательно понимать статистику и вообще глубоко разбираться в математике.

Еще менее обязательно для лиц, работающих непосредственно на ЭВМ, понимать существо геологических задач. В результате нередко геологов обслуживают люди, недостаточно сведущие в геологии или в прикладной математической статистике, а литература наводняется результатами и рекомендациями, которые в лучшем случае являются сомнительными, а иногда и грубо ошибочными. Вообще современная ситуация характеризуется тем, что для геолога стало легче получить численный результат, нежели понять, что этот результат означает.

Несмотря на перечисленные отрицательные моменты, популярность математических методов возрастает, ибо применения их настоятельно требует практика. В настоящее время в распоряжении геологов имеется несколько монографий, посвященных статистическим методам. Прежде всего отметим книгу И. П. Шарапова «Применение математической статистики в геологии», недавно вышедшую вторым изданием. Эта книга весьма полезна для начального ознакомления со статистическими методами; ее рекомендуется проработать неподготовленному читателю до того, как он обратится к более специальным трудам. Наиболее полной сводкой с ценной библиографией зарубежных работ, мало знакомых нашему читателю, является переведенная на русский язык книга Р. Л. Миллера и Дж. С. Кана «Статистический анализ в геологических науках» (1965). К сожалению, многие места этой книги слишком трудны для восприятия читателя-геолога. Изложение материала ведется конспективно и часто может служить лишь справкой о существовании описываемых методов; охватив значительный материал, авторы не везде уделили должное внимание анализу принципиальных вопросов, без понимания которых применение статистических методов может оказаться неэффективным. Недавно в США опубликована монография Дж. С. Коха и Дж. Р. Ф. Линка «Статистический анализ геологических данных» (Koch, Link, 1970). Она посвящена более узкому кругу вопросов, чем книга Миллера и Кана, но зато они освещены детальнее. Перевод этой книги на русский язык представляется весьма желательным. Хорошим пособием по линейному корреляционному анализу и смежным вопросам может служить переведенная на русский язык монография У. Крамбейна и Ф. Грейбилла «Статистические модели в геологии» (1969), тогда как вышедшая у нас позднее книга Дж. Гриффитса «Научные методы изучения осадочных пород» (1971) дает мало нового по сравнению с первой.

Наконец, для статистической библиотечки геохимика можно рекомендовать следующие четыре прекрасных пособия: В. Ю. Урбаха (1964), В. В. Налимова (1960), К. Доерфеля (1969), А. Н. Зайделя (1968). Первое из них является, по-видимому, лучшим из всех руководств по математической статистике, написанных для специалистов таких естественных наук, как биология, медицина, геология, сельское хозяйство и т. д. По широте охвата и ясности

изложения эта книга, на наш взгляд, не имеет себе равных. Работы В. В. Налимова и К. Доерфеля посвящены проблемам обработки химико-аналитических данных. Первая из них значительно глубже и шире охватывает материал, зато вторая годится и для совсем неподготовленного читателя. Это же замечание относится и к превосходной брошюре А. Н. Зайделя.

Все перечисленные книги относились к разряду общих руководств. Имеется также ряд специальных работ, посвященных применению и разработке отдельных статистических методов в геологии и геохимии. Это книги Д. А. Родионова «Функции распределения содержаний элементов и минералов в изверженных горных породах» (1964а) и «Статистические методы разделения геологических объектов по комплексу признаков» (1968). В первой из них предложены удачные модели, объясняющие возможность возникновения логарифмически-нормального распределения концентраций минералов и химических элементов, вторая же, на наш взгляд, скорее поставила задачи, чем разрешила их.

Из других книг, содержащих ценные геохимические материалы, отметим работы В. Н. Бондаренко (1967, 1970), И. М. Остафийчук, М. И. Толстого (1972), В. Я. Воробьева (1970) и В. А. Кутолина (1972). Все они посвящены конкретным вопросам; их авторы успешно применяют уже апробированные статистические методы, редко затрагивая принципиальные методические проблемы.

Цель настоящей книги в том, чтобы более подробно изложить ряд важных для практики вопросов, которые недостаточно, а иногда и ошибочно были освещены в литературе. Хотелось бы остановить внимание читателя на следующих особенностях работы.

1. Книга не является руководством по математической статистике. Для понимания основной части текста вполне достаточно элементарных знаний по теории вероятностей. Тем не менее в соответствующих местах книги читатель найдет все основные свойства случайных величин и выражающие их формулы.

2. Являясь преимущественно обзорно-методической работой, книга содержит и ряд авторских разработок или трактовок. Этот материал содержится во всех главах в разной пропорции.

3. Книга адресуется прежде всего геохимикам и геологоразведчикам, но может оказаться полезной и химикам-аналитикам (гл. 2).

4. Мы намеренно не затрагиваем проблем, которые связаны с пространственным характером большинства величин, изучаемых в геологии и геохимии, — они являются предметом геостатистики, разработанной в основном трудами Ж. Матерона (1968). Все же в необходимых случаях в гл. 3 (о распределениях) мы вынуждены были коснуться геостатистических вопросов.

Вся математическая часть и значительная доля текста принадлежат первому автору. Второму автору принадлежит замысел работы и участие в составлении большинства глав.

Авторы считают своим приятным долгом отметить помощь со стороны директора Института геологии Коми Филиала Академии Наук СССР М. В. Фишмана, оказавшего энергичную поддержку идее создания данной книги. В процессе работы авторы получили также много указаний от Ю. В. Рощина, Н. П. Юшкина, В. С. Никифорова, А. И. Таскаева, А. И. Урнышева, М. П. Кетрис, Р. И. Пименова, Н. Г. Фридлендера, взявших на себя труд познакомиться с рукописью в целом или просмотревших отдельные части. Значительную работу выполнили научные редакторы книги: канд. геол.-минер. наук Ю. Н. Приходько и канд. физ.-мат. наук В. Ф. Бушуев. Первый помог улучшить геологическую часть работы, второй — проконтролировал корректность математических выкладок. Особо следует отметить помощь, оказанную авторам И. В. Рязановым (Печорская геофизическая экспедиция, г. Воркута). Им сделано множество замечаний по тексту рукописи, большинство из которых оказалось весьма полезным.

Авторы рады выразить свою признательность всем перечисленным лицам.

Глава 1

НЕКОТОРЫЕ СВЕДЕНИЯ ИЗ ТЕОРИИ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

§ 1. Случайное событие и вероятность

Объектом изучения теории вероятностей являются случайные события, случайные величины и их вероятности. В отношении *случайного события* должно выполняться только одно условие: при реализации фиксированного комплекса условий оно может совершиться или не совершиться. Например, при сокращении минералогической пробы событием можно обозначить попадание частицы намеченного минерала в навеску. События обозначаются большими буквами латинского алфавита: A, B, C. . . и т. д. Если A — событие, то $\bar{A}$ — также событие, противоположное первому и заключающееся в том, что A не совершилось. Если A означает, что содержание в навеске в процессе сокращения отклонилось от содержания в исходной пробе не более чем на $\alpha\%$, то $\bar{A}$ означает, что отклонение составило более $\alpha\%$. Символом $A \cap B$ или AB обозначают события, заключающиеся в одновременном осуществлении обоих событий, а символом $A \cup B$ — осуществление хотя бы одного из них. Каждому событию в соответствие ставится некоторое число, которое называют *вероятностью* этого события $\mathbf{P}(A)$. Вероятность достоверного события равна 1, вероятность невозможного события — 0. Таким образом, $0 \leqslant \mathbf{P}(A) \leqslant 1$.

Сложным событиям, представляющим собой комбинации других событий, приписываются вероятности по следующим правилам: $\mathbf{P}(AB)=\mathbf{P}(A)\cdot\mathbf{P}(B/A)=\mathbf{P}(B)\,\mathbf{P}(A/B)$, где (B/A) означает появление события B при условии, что событие A уже совершилось. Таким образом, вероятность одновременного осуществления двух событий равна произведению вероятности одного из них на вероятность другого, определенную при условии осуществления первого события. Если события A и B независимы, то $\mathbf{P}(B/A)=$

$=\mathbf{P}(B)$ и $\mathbf{P}(A/B)=\mathbf{P}(A)$, т. е. появление одного из независимых событий не изменяет вероятности появления другого. Следовательно, вероятность одновременного осуществления независимых событий A и B определяется равенством

$$\mathbf{P}(AB)=\mathbf{P}(A)\cdot\mathbf{P}(B).$$

Этот вывод легко распространяется на любое число событий.

Вероятность другого сложного события $A\cup B$ определяется по следующему правилу:

$$\mathbf{P}(A\cup B)=\mathbf{P}(A)+\mathbf{P}(B)-\mathbf{P}(AB),$$

т. е. вероятность осуществления хотя бы одного из двух событий (неважно, которого из них) равна сумме вероятностей этих событий без вероятности их совместного осуществления. Если события несовместимы, то $\mathbf{P}(AB)=0$ (как вероятность невозможного события) и $\mathbf{P}(A\cup B)=\mathbf{P}(A)+\mathbf{P}(B)$. Понятие вероятности в общем случае не поддается определению, и для приложения теории к геохимическим выводам в этом и нет необходимости. Для нас вполне удовлетворительным будет определение ее как отношения числа появлений интересующего нас события к общему числу равновозможных, единственно возможных, несовместимых событий: $\mathbf{P}(A)=m/n$, где m — число случаев, в которых событие A появилось, n — общее число возможных случаев.

§ 2. Случайная величина, функция распределения

Понятие «случайное событие» порождает понятие *случайной величины*. Например, случайной величиной является число появлений события A в некоторой серии испытаний. Интересующим нас событием во многих случаях может быть, например, принятие случайной величиной ξ одного из значений $-\infty<\xi\leqslant x$.

Функцией распределения называется такая функция F, которая равна вероятности того, что случайная величина ξ примет одно из значений $-\infty<\xi\leqslant x$:

$$F(x)=\mathbf{P}(\xi\leqslant x).$$

Очевидно, $F(x)$ может принимать значения $0\leqslant F(x)\leqslant 1$.

Производная этой функции, называемая *плотностью вероятности*, или плотностью распределения, дает представление о том, какова вероятность того, что случайная величина примет одно из значений в интервале $x+\Delta x$:

$$F'(x)=f(x);$$

обратно: $\int_{-\infty}^{x} f(\xi)\,d\xi = F(x)$ и $\int_{-\infty}^{+\infty} f(x)\,d(x) = 1$. Соотношение между этими функциями показано на рис. 1. Если случайная величина принимает только некоторые значения, например целочисленные, то по аналогии определяется *функция дискретного распределения* $F(x_n) = \sum_{i=1}^{n} p_i$, где p_i — вероятность появления значения x_i. Она представляет собой ломаную линию, совершающую скачки в точках с абсциссами x_i на величину p_i (рис. 2). *Дискретная плотность вероятности*, или, короче, функция вероятности соответственно равна

$$f(x_i) = \begin{cases} p_i \text{ для всех возможных значений } x_i \\ 0 \text{ для всех остальных значений } x. \end{cases}$$

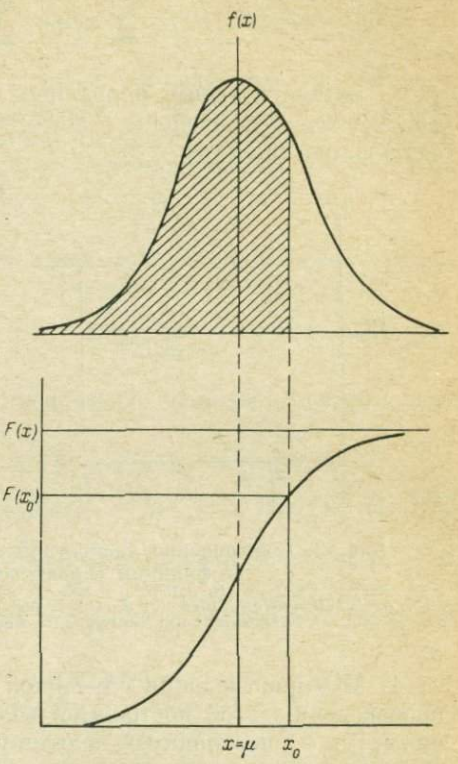


Рис. 1. Соотношение между функцией распределения $F(x)$ и плотностью вероятности $f(x)$.

Заштрихованная площадь под кривой $f(x)$ до точки $x=x_0$ численно равна ординате кривой $F(x)$ в точке x_0, μ — математическое ожидание величины x (см. § 3).

§ 3. Некоторые характеристики функций распределения и их свойства

Если случайная величина X непрерывно распределена с плотностью вероятности $f(x)$, то

$$\mathrm{M}(X) = \int x f(x)\,d(x) \qquad (1.3.1)$$

называется *математическим ожиданием* этой величины. Аналогично дискретно распределенная случайная величина имеет математическое ожидание

$$\mathrm{M}(X) = \sum_{i=1}^{n} x_i p_i, \qquad (1.3.2)$$

т. е. представляет собой средневзвешенное значение случайной величины, где весами служат вероятности различных значений.

Очевидно, что сумма весов $\sum p_i = 1$. В частности, если имеется n наблюденных значений случайной величины $x_1, x_2 \ldots x_i \ldots x_n$, то

$$\mathbf{M}(X) = \sum x_i p_i = \sum x_i \frac{1}{n} = \frac{1}{n} \sum x_i, \qquad (1.3.3)$$

т. е. математическим ожиданием в выборке является *среднее арифметическое* значение. Математическое ожидание (МО) имеет следующие свойства.

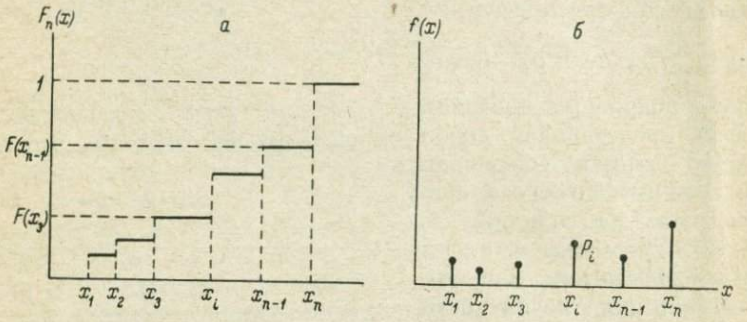


Рис. 2. Соотношение дискретной функции распределения $F_n(x)$ и функции вероятности $f(x_n) = p(x_n)$.

а — ступенчатая ломаная $F_n(x)$ с высотой ступени p_i; *б* — вертикальные отрезки, характеризующие вероятности p_i величины x_i.

1. МО произведения случайной величины на постоянную равно произведению этой постоянной величины на МО случайной величины, т. е. постоянную величину можно вынести за знак МО. Доказательство приводим только для дискретно распределенной величины:

$$\mathbf{M}(aX) = \sum a x_i p_i = a \sum x_i p_i = a\mathbf{M}(X). \qquad (1.3.4)$$

2. МО суммы случайных величин равно сумме их МО:

$$\mathbf{M}(X+Y) = \sum (x_i + y_i) p_i = \sum (x_i p_i + y_i p_i) =$$
$$= \sum x_i p_i + \sum y_i p_i = \mathbf{M}(X) + \mathbf{M}(Y). \qquad (1.3.5)$$

Заметим, что для доказательства не было необходимости считать случайные величины независимыми.

3. МО произведения независимых (!) случайных величин равно произведению их МО:

$$\mathbf{M}(X \cdot Y) = \sum_{i,j} x_i y_j k_{ij} = \sum_{i,j} x_i p_i y_j q_j =$$
$$= \sum_i \sum_j x_i p_i y_j q_j = \left(\sum x_i p_i\right)\left(\sum y_j q_j\right) = \mathbf{M}(X)\,\mathbf{M}(Y). \qquad (1.3.6)$$

Для доказательства этого свойства необходимо предположение независимости, так как только в этом случае вероятность k_{ij} появления пары величин $x_i y_j$ равна произведению вероятностей появления соответствующих значений случайных величин — $p_i q_j$.

4. МО постоянной величины[1] равно этой величине — свойство очевидное.

Для оценки степени разброса случайной величины относительно ее математического ожидания служит *дисперсия* $\mathbf{D}(X)$, которая определяется как математическое ожидание квадрата разности между случайной величиной и ее математическим ожиданием:

$$\mathbf{D}(X) = \mathbf{M}[X - \mathbf{M}(X)]^2. \qquad (1.3.7)$$

По определению, дисперсию выборки можно находить как среднее арифметическое квадратов уклонений x от $\mathbf{M}(X)$:

$$\mathbf{D}(X) = \frac{1}{n} \sum [x_i - \mathbf{M}(X)]^2. \qquad (1.3.8)$$

В некоторых случаях полезно представить дисперсию следующим образом:

$$\mathbf{D}(X) = \mathbf{M}[X - \mathbf{M}(X)]^2 = \mathbf{M}[X^2 - 2X\mathbf{M}(X) + \mathbf{M}^2(X)] =$$
$$= \mathbf{M}(X^2) - 2\mathbf{M}(X)\mathbf{M}(X) + \mathbf{M}^2(X) = \mathbf{M}(X^2) - \mathbf{M}^2(X). \qquad (1.3.9)$$

Поэтому можно сказать, что дисперсия равна МО квадрата случайной величины без квадрата ее МО. Такое представление дисперсии позволяет легко получить ее свойства из соответствующих свойств МО.

1. Дисперсия постоянной равна нулю:

$$\mathbf{D}(C) = \mathbf{M}(C^2) - \mathbf{M}^2(C) = C^2 - C^2 = 0. \qquad (1.3.10)$$

2. Постоянный множитель можно вынести за знак дисперсии, удвоив показатель его степени:

$$\mathbf{D}(CX) = \mathbf{M}(C^2X^2) - \mathbf{M}^2(CX) = C^2\mathbf{M}(X^2) -$$
$$- C^2\mathbf{M}^2(X) = C^2[\mathbf{M}(X^2) - \mathbf{M}^2(X)] = C^2\mathbf{D}(X). \qquad (1.3.11)$$

3. Дисперсия суммы независимых случайных слагаемых равна сумме их дисперсий:

$$\mathbf{D}(X + Y) = \mathbf{D}(X) + \mathbf{D}(Y). \qquad (1.3.12)$$

Отсюда следует, что дисперсия случайной величины, сложенной с постоянной, равна дисперсии самой случайной величины, так как дисперсия постоянной равна нулю:

$$\mathbf{D}(X + C) = \mathbf{D}(X) + \mathbf{D}(C) = \mathbf{D}(X). \qquad (1.3.13)$$

---

[1] Точнее: величины, принимающей одно и то же значение, ибо МО имеет смысл только для м н о ж е с т в а значений.

В геологии в качестве меры рассеяния признака часто используется коэффициент вариации

$$V = \frac{\sqrt{D(X)}}{M(X)} \cdot 100\%. \qquad (1.3.14)$$

Теоретические значения МО и дисперсии случайной величины X, т. е. их значения в изучаемой генеральной совокупности, обычно обозначают μ_x и σ_x^2, а выборочные значения — соответственно $\bar{x}$ и s_x^2. Известно, что $\mathbf{M}(\bar{x}) = \mu_x$, где

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad (1.3.15)$$

т. е. МО среднего арифметического из выборки (или, лучше, — из выборок объемом по n элементов каждая) равно МО случайной величины в генеральной совокупности. Этот вывод кажется настолько естественным, что по аналогии напрашивается обобщение его и для выборочной дисперсии. Однако такое обобщение ошибочно. Можно доказать, что МО выборочной дисперсии $\tilde{S}_x^2$ всегда меньше дисперсии генеральной совокупности. Так как $\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_i + \ldots + x_n)$, то из свойства дисперсии (1.3.12) получаем $\sigma_{\bar{x}}^2 = \frac{1}{n^2}(\sigma_{x_1}^2 + \sigma_{x_2}^2 + \ldots + \sigma_{x_n}^2) = \frac{1}{n^2} n\sigma_x^2 = \frac{1}{n}\sigma_x^2$. Выборочную дисперсию $\tilde{s}_x^2 = \sum (x_i - \bar{x})^2 \frac{1}{n}$ можно представить следующим образом: $n\tilde{s}_x^2 = \sum [(x_i - \mu) - (\bar{x} - \mu)]^2 = \sum (x_i - \mu)^2 - 2\sum (x_i - \mu)(\bar{x} - \mu) + n(\bar{x} - \mu)^2 = \sum (x_i - \mu)^2 - 2n(\bar{x} - \mu)^2 + n(\bar{x} - \mu)^2 = \sum (x_i - \mu)^2 - n(\bar{x} - \mu)^2$, т. е. $n\tilde{s}_x^2 = n\sigma_x^2 - n\sigma_{\bar{x}}^2$, или $n\tilde{s}_x^2 = n\sigma_x^2 - \sigma_x^2$. Окончательно имеем

$$\mathbf{M}(\tilde{s}_x^2) = \frac{n-1}{n}\sigma_x^2. \qquad (1.3.16)$$

Это происходит потому, что выборочная дисперсия рассчитывается по отношению к выборочному среднему $\bar{x}$, а не к истинному μ_x.

Изложенное является ответом на частый вопрос геологов: в каком случае при расчете дисперсии сумму квадратов делить на n, а в каком — на $n-1$. Если известно генеральное среднее μ_x (что случается редко), то сумму квадратов отклонений от него необходимо делить на число элементов выборки n. Если при расчете дисперсии используется выборочное среднее $\bar{x}$, то для оценки σ_x^2, как это следует из (1.3.16), сумму квадратов отклонений надо делить на $n-1$. Однако если выборочная дисперсия не используется как оценка дисперсии генеральной совокупности,

а служит лишь для свертки информации о рассеянии выборки, то делят на n; эта величина (обозначенная нами $\bar{s}_x^2$) используется редко.

Важную роль в статистике играет также величина $\sqrt{\sigma^2} = \sigma$, которая называется средним квадратичным (или квадратическим) отклонением. Широко используются синонимы: «квадратичная ошибка», «стандартное отклонение», наконец, просто «стандарт».

§ 4. Некоторые важные в геологии и геохимии дискретные распределения

Гипергеометрическое распределение. Это дискретное распределение имеет важное значение в определении ошибок сокращения материала геохимических и минералогических проб, состоящего из смеси различных частиц. Пусть генеральная совокупность, например исходная проба, состоит из N элементов (здесь — частиц). Пусть в ее составе имеется K рудных частиц. Из этой генеральной совокупности сделаем *безвозвратную выборку* — навеску, состоящую из n частиц. Если выборка сделана правильно (все возможные комбинации частиц в выборке равновероятны), то вероятность появления x рудных частиц среди n отобранных равна:

$$p(x, n, K, N) = \frac{C_k^x \cdot C_{N-K}^{n-x}}{C_N^n}. \qquad (1.4.1)$$

Распределение такого рода обозначим $g(x, n, K, N)$, где g — наименование распределения, x — случайная переменная, другие буквы в скобках — уже известные нам параметры. Дискретная плотность вероятности этого распределения равна:

$$g(x) = \begin{cases} \dfrac{C_K^x C_{N-K}^{n-x}}{C_K^n} & \text{для } x = 0, 1, 2, 3 \ldots K, \\ 0 & \text{для остальных значений } x. \end{cases} \qquad (1.4.2)$$

Математическое ожидание величины x равно

$$\mathbf{M}(X) = nK/N. \qquad (1.4.3)$$

Величина K/N есть вероятность появления рудной частицы в выборке. Обозначив ее через p, будем иметь $\mathbf{M}(X) = \mu = np$. Полученный результат подтверждает интуитивное представление о «наиболее ожидаемом» числе элементов x в выборке объемом n.

Дисперсия числа появлений интересующего нас элемента равна:

$$\mathbf{D}(X) = npqN - n/(N-1) \approx npq(1 - n/N), \quad q = 1 - p. \qquad (1.4.4)$$

Пример. Из пробы, содержащей 15 кусков породы и 5 кусков руды, взято наугад (делением пополам) 10 кусков. Изобразить функцию дискретной плотности вероятности числа кусков руды в этой выборке, найти математическое ожидание и дисперсию числа рудных кусков. В принятых обозначениях имеем: $N=20$, $K=5$, $n=10$, $p=5/20=0.25$, $q=0.75$. Решение: $M(X)=10\cdot 0.25=2.5$; $\mathbf{D}(X)=0.94\approx 1$.

График соответствующей дискретной плотности вероятности приведен на рис. 3.

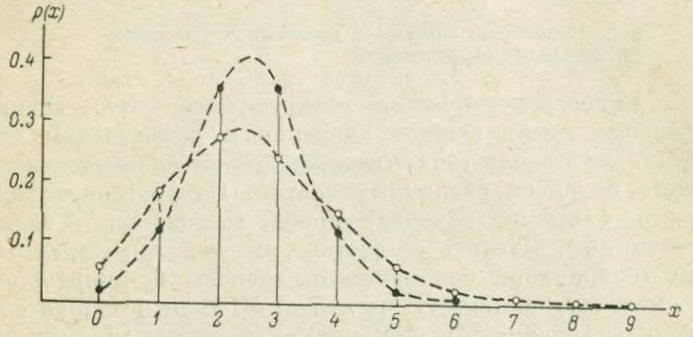


Рис. 3. График функции вероятности появления x рудных кусков в безвозвратной выборке (*темные кружки*).

Для сравнения *светлыми кружками* показаны аналогичные вероятности для выборки с возвращением. Прерывистые кривые приведены только для наглядности и не являются «кривыми распределения».

Более показательны параметры, относящиеся не к числу появлений элементов, а к наблюдаемой *частости* их появления, т. е. отношения числа появлений элементов к числу испытаний x/n. Исходя из свойств математического ожидания и дисперсии, получаем

$$\mathbf{M}\left(\frac{x}{n}\right)=\mu=\frac{1}{n}\mathbf{M}(x)=\frac{1}{n}np=p, \qquad (1.4.5)$$

т. е. математическим ожиданием частости является вероятность появления события, и

$$\mathbf{D}\left(\frac{x}{n}\right)=\sigma^2_{x/n}=\frac{1}{n^2}\mathbf{D}(x)=\frac{pq}{n}\cdot\frac{N-n}{N-1}\approx\frac{pq}{n}\left(1-\frac{n}{N}\right)=$$

$$=pq\left(\frac{1}{n}-\frac{1}{N}\right). \qquad (1.4.6)$$

Биномиальное распределение. Когда выборка производится с возвращением каждого взятого элемента или объем генеральной совокупности настолько велик, что изъятие каких-нибудь элементов в выборку практически не изменяет вероятности появления этого элемента, то число элементов x в выборке оказывается распределенным по биномиальному закону $b(x, n, p)$. Здесь b —

наименование закона, x — переменная, n и p — параметры распределения. Функция вероятности биномиального распределения равна:

$$b(x) = \begin{cases} \dfrac{n!\,p^x q^{n-x}}{x!\,(n-x)!} & \text{для } x = 0,\ 1,\ 2,\ 3 \ldots n \\ 0 & \text{для остальных значений } x. \end{cases} \qquad (1.4.7)$$

Математическое ожидание биномиального распределения можно получить как математическое ожидание суммы n слагаемых (по числу элементов выборки), принимающих значение единицы с вероятностями $p_i = \text{const} = p$:

$$\mathbf{M}(X) = \sum_{i=1}^{n} 1 \cdot p_i = n\mathbf{M}(X_i) = np. \qquad (1.4.8)$$

Аналогично получим дисперсию распределения:

$$\mathbf{D}(X_i) = \mathbf{M}(X_i^2) - \mathbf{M}^2(X_i) = 1^2 \cdot p + 0^2 \cdot q - p^2 =$$
$$= p - p^2 = p(1-p) = pq, \qquad (1.4.9)$$

$$\mathbf{D}(X) = \mathbf{D}\sum_{i=1}^{n} x_i = \mathbf{D}(X_1) + \mathbf{D}(X_2) + \ldots + \mathbf{D}(X_n) = npq.$$

Математическое ожидание относительной частоты или частости при биномиальном распределении получается переходом к новой случайной величине x/n. Тогда $\mathbf{M}\left(\frac{x}{n}\right) = \frac{1}{n}\mathbf{M}(x) = np/n = p$ и $\mathbf{D}(x/n) = \frac{1}{n^2} npq = pq/n$.

Биномиальное распределение можно рассматривать как приближение гипергеометрического с объемом генеральной совокупности, стремящимся к бесконечности: $g(x, n, K, N) \underset{N\to\infty}{\to} b(x, n, K/N)$. Их математические ожидания равны, если равны доли «рудных» частиц в генеральной совокупности, а дисперсии при большом N или N/n также равны: $\lim\limits_{N\to\infty} pq(1/n - 1/N) = pq/n$. Вообще, при большом N/n эти распределения становятся практически неразличимыми.

Пример. Видоизменим предыдущий пример таким образом, чтобы выборка производилась с возвращением. Тогда распределение числа частиц будет таким, как изображено на рис. 3, — светлыми кружками.

Полиномиальное распределение. Биномиальное распределение легко обобщается на случай, когда число различимых элементов в выборке больше двух:

$$\text{pl}(x_1, x_2, x_3 \ldots x_n) = \frac{n!\,p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}}{x_1!\,x_2! \ldots x_k!\,(n-x_1)!\,(n-x_2)! \ldots (n-x_k)!},$$

где $x_1, x_2 \ldots x_k$ — числа появления элементов, например минера-

лов сорта 1, 2, 3 ... k, а p_1, $p_2 \ldots p_k$ — вероятности их появления. Это распределение называется полиномиальным (pl).

Заметим, что совокупность, состоящую из любого числа «сортов» минералов, можно рассматривать как состоящую из двух «сортов» — интересующих нас и остальных, имеющих вероятности появления: p и $q = 1 - \sum_{i=2}^{k} p_i$. Полиномиальным распределением следует пользоваться только тогда, когда нас интересуют *вероят-*

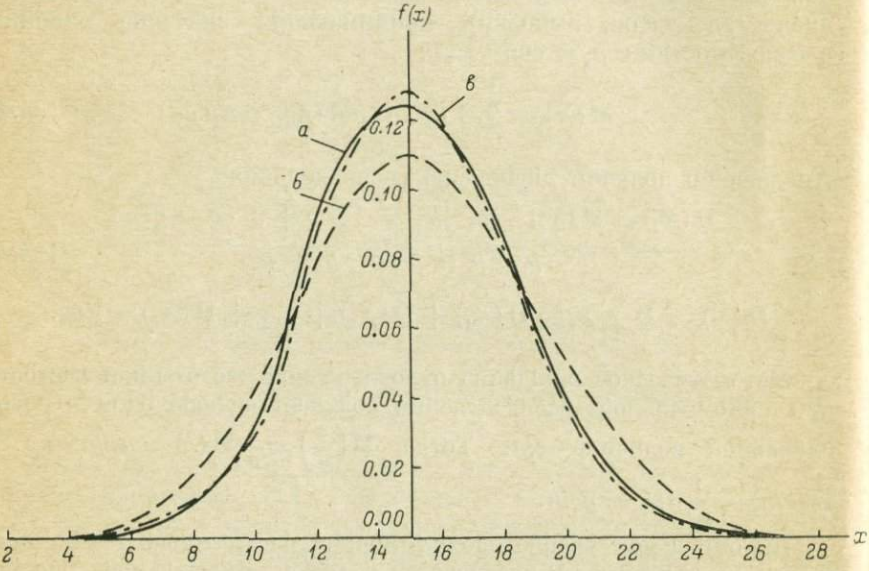


Рис. 4. Сравнение графиков плотности вероятности нескольких распределений.

а — фактическое распределение числа рудных частиц в выборке (гипергеометрическое); *б* — приближение его нормальным распределением; *в* — то же, дисперсия вычислена с поправочным коэффициентом $(1-n/N)$.

ности совместного появления заданного числа минералов каждого сорта. Очевидно, что по-прежнему $\mathbf{M}(X_i) = np_i$, а

$$\mathbf{D}(X_i) = np_i\left(1 - \sum_{i=1}^{k-1} p_i\right) = np_iq. \qquad (1.4.10)$$

§ 5. Нормальное распределение

Благодаря важности и обширности приложений, исключительно большое значение имеет так называемое *нормальное* (Гауссово) распределение вероятностей. Плотность вероятности нормального закона $N(\mu, \sigma)$, где μ и σ — его параметры (сред-

нее, равное математическому ожиданию, и стандарт), задается следующим выражением:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \qquad (1.5.1)$$

Кривая плотности вероятности нормального распределения в сравнении с гипергеометрическим показана на рис. 4.

Функция распределения его равна

$$F(x) = \int_{-\infty}^{x} f(x)\,dx = \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx. \qquad (1.5.2)$$

Нормальное распределение с параметрами $\mu = 0$ и $\sigma = 1$, т. е. $\mathbf{N}(0, 1)$ *называется нормированным нормальным распределением.* Для того чтобы нормировать нормальное распределение с параметрами μ и σ, необходимо в качестве переменной рассматривать безразмерную величину $z = (x-\mu)/\sigma$, т. е. отклонение изучаемой случайной величины от ее математического ожидания, выраженное в долях ее стандартного отклонения. Нормированная функция нормального распределения, *имеющая теперь вид* $F(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) dz$, табулирована и приводится в большинстве руководств по математической статистике или специальных таблицах (например, Янко, 1961).

Большой интерес для нас будет представлять так называемая функция Лапласа

$$\Phi(u) = \int_{0}^{u} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) dz, \qquad (1.5.3)$$

графическое соотношение которой с функцией распределения показано на рис. 5. Функция Лапласа удобна для отыскания вероятности того, что случайная величина отклонится от своего среднего значения не более чем на заданную величину $\mathbf{P}\{x_1 \leqslant X \leqslant x_2\}$:

$$\mathbf{P}\{x_1 \leqslant X \leqslant x_2\} = \Phi((x_2-\mu)/\sigma) - \Phi((x_1-\mu)/\sigma), \qquad (1.5.4)$$

или, переходя к нормированной переменной, $\mathbf{P} = \Phi(z_2) - \Phi(z_1)$. Если x_1 и x_2 симметричны относительно среднего, то

$$\mathbf{P} = \Phi(z_2) - \Phi(-z_2) = \Phi(z_2) + \Phi(z_2) = 2\Phi(z_2). \qquad (1.5.5)$$

Функция Лапласа помогает наглядно представить себе связь между дисперсией и вероятностью, с которой ошибка выборочного значения (разность между выборочным значением и математическим ожиданием) не превысит заданную величину. Так, вероятность того, что ошибка не превысит стандартного отклонения

$\pm\sigma$, равна 0.68269, или около 68.27%, т. е. $\mathbf{P}(|x-\mu|\leqslant\sigma)=0.68269\approx 68.27\%$. Соответственно:

$$\mathbf{P}(|x-\mu|\leqslant 2\sigma)=0.9545\approx 95.4\%,$$

$$\mathbf{P}(|x-\mu|\leqslant 3\sigma)=0.9973\approx 99.7\%.$$

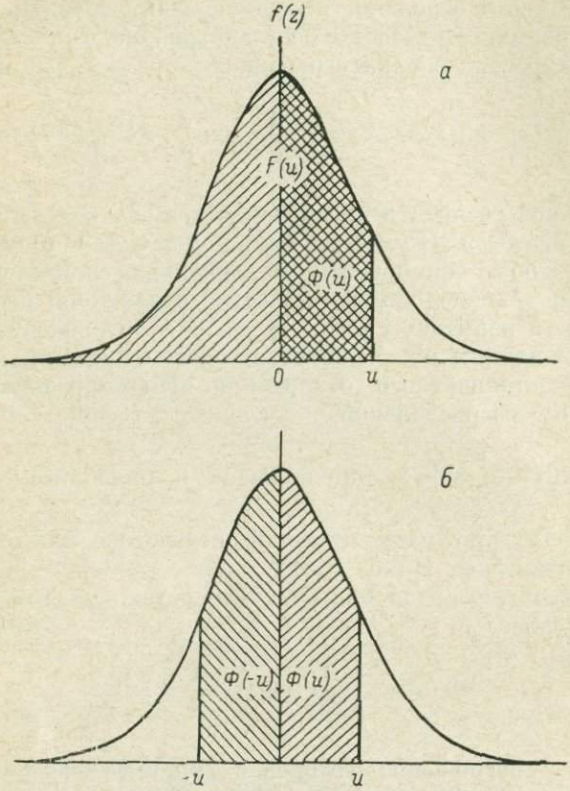


Рис. 5. Сравнение плотности вероятности нормированного нормального распределения, функции распределения и функции Лапласа.

а — заштрихованная площадь под кривой является функцией распределения $F(u)$, дважды заштрихованная площадь — функция Лапласа $\Phi(u)$; *б* — к свойству функции Лапласа $\Phi(-u)=-\Phi(u)$ площадь под кривой $f(Z)$ в интервале $(-\infty, 0)$ считается отрицательной.

На практике событие $|x-\mu|<3\sigma$ считают достоверным, а превышение «трехсигмовых» пределов практически невозможным событием. При неизвестных распределениях, сильно отличающихся от нормального, дисперсия дает недостаточно информации для установления пределов возможных погрешностей.

Значительная часть статистических выводов так или иначе связана с расчетом вероятности попадания случайной величины

в намеченные интервалы; как видим, эти расчеты весьма просты для нормального распределения. Эта простота заставляет геологов и геохимиков явно или неявно постулировать нормальное распределение даже там, где оно не имеет места. Тем не менее нормальное или близкое к нему распределение чаще других встречается в геохимической практике. Таково распределение погрешностей анализа (или их логарифмов), распределение концентраций многих элементов в горных породах и минералах. Наконец, распределение средних значений больших массивов случайных величин любой природы, согласно центральной предельной теореме, с увеличением числа проб или массивов очень быстро приближается к нормальному.

Нормальное распределение может служить хорошим приближением биномиального и гипергеометрического при безграничном увеличении объема выборки. Это избавляет от необходимости вычисления факториалов больших чисел при оценке доверительных интервалов для x. Нормирование этой величины производится по формулам

$$z=\frac{x-np}{\sqrt{npq\left(1-\frac{n}{N}\right)}}=\frac{\frac{x}{n}-p}{\sqrt{\frac{pq}{n}\left(1-\frac{n}{N}\right)}};\quad z=\frac{x-np}{\sqrt{npq}}=\frac{\frac{x}{n}-p}{\sqrt{\frac{pq}{n}}},\qquad (1.5.6)$$

с помощью которых, пользуясь функцией Лапласа, легко определить вероятность попадания x или x/n в любой заданный интервал.

Пример. Пусть $N=1500$, $n=500$, $p=0.25$, $q=0.75$, $pq=0.188$, $pq/n=3.76\cdot10^{-4}$, $\sigma=\sqrt{\frac{pq}{n}\left(1-\frac{n}{N}\right)}=1.58\cdot10^{-2}$. Вероятность того, что относительная частота p будет заключена в пределах $0.2<p<0.3$, равна 0.9984. Если бы выборка была с возвращением, то стандартное отклонение необходимо вычислять по формуле $\sigma=\sqrt{\frac{pq}{n}}$, и соответствующая вероятность была бы ощутимо ниже (0.992). Рис. 4 иллюстрирует хорошее приближение гипергеометрического распределения (с параметрами $N=3000$, $n=300$) нормальным.

Фактически любое геохимическое опробование или опробование месторождений является выборкой без возвращения, ибо одна и та же порция вещества не может попасть в пробу более одного раза.[1] Поэтому фактически оценка среднего в генеральной совокупности (массиве, толще, месторождении) происходит с дисперсией, меньшей, чем теоретическая для выборки с возвращением σ^2/n, а именно с дисперсией

$$\left(1-\frac{n}{N}\right)\frac{1}{n}\sigma^2. \qquad (1.5.7)$$

[1] Это обстоятельство было подчеркнуто Ж. Матероном (1968, стр. 10). Геостатистический аспект проблемы здесь нами не затрагивается.

Формула применима независимо от того, подчиняется ли распределение случайной величины нормальному закону. Полученный результат часто будет использоваться в гл. 5 и 6.

Заметим, наконец, что график плотности вероятности нормального распределения представляет собой колоколообразную «кривую Гаусса», симметричную относительно математического ожидания μ. Вообще асимметрия распределения характеризуется параметром $\nu_1 = \frac{1}{\sigma^3} M(x-\mu)^3$. Асимметрию называют отрицательной $\nu_1 < 0$, если $f(x)$ вытянута влево от μ, положительной — если она вытянута вправо. Нормальное распределение дает $\nu = 0$, на чем основан один из методов проверки этого закона.

Для характеристики «островершинности» одномодальных (одновершинных) распределений применяется показатель *эксцесса*

$$E = \frac{1}{\sigma^4} M(x-\mu)^4 - 3.$$

Для нормального распределения он также равен нулю, поэтому нормальная кривая принимается как бы за эталон островершинности распределений при равных дисперсиях.

§ 6. Некоторые другие важные распределения случайных величин

Рассмотрим некоторые другие распределения, которые по разным причинам будут представлять интерес для дальнейшего изложения.

Равномерное (прямоугольное) распределение. Это распределение непрерывной случайной величины имеет плотность вероятности

$$f(x) = \begin{cases} \frac{1}{\alpha - \beta} & \text{при } \alpha \leqslant x \leqslant \beta \\ 0 & \text{при всех других значениях } x. \end{cases} \qquad (1.6.1)$$

Его график представляет собой прямую, параллельную оси абсцисс и отстоящую от нее на расстоянии $1/(\beta-\alpha)$. Математическое ожидание равномерно распределенной случайной величины равно середине интервала $(\beta+\alpha)/2$. Дисперсия этого распределения $\sigma^2 = (\beta-\alpha)^2/12 = h^2/12$ важна для нас тем, что объясняет происхождение так называемой поправки Шеппарда, которую необходимо учитывать как составную часть дисперсии при округлении данных, где h — интервал округления.[1] Применение поправки Шеппарда, вообще говоря, справедливо только при малых h, так как с увеличением ширины интервала распределение в его пределах уже нельзя будет считать равномерным.

[1] Или предельное деление шкалы измеряющего прибора.

Распределение Пуассона. Это весьма замечательное дискретное распределение является частным случаем биномиального с очень малым значением вероятности p. Если вероятность появления некоторого события в малый промежуток времени прямо пропорциональна этому промежутку, а появления двух событий в течение этого промежутка равна нулю, то вероятность числа событий будет подчиняться закону Пуассона

$$\mathbf{P}(X=x, \lambda)=\lambda^x e^{-\lambda}/x! \qquad (1.6.2)$$

с одним параметром λ, характеризующим как дисперсию, так и среднее значение

$$\mu_x=\lambda, \quad \sigma_x^2=\lambda. \qquad (1.6.3)$$

Распределение Пуассона используется для описания числа радиоактивных распадов, числа зерен редко встречающихся минералов в породе или раздробленной пробе, числа жил, редких палеонтологических остатков или других объектов в единице площади или объема. В частности, отклонение закона от пуассоновского будет свидетельствовать о неслучайном (закономерно скученном или слишком равномерном) характере распределения изучаемых объектов. Распределение Пуассона является как бы эталоном случайной неравномерности в расположении точек по разрезу, площади или объему.

Очень любопытна двойная пуассоновская модель, которая была применена (Ткачев и др., 1970) к изучению месторождений горного хрусталя. Было установлено, что число полостей-гнезд с кристаллами распределено по единичным объемам, на которые условно разбито месторождение, по закону Пуассона. Количество кристаллов (в условных «порциях») по отдельным гнездам также распределено по закону Пуассона. Тогда количество кристаллов z по единичным объемам месторождения распределено по двойному пуассоновскому закону:

$$\mathbf{P}_z=\sum_{r=0}^{\infty}\frac{b^r e^{-b}}{r!}\,\frac{(ar)^z e^{-ar}}{z!}, \qquad (1.6.4)$$

где b — среднее число полостей в единичном объеме, a — среднее количество кристаллов в полости, r — переменная суммирования.

Гамма-распределение. Плотность вероятности этого распределения, используемого в геологии пока что редко, равна

$$G(x, r, \beta)=x^{r-1}e^{-x/\beta}/\Gamma(r)\,\beta^r. \qquad (1.6.5)$$

При $\beta=1$ кривые этого распределения весьма напоминают «кривые» распределения Пуассона. Это и понятно, так как гамма-функция является обобщением факториала (ср. формулу (1.6.2)) для непрерывных величин. Как и распределение Пуассона, функция (1.6.5) подходит для описания резко правоасимметричных распре-

делений величин, среди которых определенная доля имеет нулевые значения. Именно вследствие их присутствия для аппроксимации здесь не может быть применено обычное для правоасимметричных гистограмм логнормальное распределение. Гамма-распределение несомненно найдет более широкое применение в геохимии, например, как составная часть двойной пуассоновской модели: когда число объектов распределено в пространстве по Пуассону, а значение случайной величины в объекте — по гамма-распределению.

Логарифмически-нормальное распределение. Если по нормальному закону распределены не сами величины, а их логарифмы, то получается распределение вида

$$\mathbf{L}(x, \nu, \theta) = \frac{1}{\theta\sqrt{2\pi}} \exp\left(-\frac{(\log x - \nu)^2}{2\theta^2}\right). \qquad (1.6.6)$$

Как и нормальное распределение, оно характеризуется двумя параметрами — средним значением логарифма $\mu_L = \nu$ и логарифмической дисперсией $\sigma_L^2 = \theta^2$. Среднее значение случайной величины x и ее дисперсия σ_x^2 связаны с μ_L и σ_L^2 следующими соотношениями:

$$\mu_x = \exp\left(a\mu_L + \frac{1}{2}a^2\sigma_L^2\right), \qquad (1.6.7)$$

$$\sigma_x^2 = \exp(2a\mu_L + a^2\sigma_L^2) \cdot \exp(a^2\sigma_L^2 - 1), \qquad (1.6.8)$$

где $a = \log_b e$ (b — основание логарифмов). Эти же соотношения в натуральных логарифмах:

$$\mu_x = \exp\left(\mu_L + \frac{1}{2}\sigma_L^2\right),$$

$$\sigma_x^2 = \exp(2\mu_L + \sigma_L^2) \cdot \exp(\sigma_L^2 - 1).$$

Логнормальное распределение широко применяется в геохимии для аппроксимации содержаний редких химических элементов, размеров зерен осадочных пород, зольности углей и т. д. Использование этого распределения и некоторых его разновидностей более подробно освещено в гл. 3 и 4.

Распределение Вейбулла — недавно предложенное весьма универсальное распределение для непрерывной случайной величины с плотностью вероятности

$$f(x) = \frac{m(x - x_0)^{m-1}}{C_0} \cdot \exp\left(-\frac{(x - x_0)^m}{C_0}\right) \qquad (1.6.9)$$

и функцией распределения

$$F(x) = \int_{x_0}^{x} f(x)\,dx = 1 - \exp\left(-\frac{(x - x_0)^m}{C_0}\right), \qquad (1.6.10)$$

где x — случайная переменная, равная или большая x_0; x_0 — на-

чало отсчета (параметр положения); C_0 — параметр масштаба (приблизительная аналогия стандартного отклонения в нормальном распределении); m — параметр формы. Для большинства геохимических переменных, например содержаний химических элементов, $x_0 = 0$; тогда функция распределения Вейбулла примет вид

$$F(x) = 1 - \exp\left(-\left(\frac{bx}{\mu}\right)^m\right), \qquad (1.6.11)$$

где b и m — параметры распределения, определяемые в зависимости от коэффициента вариации V; μ — математическое ожидание изучаемой переменной.

Экспоненциальное распределение можно рассматривать как частный случай распределения Вейбулла при $m = 1$ и $x_0 = 0$. Тогда плотность его вероятности будет равна

$$f(x) = \frac{1}{C_0}\exp\left(-\frac{1}{C_0}x\right). \qquad (1.6.12)$$

Если обозначить параметр $1/C_0$ через β, то

$$f(x) = \beta e^{-\beta x}. \qquad (1.6.13)$$

Экспоненциальное распределение широко распространено в физике и термодинамических разделах геохимии, имеются примеры его применения и в геологии.

Для облегчения чтения геологами и геохимиками статистической литературы, в том числе и данной книги, добавим несколько слов по поводу терминологии и обозначений. Понятия математического ожидания и дисперсии обозначаются большими буквами **M** и **D** с указанием в скобках множества величин, к которым относятся эти понятия. Например $\mathbf{D}(X+C)$, где $X = \{x_1, x_2, \ldots x_i, \ldots, x_n\}$, обозначает дисперсию множества случайных величин $\{x_1+C, x_2+C, \ldots, x_n+C\}$. Параметры изучаемой генеральной совокупности обозначаются греческими буквами. Так, математическое ожидание случайной величины в данной генеральной совокупности обозначается обычно μ_x, а дисперсия — σ_x^2 или просто μ и σ^2, если ясно, о какой случайной величине идет речь. Необходимо различать также математическое ожидание и дисперсию случайной величины данной выборки и *выборочную оценку* математического ожидания и дисперсии той генеральной совокупности, которую характеризует эта выборка. Для первой пары из этих величин нет общепринятых обозначений, и они не имеют самостоятельного значения. Их роль, как будет показано в дальнейшем, заключается в том, что с помощью тех или иных поправок из них получают оценки соответствующих параметров генеральной совокупности. Мы их будем обозначать $\bar{x}$ и $\bar{s}^2$. Выборочные оценки математического ожидания и дисперсии

генеральной совокупности обозначают соответствующими латинскими буквами: m_x и s_x^2. Аналогичных принципов придерживаются и при обозначениях других параметров.

§ 7. Статистические выводы и критерии. Статистические критерии точечных оценок параметров

Под термином *статистический вывод* понимается составление заключений относительно генеральной совокупности на основании информации, получаемой из выборок. Бóльшая часть статистических выводов сводится к следующим типам: а) *точечные оценки* параметров, когда находят «лучшую» оценку $\hat{\Theta}$ некоторого параметра Θ; б) *интервальные оценки* параметров (верхний и нижний предел «правильного» интервала называют доверительными пределами); в) *проверка гипотез*, заключающаяся в ответе на вопрос: принять ли эту гипотезу или отвергнуть ее, отдав предпочтение некоторой альтернативной гипотезе.

Понятие «лучшая» оценка параметра не является абсолютным; оно зависит и от исследуемого материала и в основном — от целей исследования, если имеются в виду те последствия, которые повлечет за собой ошибка в оценке параметра в бóльшую или меньшую сторону. То же самое относится и к понятию «правильного» интервала. Кроме этого, в геологии и геохимии очень остро стоит вопрос о рациональном выборе *параметра* или какой-либо другой статистики, которые наилучшим образом подходили бы для решения поставленной задачи. В ряде работ обсуждается, например, вопрос о преимуществе среднего взвешенного или среднего арифметического содержания, среднего значения логарифмов или медианы для оценки фоновых значений и т. д.

Состоятельность. Оценка является состоятельной, если при неограниченном возрастании объема выборки она неограниченно приближается к значению параметра. Очевидно, что такая *статистика* [1] как среднее значение выборки $\bar{x}$ является состоятельной оценкой математического ожидания.

Несмещенность. Если математическое ожидание оценки совпадает с оцениваемым параметром $M(\hat{\Theta}) = \Theta$, то такая оценка называется несмещенной. Среднее арифметическое по выборке является несмещенной оценкой генерального среднего. В разделе 1.3 мы убедились, однако, что статистика $\tilde{s}_x^2 = \Sigma(x_i - \bar{x})^2/n$ является смещенной и лишь преобразование $s_x^2 = \tilde{s}_x^2 n/n-1$ приводит к несмещенной оценке

$$M(\tilde{s}_x^2)\, n/n - 1 = M(s_x^2) = \sigma_x^2.$$

---

[1] Статистикой называют какую-либо функцию выборочных (наблюденных) значений случайных величин, используемую в основном для оценки параметров и гипотез, а также и в других целях.

При расчете дисперсии как половины среднего значения квадратов разностей всех возможных пар элементов выборки (такой способ часто используется при расчете аналитической дисперсии по двукратным анализам — см. гл. 2) $\tilde{s}^2 = \frac{1}{2}\sum (x_i - x_j)^2/n$ несмещенность можно устранить делением суммы не на $2n$ (двойное число суммируемых разностей), а на $2(n-1)$.

Необходимо предупредить читателя, что, хотя $M(s^2) = \sigma^2$, величина $M(s) \leqslant \sigma$.[2] Это надо понимать так, что нельзя находить среднее арифметическое значение стандартного отклонения из ряда выборок, а надо найти соответствующие дисперсии, усреднить их и лишь затем при необходимости извлечением квадратного корня определить соответствующее среднее квадратичное отклонение (стандарт).

Эффективность. Статистика называется эффективной в абсолютном смысле, если ее дисперсия относительно оцениваемого параметра не больше дисперсии другой статистики. Обычно говорят об *относительной эффективности* статистики (как оценки) какого-либо параметра, подразумевая при этом отношение дисперсии данной оценки $M(\hat{\Theta} - \Theta)^2$ к дисперсии сравниваемой оценки. Значение $\bar{x}$ является эффективной оценкой математического ожидания в абсолютном смысле при любом объеме выборки, но только для нормально распределенной случайной величины. И только в этом случае дисперсия статистики $\tilde{s}_x^2$ меньше, чем любой другой оценки величины σ^2, в том числе и статистики s_x^2.

Однозначного решения о том, какая статистика является наиболее эффективной, априори не существует даже для оценки такого параметра, как μ_x. Выбор нужной статистики должен диктоваться сведениями или хотя бы соображениями о характере распределения.

Достаточность. Если статистика использует всю необходимую информацию, заключенную в выборке, то она называется достаточной. Так, для нормальной генеральной совокупности $\bar{x}$ — достаточная оценка μ_x, тогда как медиана ею не является, поскольку использует лишь ранги содержаний, но не их значения. Не является достаточной и середина размаха, так как использует лишь два (!) из n значений выборки. Очень важно, что достаточная статистика является асимптотически эффективной (эффективной в абсолютном смысле при безграничном увеличении выборки). Кроме того, если достаточная статистика является несмещенной, то любая другая не достаточная статистика будет менее эффективной.

Редко одна и та же статистика обладает всеми перечисленными «хорошими» свойствами, как например $\bar{x}$ (да и то лишь для нормаль-

---

[2] Возведением в квадрат неравенства $M(s) \leqslant \sigma$ получаем $M^2(s) \leqslant \sigma^2$, откуда $M^2(s) \leqslant M(s^2)$.

но распределенной генеральной совокупности). Обычно, обладая одними хорошими свойствами, она не обладает другими, присущими иной статистике. Например, величина $\tilde{s}_x^2$ является, как отмечалось, смещенной оценкой σ_x^2, но зато более эффективной, чем s_x^2. Поэтому при выборе статистик необходимо руководствоваться тем, какие из свойств ценнее для решения данной конкретной задачи.

§ 8. Основы проверки статистических гипотез

Конечной целью статистической обработки геохимических данных (но не конечной целью геохимических исследований!) является получение определенных заключений типа $A > B$, $A = B$, A зависит от B, X распределено по нормальному закону и т. д. Например, верно ли, что содержание свинца уменьшается по падению жилы; что зольность угля не изменяется с изменением степени углефикации; что содержание тория в этом интрузивном комплексе больше, чем в другом, и т. п. Заключения и формулировки такого рода не являются еще содержательными геохимическими гипотезами (хотя и могут быть практически полезными сами по себе), они относятся к так называемым статистическим гипотезам и подлежат проверке.

Процедура проверки гипотез включает следующие операции.

1. Формулировка *нулевой гипотезы*, обозначаемой $\mathbf{H}_0$. Проверяемые гипотезы могут касаться видов распределений, значений искомых параметров, их равенств или различий, а также других характеристик, не являющихся параметрами распределений. В этом последнем случае они называются непараметрическими. Наиболее частыми примерами гипотез, касающихся распределений, являются такие: распределение величины x следует нормальному закону (записывается $\mathbf{H}_0 : F(x) = \mathbf{N}$), логнормальному закону и т. д., или более общие: распределение $F(x)$ есть распределение $G(x)$ (записывается $\mathbf{H}_0 : F(x) = G(x)$). В большинстве случаев в геологии и геохимии альтернативными гипотезами являются *сложные гипотезы*, заключающиеся в том, что распределение может быть одним из многих других, за исключением распределения, составляющего нулевую гипотезу.

Гипотезы о параметрах могут быть, например, следующие: математическое ожидание случайной величины x равно нулю ($\mathbf{H}_0 : \mu_x = 0$) либо величине a ($\mathbf{H}_0 : \mu_x = a$), либо — математическому ожиданию величины y ($\mathbf{H}_0 : \mu_x = \mu_y$); коэффициент корреляции между изучаемыми переменными равен нулю ($\mathbf{H}_0 : \rho = 0$); дисперсии случайных величин данных совокупностей равны ($\mathbf{H}_0 : \sigma_1^2 = \sigma_2^2$) и т. д.

2. Выбор статистики, с помощью которой проверяется гипотеза. Под статистикой, как отмечалось, понимается некоторая величина, являющаяся функцией выборочных значений изучае-

мой переменной, т. е. вычисляемая из выборки. Статистики, с помощью которых проверяются гипотезы, называют также *критериями проверки гипотез*. Статистика, чтобы служить критерием проверки гипотез, должна обладать следующими свойствами: а) должен быть известен закон ее распределения в условиях справедливости нулевой гипотезы; б) ее распределение в указанных условиях не должно зависеть от неизвестных (проверяемых, искомых или других) параметров. Распределение наиболее употребительных критериев обычно табулировано.

Таблица 1

Соотношение ошибок первого и второго рода

Истинное положение	Нулевая гипотеза не отвергается	Нулевая гипотеза отвергается
Нулевая гипотеза истинна	Правильное заключение делается с вероятностью $1-\alpha$ (*статистическая достоверность*)	Ошибка первого рода (ее вероятность обозначают α и называют *уровнем значимости*)
Нулевая гипотеза ложна	Ошибка второго рода (ее вероятность обозначают β)	Правильное заключение делается с вероятностью $1-\beta$ (*мощность критерия*)

3. Выбор уровня значимости. Известно, что статистические выводы теоретически никогда не могут быть абсолютными. Соотношение между истиной и результатом проверки гипотезы сведено в табл. 1. Для ее легкого усвоения полезно такое мнемоническое правило: ошибка *первого рода* (истина отвергается, невиновный осуждается) есть «болезнь диктатора», ошибка *второго рода* (ложная гипотеза не отвергается, виновный не осуждается) — «потеря бдительности».

Вероятность ошибки первого рода (α) указывает на вероятность отвергнуть гипотезу, когда она верна; другими словами, — на относительную частоту, с которой будет приниматься альтернативная гипотеза, когда на самом деле верна нулевая.

Выбор уровня значимости при проверке гипотезы диктуется соображениями нестатистического характера. Например, если ошибка первого рода повлечет за собой лишь некоторый объем дешевого дополнительного опробования, то даже уровень значимости в 10% может считаться приемлемым. Но если в результате ошибки первого рода будет построено горнорудное предприятие стоимостью в миллиард рублей, то и 0.1%-й уровень значимости ($\alpha=0.001$) не может считаться достаточным.

При проверке гипотезы мы можем контролировать (назначить) лишь вероятность ошибки первого рода. Вероятность ошибки второго рода мы можем только констатировать и то лишь

по отношению к какой-либо конкретной простой альтернативной гипотезе: она зависит от неизвестных параметров и видов распределений, которые имеют место в действительности. Чрезвычайно важен вопрос, является ли доказанной нулевая гипотеза, если в результате статистической проверки она не отвергнута. В общем случае на этот вопрос надо ответить отрицательно. По аналогии с юриспруденцией такое положение можно назвать «презумпцией альтернативной гипотезы». Отвержение нулевой гипотезы, напротив, вполне доказывает справедливость альтернативной гипотезы, но поскольку, как мы видели, она почти всегда слишком всеобъемлюща, то практически этим мало что доказывается. Исключение составляют случаи, когда альтернативная гипотеза является простой. Приведем довольно искусственный пример: сфалерит на месторождении представлен только двумя генерациями, которые характеризуются устойчивым комплексом элементов-примесей. Определив их содержание в исследуемом зерне, с помощью конкретного статистического теста можно отвергнуть гипотезу о принадлежности зерна к первой генерации. Тем самым доказывается его принадлежность ко второй генерации (и то только с вероятностью $1-\beta$). Наиболее распространенная ошибка в толковании геологами результатов проверки гипотез как раз заключается в нарушении «презумпции альтернативной гипотезы».

§ 9. Специальные распределения, используемые при проверке статистических гипотез, их применение

Простейшим из таких распределений является распределение χ^2 («хи-квадрат»):

$$\chi^2 = \sum_{i=1}^{n} z_i^2, \qquad (1.9.1)$$

где z_i — независимые центрированные и нормированные, нормально распределенные (по закону N (0.1)) случайные величины $z_i = (x_i - \mu)/\sigma$. Распределение этой суммы квадратов нормированных нормальных величин зависит только от числа независимых слагаемых n, именуемого степенями свободы. Подробные таблицы распределения χ^2 с объяснениями к ним и примерами приведены у Я. Янко (1961). Если образовать величину $z_i = (x_i - \bar{x})/\sigma$, то в сумме (1.9.1) будут независимы только $f = n-1$ слагаемых. Одна степень свободы как бы «израсходована» на оценку $\bar{x}$. Применяя упоминаемое на стр. 31 мнемоническое правило определения числа степеней свободы, находим, что при одном ($n=1$) наблюдаемом значении случайной величины $\bar{x} = x_1$ разность $x_i - \bar{x}$ образовать нельзя, следовательно, должно быть $f = 0 = n-1$. Таким образом, действительно в этом случае $f = n-1$. Распределение χ^2 применяется при проверке гипотез очень широко. Рассмотрим два важных примера.

1. Выборочную оценку дисперсии можно преобразовать следующим образом:

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n-1}\,\sigma^2\sum\left(\frac{x_i-\bar{x}}{\sigma}\right)^2=\frac{\sigma^2\chi^2}{n-1}, \quad (1.9.2)$$

или

$$\sigma^2=\frac{s^2(n-1)}{\chi^2}=\frac{s^2}{\chi^2/f}.$$

Математическое ожидание χ^2/f равно единице, что еще раз иллюстрирует несмещенность оценки s^2, но оцениваемый параметр выступает теперь (при заданном выборочном s^2) как случайная величина, определяемая распределением χ^2. Отсюда легко получить для нее доверительный интервал

$$\frac{s^2(n-1)}{\chi^2_{\alpha/2}}<\sigma^2<\frac{s^2(n-1)}{\chi^2_{1-\alpha/2}}. \quad (1.9.3)$$

Искомый параметр лежит в нем с вероятностью $\mathbf{P}=1-\alpha$. Критические значения $\chi^2_{\alpha/2}$ и $\chi^2_{1-\alpha/2}$ для указанных уровней значимости и числа степеней свободы находят в таблицах (например, Янко, 1961, стр. 115).

2. Широко известно применение распределения χ^2 для проверки гипотез о принадлежности (непротиворечивости!) эмпирического распределения выбранным теоретическим — так называемый *критерий согласия Пирсона*. Он основан на сравнении эмпирических частот случайной величины в данных классах (интервалах) с теоретическими. Другими словами, сравнивается высота «столбиков» эмпирической гистограммы с высотой «столбиков» теоретической гистограммы, вытекающей из «подозреваемого» распределения. Если гистограмма состоит из k классов, то

$$\chi^2=\sum_{i=1}^{k}\frac{(n_i-np_i)^2}{np_i}, \quad (1.9.4)$$

где n — общее число наблюдаемых величин, n_i — их число в соответствующих классах i, p_i — теоретические вероятности попадания величины в заданные классы (вычисляются различными способами в зависимости от вида теоретического распределения). Нетрудно убедиться, что формула (1.9.4) является суммой квадратов нормально распределенных величин с нулевым математическим ожиданием $\mathbf{M}(n_i-np_i)=0$ и почти единичной дисперсией. Такое значение дисперсии определяется из биномиального распределения, которое здесь имеет место, и равно $np_iq_i=np_i(1-p_i)$, но при малых p_i, как и в законе Пуассона, дисперсия принимается равной среднему np_i.

Число степеней свободы f здесь равно $k-1$, так как частота в одном из классов не является независимой ввиду равенства $\Sigma n_i = n$. Кроме того, если мы по этим же эмпирическим данным оценим l параметров того распределения, с которым сравниваем свое, то число степеней свободы необходимо уменьшить на l. Итак, $f = k-1-l$. Сравнивая полученное значение χ^2 с табличным χ^2_α, заключаем о непротиворечивости аппроксимации с вероятностью $1-\alpha$, если $\chi^2 \leqslant \chi^2_\alpha$.

Распределение χ^2 применяется также для сравнения двух вероятностей по наблюденным частотам, существенности разности двух частот, равенства дисперсий и т. д.

Распределение Стьюдента (t-распределение). Известно, что линейные преобразования случайной величины не изменяют закона ее распределения, так что нормированная величина $z = (x-\mu)/\sigma$ или величина $z\sqrt{n}$ также распределена по нормальному закону, если величина x распределена нормально. Это позволяло по формуле (1. 5. 4) рассчитать для нее доверительные интервалы. Иначе обстоит дело с величиной t, нормированной по выборочной оценке стандартного отклонения: $t = (x-\bar{x})/s$. Используя равенство (1. 9. 2), получим

$$t = \frac{x-\bar{x}}{\sqrt{\sigma^2\chi^2/n-1}} = \frac{x-\bar{x}}{\sigma^2} : \sqrt{\frac{\chi^2}{n-1}} = \frac{z}{\sqrt{\chi^2/n-1}} = \frac{z}{\sqrt{\chi^2/f}}. \qquad (1.9.5)$$

Таким образом, величина t есть отношение величины, распределенной по $N(0, 1)$, к частному $\sqrt{\chi^2/n-1}$, где n — число значений, по которым определена выборочная дисперсия. Читателю уже ясно, что если вместо $\bar{x}$ использовать известное значение μ, то $f = n$. Несколько упрощая, можно сказать, что распределение t есть распределение нормальной случайной величины, нормированной по выборочному стандарту. Распределение величины (1. 9. 5) табулировано (Янко, 1961, стр. 121—123) и широко используется при проверке гипотез о средних значениях, на которых мы остановимся подробнее.

Простейшей процедурой является проверка гипотезы о среднем значении μ нормально распределенной генеральной совокупности с известной дисперсией σ^2_x. Этот случай редок в геохимической практике, так как трудно представить себе природное распределение, дисперсия которого была бы известна заранее. Соответствующая нулевая гипотеза $H_0 : \mu = \mu_0$ заключается в том, что исследуемая выборка принадлежит к нормальной совокупности с математическим ожиданием, равным μ_0. Для проверки необходимо вычислить статистику

$$z = \frac{\bar{x}-\mu_0}{\sigma_{\bar{x}}} = \frac{\bar{x}-\mu_0}{\sigma_x}\sqrt{n}. \qquad (1.9.6)$$

Величина z (если только нулевая гипотеза справедлива!) распределена нормально с нулевым средним и единичной дисперсией.

Задавшись определенным уровнем значимости α, по таблицам нормального распределения (лучше — по таблицам функции Лапласа) определяют такое (критическое) значение z_α, что вероятность появления большего абсолютного значения равна α. Если при этом $|z| \geqslant z_\alpha$, нулевая гипотеза отклоняется.

При проверке аналогичной гипотезы, но при заранее неизвестной дисперсии генеральной совокупности используется выборочная оценка дисперсии s_x^2 и рассчитывается почти аналогичная статистика

$$t = \frac{\bar{x} - \mu_0}{s_x} = \frac{\bar{x} - \mu_0}{s_{\bar{x}}} \sqrt{n}, \qquad (1.9.7)$$

распределение которой, как мы видели [9](1.9.5), отличается от нормального и тем больше, чем меньше объем выборки n. Для заданного α находят критическое значение $t_{\alpha, f}$. Существенное отличие от предыдущего случая заключается в том, что t_α различно для разных степеней свободы.

Правила для определения числа степеней свободы f при проверке различных гипотез разные. Здесь мы отметим общий принцип, заключающийся в том, что f должно обратиться в нуль, если число наблюдений (именно оно определяет в основном f) не позволяет вычислить параметр, входящий в статистику для проверки гипотезы. В данном случае таким параметром является $s = \sqrt{s^2}$. Минимальное число наблюдений, при котором еще можно вычислить этот параметр, равно двум, откуда $f = n - 1$, так как при $n = 1$ f по необходимости должно обратиться в нуль.

Рассмотрим построение доверительных интервалов для значения μ в последнем описанном случае. Требуется найти такое значение Δ_α, чтобы разность $\bar{x} - \mu$ не превышала его по абсолютной величине с вероятностью $1 - \alpha$ (т. е. превышала бы его с вероятностью, не большей α). Обозначив $\bar{x} - \mu = \Delta$, из формулы (1.9.7 найдем

$$\Delta_\alpha = \frac{t_{\alpha, f} s}{\sqrt{n}}. \qquad (1.9.8)$$

По известной заданной вероятности $\mathbf{P} = 1 - \alpha$ и числу степеней свободы f определяем значение $t_{\alpha, f}$, а затем и Δ_α. При этом можно утверждать, что $\mathbf{P}\{|\bar{x} - \mu| < \Delta_\alpha\} \geqslant 1 - \alpha$.

Иными словами, $\bar{x} \pm \Delta_2$ определит интервал возможных значений μ с вероятностью $\mathbf{P}$. Здесь нам представляется случай заметить, что построение доверительного интервала для μ, впрочем как и для большинства других параметров, есть та же проверка гипотезы, но производимая «в обратном» порядке. «Прямой» порядок предлагает задаться некоторым значением параметра $\Theta_0 = \hat{\Theta} + \Delta\Theta$ и определить статистику

$$t = \varphi(\hat{\Theta}, \Theta_0) \qquad (1.9.9)$$

с последующим сравнением ее значения с табличным $t_{\alpha, f}$ для данного уровня значимости α и числа степеней свободы f. При «обратном» порядке мы для данных α и f получаем табличное $t_{\alpha, f}$, а по нему с помощью функции, обратной (1. 9. 9), т. е., например, в нашем случае — по уравнению (1. 9. 8), определяем граничные (критические) значения параметра, которые он еще может принимать при данном уровне значимости.

Возможны еще две «обратные» задачи. Это определение α — уровня значимости, при котором мы еще можем считать приемлемой нулевую гипотезу, или как бы определение «критического» уровня значимости. Такой подход довольно часто встречается в геологических работах, хотя его нельзя считать вполне корректным. Например, нередко можно читать, что гипотеза о логнормальном распределении урана в гранитах «была бы справедлива вплоть до уровня значимости 0.05». Впрочем, такие «прикидки запаса прочности» гипотезы иногда могут быть полезными. Наконец, полезной обратной задачей является определение числа степеней свободы, при котором предположительно можно было бы проверить (подтвердить или отклонить) гипотезу на данном уровне значимости. В этой задаче проверка гипотез примыкает к проблеме определения минимального необходимого числа наблюдений для получения определенного заключения на заданном уровне значимости.

Распределение Фишера F является распределением отношения

$$F = \chi_1^2/f_1 : \chi_2^2/f_2. \qquad (1.9.10)$$

Математическое ожидание этого отношения, как и математические ожидания делимого и делителя, равно единице, так что распределение зависит только от числа степеней свободы (числа слагаемых, образующих χ_1^2 и χ_2^2). Оно подробно табулировано. Запишем снова выражение (1. 9. 2) для двух независимых оценок дисперсий (например, из разных выборок численностью n_1 и n_2)

$$s_1^2 = \frac{\sigma_1^2\chi_1^2}{n_1 - 1}, \qquad s_2^2 = \frac{\sigma_2^2\chi_2^2}{n_2 - 1}.$$

Если нулевая гипотеза заключается в том, что $\sigma_1^2 = \sigma_2^2$, то отношение

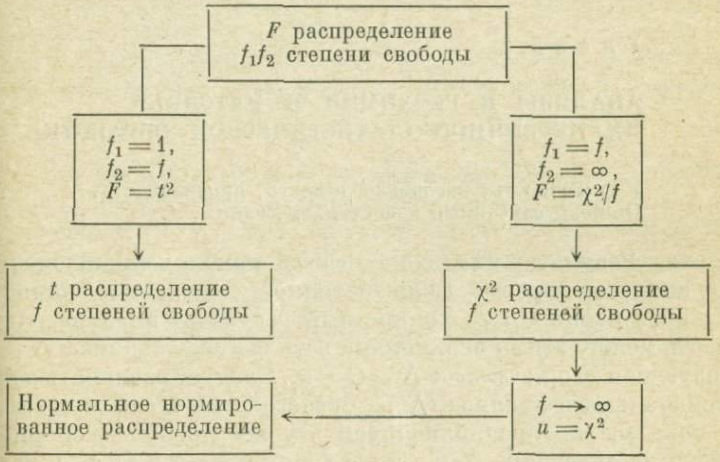
$$\frac{s_1^2}{s_2^2} = \frac{\sigma_1^2\chi_1^2}{n_1 - 1} : \frac{\sigma_2^2\chi_2^2}{n_2 - 1} = \frac{\chi_1^2}{n_1 - 1} : \frac{\chi_2^2}{n_2 - 1} = \frac{\chi_1^2}{f_1} : \frac{\chi_2^2}{f_2} = F, \qquad (1.9.11)$$

как видим, распределено по Фишеру.

Находим по таблицам (Янко, 1961, стр. 135—146) критическое значение $F_{f_1 f_2, \alpha}$ для заданного уровня значимости, и в случае $s_1^2/s_2^2 > F_{\text{кр}}$ *гипотеза о равенстве дисперсий отвергается* (бо́льшая дисперсия всегда в числителе!). Из (1. 9. 11) ясно также, что анало-

гичным образом можно проверять гипотезы типа $\sigma_1^2 : \sigma_2^2 = a$. Если выборочные дисперсии определялись в условиях известных математических ожиданий μ и μ_2, то $f_1 = n_1$ и $f_2 = n_2$.

В заключение представим распределения t, χ^2 и N (0, 1) как частные случаи распределения F. Действительно, если в (1.9.10) положить $f_1 = 1$, то окажется, что t^2 распределено как F с $f = f_2$ степенями свободы. Если $f_2 \to \infty$, то $\chi_2^2/f_2 = 1$ и окажется, что χ^2 распределено как $F \cdot f$ или $F = \chi^2/f$. Положив $f \to \infty$ и $u = \chi^2$, обнаружим, что u распределено по N (0, 1). Указанные соотношения сведены в следующую схему (по Доерфелю, 1969, стр. 63):



Глава 2

АНАЛИЗЫ В ГЕОХИМИИ И МЕТОДИКА ИХ ПЕРВИЧНОЙ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ

§ 1. Точность, воспроизводимость, правильность. Ошибки случайные и систематические

Результат анализа — всегда приближенная оценка состава вещества пробы. Если истинное содержание компонента в i-й пробе равно C_i, то конкретный результат анализа равен x_i. Разность между этими величинами есть ошибка анализа Δ_i, характеризующая его *качество*: $\Delta_i = C_i - x_i$. Следует различать качество индивидуального анализа (Δ_i) — своего рода «вещь в себе» до тех пор, пока мы не проанализируем эту же пробу бесспорно безошибочным способом, и качество метода, т. е. качество «среднего» анализа. В литературе сосуществуют термины, применяемые для характеристики качества анализов, в которые разными авторами вкладывается неодинаковый смысл: точность, воспроизводимость, правильность, а также заимствованное из англо-американской литературы понятие «повторяемость» и некоторые др. В дальнейшем мы будем считать, что а) точность анализа — мера общей его ошибки, т. е. воспроизводимость + правильность; б) воспроизводимость — мера отклонений отдельных результатов от среднего значения в серии повторных анализов одной и той же пробы (как правило, такой мерой является стандартное отклонение или дисперсия); в) правильность анализа — мера отклонений математического ожидания его результата от истинного содержания. В качестве «истинного» обычно используют либо искусственно приготовленное содержание в эталоне, либо его находят значительно более точным методом.

По смыслу ошибка воспроизводимости для геохимика является случайной. Ее распределение может быть различным, но чаще всего нормальным или логнормальным. Ошибка правильности наиболее часто выступает как ошибка систематическая. Однако разделение ошибок на случайные и систематические в значительной

степени условно и зависит, согласно В. В. Налимову (1960), от организации данного эксперимента. Например, если группа проб анализируется одним исполнителем в стандартных условиях, ошибки воспроизводимости можно отнести к случайным. Если же анализы выполняются двумя (несколькими) аналитиками или с использованием разных партий реактивов, можно говорить о систематических ошибках в одной группе анализов по сравнению с другой как о случайных. При постоянстве условий анализа систематическая ошибка — величина постоянная. Отклонения от истинного значения в каждом конкретном результате являются случайными величинами только потому, что упомянутая постоянная величина всякий раз складывается со случайной ошибкой воспроизводимости.

Таким образом, общую ошибку индивидуального анализа можно записать как сумму случайной и систематической компоненты:

$$\Delta_{\text{общ}} = \Delta_{\text{воспр}} + \delta, \qquad (2.1.1)$$

где $\Delta_{\text{воспр}}$ — случайная ошибка воспроизводимости, δ — постоянная величина систематического смещения.

Приведенное выше значение индивидуальной ошибки Δ_i обычно заменяют некоторой средней величиной $\bar{\Delta}$. Эта величина называется средней абсолютной ошибкой и получается усреднением всех индивидуальных отклонений Δ_i, взятых по абсолютной величине:

$$\bar{\Delta} = \frac{1}{n}\sum |\Delta_i|. \qquad (2.1.2)$$

На практике значительно удобнее пользоваться другой мерой ошибки — средней квадратичной ошибкой s_Δ, или дисперсией ошибки s_Δ^2. Поскольку Δ представляет собой сумму, s_Δ^2 выразится, согласно формуле (1.3.12), как $s_\Delta^2 = s_{\Delta\,\text{воспр}}^2 + s_\delta^2$. Величина каждого слагаемого будет равна:

$$s_{\Delta\,\text{воспр}}^2 = s_{\text{воспр}}^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2,$$
$$s_\delta^2 = \delta^2, \qquad (2.1.3)$$

где $\bar{x}$ — среднее в серии анализов одной пробы. Величина s_Δ^2 имеет существенно разный смысл в зависимости от того, вокруг какого центра отсчитывают отклонения Δ. Возможны два случая: 1) отклонения отсчитывают от среднего в серии n анализов одной пробы: $s_\Delta^2 = s_{\text{воспр}}^2$; 2) отклонения отсчитывают от «истинного» содержания — например, при многократных анализах эталона с некоторым номинальным содержанием C:

$$s_\Delta^2 = \frac{1}{n-1}\sum (x_i - C)^2. \qquad (2.1.4)$$

Поскольку $C = \bar{x} + \delta$, то $s_\Delta^2 = s_{\text{воспр}}^2 + \delta^2$. (2.1.5)

Таким образом, всегда выдерживается условие $s_{\Delta}^2 \geqslant s_{\text{воспр}}^2$, причем знак равенства имеет место только при отсутствии систематической ошибки ($\delta = 0$). Представление общей ошибки в форме (2.1.5) дает больше информации, чем в форме (2.1.4).

Квадрат величины смещения δ^2 по свойству аддитивности аналогичен дисперсии, что оказывается весьма удобным. Действительно, на практике величина смещения δ для данной лаборатории и данной методики анализа обычно остается неизвестной, она зависит от ряда факторов. Поэтому при анализах различных проб в разных лабораториях величина δ является случайной. В данном случае выражение (2.1.5) показывает, что величина дисперсии полной ошибки есть сумма двух дисперсий: дисперсии внутрилабораторной ошибки воспроизводимости $s_{\text{воспр}}^2$ и дисперсии межлабораторной ошибки δ^2. Далее будет показано, что вторая дисперсия, как правило, гораздо больше первой.

Таким образом, при обработке анализов одной лаборатории для статистического обоснования геохимических выводов необходимо использовать внутрилабораторную дисперсию воспроизводимости, подразумевая возможность присутствия некоторого постоянного смещения; при работах с использованием данных различных лабораторий — сумму внутрилабораторной и межлабораторной дисперсий. Это служит теоретическим обоснованием весьма полезной рекомендации: в процессе работы по одной теме пользоваться услугами одной и той же лаборатории.

При анализе геохимических объектов могут встретиться две ситуации. Одни аналитические методы дают ошибку воспроизводимости, которая зависит от концентрации определяемого компонента. Переход к относительной ошибке воспроизводимости, т. е. к величине $\Delta_{\text{воспр}}/x$, часто (например, в спектральном анализе) приводит к тому, что она становится постоянной для всего интервала концентраций. Воспроизводимость других методов, например весового химического анализа, не зависит от концентрации; в этом случае от нее зависит относительная ошибка. Так, кларковые содержания многих элементов-примесей определяются химическим анализом с относительными ошибками до 100%, а повышенные — с ошибками всего 10—15%. Следовательно, в большинстве случаев при анализе определенной группы проб имеет смысл говорить лишь о средней ошибке воспроизводимости. Для этого необходимо сначала проанализировать несколько серий проб с различными концентрациями и определить среднюю ошибку воспроизводимости по формуле, рекомендуемой, например, В. В. Налимовым (1960):

$$s_{\text{воспр}}^2 = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} (x_{ij} - \bar{x}_j)^2}{m(n-1)}, \qquad (2.1.6)$$

где n — число повторных анализов каждой пробы, m — число

проб, $\bar{x}_j$ — среднее значение по каждой пробе. Как видим, здесь вычисляется среднее арифметическое из дисперсий по отдельным сериям. Дисперсии по сериям получаются различными как раз оттого, что концентрации определяемого компонента в пробах не равны. Для правильного применения формулы (2. 1. 6) нужно выбирать концентрации по сериям так, чтобы отразить их распределение в партии анализируемых проб. Это не совсем удобно при анализе сильно отличающихся партий.

В геохимической практике для определения воспроизводимости обычно выборочный ряд проб анализируют дважды и дисперсию рассчитывают по формуле

$$s^2_{\text{воспр}} = \frac{1}{2(m-1)} \sum_{i=1}^{m} (x_{i1} - x_{i2})^2, \qquad (2.1.7)$$

где x_1, x_2 — результаты основного и повторного анализов, m — число двукратно анализированных проб. Если выполнено оговоренное условие — правильное отражение распределения частот разных концентраций, то формулы (2. 1. 6) и (2. 1. 7) будут давать одинаковые результаты. В то же время дисперсия воспроизводимости, рассчитанная по результатам многократного анализа одной пробы (см. (2.1. 3)), будет точной лишь настолько, насколько эта проба «типична» для всей совокупности.

Необходимо указывать, к а к о й п р о м е ж у т о к в р е м е н и характеризуют ошибки воспроизводимости, так как с течением времени условия анализов изменяются (табл. 2).

Таблица 2

Содержание ртути (в г/т), определенное двумя методами

Метод	Год анализа						
	1964	1965	1967	1968	1969	1970	1971
Нейтронно-активационный	340	—	245	120	70	—	—
Атомно-абсорбционный	—	130	—	—	—	97	80

Примечание. Составлено по данным «Mercury. . .», 1970.

Если учесть прецезионный характер примененных методов, то полученные результаты следует признать удручающими. Вместе с тем они вполне типичны для многих геохимических исследований.

§ 2. Внутренний контроль

Процедура, при помощи которой оценивают точность анализа, не прибегая к услугам других лабораторий, называется внутренним контролем. По традиции считают, что объектом внутреннего контроля являются только ошибки воспроизводимости.

Таблица 3

Содержание внутреннего контроля

Наименование	Назначение	Метод
1. Внутрилабораторный контроль:		
а) воспроизводимости	Выявление ошибки воспроизводимости, зависящей от факторов, быстро меняющихся со временем	Анализ параллельных проб
б) правильности	Выявление ошибки правильности для данного метода анализа	Периодический анализ искусственных эталонов («проб-свидетелей») и некоторые другие методы
2. Геологический — собственно внутренний контроль:		
а) воспроизводимости	Выявление ошибки воспроизводимости, зависящей от факторов, медленно меняющихся со временем	Анализ зашифрованных дубликатов проб
б) правильности	Выявление ошибки правильности для проб данного типа	Анализ зашифрованных эталонов (естественных или искусственных)

Однако в последние годы все шире используют внутренний контроль правильности анализов (то, что раньше относили только к внешнему контролю). Кроме того, как указал А. П. Прокофьев (1955), иногда геологи ошибочно считают процедуру внутрилабораторного контроля достаточной и не проводят специального геологического внутреннего контроля. Поэтому полезно привести общую типизацию задач, которые решает внутренний контроль (табл. 3). Следует помнить, что лаборатории, как правило, имеют вполне удовлетворительную сходимость параллельных определений (ошибка типа 1а), ибо методика не считается отработанной до тех пор, пока не добьются такой сходимости. Однако это нисколько не гарантирует малой ошибки типа 2а. Более того, поскольку в реальных условиях ошибка воспроизводимости для раз-

ных по составу проб неотделима от ошибки правильности (типа 2б, 1б), то внутренний контроль одной только воспроизводимости часто дает неудовлетворительные результаты: он не показывает, как повысить общую точность анализа.

Л. Е. Беренштейн и О. Б. Фалькова (1967) провели дисперсионный анализ ошибок воспроизводимости 19 спектральных и 5 химических методик в 15 лабораториях. Общая ошибка воспроизводимости $\sigma^2_{общ}$, характеризующая данную методику, расчленялась ими на три компоненты, вклад которых оценивался раздельно: $\sigma^2_{общ} = \sigma^2_в + \sigma^2_т + \sigma^2_с$, где $\sigma^2_в$ — мера рассеяния параллельных определений, выполняемых непосредственно друг за другом (источник ошибок — неконтролируемые факторы, быстро меняющиеся во времени); $\sigma^2_т$ — мера рассеяния определений по дубликатам зашифрованных проб, т. е. то, что охватывается понятием «собственно внутренний контроль» (источник ошибок — неконтролируемые факторы, медленно меняющиеся во времени); $\sigma^2_с$ — мера рассеяния определений по пробам различного валового состава (источник ошибок — несоответствие состава проб и эталонов). Оказалось, что в 73% рассмотренных случаев основной вклад в общую ошибку вносит слагаемое $\sigma^2_с$, причем существующая система внутреннего контроля не способна выявить эту ошибку. При этом, «одни и те же методики в руках разных аналитиков дали существенно различные величины ошибок. Это указывает. . . что унификация методик без метрологически грамотной и объективной системы контроля за работой лабораторий может оказаться мало эффективной» (Беренштейн, Фалькова, 1967, стр. 249).

Опытные аналитики в каждую партию проб вводят несколько лабораторных стандартов — детально изученных естественных проб. Если результаты анализов этих «проб-свидетелей» слишком отличаются от номинальных содержаний, это указывает на неконтролируемое изменение каких-либо условий анализа. Использование проб-свидетелей в течение нескольких лет, например в период работы над данной темой, позволит гораздо объективнее подойти к проблеме сопоставления анализов, полученных в разные годы. Пусть, например, в течение трех лет выполняются исследования по геохимии марганца в карбонатных породах. Первые же анализы показали исполнителю, что концентрации марганца лежат в широком интервале — от 0.001 до 1%. Тогда есть смысл отобрать десять проб, характеризующих весь интервал концентраций, тщательно проанализировать их точными методами, а затем регулярно включать в каждую партию анализируемых проб. Тогда к моменту обобщения данных накопится ряд анализов каждой пробы-свидетеля x_i, позволяющих по формуле (2. 1. 3) найти дисперсию воспроизводимости. Рассчитать общую дисперсию воспроизводимости, характеризующую всю изученную совокупность карбонатных пород или какую-либо партию проб, можно будет по формуле (2. 7. 2).

Иллюстрацией высказанных положений служит опыт работы Тематической партии Управления геологии Эстонской ССР (Орлова, Кивисилла, 1971). С помощью системы стандартных образцов осуществлялся постоянный внутренний контроль. Было установлено, что важным фактором, ухудшающим внутрисерийную воспроизводимость, является низкое качество спектральных фотопластинок. Следует заметить, что воспроизводимость анализа, определенная с помощью рекомендованной процедуры, получается несколько приукрашенной, ибо относится не к рядовым пробам, а к стандартам, которые как правило, более тщательно подготовлены к анализу.

Известно, что даже в случае полного соответствия состава пробы и эталонов воспроизводимость спектрального анализа остается довольно низкой. Причина заключается в том, что при фотографировании спектры проб и эталонов возбуждаются в разное время. Кроме того, они попадают на разные места фотопластинки, и вследствие неоднородности фотоматериала возникают дополнительные ошибки. Чтобы избежать их, В. Г. Тепляков (1964) разработал способ попеременного фотографирования спектров исследуемого вещества и эталона в процессе одной экспозиции и сконструировал аппаратуру, позволяющую автоматически получать спектрограммы, на которых каждый спектр пробы сфотографирован рядом со спектром эталона. Однако, как указывают Н. Г. Фридлендер и Г. Е. Юшкова (1972), недостатком этого метода является сложность применяемой аппаратуры, доступной не всем лабораториям.

Очевидно, что использование проб-свидетелей, позволяющее контролировать ошибку воспроизводимости, одновременно дает возможность контролировать и правильность анализа. Если разброс результатов анализа пробы-свидетеля симметричен вокруг номинального содержания, то можно считать, что анализ выполняется правильно; если же в какой-то серии анализов наблюдается закономерное занижение или завышение номинального содержания, то имеется ошибка правильности. Как уже указывалось, бóльшая доля ошибок воспроизводимости, а вместе с тем и основная часть ошибок правильности возникают вследствие несоответствия валового химического состава анализируемых проб (в том числе и проб-свидетелей) валовому химическому составу эталонов, или же несоответствия формы нахождения определяемого элемента в пробах и эталонах, по которым построен градуировочный график. Аналитики для улучшения правильности анализа в пределах данной лаборатории (не привлекая межлабораторные стандарты и внешний контроль) идут обычно следующим путем: а) изготавливают как можно больше эталонных коллекций («на все случаи жизни»), которые по валовому составу и молекулярной форме определяемого элемента идентичны с анализируемыми пробами; б) применяют (в спектральном ана-

лизе) сильное буферирование анализируемого вещества, что подавляет особенности индивидуального состава, и специальной предварительной обработкой (сплавлением) переводят все элементы в одинаковые молекулярные формы (Кибисов и др., 1961); в) по результатам анализов проб-свидетелей сдвигают параллельно самому себе градуировочный график, добиваясь, чтобы он давал номинальное содержание для пробы-свидетеля; [1] г) применяют одну серию эталонов, но с помощью особой процедуры метода добавок определяют величину поправки, которую нужно внести в результаты анализов данной партии проб (Огнева и др., 1972).

Все эти методы имеют свои достоинства и недостатки. Так, метод (а) весьма трудоемок, он хорош только для лабораторий узкого профиля, например рудничных, имеющих дело с небольшим числом основных типов проб. Метод (б) для спектрального анализа универсален, но связан с сильной потерей чувствительности вследствие разбавления пробы буфером. Он годится лишь для элементов, содержание которых не слишком мало, вдобавок, весьма трудоемок. Метод (в) опасен тем, что не всегда есть уверенность в возможности параллельного сдвига графиков, ибо при изменении состава проб (т. е. условий анализа) может измениться не только начало отсчета на графике, но и его угол наклона. Этим же недостатком отличается весьма остроумный метод (г). В целом можно отметить, что методы внутреннего контроля точности анализов находятся в стадии бурного развития. Пока лучше разработаны методы контроля воспроизводимости; для них имеются специальные инструкции, знание которых весьма полезно геохимику. Ниже приведены краткие сведения об одной из наиболее удачных инструкций.

Научным советом по аналитическим методам при ВИМС Министерства геологии СССР в 1968 г. была выпущена «Инструкция по внутрилабораторному контролю точности (воспроизводимости) результатов количественных анализов рядовых проб полезных ископаемых, выполняемых в лабораториях Министерства геологии СССР». Приведем основные положения названной инструкции, которые применимы, разумеется, не только к полезным ископаемым, но и к любым геохимическим объектам.

В основу оценки воспроизводимости анализов положено допущение о нормальном распределении ошибок воспроизводимости с математическим ожиданием, равным нулю. Критерием для браковки или приема партии анализов является величина *допуска* — числа проб этой партии, в которых расхождения между основным и контрольным анализами превышают заранее назначенную величину. Такие расхождения далее называются *промахами*.

[1] Отметим, что вообще статистически обоснованное использование градуировочных графиков — это самостоятельная проблема, более детально она разбирается нами в гл. 9, § 2.

Величину расхождения рассчитывают по формуле

$$П = \frac{(X_1 - X_2)}{\frac{1}{2}(X_1 + X_2)} \cdot 100\%, \qquad (2.2.1)$$

где X_1 и X_2 — результаты основного и контрольного определений.

Предельные величины относительных расхождений для каждого элемента и каждого интервала концентраций табулированы: они приближенно (при больших выборках) равны удвоенному относительному стандарту разности основных и контрольных определений $\left(2 \cdot \frac{s_\Delta}{x} \cdot 100\%\right)$ — удвоенному коэффициенту вариации этих разностей ($2V_\Delta\%$). Учитывая, что $s_\Delta^2 = 2s_{воспр}^2$, получаем $s_\Delta = \sqrt{2} \times s_{воспр}$, откуда $2V_\Delta = 2V_{воспр}\sqrt{2}$, где $V_{воспр}$ — коэффициент вариации, или относительная ошибка воспроизводимости. Например, если известно, что $V_{воспр}$ данного метода равна 10%, предельная величина расхождения между основным и контрольным анализами не должна превосходить $2\sqrt{2} \cdot 10\% = 28\%$.

Партия проб считается проанализированной правильно и принимается только тогда, когда число промахов в ней равно или меньше табулированного *приемочного числа*, полученного из биномиального распределения при заданном объеме контрольной выборки.

§ 3. Внешний контроль анализов при геохимическом опробовании и в геологоразведке

Задача внешнего контроля — выявление систематических ошибок в результатах анализов основной (контролируемой) лаборатории. Для этого некоторое число проб, по возможности представляющих весь интервал содержаний, из проанализированной в основной лаборатории партии (например, 20 из 130) направляют в другую лабораторию для анализа более точным методом. Таким образом, при внешнем контроле «истинным» (за исключением спорных случаев [1]) считают результат контрольного анализа. При проведении внешнего контроля можно выделить две ситуации: а) анализы контрольным методом отличаются от основных не только отсутствием систематической ошибки, но и пренебрежимо малой дисперсией воспроизводимости; б) контрольные анализы при несомненном отсутствии систематической ошибки имеют воспроизводимость одного порядка с основными. Ниже обсуждается преимущественно случай (а). Методы обработки случая (б) существенно сложнее. Отметим, что внешний контроль

[1] Пробы направляют на арбитражный анализ в третью лабораторию, которая считается более компетентной, чем контролирующая.

как средство исключения *систематической ошибки не обязательно* осуществляется в другой лаборатории.

В 1950—1960-х годах среди геологоразведчиков оживленно обсуждалась проблема математической обработки данных внешнего контроля (Юфа, 1951, 1954, 1958; Прокофьев, 1955, 1962; Смирнов и др., 1960; Раевский, Шурубор, 1958; Шашкин, 1956; Шарапов, 1954; Королев, 1969; Кельин, Михайлович, 1969; Родионов, 1964б, и др.). Предметом дискуссии были два вопроса: 1) как по данным внешнего контроля выявить наличие систематической ошибки; 2) если ошибка выявлена, как внести поправку в результаты основных анализов. В свое время Н. В. Барышев (1948) предложил использовать известную в математической статистике формулу критерия t (Стьюдента) для оценки существенности разности средних двух выборок с попарно сопряженными нормально распределенными случайными величинами:

$$t = \frac{\bar{x} - \bar{y}}{s_{\bar{\Delta}}} = \frac{\bar{x} - \bar{y}}{\sqrt{s_{\bar{x}}^2 + s_{\bar{y}}^2 - 2 s_{\bar{x}} s_{\bar{y}} \cdot r_{xy}}}, \qquad (2.3.1)$$

где $\bar{x}$ — среднее значение основных анализов n проб; $\bar{y}$ — то же, контрольных анализов; $s_{\bar{x}}^2$, $s_{\bar{y}}^2$ — дисперсии содержаний в контролируемой группе проб соответственно по основным и контрольным анализам; r_{xy} — коэффициент корреляции между рядами основных и контрольных анализов.[2] Этот метод выявления систематических ошибок вполне корректен, за исключением редких случаев, на которые будет указано далее. Кроме того, Н. В. Барышев предлагал (по-видимому, для простоты) систематическую ошибку считать незначимой, если $t < 2$. Правильно было бы сравнивать значение t со значением $t_{\alpha, f}$ для заданного α-процентного уровня значимости, $f = n - 1$ степеней свободы, находимым по таблицам распределения Стьюдента, и делать такое заключение при $t < t_{\alpha, f}$.

Значительная часть более поздних методов либо тождественны указанному, либо хуже. Так, Д. А. Родионов (1964б) предложил определять наличие систематической ошибки по критерию $t = \bar{\Delta}/s_{\bar{\Delta}}$, где $\bar{\Delta} = \frac{1}{n}\sum \Delta_i$, $\Delta_i = x_i - y_i$. Тождество с (2.3.1) очевидное,[3] но вычисление $s_{\bar{\Delta}}$ непосредственно по эмпирическим разностям безусловно более просто. К табличным значениям $t_{\alpha, f}$ Д. А. Родионов здесь также не обращается.

---

[2] При $r = 0$ формула (2.3.1) превращается в обычную для критерия t (однако в этом случае один из анализов или оба обладают недопустимо малой воспроизводимостью!).

[3] $$\bar{x} - \bar{y} = \frac{1}{n}\sum x_i - \frac{1}{n}\sum y_i = \frac{1}{n}\left(\sum x_i - \sum y_i\right) = \frac{1}{n}\sum (x_i - y_i) = \frac{1}{n}\sum \Delta_i = \bar{\Delta}. \qquad (2.3.2)$$

И. П. Шарапов (1954) предложил вычислять «более простую» статистику

$$k = \frac{\sum (y_i - x_i)}{\sum |y_i - x_i|}, \qquad (2.3.3)$$

в наших обозначениях равную $k = \bar{\Delta}/|\bar{\Delta}|$, где $|\bar{\Delta}|$ — среднее арифметическое значение абсолютных отклонений. Он предлагал считать, что систематическая ошибка есть при k от ± 0.6 до ± 1 и отсутствует при $k < \pm 0.6$. Позже В. П. Королев (1969) отметил, что при нормальном распределении разностей выполняется известное статистическое равенство $s_\Delta = 1.25\,|\bar{\Delta}|$. Заменяя в (2. 3. 3) величину $|\bar{\Delta}|$ величиной $s_\Delta/1.25$, он показал тем самым, какие значения k следует считать значимыми, т. е. превратил критерий k в критерий $t : k = 1.25\,t$. «Упрощение» метода И. П. Шарапова с усовершенствованием В. П. Королева по сравнению с методом Н. В. Барышева заключается в том, что разности Δ_i не возводятся в квадрат (!). Потеря же в статистическом смысле значительная: $s_\Delta = 1.25\,|\bar{\Delta}|$ является не точным, а статистическим равенством. Это ухудшает распределение статистики k по сравнению со статистикой t, что особенно сказывается при распределениях, отличающихся от нормального, так как при этом указанное равенство не выполняется. Некоторые другие авторы, в основном работники Государственной комиссии по запасам (ГКЗ), вообще высказались против статистических расчетов при выявлении систематической ошибки: «В подавляющем большинстве случаев при наличии систематической погрешности. . . видно явное преобладание одних знаков (разностей, — *Ю. Т., Я. Ю.*), указывающих на наличие этой ошибки, и подтверждать ее какими-либо специальными расчетами нет необходимости» (Прокофьев, 1955, стр. 28).[4] Мерой систематической ошибки они (Прокофьев, 1955, 1962, и др.) считают среднее арифметическое отклонение $\bar{\Delta}$, мерой случайной ошибки — среднее арифметическое абсолютных значений отклонений $|\bar{\Delta}|$, т. е. стоят по существу на позициях И. П. Шарапова, который оперирует отношением этих величин.

Если желательно максимальное упрощение расчетов, то для выявления систематической ошибки можно рекомендовать критерий знаков Ван дер Вардена (1960), для которого требуется найти лишь число положительных k и отрицательных $n-k$ отклонений. Нулевые разности, т. е. совпадения основного и контрольного анализов, отбрасываются и в общее число наблюдений не входят. Критические пределы k находят по таблицам (Ван дер Варден,

[4] Конечно это справедливо, если иметь в виду примеры, аналогичные ставшему классическим (благодаря его перепечатке почти всеми авторами) примеру из работы Н. В. Барышева (1948), где из 44 разностей 43 имеют знак минус!

1960, табл. 9, стр. 416). Этот критерий хорош не только предельной простотой, но и тем, что не требует никаких (кроме непрерывности) предположений о распределениях погрешностей. Однако при нормальном распределении эффективность этого критерия составляет 0.64 критерия t, т. е. для заключения с одинаковой значимостью потребуется почти вдвое больше наблюдений. Заметим, что известный из геологической литературы (Крейтер, 1960, стр. 271) так называемый способ Б. И. Галкина является ни чем иным, как критерием знаков, представленным в виде номограммы. Входами в номограмму служат: объем выборки проконтролированных проб (n) и разность между числом отклонений разного знака (s). По этим данным находится вероятность (p) того, что расхождения являются не случайными.

Полагая, что при отсутствии систематической ошибки вероятности положительных и отрицательных отклонений одинаковы, можно воспользоваться критерием согласия Пирсона χ^2:

$$\chi^2 = \sum \frac{(np_i - np)^2}{np} = \sum_{i=1}^{2} \frac{(np_i - 0.5n)^2}{0.5n} \qquad (2.3.4)$$

(использование χ^2 подробно изложено в § 1.9). В отличие от критерия t здесь не требуется нормальности распределения разностей, лишь бы вероятности положительных и отрицательных погрешностей были равны, что важно при сравнении спектральных анализов с химическими: в каком бы масштабе мы ни выражали погрешности — логарифмическом или простом, — одно из распределений погрешностей не будет симметричным, и разности не будут распределены нормально. Вообще заметим, что при большом числе контрольных анализов n средняя разность $\bar{\Delta}$ всегда будет распределена нормально.

Что касается способа введения поправки в результаты основных анализов, то Н. В. Барышев, а затем А. П. Прокофьев (1955), В. И. Смирнов и А. П. Прокофьев (Смирнов и др., 1960) предлагали умножать результаты основных анализов на коэффициент $k_a = \bar{y}/\bar{x}$. Однако И. П. Шарапов (1954) постоянно указывал, что поправки для разных классов содержаний могут иметь различную величину, и предлагал поэтому вычислять поправочный коэффициент раздельно для разных классов содержаний. В. Л. Шашкин (1956) предлагал выборку для контрольных анализов составлять не только по разным типам руд и разным классам содержаний, но каждую контрольную пробу составить из вещества многих основных проб. В этом последнем есть определенное преимущество (не отмеченное автором), заключающееся в том, что случайная погрешность основного анализа в сборной пробе (т. е. погрешность среднего содержания, рассчитанного по данным основного анализа составляющих частных проб) будет в $\sqrt{m}$ раз меньше, где m — число частных проб. Это значительно повышает эффективность

статистических критериев выявления систематической ошибки и уточняет переводной коэффициент.

Д. А. Родионов (1964б) высказался против применения коэффициента k_a, так как введение такой поправки искажает дисперсию исследуемых концентраций $s^2_{x'}=k^2_a s^2_x$, что «весьма нежелательно». Однако поскольку основные анализы содержат систематическую ошибку, параметры распределения (среднее и дисперсия) также неверны и нуждаются в исправлении. Разногласия между сторонниками умножения на переводной коэффициент и сторонниками прибавления постоянного смещения, равного $\bar{y}-\bar{x}=\bar{\Delta}$ (Д. А. Родионов), не могут быть решены перечисленными выше средствами, ибо они упираются в вопрос, постоянна ли систематическая погрешность (в этом случае надо прибавлять) или пропорциональна содержанию (надо умножать).

Эти вопросы могут быть решены с помощью исследования уравнения регрессии между результатами основных и контрольных анализов. Впервые такое предложение выдвинул Б. Я. Юфа (1951). По результатам основного x и контрольного y анализов он рассчитывал уравнение регрессии x на y и поправки в значения x (не проконтролированные) вносил непосредственно по графику регрессии x на y: по заданному значению x_i сразу определялось значение $x_{испр}$. Это предложение подвергалось неквалифицированной критике (Шарапов, 1954; Шашкин, 1956), сводившейся в основном к тому, что нельзя исправлять (корректировать по уравнению регрессии) индивидуальные значения x. Однако именно в этом и заключается достоинство метода (усреднить уже исправленные значения всегда можно!), избавляющее от необходимости строго следить за репрезентативностью контрольной выборки и корректировать k по мере изменения эмпирических частот содержаний в исследуемом объекте. Главное же преимущество методики Б. Я. Юфы заключается в выявлении *раздельно* постоянной и пропорциональной составляющих систематической погрешности $x=by+a$, где a — постоянная составляющая систематической ошибки, которую надо прибавлять, b — пропорциональная часть, на которую надо умножать! Использование графиков регрессии означает внесение обеих поправок.

В. И. Раевский и Ю. В. Шурубор (1958) усовершенствовали метод Б. Я. Юфы, предложив оценивать статистическую значимость параметров b и a в уравнении регрессии путем проверки гипотез $H_0: b=1$ и $H_0: a=0$. К сожалению, в формулах статьи имеются опечатки или ошибки, не позволяющие рекомендовать их читателю, при полной правильности идеи.[5] В 1969 г. А. М. Кельин и М. П. Михайлович, по-видимому, не знавшие работ Б. Я. Юфы и В. И. Раевского, Ю. В. Шурубора, подвергли критике применение коэф-

[5] Формулы для углового коэффициента b, для ошибки уравнения регрессии (σ) и для критериев t_a, t_b в статье В. И. Раевского и Ю. В. Шурубора ошибочны.

фициента k_a, поскольку он характеризует только переменную долю систематической ошибки, тогда как в ней имеется и некоторая постоянная доля. Авторы предложили разбивать ряд значений, предварительно ранжированный по содержаниям, на две примерно равные части и находить поправки a и b, решая систему уравнений

$$\bar{x}_1 = b\bar{y}_1 + a,$$

$$\bar{x}_2 = b\bar{y}_2 + a,$$

где $\bar{x}_1$, $\bar{x}_2$ — средние значения в первой и второй половине ряда основных анализов (контрольной выборки); $\bar{y}_1$, $\bar{y}_2$ — то же, контрольных анализов. Более подробно процедура регрессионного анализа и его применения для выявления и устранения систематических погрешностей и построения градуировочных аналитических графиков изложена в гл. 7—9.

§ 4. О соотношении внутрилабораторных ошибок воспроизводимости и межлабораторных ошибок правильности

Представление о том, что случайными являются только внутрилабораторные ошибки воспроизводимости, нанесло значительный вред и аналитикам и геохимикам, использующим их данные. «Грубое и метрологически неоправданное деление аналитических ошибок только на две категории — внутрилабораторные ошибки воспроизводимости и «систематические» ошибки — привело к тому, что объектом применения математической статистики в аналитической работе до сих пор часто оказываются внутрилабораторные ошибки воспроизводимости» (Налимов, 1960, стр. 20—21). Совокупность анализов стандартных образцов, выполненных в разных лабораториях, необходимо рассматривать как *статистический ансамбль* и применять к таким данным средства математической статистики. Х. У. Фэрберн и его группа (Fairbairn et al., 1951) опубликовали результаты изучения межлабораторных ошибок силикатного анализа горных пород. Эти результаты, как заметил Д. М. Шоу (1969, стр. 29), «оказались самыми важными для геохимии и петрологии за последние сто лет». Как отмечают В. Г. Хитров и Р. В. Кортман (1969, стр. 1), «результаты этой работы, а также специальных исследований точности и правильности силикатного анализа горных пород. . . оказались неожиданными для химиков. . . и вызвали тревогу и озабоченность геохимиков и петрографов. . .». Эту озабоченность легко понять, ибо оказалось, что точность оценки состава породы по результатам громоздкого и дорогого силикатного анализа в соответствии с результатами Фэрберна ничуть не выше, чем с помощью простого подсчета минералов в шлифе!

Надежным путем, обеспечивающим правильность анализа, является широкое внедрение в практику работ аналитиков тщательно изученных *стандартов*, отвечающих по своему составу типичным горным породам и рудам. В настоящее время в ряде стран по примеру работников геологической службы США, впервые изготовивших известные стандарты гранита *G*—1 и диабаза *W*—1, изготовлены и запущены в анализ многочисленные стандарты.

В 1965 г. ИГЕМ АН СССР разослал в 84 лучшие геологические лаборатории нашей страны семь порошкообразных стандартных образцов: 1003 — гранодиорит «Рыжик», 1001, 1005 — миаскит МИВ-1, 1004, 1006 — диабаз ДИМ-1, 1002, 1007 — перидотит ПИМ-1. Каждая лаборатория выполняла анализ на предложенные компоненты своим апробированным методом. В результате статистической обработки полученной информации В. Г. Хитров и Р. В. Кортман (1969) пришли к следующим важнейшим для геохимиков выводам.

1. Если содержания определяемых компонентов в горных породах лежат ниже 10—20 вес.%, то точность анализов в большинстве лабораторий неудовлетворительна. При содержаниях ниже 1—2% решающее значение приобретают именно м е ж л а б о р а т о р н ы е расхождения. Отклонения в большей степени зависят от концентраций анализируемых элементов, чем от различия применяемых методик и типов пород.

2. Ошибки количественных (!) анализов могут достигать 100% вследствие решающего вклада ошибок правильности. Всего 1/4 всех полученных результатов характеризуется приемлемой общей точностью (например, 30 отн.%). Остальные количественные анализы (3/4 всех результатов!) не точнее полуколичественных (ошибки 30—50 отн.% и выше). В процессе массовых (а не специальных, особо точных) анализов при сохранении существующего положения получение правильных результатов невозможно. В н у т р и л а б о р а т о р н ы й к о н т р о л ь в о с п р о и з в о д и м о с т и с е б я н е о п р а в д ы в а е т, поскольку таким путем контролируется не более 1/5—1/3 полной ошибки, а внешний контроль правильности слишком эпизодичен и поэтому явно не эффективен.

3. Полуколичественные методы, примененные грамотно и экономически оправданно, особенно если удается улучшить их воспроизводимость, — отнюдь не хуже при существующем положении дела многих «количественных» методов.

4. Для обеспечения правильности анализов необходимо обеспечить все лаборатории едиными естественными стандартами хотя бы важнейших типов горных пород.

Приведенные выводы еще раз показывают условность понятия «ошибка воспроизводимости» и его изменение с изменением рассматриваемой ситуации. Если выводы геохимика основаны

на сравнении результатов анализов, выполненных по одной методике в одной лаборатории, статистическая значимость их определяется ошибкой воспроизводимости в узком смысле. При обобщениях материалов других авторов и лабораторий, а также при обосновании выводов, основывающихся на абсолютных значениях содержаний, важнейшее значение приобретают ошибки правильности анализов.

§ 5. Особенности обработки, связанные с наличием нижнего предела определений (порога чувствительности)

Нижним пределом определений, или порогом чувствительности метода анализа, называют такую величину концентрации, меньше которой данным методом определить нельзя. При содержаниях, примерно равных порогу чувствительности, обычно выдают значение «следы», а при содержаниях ниже порога — «не обнаружено». С наличием порога чувствительности связан ряд вопросов, не получивших еще строгого решения. При этом подход аналитика и геохимика к данной проблеме не одинаков. Аналитика интересует так называемый *предел обнаружения*. Поскольку измеряемый аналитический сигнал (вес осадка, объем титрующего раствора, интенсивность спектральной линии и т. д.) имеет статистический характер, то порог чувствительности нельзя себе представить в виде минимального содержания (или количества вещества, элемента), ниже которого аналитический сигнал никогда не воспринимается. Предел обнаружения — такое наименьшее количество элемента в анализируемой пробе, при котором аналитический сигнал с заданной статистической достоверностью может быть отличен от сигнала холостой пробы (Зильберштейн, 1971, стр. 15). Если сигнал холостой пробы $a_{хол}$ характеризуется стандартным отклонением $\sigma_{хол}$, а сигнал пробы с минимальным обнаруживаемым содержанием — $\sigma_{мин}$, то приведенное выше определение предела обнаружения запишется следующим образом: $a_{мин}=a_{хол}+u_1\sigma_{хол}+u_2\sigma_{мин}$, где $a_{мин}$ — величина аналитического сигнала, соответствующего пределу обнаружения элемента; u_1 и u_2 — аргументы функции нормального распределения, соответствующие заданным вероятностям p_1 и p_2. Эти вероятности, по замыслу привлекших их к данному вопросу Х. Кайзера и Х. Зильберштейна, являются вероятностями известных в математической статистике ошибок первого и второго рода, т. е. p_1 — вероятность принять сигнал холостой пробы за аналитический сигнал, p_2 — вероятность принять аналитический сигнал за сигнал холостой пробы.

Геохимик, обрабатывающий результаты анализов, считает предел обнаружения известным; его интересуют искажения, связанные с тем, что часть проанализированных проб попадает в разряд «не содержащих» данного элемента (рис. 6).

Влияние порога чувствительности на результаты анализов вынуждает геохимиков в своих работах часто приводить не средние содержания, а процент встречаемости данного элемента. Эта величина зависит от среднего содержания и порога чувствительности: чем ниже среднее содержание, тем обычно ниже и встре-

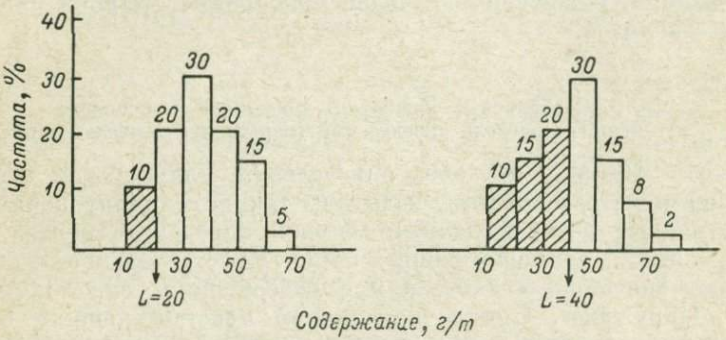


Рис. 6. Два усеченных распределения.
L — порог чувствительности.

чаемость. Однако геохимическая литература наводнена статьями, где приведены данные о низком проценте встречаемости элемента в пробах и одновременно вычислены средние содержания. Как правило, в этих случаях бывает ясно, что среднее содержание рассчитано только по тем пробам, в которых элемент был обнаружен. Такие средние не имеют ничего общего с действительностью.

Если среди выборки проб имеются «пустые» пробы, то нередко пользуются различными эмпирическими приемами при расчетах среднего и дисперсии. Чаще всего используют один из трех приемов.

1. Содержания «пустых» проб принимаются за нуль и включаются в расчеты среднего и дисперсии. Тогда дисперсия искусственно завысится, ибо замена низких содержаний нулем увеличивает сумму отклонений от среднего. Среднее значение занижается.

2. Пробам с подпороговыми и пороговыми содержаниями приписывают значение порога (например, 0.0001%) или его половины, третьей части и т. д. В этом случае дисперсия занизится, среднее содержание завысится.

3. Анализы с пороговыми и подпороговыми содержаниями исключают из подсчета, расчет ведут только по оставшимся пробам. Среднее получится завышенным, а дисперсия заниженной, поскольку размах колебаний стал меньшим.

Перечисленные эмпирические приемы (имеются и другие, менее употребительные) не обеспечивают получения несмещенных

и эффективных оценок параметров природных распределений. Для получения правильных результатов необходимо, хотя бы приближенно, знать закон распределения изучаемых концентраций. Для важного частного случая, когда этот закон не противоречит нормальному, несмещенные оценки параметров исходного природного распределения можно определить по методу Р. А. Фишера (Беус и др., 1965; Янко, 1961).

Будем различать *усеченное* и *неполно определенное* распределение (выборку). Для первой характерно не только отсутствие значений содержаний ниже порога L, но и вообще отсутствие информации о доле подпороговых проб h в выборке. Для неполно определенной выборки эта доля известна.

В случае усеченной выборки поступают следующим образом. Составляют разности между i-ми значениями содержаний в пробах и пороговыми значениями $d_i = x_i - L$, по которым вычисляют статистику $y = \frac{1}{2} N \left[\sum d_i^2 \Big/ \left(\sum d_i \right)^2 \right]$. Затем по таблицам односторонне усеченного нормального распределения (Янко, 1961, табл. 17) находят величины $z\,(y)$ и $g\,(z)$, дающие возможность найти несмещенные оценки параметров

$$s = \left(\frac{1}{N} \sum d_i \right) \cdot g(z), \qquad (2.5.1)$$

$$\bar{x} = L - z \cdot s, \qquad (2.5.2)$$

а также оценку степени усечения $h = \Phi(z)$ по таблицам нормального распределения.

Гораздо чаще в геохимии встречаются неполно определенные выборки, в которых известна доля проб с подпороговыми содержаниями. Методика оценки параметров здесь аналогична, но использование дополнительной информации (параметр h) увеличивает эффективность оценок. Значения $z = f(y, h)$ и $g(z, h)$ находят по таблице (Янко, 1961, табл. 18).

Указанные методы применимы также, если выборка усечена сверху. Н. К. Разумовский (1962) использует их для оценки параметров фоновых распределений, справедливо считая, что среди проб с содержанием выше некоторого порога могут оказаться пробы из геохимических аномалий.[1] Г. А. Вострокнутов (1969) предлагает еще один метод оценки параметров неполностью определенной выборки (которую он ошибочно отождествляет с усеченной выборкой). Зависимость между стандартным отклонением полного s и усеченного s_y нормального распределения автор метода эмпирически аппроксимирует кривой $s = s_y (0.6 + e^{0.406\, z_y})$, где z_y — нормированное значение точки усечения (порога), которое находят

[1] Можно доказать, что примененный Н. К. Разумовским метод Ли и Пирсона полностью тождествен методу Фишера.

как аргумент функции нормального распределения, соответствующий значению $F = \frac{n_y - n}{n}$ (n — общее число проб, n_y — число

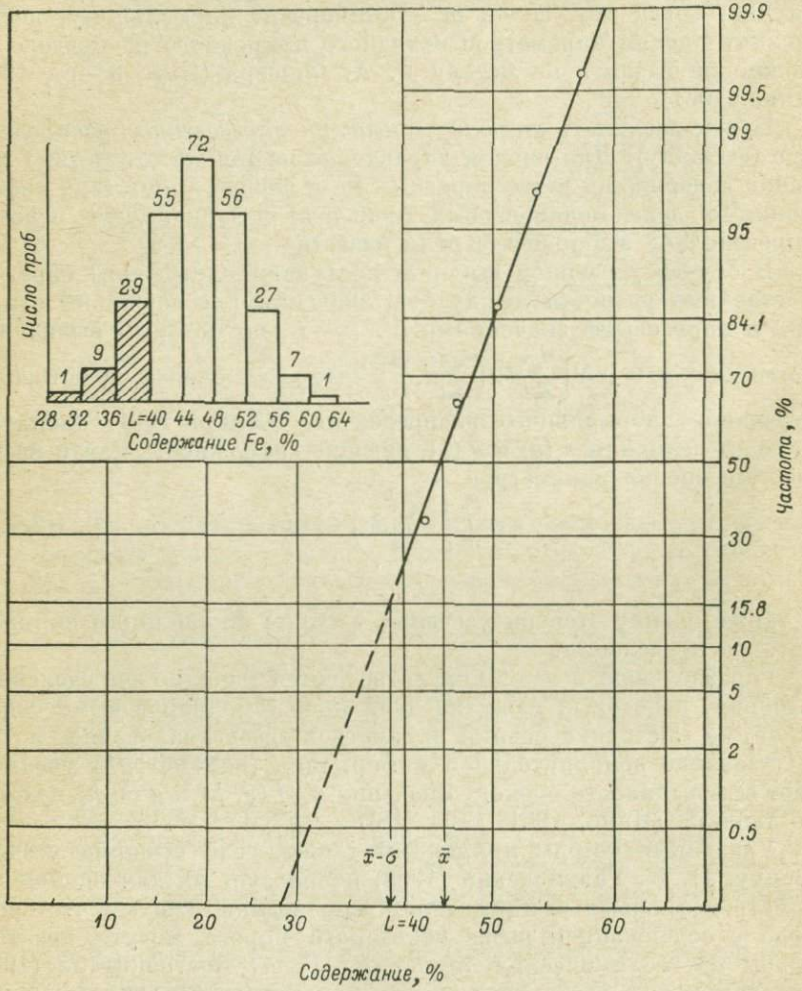


Рис. 7. Построение прямой Генри для нормального распределения.

Пример взят у И. П. Шарапова (1965, стр. 90); параметры: $\bar{x}=46$, $s=5.64$. При назначении порога, например $L=40$, значения параметров $\bar{x}$ и s, снятые с графика, хорошо согласуются с фактическими.

проб выше точки усечения). Среднее значение определяется по формуле (2. 5. 2). Эффективность оценок зависит прежде всего от того, насколько исследуемое распределение близко к нормальному. Она падает, кроме того, с увеличением степени усече-

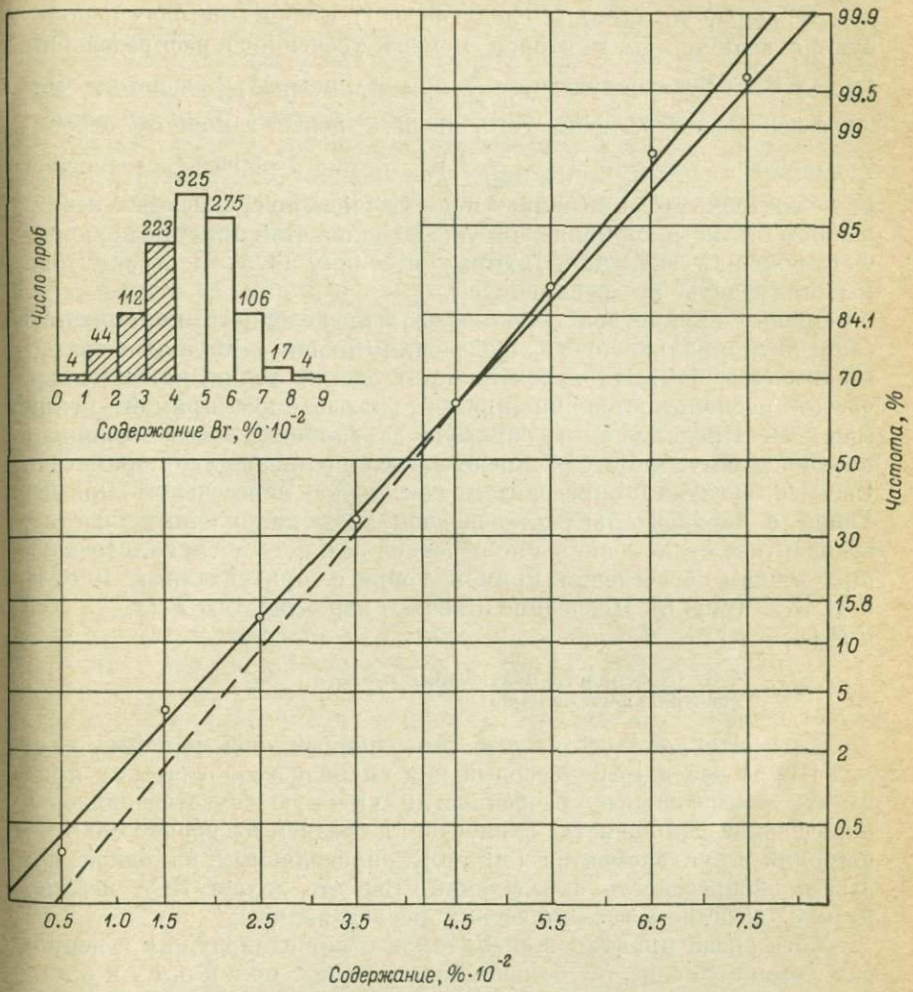


Рис. 8. Построение прямой Генри для распределения, не являющегося нормальным.

Пример взят у И. П. Шарапова (1965, стр. 89); параметры: $\bar{x}=4.363$, $s=1.357$. При назначении порога, например $L=4$, значения параметров $\bar{x}$ и s, снятые с графика, хуже согласуются с фактическими. Прерывистая линия — продолжение прямой Генри «за порог», сплошная — фактическая прямая Генри.

ния. Что касается замечания Г. А. Востроκнутова о большей эффективности предложенного им метода по сравнению с методом Фишера, то заметим следующее: 1) в обоих методах используются величина h и второй момент усеченного распределения (s_y — у Г. А. Востроκнутова, $\frac{1}{n}\sum d_i^2$ — у Фишера); 2) в методе Фишера используется, кроме того, квадрат первого момента относительно точки усечения $\left(\frac{1}{n}\sum d_i\right)^2$; 3) таким образом, статистика Г. А. Востроκнутова не является достаточной по сравнению с несмещенной статистикой Фишера и уже по одной этой причине не может быть более эффективной. Поэтому методику Г. А. Востроκнутова мы применять не рекомендуем.

Оценку параметров неполностью определенного нормального (или логнормального) распределения проще выполнять графическим способом. На вероятностном бланке точки накопленных частот располагаются на прямой, называемой прямой Генри (рис. 7). Продолжая эту прямую за «порог», восстанавливают предполагаемую форму распределения и оценивают его параметры. Среднее значение определится как точка пересечения прямой Генри с горизонталью, отвечающей 50% накопленных частот, стандартное отклонение — по половине разности абсцисс, отвечающих точкам пересечения прямой Генри с горизонталями 15.87% и 84.13% (рис. 8). Последние отвечают вероятностям $P\{x \geqslant -\sigma\}$ и $P\{x \leqslant \sigma\}$.

§ 6. Особенности полуколичественного спектрального анализа

Полуколичественный эмиссионный спектральный анализ (ПСА) как самый массовый вид анализа геологических проб имеет замечательные особенности: высокую чувствительность определения большинства элементов, в среднем на уровне 0.001%, широкий круг элементов (30—40), определяемых из одной навески, экспрессность определений. Все это делает ПСА незаменимым средством геохимических исследований.

Хотя среди практиков-аналитиков и нет единодушия в вопросах терминологии, но большинство из них под полуколичественным спектральным анализом понимает такой метод, в котором оценка интенсивности линий производится без фотометрирования и, как правило, без использования внутреннего стандарта.

Основное уравнение спектрального анализа, связывающее интенсивность спектральных линий J с содержанием элемента в пробе, имеет следующий вид:

$$J = a \cdot x^b, \qquad (2.6.1)$$

где a — параметр условий возбуждения спектра; b — параметр, зависящий от поглощения излучаемой волны в газовом облаке

источника света (Лонцих и др., 1967). Логарифмирование выражения (2. 6. 1) дает уравнение прямой: $\lg J = b \cdot \lg x + \lg a$, откуда следует, что если ошибки в оценке интенсивности спектральных линий распределены нормально, то ошибки оценок концентраций будут распределены логарифмически-нормально. С этой особенностью спектрального анализа связано завышение среднего арифметического из нескольких параллельных определений одной и той же пробы по сравнению с истинным значением концентрации в ней. Правильной оценкой при нормальном распределении ошибок логарифмов концентраций будет мода на кривой распределения логарифмов, или среднее геометрическое.

Другая особенность ПСА — дискретный характер выдаваемых результатов, обычно в форме условных единиц — баллов. Именно вследствие логарифмической формы зависимости между интенсивностями и концентрациями наиболее удобно (и сейчас общепринято) такое деление шкалы содержаний на баллы, когда последние отвечают серединам равных логарифмических интервалов. Как показал опыт, применение баллов вместо весовых процентов дает большие технические преимущества при обработке анализов (Юдович и др., 1970). Такая возможность основана на существовании линейной функциональной связи между баллами b и концентрациями x:

$$b = a \cdot \lg x + b_0, \qquad (2.6.2)$$

где a — число частей, на которое поделен один порядок концентрации («кратность шкалы»); b_0 — балл, соответствующий 1%-й концентрации. Получив параметры распределения повторных анализов — среднее $\bar{b}$ и дисперсию s_b^2 в баллах, можно теперь «вернуться» к весовым процентам, используя следующие формулы:

$$\lg x = \frac{\bar{b} - b_0}{a}, \qquad (2.6.3)$$

$$\tilde{x} = 10^{\overline{\lg x}}, \qquad (2.6.4)$$

$$s^2_{\lg} = \frac{s_b^2}{a^2}, \qquad (2.6.5)$$

где $\tilde{x}$ — среднее геометрическое.

Наконец, пользуясь баллами, можно сразу определить важнейшую характеристику анализа — относительную ошибку воспроизводимости для концентраций в обычной форме для весовых процентов:

$$V_x, \% \approx 1.5 \cdot \frac{s_b}{a} \cdot 100. \qquad (2.6.6)$$

Важной особенностью ошибки воспроизводимости ПСА является ее нелинейный характер: она указывает н е н а с к о л ь к о единиц искажено истинное содержание, а в о с к о л ь к о р а з

найденный результат отличается от истинного значения. Поэтому обычная запись, справедливая для нормального распределения ошибок ($\bar{x} \pm s_x$, $\bar{x} \pm 2s_x$ и т. д.), здесь оказывается неверной. Переходя к концентрациям (в вес.%), получим, что интервалы вправо и влево от $\bar{x}$ окажутся неравными (шире в абсолютных содержаниях справа и у́же — слева).

Оперируя с баллами, легко сравнивать результаты разных лабораторий, даже если они использовали различные варианты деления шкалы (табл. 4). Из (2. 6. 3) получаем

$$a/a' = \frac{b' - b'_0}{b - b_0}, \qquad (2.6.7)$$

где a и a' — кратности разных шкал, а b_0 и b'_0 — баллы, соответствующие содержанию 1% в этих шкалах. Если в одной лаборатории порядок делят на три части ($a=3$), а в другой — на четыре ($a=4$) и в ней получили результат 10 баллов, то «на языке» первой лаборатории это составит 7.5 балла (рис. 9). Подобные переходы в шкалах разной кратности приходится часто применять при внешнем контроле и объединении анализов разных лабораторий для расчетов средних. Для сопоставления воспроизводимости результатов, выдаваемых различными лабораториями, необходимо ошибку воспроизводимости в баллах перевести в ошибку воспроизводимости логарифмов по формуле (2. 6. 5). Из предыдущего изложения ясно, что дисперсия воспроизводимости анализа для весовых процентов (не для логарифмов!) зависит от самих определяемых концентраций:

$$s_x^2 \approx \frac{\hat{x}^2 \cdot s_{\lg}^2}{0.43} \approx \frac{\hat{x}^2 \cdot s_b^2}{0.43 \cdot a^2}, \qquad (2.6.8)$$

где s_x^2 — дисперсия для весовых процентов; $\hat{x}$ — среднее значение концентрации, для которой определяется дисперсия воспроизводимости. Основным пороком ПСА (как и количественного спектрального анализа) являются ошибки, которые можно назвать «ошибками условий анализа». Чаще всего источником весьма грубых ошибок является несоответствие валового состава или минеральных форм анализируемых проб составу эталонов, по которым строится градуировочный график.

С интервальным характером результатов ПСА связана интересная проблема «ошибок группировки» и «оптимальной кратности шкалы». Минимальная разность между двумя концентрациями, которая еще может быть замечена данным прибором или при помощи данного метода анализа, характеризует «разрешение» или «разрешающую способность» метода или прибора. При полуколичественном спектральном анализе с применением 10-ступенчатого ослабителя аналитик на глаз может измерить длину линий в спектре с точностью только до 1/3 ступеньки. Если построить градуировочный график в координатах «число ступенек—число

Таблица 4

Представление результатов в баллах при делении порядка на 2, 3, 4 и 6 частей

Порядок концентраций, %	$a=2,\ b_0=8.5$			$a=4,\ b_0=16.5$			$a=3,\ b_0=12.5$			$a=6,\ b_0=24.5$		
	логарифмические интервалы	середины интервалов	балл	логарифмические интервалы	середины интервалов	балл	логарифмические интервалы	середины интервалов	балл	логарифмические интервалы	середины интервалов	балл
				1.0—1.8	1.3	1	1.0—2.2	1.5	1	1.0—1.5	1.2	1
10^{-4}	1.0—3.2	1.8	1							1.5—2.2	1.8	2
				1.8—3.2	2.4	2	2.2—4.7	3.2	2	2.2—3.2	2.7	3
	3.2—10.0	5.6	2	3.2—5.6	4.2	3				3.2—4.7	3.9	4
							4.7—10.0	6.8	3	4.7—6.8	5.6	5
				5.6—10.0	7.5	4				6.8—10.0	8.2	6
				1.0—1.8	1.3	5	1.0—2.2	1.5	4	1.0—1.5	1.2	7
10^{-3}	1.0—3.2	1.8	3							1.5—2.2	1.8	8
				1.8—3.2	2.4	6	2.2—4.7	3.2	5	2.2—3.2	2.7	9
								6.8	6	3.2—4.7	3.9	10
	3.2—10.0	5.6	4	3.2—5.6	4.2	7	4.7—10.0			4.7—6.8	5.6	11
				5.6—10.0	7.5	8				6.8—10.0	8.2	12
				1.0—1.8	1.3	9				1.0—1.5	1.2	13
10^{-2}	1.0—3.2	1.8	5				1.0—2.2	1.5	7	1.5—2.2	1.8	14
				1.8—3.2	2.4	10				2.2—3.2	2.7	15
	3.2—10.0	5.6	6	3.2—5.6	4.2	11	2.2—4.7	3.2	8	3.2—4.7	3.9	16
				5.6—10.0	7.5	12	4.7—10.0	6.8	9	4.7—6.8	5.6	17
										6.8—10.0	8.2	18
				1.0—1.8	1.3	13	1.0—2.2	1.5	10	1.0—1.5	1.2	19
	1.0—3.2	1.8	7							1.5—2.2	1.8	20
10^{-1}				1.8—3.2	2.4	14				2.2—3.2	2.7	21
				3.2—5.6	4.2	15	2.2—4.7	3.2	11	3.2—4.7	3.9	22
	3.2—10.0	5.6	8	5.6—10.0	7.5	16	4.7—10.0	6.8	12	4.7—6.8	5.6	23
										6.8—10.0	8.2	24
10^{0}		1.0	8.5		1.0	16.5		1.0	12.5		1.0	24.5

баллов концентраций» и выбрать кратность шкалы, равную, например, четырем, то «ошибка разрешения», как было экспериментально найдено при исследовании обычной методики ПСА

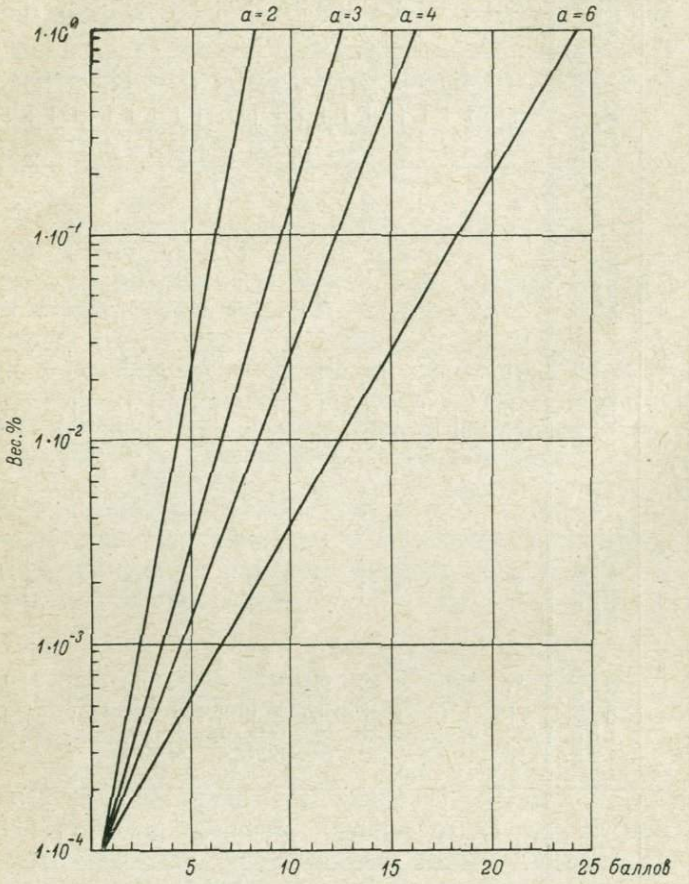


Рис. 9. Функциональная зависимость между баллами и логарифмированными значениями содержаний (в вес.%).

График позволяет перевод из одной шкалы в другую. Например, 8 баллов в шкале $a=4$ составит 4,6, 11,5 баллов соответственно в шкалах с a, равным 2, 3, 6. Аналогично любое содержание можно перевести в баллы. Например, 0.01% выразится в шкалах с a, равным 2, 3, 4, 6, как 4.5, 6.5, 8.5, 12.5 *соответственно*.

(испарение из кратера), составит ± 0.5 балла (Юдович и др., 1970). Это значит, что две концентрации, отличающиеся меньше чем на 0.5 балла, например 10.6 и 10.9, данным методом различить нельзя. При равномерном распределении внутри интервалов среднее квадратичное отклонение отдельных значений вокруг центра

интервала $\sigma = h/2\sqrt{3}$, или $\sigma^2 = h^2/12$ (т. е. дисперсия равна поправке *Шеппарда*). *При достаточно малом* h по сравнению с диапазоном анализируемых содержаний распределение внутри интервалов незначительно отличается от равномерного. Следовательно, если интервал группировки *равен 1 баллу, то средняя* квадратичная ошибка, вносимая группировкой, составит около 0.3 балла.

При нормальном распределении измеряемой величины максимальное расхождение $|\alpha|$ математических ожиданий сгруппированных и несгруппированных данных (величина смещения среднего) составит, как показал Л. Д. Мешалкин (1964),

$$|\alpha| = 0.04 \cdot h^3/\sigma^2. \qquad (2.6.9)$$

Например, если распределение баллов некого компонента нормально, причем $h=1$ баллу, а $\sigma=2$ баллам, максимальное смещение среднего составит: $\alpha=0.04 \cdot 1/4=0.01$ балла. Это значит, что при условии бесконечно большой выборки группировка внесет в средний результат ошибку не более ± 0.01 балла. Если выразить интервал h в долях квадратичной ошибки σ, то для нормально распределенной величины получим максимальные значения разности сгруппированных и несгруппированных результатов:

Интервал h в долях стандартного отклонения . .	0.5	1.0	1.5	2.0	3.0	4.0
Максимальная относительная ошибка в долях стандартного отклонения $\lvert\beta\rvert = \lvert\alpha\rvert/\sigma$	0.005	0.04	0.11	0.32	1.1	2.56

Из этих данных видно, что влияние группирования на средний результат начинает ощущаться только при чрезвычайно грубой группировке ($h/\sigma=2$ и более), а при h, близких или меньших σ, оно пренебрежимо мало. В нашем исследовании полуколичественного спектрального анализа σ была экспериментально оценена для большой группы элементов (Юдович и др., 1970). При делении порядка на четыре части среднее значение σ оказалось равным около 1 балла. Значит, в этом случае и ошибка, вносимая в результат принятой дискретностью шкалы, не выше чем ± 0.04 балла.

Л. Д. *Мешалкин полагал*, что в полученной им формуле (2.6.9) $|\alpha|$ означает максимальную погрешность индивидуальных результатов анализа (вызываемую группировкой); на самом деле им *найдена средняя погрешность*. Индивидуальная погрешность, что очевидно, равна максимально $h/2$.

Итак, мы оценили возможную величину смещения среднего результата, порождаемую группировкой *аналитических данных*. Существует и другой аспект проблемы группировки, связанный с ошибкой воспроизводимости. Известно (Лонцих, Мешалкин,

1964; Юдович и др., 1970), что чем меньше интервал группирования, тем больше (в единицах группирования, т. е. в баллах!) дисперсия ошибки воспроизводимости:

$$a_1^2/a_2^2 = s_1^2/s_2^2. \qquad (2.6.10)$$

Это означает, что если лаборатория выдавала анализ, например в системе $a=2$ (два балла на порядок), а затем, не изменяя методики анализа, стала выдавать результаты в системе $a=3$ (три балла на порядок), то дисперсия воспроизводимости возрастает в 1.5 раза. В связи с этим аналитики иногда задаются вопросом выбора «оптимальной» шкалы, при которой дисперсия воспроизводимости удовлетворяла бы каким-то оговоренным заранее условиям. Например, О. Б. Фалькова и сотр. (1971) подбирают шкалу баллов такой, чтобы дисперсия воспроизводимости была бы равна единице (т. е. одному баллу). В этом случае при нормальном распределении баллов, вероятность того, что данное содержание будет оценено одним и тем же баллом, составит 68%. Другой подход к «оптимизации» шкалы — требование минимальности ошибки воспроизводимости в этой шкале. Однако очевидно, что чем грубее шкала, тем хуже абсолютная воспроизводимость результатов, но лучше «искусственная» воспроизводимость, т. е. повторение в анализе одного и того же балла! Например, если принять систему, в которой на порядок приходится всего один балл, то ошибка воспроизводимости в баллах может стать вообще нулевой, хотя ясно, что анализ дает громадную ошибку. Величина занижения «истинной» дисперсии воспроизводимости соответствует известной поправке Шеппарда. Если хотят устранить влияние группировки данных на величину дисперсии воспроизводимости, то к найденной дисперсии прибавляют поправку Шеппарда:

$$s_0^2 = s_b^2 + h^2/12, \qquad (2.6.11)$$

где s_0^2 — «истинная» дисперсия (несгруппированных значений); s_b^2 — дисперсия в баллах (сгруппированных значений); h — ширина интервала группировки (в данном случае — 1 балл).

Исследуя вопрос о применении поправки Шеппарда, Р. И. Дубов (1972) нашел, что внести ее можно только при $h/s_0 \leqslant 2.5$; если величина этого отношения больше, то внести поправку Шеппарда нельзя, ибо значение ее становится слишком неопределенным. Для практических целей можно принять удовлетворительным такое h, при котором величина s_0^2 превосходит величину поправки Шеппарда ($h^2/12$) в несколько раз. Часто из чисто экономических соображений интервал h можно принять достаточно большим в тех случаях, когда переход к другой, более точной методике анализа, стоит дешевле, нежели старый анализ, но с уменьшенной величиной h. Например, как показали С. В. Лонцих и сотр. (1967), при анализе металлометрических проб есть смысл делить порядок не более чем на шесть частей. Впоследствии

С. В. Лонцих с сотр. (1969, стр. 46) отметили, что «дальнейшее уменьшение величины интервала . . . не имеет никакого метрологического смысла, так как оно, не приводя к сколько-нибудь существенному увеличению точности анализа, значительно затрудняет интерпретацию спектрограмм, снижая экспрессность определений».

§ 7. Некоторые выводы и рекомендации

Не все из проблем, затронутых в этой главе, одинаково хорошо разработаны. Для одних уже можно дать рекомендации, другие еще обсуждаются, третьи только поставлены.

Одним из условий получения надежных геохимических выводов является правильный и всесторонний контроль работы аналитической лаборатории. При этом должны быть выявлены и учтены систематические ошибки и установлен уровень случайных ошибок воспроизводимости. И те, и другие могут не оставаться постоянными во всем диапазоне определяемых концентраций. Если рассматривать только «прогрессивные» и «постоянные»[1] ошибки, то, например, полную систематическую ошибку можно представить в виде суммы $\delta = \delta_1 + \delta_2(c_i)$, где δ_2 — доля систематической ошибки, зависящая от концентрации c_i. На практике мы должны определить некую среднюю ошибку, используя серии проб с различными концентрациями, но такими, которые нам уже наперед известны. Представительность среднего значения ошибки будет зависеть от подбора проб для контроля, т. е. в статистическом смысле — от репрезентативности контрольной выборки. Если бы нам было известно распределение компонента во всей изучаемой совокупности (например, вероятности встречи определенных концентраций марганца в карбонатных породах), то составление контрольных выборок для изучения погрешностей не представило бы затруднений. На практике изучаемые природные распределения известны далеко не всегда. Кроме того, систематическую ошибку обычно определяют не для каждой партии проб, а всего один-два раза для данной методики. В такой ситуации рекомендуем поступать следующим образом. Определить ошибки воспроизводимости и систематические ошибки правильности для проб, равномерно отобранных по всему интервалу концентраций. При расчете средних значений случайных и систематических ошибок для данной партии проб ошибки для каждого интервала концентраций взвесить на относительные частости данного интервала концентраций в анализируемой партии:

$$\bar{\delta}^2 = \sum_{i=1}^{k} \delta_i^2 p_i, \qquad (2.7.1)$$

[1] Термины, применяемые в метрологии (Маликов, Тюрин, 1966).

$$s^2_{\text{воспр}} = \sum_{i=1}^{k} s_i^2 p_i, \qquad (2.7.2)$$

где k — число интервалов концентраций, которое надо выбирать тем больше, чем сильнее зависит ошибка от концентрации; p_i — относительные частости данного интервала концентраций.

Анализ поведения систематической ошибки во времени для заданного интервала концентраций можно провести с помощью критерия Аббе. Для этого все анализы располагают в хронологической последовательности, от 1 до n и рассчитывают две дисперсии: s_1^2 — обычная дисперсия для всего ряда анализов, s_2^2 — рассчитывается по так называемым последовательным разностям

$$s_1^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2, \qquad (2.7.3)$$

$$s_2^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n} (x_i - x_{i-1})^2. \qquad (2.7.4)$$

Величина s_1/s_2 имеет распределение F, и при отсутствии направленного изменения величины систематической ошибки обе дисперсии будут равны с точностью до случайной флуктуации.

Изучение систематических и случайных погрешностей обязательно должно завершаться организацией приемочного контроля, который рекомендуется проводить в соответствии с инструкцией ВИМСа (Инструкция по внутрилабораторному контролю. . ., 1961).

Наилучшим универсальным методом внесения поправок в данные основных анализов является применение уравнения регрессии.

Вследствие неудовлетворительной правильности большинства применяемых методов не следует пренебрегать дешевыми и экспрессными полуколичественными спектральными анализами (ПСА).

При выборе кратности шкалы ПСА (числа баллов на порядок содержаний) следует руководствоваться метрологическим принципом соответствия точности полученных результатов и формы их записи.

Ни один из методов обработки данных неполно определенных выборок (содержащих часть «пустых» проб) не дает хороших результатов, если неизвестен закон распределения концентраций в изучаемой совокупности. Если же имеются данные о том, что это нормальный (или логнормальный) закон, то лучшие результаты дает простой метод Фишера. Приближенный результат можно очень просто получить с помощью прямой Генри.

Глава 3

ЧАСТОТНЫЕ РАСПРЕДЕЛЕНИЯ В ГЕОХИМИИ

§ 1. Вводные замечания

Отложив по оси x значения признака, сгруппированные в разряды (классы), а по оси y — частоту появления данного класса, получают частотный график — простейший, но весьма эффективный способ статистического исследования. Полученный график выражает частотное распределение изучаемого признака и с формальной стороны ничем не отличается от дифференциальной кривой распределения случайных величин. Это внешнее подобие иногда скрывает глубокое различие между изучаемым признаком и «случайной величиной». Эта тонкость еще недостаточно осознана геологами, широко применяющими частотные распределения.

Можно указать следующие задачи, которые решают с помощью изучения частотных распределений.

1) В наглядной и компактной форме рассмотреть числовую совокупность, выявив в ней модальные (наиболее часто встречающиеся) и более редкие значения признака.

2) Произвести свертку информации, позволяющую давать компактную характеристику объекта, подобрав вид распределения и определив значение его параметров, число которых не превышает обычно двух-трех. Такая свертка позволяет иногда классифицировать объекты в узких целях. Например, Ю. А. Ткачевым и сотр. (1970) распределение пьезооптического кварца гидротермальных месторождений было аппроксимировано законом Пуассона, позволяющим основные особенности распределения, которые влияют на методику опробования и разведки, свести к двум параметрам и построить простую классификацию месторождений для целей геологоразведки.

3) Обосновать применимость статистических тестов и выбрать наиболее эффективные оценки средних и дисперсий.

4) Произвести сравнение эмпирических распределений с «модельными», т. е. распределениями, вытекающими из предполагаемых моделей геохимических процессов.

§ 2. Частотные распределения в геохимии. Аналитический обзор

Сравнительно недавно В. А. Кутолин (1972) в своей монографии о базальтах дал хороший обзор проблемы частотных распределений применительно к геологии; это позволяет нам здесь сосредоточить внимание лишь на более важных работах, поскольку литература, в которой затрагиваются частотные распределения, стала в настоящее время практически необъятной.

Ф. Ю. Левинсон-Лессинг один из первых использовал ряд статистических параметров для целей петрографической классификации. Затем в течение ряда десятилетий у геологов не наблюдалось интереса к данным вопросам.

Важным было появление статей Н. К. Разумовского (1940, 1948), который впервые указал, что распределение ряда элементов в рудных месторождениях близко к логнормальному закону. Позднее известный геохимик Л. Аренс (Ahrens, 1953, 1954, 1957), изучая частотные распределения 20 химических элементов в горных породах, заключил, что эти распределения являются логарифмически-нормальными. Этот вывод был сделан им на основании внешнего сходства частотных графиков с теоретическими логнормальными кривыми и не был обоснован статистическими тестами. Статьи Аренса, подкупавшие новизной темы и широтой обобщения, получили широкий резонанс в геохимической литературе. Появилась целая серия работ, развивающих его идею или дающих ей отрицательную оценку (с разных позиций).

Ф. Чейз (Chayes, 1954) отметил, что заключение Аренса, не подкрепленное статистическими тестами, не может считаться достоверным. Р. Миллер и Е. Голдберг (Miller, Goldberg, 1955) подвергли критике не только методику Аренса, но и его обобщение логнормального закона как фундаментального закона геохимии. На вид функции распределения элемента в геологических объектах влияет целый ряд генетических факторов, которые можно считать независимыми, а также факторов, связанных с методикой опробования и ошибками анализа. Исходя из этого, они заключили, что многообразие наблюдаемых распределений не может быть описано в рамках какого-то одного закона. Затем с развернутой критикой концепции Аренса выступил К. Обри (Aubrey, 1956), который указал, что распределение породообразующих компонентов в горных породах даже в идеальном случае не может удовлетворить логнормальной модели. Например, если порода состоит только из двух минералов и один из них распределен

логнормально, то кривая распределения второго минерала будет зеркальным ее отражением, т. е. будет иметь отрицательную асимметрию. Если породообразующих компонентов несколько, то по крайней мере один из них (чаще всего кремнезем) должен обладать отрицательной асимметрией распределения, тогда как малые элементы обычно имеют правоасимметричные распределения, нередко хорошо аппроксимируемые логнормальным законом. Позже концептуальные построения К. Обри были более строго обоснованы в статье Дж. Роджерса и Дж. Адамса (Rodgers, Adams, 1963).

Наиболее известной отечественной работой, посвященной изучению природных распределений минералов и химических элементов в изверженных горных породах, является монография Д. А. Родионова (1964а). Им вслед за Л. Аренсом было эмпирически установлено, что распределение элементов в минералах и самих минералов в породах в большинстве случаев не противоречит логнормальному закону. Если элемент содержится только в одном минерале, то $y=z\cdot x$ (x — содержание элемента в минерале, y — в породе; z — содержание минерала в породе), из чего следует, что величина y как произведение двух логнормальных величин также должна иметь логнормальное распределение. Если же элемент содержится в m минералах, то его содержание будет суммой: $y=z_1x_1+z_2x_2+\ldots+z_mx_m$. Если слагаемые суммы примерно сопоставимы по величине и независимы, то, согласно центральной предельной теореме, при большом числе слагаемых возникает нормальное распределение величины y. Однако в реальных условиях число слагаемых (минералов) невелико и даже приближенное равенство их вкладов не выполняется (ибо это означало бы отсутствие в породе минералов-носителей, вносящих главный вклад в общее содержание элемента). Поэтому для таких ситуаций естественно ожидать отклонений от нормального закона в сторону асимметричных (в частности, правоасимметричных) распределений, близких к логнормальным. Следовательно, если элемент находится в породе только в одном минерале, то может возникнуть логнормальное распределение, а если в нескольких, то распределение будет тем ближе к нормальному, чем больше минералов-носителей и чем менее контрастно распределение элемента между ними.

Изложенные представления Д. А. Родионова быстро завоевали широкое признание вследствие ясности геохимической интерпретации типа распределения и удачного применения вполне правдоподобной концептуальной модели (элементы в минерале, минералы в породе).

Дальнейшее изучение проблемы существенно уточнило концепцию Д. А. Родионова. Например, Ю. П. Трошин и сотр. (1966), построив физико-химическую модель распределения элемента-примеси в минералах гидротермальных жил, получили, что функ-

ция распределения концентрации зависит от одного параметра D — коэффициента распределения элемента между раствором и твердой фазой. В частности, при значениях $2 < D < 1$ получается положительно-асимметричное распределение, при $1 < D < 2$ — отрицательно-асимметричное, при $D = 2$ распределение вырождается в равномерное на некотором интервале; наконец, при $D = 1$ концентрация элемента становится константой. Сходные распределения были получены И. Д. Рябчиковым в 1960 г.

Иногда можно возразить также против допущения о независимости величин типа $z_m \cdot x_m$ (доля элемента, вносимая данным минералом). При некотором фиксированном количестве («запасе») элемента в данном объеме кристаллизующейся породы между минералами-носителями могут возникать определенные корреляционные связи (отрицательные и положительные). Например, для метаморфических пород было обнаружено (Петров, 1970), что при наличии двух минералов-конкурентов, содержащих один и тот же элемент-примесь, концентрация этого элемента в минерале зависит от количества другого минерала. Например, биотиты оказались обеднены цинком при наличии в породе ставролита, сильно концентрирующего цинк. Следовательно, в данном случае доли цинка, вносимые в породу биотитом и ставролитом, уже не являются независимыми.

А. Б. Вистелиус (Vistelius, 1960) развивал иной подход к проблеме природных распределений концентраций. Он полагал, что многообразие факторов, так или иначе влияющих на распределение концентраций, должно скорее всего приводить к нормальному распределению, а не к логнормальному; в течение одной фиксированной стадии геохимического процесса распределение элементов является нормальным, а положительная асимметрия возникает при наложении результатов процессов ряда стадий. Позднее (Вистелиус, 1963б) он предложил математическую модель, согласно которой распределение является «локально нормальным», но его параметры (среднее и дисперсия) меняются от точки к точке; такая модель также может обусловить возникновение асимметричных распределений, напоминающих логнормальные.

Помимо перечисленных главных геохимических работ, имеется также значительное число литологических исследований, посвященных изучению распределения размеров зерен в осадках. Хороший обзор работ, содержание которых лежит в стороне от нашей темы, читатель найдет в статье Г. В. Миддлтона (1968), а также в монографиях У. Крамбейна, Ф. Грейбилла (1969, стр. 88—93) и Дж. Гриффитса (1971, стр. 227—233). Здесь мы отметим только, что литологи, использующие логнормальные аппроксимации распределения зерен в осадках, обычно исходят из той или иной модели процесса: они принимают логнормальное распределение не только потому, что оно хорошо сглаживает эмпирические частоты, но и потому, что имеют дополнительные

основания для выбора именно логнормального распределения среди ряда похожих правоасимметричных распределений.

Однако в геохимической литературе не так много примеров, когда выбор той или иной гипотезы о виде функции распределения, сделанный на основании изучения эмпирических частот, подкреплялся какими-то дополнительными (генетическими и пр.) соображениями. Еще менее типичен такой подход для геофизиков, которые чаще геологов пользуются различными аппроксимациями. В большинстве случаев геолог выбирает ту кривую частотного распределения, которая просто лучше, чем другие, аппроксимирует эмпирические частоты. Этот подход особенно характерен для практики геологоразведки. Так, в монографии П. А. Рыжова и В. М. Гудкова (1966) активно пропагандируется гамма-распределение. Хорошим примером подбора аппроксимирующей функции является исследование Д. Г. Криге (1968), посвященное распределению золота в рудниках Южной Африки. Огромная практическая важность задачи (знание закона распределения позволяет обоснованно подсчитывать запасы по результатам опробования) требовала выполнения только одного условия: как можно большей сходимости теоретически предсказанных вероятностей с эмпирическими частотами. Такому требованию удовлетворило трехпараметрическое логнормальное распределение, где исходные величины z (содержание золота в руде) предварительно преобразуются к виду $x=\ln(z+a)$. Параметрами этого распределения служат $\overline{\ln x}$, $\sigma=\sigma_{\ln x}$ и a — эмпирическая постоянная, которая определяется либо по региональному фону для рудника или поля, либо рассчитывается непосредственно по большой выборке.

Поскольку подобные процедуры, вообще говоря, не связаны с генетическими или другими концепциями, постольку будут оправданы любые другие распределения, позволяющие удачно аппроксимировать эмпирические частоты. Например, иркутские геохимики предложили использовать для целей аппроксимации распределение Вейбулла. Авторы (Шиманский, Базанов, 1966, стр. 61) пишут: «. . . при использовании логарифмически-нормальной функции нередко возникают трудности, связанные с расчетами и произвольным выбором более общего типа преобразующих функций, таких как $q(c)=\lg(c+a)$, $q(c)=\lg(c-a)$ и других, что делает логнормальную функцию неудобной. Авторами . . . была принята функция Вейбулла, как не требующая никаких преобразований, отличающаяся простотой расчета и дающая в большинстве случаев более удовлетворительное решение, чем логнормальная функция». Другим примером удобной аппроксимирующей функции может служить предложенное С. Уиксвеллом, изученное Дж. Ачисоном и И. Брауном (Aitchison, Brown, 1957) и примененное на практике Н. Н. Боровко (1964) четырехпараметрическое логнормальное распределение величины $\xi=\lg\frac{x-a}{b-x}$. Оно харак-

теризуется большой общностью и при $x \to a$ или $x \to b$ превращается в трехпараметрическое, описанное Д. А. Родионовым (1964а) и Д. Г. Криге (1968).

В последние годы наблюдается тенденция геологоразведчиков использовать как можно более универсальные функции, описывающие природные распределения такие, чтобы целый ряд конкретных распределений получался из этой функции как частный случай. Это стремление наиболее отчетливо отражено в статье И. В. Францкого (1970). «Нам представляется, — пишет автор, — что пришла пора как-то обобщить. . . приемы и дать рекомендации математического описания распределения показателей, основанные на геологической природе месторождений. . .» (стр. 230). В качестве универсальной функции И. В. Францкий рекомендует функцию $F(x)=1-\exp(-bx/\bar{x})$, где $b=f(x)$ есть некоторый коэффициент, зависящий от x. Проанализировав эмпирические данные, автор заключил, что величина b для большинства месторождений может быть представлена степенной функцией вида $b=A\cdot x^n$, и тем самым процедура подбора нужной функции распределения $F(x)$ сводится к нахождению A и n. Последние находятся эмпирически, путем построения зависимости $n, A=f(V_x)$, где V_x — коэффициент вариации показателей, изученный автором в диапазоне от 55 до 144%.

Трудно судить о том, насколько универсальна рекомендуемая И. В. Францким функция (в частности, автором не изучены распределения с V_x, бóльшими 144%, хотя они тоже не редки), но в основе и функции, рекомендуемой И. В. Францким, и по сути аналогичного ему распределения Вейбулла, заложена, на наш взгляд, весьма плодотворная идея: использование коэффициента вариации V_x в качестве фундаментального геохимического параметра.

В геохимии и геологоразведочном деле уже накоплена весьма обширная информация, позволяющая рассматривать коэффициент вариации как важнейшую характеристику геологических объектов. В книге Дж. Коха и Дж. Линка (Koch, Link, 1970) на материалах по нескольким сотням месторождений США показано, что между коэффициентом вариации и средним значением содержания полезного компонента в руде существует сильная отрицательная корреляция. Малые и редкие элементы, и только они, имеют коэффициенты вариации, достигающие 200—250%, тогда как полезные компоненты, содержание которых в рудах составляет десятки процентов, имеют низкие коэффициенты вариации — до 20%. На графике в координатах «среднее содержание—коэффициент вариации» (Koch, Link, 1970, стр. 296) можно условно выделить два поля. Первое характеризуется V_x, равным 50—20%, $\bar{x}$ — от 5% и выше. Для таких месторождений возможно (но не обязательно) нормальное распределение содержаний компонента в руде. Второе поле характеризуется V_x, превышающими 50%,

$\bar{x}$ — от 20% и ниже (чаще всего ниже и 10%). В этом поле распределение содержаний не может быть нормальным. Указанные эмпирические закономерности получены по данным более 50 000 анализов руд на 484 месторождениях США (Hazen, Meyer, 1966).

Вся рассмотренная выше литература по проблеме частотных распределений имеет одну общую черту: авторы (одни явно, другие неявно) допускали, что изучаемые частотные распределения относятся к таким признакам, которые подчиняются определению независимой случайной величины.[1] Однако геологические характеристики, такие как содержание полезного компонента, мощность рудного тела, линейный запас и т. д., как правило, не отвечают данному определению: они зависят от координат пространства, и их частоты, строго говоря, не могут быть сопоставлены с вероятностями. Ж. Матерон (1968, стр. 10) указывает: «. . . если, например, испытание состоит в отборе пробы в точке с координатами x, y, z, то содержание металла в такой пробе будет единственным, физически определенным и ни в коей мере не случайным». Таким образом, если в рамках других проблем, затронутых на страницах нашей книги, можно было «договориться» об исключении из рассмотрения пространственных переменных и ограничиться областью традиционной статистики, то в проблеме распределений этого уже сделать нельзя. Еще более существенным оказывается, что нельзя абстрагироваться от способа, которым получено значение изучаемой переменной, т. е. в случае распределения содержаний — от способа опробования. Нельзя говорить о распределениях в о о б щ е, без указания, на каких пробах (или, по Матерону, — на какой «геометрической базе») они получены. Таким образом, частотные распределения — специфическая область прикладной статистики в геохимии; специфика состоит в том, что проблема частотных распределений не отделима от проблем геостатистики. Элементы геостатистического подхода к частотным распределениям имеются в ряде интересных работ Л. И. Четверикова (1964), А. В. Канцеля (1966), Б. И. Белова (1966), М. В. Рача (1968), однако впервые наиболее полно вопрос освещен самим Ж. Матероном (1968, стр. 296—327).

Для нас особый интерес представляет подход Ж. Матерона к проблеме логнормального распределения. Практика геологоразведки показала, что логнормальное распределение обладает некоторыми исключительными свойствами. Например, как указывает Э. Карлье (1966), для месторождений урана попытки применения других распределений оказались неудачными. Ж. Матерон (1968, стр. 306—307) указывает на общий недостаток аппроксимации такими распределениями, как гамма, бета и бино-

[1] Это замечание не относится к работе Д. Г. Криге (1968).

миальное, состоящий в том, что все они базируются на математической модели «независимых испытаний» (модель Бернулли): «. . .при этом рудные и безрудные частицы в той или иной степени всегда уподобляются черным и белым шарам, которые извлекаются из урны независимо и случайно. Поскольку для реальных геологических полей, где результаты отдельных «испытаний» (то есть, например, измерений концентраций в пробах) нельзя считать полностью независимыми, постольку и эти аппроксимации плохо описывают геологические поля». Логнормальный закон свободен от этого недостатка, поскольку он основан на математической модели «пропорциональных эффектов» (модель Каптейна), которая хорошо согласуется с автокоррелированностью геологических признаков.

Однако Ж. Матерон указывает и на два принципиальных ограничения логнормального закона в геологии, вызванных а) эффектом «объемного насыщения» и б) эффектом «нулевых содержаний». В тех случаях, когда среднее содержание полезного компонента близко к его максимально возможному, логнормальное распределение не имеет места. «В этом нет ничего удивительного, так как логнормальный закон представляет распределение величин, изменяющихся от 0 до бесконечности, а содержания полезного компонента заключены между 0 и x_0, соответствующим содержаниям в пустой породе и рудном минерале. Когда среднее содержание мало, наличие верхней границы неощутимо сказывается на виде функции распределения, но по мере того как среднее содержание увеличивается и приближается к x_0, следует ожидать появления объемного насыщения, которое проявляется в накоплении частот, соответствующих значениям, близким x_0» (Матерон, 1968, стр. 321). Напротив, в тех случаях, когда «размер проб перестает быть очень большим по сравнению с размерами рудных частиц, появляются пробы с нулевыми содержаниями. Но таким содержаниям, соответствующим бесконечным значениям в логарифмическом масштабе, при логнормальном законе распределения вероятностей отвечает нулевая вероятность» (там же). Заметим, что именно в этом заключена причина появления «гиперболических» частотных распределений, которые В. И. Богацкий (1963) именует «болдыревскими». Они весьма характерны для месторождений золота, горного хрусталя, алмазов. Ж. Матерон заключает, что хотя оба указанных ограничения логнормального закона действуют одновременно, практически они исключают друг друга: «. . . при обычных размерах проб можно считать нереальным как получение нулевого содержания в пробе на месторождении железа, так и содержания, равного 100% на месторождении алмазов» (стр. 322).

В связи с логнормальным распределением возникает замечательный парадокс, детальным разбором которого мы также обязаны Ж. Матерону. В самом деле, следуя А. В. Канцелю (1966),

можно каждую пробу из рудного тела рассматривать как выборку некоторого числа более мелких проб, на которую можно было бы поделить эту пробу. Поскольку операция дробления и перемешивания приводит к усреднению состава пробы во всем ее объеме, то эту процедуру можно рассматривать как получение среднего арифметического из содержаний по мелким пробам. Но распределение средних независимо от распределения самого признака сходится к нормальному, притом довольно быстро (см. гл. 4). В таком случае, укрупняя пробы, т. е. моделируя увеличение объема выборки, мы должны получать распределения, все более близкие к нормальным. На самом деле это не так! Логнормальное распределение оказывается удивительно «устойчивым» и сохраняется даже в пробах крупного объема. Объяснение этого парадокса оказывается возможным только на основе геостатистической теории.

§ 3. Оценочный и эвристический подход к аппроксимации природных распределений. Модели

Из приведенного выше материала легко усмотреть два существенно различных подхода к оценкам природных распределений по выборочным данным.

Первый подход — оценочный. Он характеризуется подбором функции, которая удовлетворяет требованию наилучшего сглаживания наблюденных распределений частот. На эту функцию не накладывают никаких принципиальных ограничений, кроме некоторых формально-математических требований, вытекающих из постановки решаемой задачи (например, требование неотрицательности переменных, непрерывности и дифференцируемости функции и т. д.). Из множества «похожих» функций предпочтение отдают более общим или более простым в вычислительном отношении (впрочем, в век ЭВМ это условие стало не обязательным), а также таким, параметрами которых могут служить простые выборочные статистики, например, среднее и дисперсия. В настоящее время в математической статистике имеется детально разработанный раздел, посвященный аппроксимации выборочных распределений семейством весьма общих функций, предложенных Пирсоном, Джонсоном и другими математиками. Все эти стандартные распределения табулированы, выбор одного из них делается на основании тех или иных соотношений между выборочными статистиками (см. хорошую сводку: Бостанджиян, 1971). Хорошая аппроксимация может иметь большое практическое значение: она дает возможность предсказать вероятность значения признака в промежутках между полученными числовыми значениями, более правильно вычислить среднее, построить прогноз по небольшому числу выборочных данных. К сожалению, с аппроксимацией выборочных распределений связано немало

бессодержательных статей, в которых единственным результатом является отнесение эмпирического распределения к одному из двух типов: нормальному или логнормальному. Геохимическая литература последних десятилетий буквально наводнена такими статьями, где аппроксимация выборочных распределений делается без всякой цели. Это было еще оправдано в период бурного и хаотического накопления эмпирического материала. К примеру, Я. Э. Юдовичем (1964) было впервые установлено логнормальное распределение зольности углей некоторых бассейнов. Сам по себе важный результат позволил автору всего лишь сослаться на возможность его применения для предсказания вероятности встречи проб с заданной зольностью. Однако распространена и другая крайность, когда, приняв статистическую гипотезу, ей без достаточных оснований дают широкую геологическую интерпретацию. Например, Э. Ф. Бачурин и сотр. (1970), придя к заключению, что распределение зольности углей Кизеловского бассейна может аппроксимироваться и логнормальным, и нормальным законами (численных результатов статистической проверки не приведено), сделали вывод о сложности палеогеографических условий бассейна, об отличиях Кизеловских углей от углей Восточной Сибири и Англии (!), о числе влияющих на углеобразование факторов и т. д. Эти далеко идущие интерпретации не находятся ни в какой связи с принятыми статистическими гипотезами, поэтому их ценность весьма сомнительна.

Второй подход — эвристический. Здесь предположение о виде функции распределения не является формальной процедурой сглаживания эмпирических частот, а исходит из такой модели природного процесса, которая должна привести к некоторому теоретическому природному распределению. Эта модель может быть весьма простой — понятийной (концептуальной) или очень сложной, например, физико-химической. Предметная модель должна быть формализована и записана на языке математики — в виде математической модели, которая и позволяет предсказать (вывести) тот или иной закон распределения. Иногда математическая модель не строится, суждение выносится на качественном уровне, но, вообще говоря, без математической модели предсказать теоретическое распределение невозможно. Известны и другие примеры, когда строится только математическая модель, которой могут удовлетворить целый ряд предметных — механических, физических, физико-химических и других моделей.

Эвристический подход позволяет познавать природные явления и процессы, ибо как подтверждение построенных моделей эмпирическим материалом, так и опровержение их имеют объективную познавательную ценность. Например, Е. М. Камерон (Cameron, 1971) сообщает о широком применении Геологической службой Канады частотных распределений меди для целей геохимических поисков. В случае логнормального распределения

геохимик ориентируется на поиски вкрапленных сульфидных руд, связанных с вулканическими породами, тогда как для территорий, характеризующихся нормальным распределением меди, следует ориентироваться на поиски массивных руд, связанных с трубками взрыва (диатремами). Такой вывод основан на том, что в первом случае рудный процесс является как бы частью кларкового процесса, во втором резко отличен от него. Соответственно этому планируется и методика геохимических поисков. В первом случае достаточно редкой сети данных и гидрохимических проб, во втором при такой сети была бы обнаружена лишь каждая десятая трубка.

1. **Математическая модель.** Она является обязательным этапом при использовании других моделей. Особенность математической модели — ее абстрактность. Например, математическая модель, лежащая в основе нормального закона распределения, записывается в форме

$$x = \sum_{i=1}^{n} x_i, \qquad (3.3.1)$$

где случайны, независимы и примерно соизмеримы по величине величины x_i. При большом n величина x будет иметь нормальное распределение, причем несущественно, какой физический смысл вкладывается в величины x_i. Кроме того, x_i могут представлять собою любые линейные функции от целого ряда других переменных z_i вида $x_i = a_0 + a_1 z_1 + a_2 z_2 + \ldots + a_n z_n$. Для возникновения логнормального распределения величины x достаточно требования

$$x = \prod_{i=1}^{n} x_i, \qquad (3.3.2)$$

или, что то же самое,

$$\log x = \sum_{i=1}^{n} \log x_i; \qquad (3.3.3)$$

при этом достаточно только, чтобы величины x_i были положительными и несущественно, что они представляют собой в предметном смысле. Другая известная модель, генерирующая логнормальное распределение, — так называемая модель пропорциональных эффектов Каптейна, требующая лишь того, чтобы разности $x_i - x_{i+1}$ были пропорциональны i-й случайной величине:

$$x_i - x_{i+1} = k_i x_{i+1}, \qquad (3.3.4)$$

где k_i — независимые случайные коэффициенты пропорциональности (Aitchison, Brown, 1957, стр. 22, 23 — цит. по: Миддлтон, 1968, стр. 41). Классическим примером математической модели является также построенная акад. А. Н. Колмогоровым (1941)

теория логнормального распределения размеров частиц при дроблении. Она опирается на очень простые предпосылки: а) вероятности частиц быть раздробленными на более мелкие в единицу времени независимы от их размеров; б) процесс дробления непрерывен в течение длительного времени.

Среди математических моделей можно выделить детерминированные, «жестко» характеризующие процесс, распределение вероятностей в которых является всего лишь отражением действия искажающих «шумов» (в частном, но распространенном случае — погрешностями наблюдений). Более сложны, но значительно содержательнее стохастические модели, сами по себе задающие распределение вероятностей, например, приведенная выше модель А. Н. Колмогорова (1941), или модель О. В. Сарманова и А. Б. Вистелиуса (1947). Подчеркнем во избежание недоразумений, что излагаемое здесь подразделение математических моделей нисколько не претендует на ранг классификации. К тому же очень часто математическая модель, уже по самим своим начальным условиям, неотделима от модели концептуальной. Так, предпосылки построения модели А. Н. Колмогорова в сущности — элементы концептуальной модели.

2. **Концептуальная модель.** По существу это простая логическая схема. Примеры таких моделей весьма многочисленны (Родионов, 1964; Вистелиус, 1960, 1963а, 1963б; Aubrey, 1956; Канцель, 1966, и др.). Очень много таких моделей можно найти в корреляционном анализе (Chayes, 1948; Юдович, Шасткевич, 1966), в анализе поверхностей тренда (Крамбейн, Грейбилл, 1969) и т. д.

Необходимо заметить, что концептуальные модели наиболее близки геологам с их профессиональной ориентацией на качественное (а не количественное) описание геологических явлений. При этом переход от концептуальной модели к математической нередко представляет огромные трудности вследствие необходимости формализации множества интуитивных представлений.

Примером одной из самых наглядных и простых концептуальных моделей служит модель распределения данного химического элемента в горной породе. Она нашла широкое применение в проблеме частотных распределений. Подробное ее изложение и применение в корреляционном анализе дано в гл. 8 и 9.

Чрезвычайно важную для геологической разведки концептуальную модель предложил де Вейс (цит. по: Матерон, 1968). Она основана на принципе подобия, согласно которому при делении объекта (месторождения) или любой его части (блока) пополам нормированная разность содержаний в этих половинах x_1 и x_2 пропорциональна некоторой величине

$$\frac{x_1 - x_2}{x_1 + x_2} = d, \qquad (3.3.5)$$

где d — так называемый *параметр де Вейса*, или относительный

градиент содержания. Из равенства (3.3.5) получают

$$x_1 = (1 - d)m,$$

$$x_2 = (1 + d)m,$$

где m — содержание в блоке до деления, $m = x_1/(1 - d) = x_2/(1 + d)$. Параметр де Вейса представляет собой случайную переменную пространственного типа. Оказывается, что для определенных типов месторождений характерно определенное среднее значение этого параметра. Продолжая деление получающихся «блоков» k раз, увидим, что вероятность для случайно выбранного блока иметь содержание

$$x = (1 + d)^{k-r}(1 - d)^r m \qquad (3.3.6)$$

равна

$$P(x) = \frac{C_k^n}{2^k}. \qquad (3.3.7)$$

Логарифмирование (3.3.6) показывает, что логарифм содержания распределен по биномиальному закону, стремящемуся при увеличении k к нормальному. Следовательно, содержания распределены логнормально с логарифмической дисперсией, зависящей только от k и d:

$$\sigma^2 = \frac{k}{4}\left(\ln\frac{1-d}{1+d}\right)^2. \qquad (3.3.8)$$

Значение k в свою очередь является отношением объема месторождения V к объему конечного блока (пробы) v:

$$\ln\frac{V}{v} = k\ln 2, \quad k = \frac{\ln\frac{V}{v}}{\ln 2},$$

откуда

$$\sigma^2 = \frac{1}{4\ln 2}\left(\ln\frac{1-d}{1+d}\right)^2 \ln\frac{V}{v} = \alpha\ln\frac{V}{v};$$

здесь α — важнейшая характеристика неравномерности распределения, так называемый *коэффициент абсолютного рассеивания*.

Нам представляется, что ввиду широкого распространения логнормального распределения в геохимии и его предположительной устойчивости (специальных исследований на этот счет не проводилось) коэффициент абсолютного рассеивания может иметь в геохимии фундаментальное значение.

3. **Механическая, физическая, физико-химическая и термодинамическая модели.** Их отличие от концептуальных моделей заключается в использовании (вместо частных концепций) фундаментальных законов природы. Известным примером является модель возникновения логнормального распределения скоростей потока (воды, воздуха, льда). Для ламинарного потока связь между v (скорость) и r (радиус частицы) описывается (Гриффитс,

1971, стр. 218) законом Стокса, а для турбулентного потока ($v=cr^2$, $\log v=\log c+2\log r$) и для грубых частиц ($v=c\cdot r$, $\log v=\log c+\log r$) — законом столкновений.

Из этих выражений (c — константа) видно, что если логарифмы скорости потока распределены нормально (а это доказано экспериментально), то и линейная функция от них — логарифмы радиусов частиц — *также будет иметь нормальное распределение* (логнормальное распределение самих радиусов). Другой механической моделью логнормальности является построенная Г. В. Миддлтоном (1968) «теория сортирующих событий», позволяющая в конечном счете свести воздействие сортирующих событий на размер частицы к схеме умножения (3. 3. 2), которая генерирует логнормальное распределение.

Весьма интересная (физическая) модель распределения древнейших радиогеологических дат была разработана М. Б. Соколовым (1968). Вначале автор предлагает правдоподобную концептуальную модель: «. . . одновременно с распадом материнского и накоплением дочернего изотопа неизбежно высвобождается существенное количество энергии. По-видимому, оценкам дат по радиоактивному распаду мешает сам радиоактивный распад, вкладом которого в энергетический баланс геологических (возможно, наложенных?) процессов вряд ли допустимо пренебречь» (стр. 11). Затем строится математическая модель процесса такого, «самотормозящего», распада в виде системы дифференциальных уравнений. По мнению М. Б. Соколова, параметры распределения τ_0 (модальная дата) и N_0 (модальная плотность вероятности), например, для метеоритов они равны соответственно $3.1\cdot10^9$ лет и 0.478, могут служить некоторыми фундаментальными константами планетарного масштаба. Обнаружилось также существование верхнего предела модальных дат в совокупности оценок древних дат ($T=6.9\cdot10^9$ лет), который, быть может, также имеет фундаментальное значение.

Для нас здесь интересна не столько достоверность оригинальной концепции М. Б. Соколова (которая, по его замечанию, не была ясна и самому автору), сколько сам метод построения концептуальной и математической моделей. При этом математическая модель в силу своей абстрактности, оказывается, может описывать и процессы, весьма далекие от радиогеологии: «. . . бактерии, размножаясь, вырабатывают яд, их же истребляющий. При этом рост биомассы бактерий и количества антиростового препарата — яда — пропорциональны наличествующему количеству бактерий» (Соколов, 1968, стр. 28). Хорошим примером физико-химических моделей распределения являются построенные И. Д. Рябчиковым (1960) и Ю. П. Трошиным (Трошин и др., 1966) модели распределения концентраций элемента-примеси, содержащегося в твердой фазе, которая выпадает из раствора или из расплава.

Однако построение обоснованных физико-химических моделей — трудная задача как по существу (необходимо иметь целый ряд априорных или экспериментальных данных о состоянии данной геохимической системы в отдельные моменты времени и т. д.), так и по технике, ибо требует глубокого знания существа моделируемого процесса. Поэтому в литературе нередко встречается подмена сложной физико-химической модели простой концептуальной моделью. Например, А. А. Шиманский и др. (1970), обрабатывая данные анализов 250 мономинеральных фракций микроклина и мусковита на калий, литий, рубидий, цезий, нашли, что распределение калия не противоречит как нормальному, так и логнормальному закону (V_x мал — менее 13%), тогда как для редких щелочей гипотеза нормального распределения отвергается. Авторы предположили «независимость количества калия, переходящего из жидкой фазы в кристаллическую, от содержания его в последней. . . Логнормальный тип распределения редких щелочей объясняется, возможно, тем, что при переходе этих элементов в кристаллическую фазу на количество переходящего в твердую фазу элемента существенно влияет количество его, выделившееся перед этим» (стр. 202—203). Такое объяснение, преследующее очевидную цель — удовлетворить требованиям уравнений (3. 3. 1) для калия и (3. 3. 4) — для редких щелочей — мало что объясняет, ибо остается неясным, п о ч е м у элемент должен распределяться между фазами так, а не иначе. Ответ на такой вопрос могла бы дать лишь корректная физико-химическая модель кристаллизации пегматитового расплава.

Подводя итог обзору частотных распределений, сделаем несколько замечаний общего характера. Независимо от нашего желания избежать в книге геостатистических аспектов, сам предмет этой главы заставляет хотя бы в общих чертах остановиться на них. Это тем более важно, что в большинстве цитированных работ этому вопросу уделяется недостаточное внимание.

Объектом статистического исследования частотных распределений в геохимии являются геохимические поля, или поля концентраций химических элементов. Главной их особенностью, в отличие от объектов традиционной теории вероятностей, является отсутствие реально существующего «элемента» генеральной совокупности. Отбираемые геологами образцы и пробы не являются объективно изолированными друг от друга, существующими независимо от исследователя объектами. Суждение о функциях распределения концентраций без указания о выбранных размерах, форме проб и особенностях их размещения в изучаемом теле просто не имеет смысла.

Не выполняется и другое, может быть, самое важное условие, необходимое для статистических выводов: «элементы» выборки, а точнее — значения изучаемого признака в них, нельзя считать

независимыми. Эта зависимость возрастает с уменьшением расстояний между пробами. Далее, поле концентраций не является вероятностной категорией, оно физически определено, и в этом смысле можно говорить лишь о «проценте встречаемости» различных содержаний, т. е. о «функции встречаемости», но не о плотности вероятности в статистическом смысле.

Неопределенность в выборе формы и размера проб, казалось бы, должна сводить на нет результаты исследований частотных распределений. Тем более удивительно, что это не так, и некоторые виды распределений обладают определенной устойчивостью по отношению к величине проб, однако же с тем условием, чтобы и расстояния между ними увеличивались в приблизительных пропорциях с размерами. Попытки половинчатого подхода к решению, хотя и привели к определенному прогрессу и частным интересным решениям, но не исчерпали проблемы в целом. Так, М. В. Рац (1968) вводит понятие «элемент неоднородности» — наибольший объем горной породы, который при данном масштабе исследования может рассматриваться как внутренне однородный и отличающийся от смежных с ним объемов. По отношению линейного размера пробы P к размеру элемента неоднородности C автор выделяет следующие уровни: а) неоднородность высшего порядка — $P \gg C$; б) эффективная неоднородность — $P > C$; в) неоднородность низшего порядка — $P < C$.

Нам представляется, что назрела необходимость ввести в геохимическую практику понятие о законах распределения концентраций на разных уровнях организации геохимических систем. Изложенное в гл. 6 представление о кларках разных уровней есть не что иное, как отражение упомянутого понятия о параметрах распределений на этих уровнях. Непременным условием содержательной интерпретации частотных распределений является однородность изучаемой совокупности. Хотя ее и называют «статистической однородностью» и применяют для ее проверки статистические тесты, это понятие (в связи с изложенными выше замечаниями о нестатистической природе объекта изучения) заключает, скорее, геологический смысл. Объединение различных по природе объектов (разнородных массивов и т. д.) приводит, как правило, к полимодальным распределениям, для которых отыскание аппроксимирующей функции распределения лишено смысла. Например, В. А. Кутолин (1972, стр. 57) замечает по поводу полученных им частотных распределений окислов в базальтах: «. . . наши выборки составлены из анализов, относящихся к разным телам, а в объединенных выборках по формациям сведены данные по разным районам. . .». Неудивительно, что на таком материале были получены самые разнообразные частотные распределения, не обнаруживающие каких-либо закономерностей, и автор вынужден был вообще отказаться от геохимической интерпретации полученных гистограмм.

Если механизм образования полимодальных или «неправильных» распределений давно ясно осознавался геологами в терминах «статистически неоднородная совокупность», то «геометрической базе» проб до последнего времени в геохимической литературе не уделялось должного внимания.

Наконец, здесь же стоит упомянуть еще о двух осложняющих особенностях в изучении геохимических распределений — замкнутости (в пределах 100%) системы процентных единиц, каковыми являются содержания (Вистелиус, 1968), и искажении распределений, вносимых часто высокими погрешностями анализов.

§ 4. Влияние погрешностей анализа на функции распределения. Композиция распределений

Рассматривая сумму независимых случайных величин, в первой главе мы показали, что ее математическое ожидание и дисперсия равны соответственно сумме математических ожиданий и дисперсий слагаемых. Но эти параметры не исчерпывают характеристики распределения. Пусть в породе имеется два минерала-носителя данного элемента, причем распределение каждого из них известно. Важной задачей является решение вопроса о распределении суммы вкладов этих минералов, составляющих наблюдаемое содержание элемента в породе. Моделируя систему, исходя при этом из некоторых законов распределения концентраций в минералах породы, важно уметь выводить из них «итоговые» распределения. Частным, но очень важным случаем этой задачи является нахождение распределения среднего из n проб (распределение по «большим» пробам при известном распределении по малым пробам). Эта задача также сводится к распределению суммы (которую надо затем разделить на n для возвращения к исходному масштабу измерения).

Однако, может быть, еще важнее в этой проблеме вывести распределение суммы «природного» (истинного) содержания и ошибки анализа. Точнее, практическую ценность представляет обратная задача — восстановление природного распределения по наблюдаемому распределению результатов анализов и по известному (из экспериментов с эталонными пробами) распределению аналитических погрешностей. Может показаться, что эта задача слишком абстрактна и не имеет отношения к запросам геохимической практики. Однако после периода общего увлечения логнормальными аппроксимациями природных распределений стали замечать, что в ряде случаев распределения, построенные на данных спектральных (в особенности полуколичественных!) анализов, отлично аппроксимировались логнормальной функцией, тогда как распределение того же компонента, полученное по данным анализов химических, не согласовывалось с логнормальным законом, а отвечало скорее нормальному или иному закону (см.

напр., Мищенко, 1965). Подобные примеры в настоящее время довольно многочисленны. Интуитивно не вызывает сомнения, что при использовании малоточного метода, обладающего значительной дисперсией, распределение результатов анализа будет находиться в сильной зависимости от вида распределения погреш-

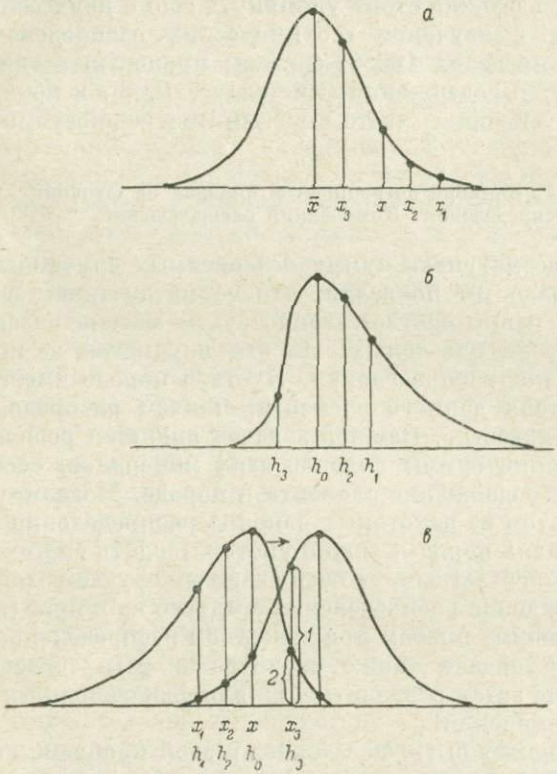


Рис. 10. Иллюстрация графо-аналитического способа композиции распределений.

Компоненты распределения: *а* — первая (нормальное распределение «истинных концентраций»); *б* — вторая (логарифмически нормальное распределение погрешностей *анализов*); *в* — *совмещение одной из компонент с зеркальным отражением другой*. *1* — $f_1(x-h_3)$; *2* — $f_2(h_3)$.

ностей.[1] В гл. 2 было указано, что ошибки спектрального анализа распределяются по логнормальному закону, ибо при этом выдерживается постоянство относительной, а не абсолютной погрешности. Например, если стандартная ошибка спектрального ана-

[1] Практика показывает, что серьезное влияние распределений погрешности анализа начинается уже с величин их стандартных отклонений порядка 10—15 отн.%.

лиза равна 20%, то для концентрации 0.001% ее абсолютная величина составит 0.0002%, а для концентрации 0.1% — 0.02%, т. е. в 100 раз больше. Поэтому при использовании данных спектрального анализа всегда есть опасность, что логнормальный характер распределения обусловлен исключительно подавляющим влиянием распределения аналитической погрешности. Возможна и другая ситуация. Пусть распределение ошибки нормально, но дисперсия его велика. Тогда даже в том случае, если природное распределение было логнормальным (с небольшой дисперсией), аналитическое при композиции «забьет» природное, и итоговое распределение будет выглядеть как нормальное.

Решение задачи достигается с помощью *композиции распределений*.

Пусть $f_1(x)$ — кривая плотности вероятности концентраций в изучаемом объекте, а $f_2(x)$ — погрешностей анализа (рис. 10, *а*, *б*). Какими способами можно получить в результате анализа значение x? Один из способов заключается в следующем: пусть истинное содержание в пробе составляет x_0, а погрешность анализа $h_0=x-x_0$. Сумма этих величин составит x. Вероятность этого «способа появления» x равна произведению вероятностей

$$p_1(x_0)\cdot p_2(h_0)=p_1(x-h_0)\cdot p_2(h_0). \qquad (3.4.1)$$

Однако таких «способов» много; например, истинное содержание может составлять x_1, x_2, x_3 и т. д. Чтобы в сумме получилось x, погрешность должна составлять соответственно $h_1=x-x_1$, h_2, h_3 и т. д. Суммируя вероятность типа (3. 4. 1), получим полную вероятность появления значения x. Так как наши распределения непрерывные, то вместо (3. 4. 1) необходимо записать $f_1(x_0)\cdot f_2(h_0)=f_1(x-h_0)\cdot f_2(h_0)$, а вместо суммирования вероятностей произвести интегрирование плотностей вероятностей:

$$\varphi(x)=\int_{-\infty}^{+\infty} f_1(x-h)\cdot f_2(h)\cdot dh, \qquad (3.4.2)$$

где $\varphi(x)$ — итоговая плотность вероятности (композиция плотностей). Интеграл (3. 4. 2) известен как *преобразование Фурье*. Оно может быть легко выполнено на ЭВМ или (значительно сложнее) вручную. Для этого удобно одну из исходных кривых, например $f_2(h)$, перенести на тот же график, что и $f_1(x)$, предварительно получив ее «зеркальное отражение» от вертикальной оси (это легко сделать, сняв копию на прозрачную бумагу и перевернув ее вниз лицевой стороной чертежа). Для любого заданного x эта отраженная кривая устанавливается так, чтобы точка x совпала с $h=0$ (рис. 10, *в*). Разбив ось абсцисс на достаточное количество интервалов (для практических целей 10—20), выписы-

вают произведения ординат кривых в точках разбиения:

$$\begin{aligned} &+ f_1(x - h_1) \cdot f_2(h_1) \\ &+ f_1(x - h_2) \cdot f_2(h_2) \\ &\quad \cdots\cdots\cdots \\ &+ f_1(x - h_n) \cdot f_2(h_n). \end{aligned} \qquad (3.4.3)$$

Эту сумму умножают на ширину интервала разбиения. Получив значение $\varphi(x)$, передвигают кривую $f_2(h)$ в соседнее положение, соответствующее новому значению x. Кривую $\varphi(x)$ получают в достаточно большом интервале плотности (например, в пределах $\pm 3\sigma$). Результаты удобно оформлять в виде таблицы (табл. 5).

Таблица 5

Схема вычислений при композиции распределений

Значение x	Плотности вероятности	Номера точек разбиения					Сумма	$\Delta h \cdot \Sigma$
		1	2	3	.....	n		
	f_1							
x_1	f_2							
	$f_1 \cdot f_2$						$\Sigma f_1 \cdot f_2$	$\Delta h \cdot \Sigma_1$
	f_1							
x_2	f_2							
	$f_1 \cdot f_2$						$\Sigma f_1 \cdot f_2$	$\Delta h \cdot \Sigma_2$
. . . . .								и т. д.

Контроль: $\Sigma \varphi(x) = 1$

Мы подробно описали схему вычислений потому, что она может быть положена в основу алгоритма расчета на ЭВМ.

Получение распределения среднего из двух проб ничем не отличается от описанного выше, за исключением того очевидного обстоятельства, что одна из двух кривых на рис. 10, *в* является зеркальным отражением другой, а параметры полученного распределения необходимо разделить: математическое ожидание — на 2, дисперсию — на 4 (что соответствует изменению масштаба измерения в 2 раза). Распределение среднего из трех $\varphi_3(x)$ получается композицией кривых $\varphi_2(x)$ и $\varphi_1(x)$, из четырех — $\varphi_2(x)$ и $\varphi_2(x)$ и т. д.

При наличии ЭВМ высокого быстродействия композицию распределений можно делать принципиально иным способом — с помощью метода Монте-Карло, используя датчики случайных чисел с заданными распределениями.

Если исходные функции распределения имеют известные аналитические выражения (для этого предварительно производят аппроксимацию частотных распределений), то композиция распределений достигается применением так называемых *характе-*

ристических функций. С помощью этих функций легко решаются многие статистические задачи, и нам кажется полезным познакомить читателей с их свойствами.

Пусть $F(x)$ — функция распределения величины ξ, т. е.

$$F(x) = \int_{-\infty}^{+\infty} \varphi(\xi)\, d(\xi) = P\{\xi \leqslant x\}. \qquad (3.4.4)$$

Обратно $\varphi(\xi) = \varphi(x) = dF(x)$. Характеристической функцией $f(t)$[2] называют функцию, обладающую свойством

$$f(t) = \int e^{itx} \varphi(x)\, d(x) = Me^{itx}. \qquad (3.4.5)$$

Значения $f(t)$ для каждого значения вещественного аргумента t равны попросту математическому ожиданию величины e^{itx}, где i — мнимая единица. Из (3.4.5) видно, что $f(0) = Me^0 = 1$, $f(t) \leqslant 1$, $(-\infty < t < \infty)$.

Основные полезные для целей композиции распределений свойства характеристических функций заключаются в следующем.

1. Если $\eta = a\xi + b$, где a и b — постоянные, ξ — случайная переменная, то

$$f_\eta(t) = f_\xi(at)\, e^{itb}, \qquad (3.4.6)$$

где f_η и f_ξ — характеристические функции случайных величин η и ξ соответственно. Доказательство очень простое и вытекает из свойств математического ожидания: $f_\eta(t) = \mathbf{M}e^{it\eta} = \mathbf{M}e^{it(a\xi+b)} = \mathbf{M}\{e^{ita\xi} \cdot e^{itb}\} = \mathbf{M}e^{i(at)\xi} \cdot \mathbf{M}e^{itb} = f_\xi(at) \cdot e^{itb}$.

2. Если η и ξ — независимые случайные величины, $x = \eta + \xi$, то $f(x) = f_\eta \cdot f_\xi$. Доказательство: $f(x) = \mathbf{M}e^{itx} = \mathbf{M}e^{it(\eta+\xi)} = \mathbf{M}e^{it\eta} \times \times e^{it\xi} = \mathbf{M}e^{it\eta}\mathbf{M}e^{it\xi} = f_\eta \cdot f_\xi$.

Сложение независимых случайных величин приводит, как мы видели, к весьма сложной операции — композиции функций распределения слагаемых. Свойство 2 показывает, что эта операция сводится к умножению характеристических функций. Композицию распределений (например, аналитического $F_а$ и геохимического $F_г$) обозначают выражением $F_а * F_г = F_н$, где $F_н$ — наблюдаемое распределение, являющееся композицией первых двух. Это выражение обозначает следующую последовательность операций:

$$\begin{array}{ccc} F_а & F_г & F_н \\ \downarrow & \downarrow & \uparrow \\ f_а \cdot f_г & = & f_н. \end{array} \qquad (3.4.7)$$

По аналогии выделение одной из компонент распределения (важная

[2] Именно этим знаком мы будем их обозначать, оставляя здесь знак φ за функцией плотности вероятности.

геохимическая задача!) по наблюдаемому распределению и другой его компоненте обозначим $F_г = F_н * F_а^{-1}$, подразумевая при этом, что $f_г = f_н/f_а$. Знание характеристической функции позволяет легко находить *моменты распределения*:

$$f^{(k)}(0) = i^k \mathbf{M}\xi^k, \qquad (3.4.8)$$

где $f^{(k)}(0)$ — k-я производная характеристической функции в точке $t = 0$; $\mathbf{M}\xi^k$ — математическое ожидание k-й степени случайной величины. В частности, первая производная ($k = 1$) дает математическое ожидание изучаемой величины, вторая ($k = 2$) — $M\xi^2$, что позволяет найти дисперсию $\mathbf{M}\xi^2 - \mathbf{M}^2\xi = \mathbf{D}\xi$. Ниже (табл. 6) мы приводим характеристические функции некоторых важных распределений (Феллер, 1967).

Таблица 6

Характеристические функции некоторых распределений

Наименование распределения	Плотность вероятности	Область определения	Характеристическая функция
Нормальное	$\frac{1}{\sqrt{2\pi}} e^{-1/2x^2}$	$-\infty < x < \infty$	$e^{-1/2t^2}$
Равномерное	$\frac{1}{a}$	$0 < x < a$	$\frac{e^{iat} - 1}{iat}$
Треугольное	$\frac{1}{a}\left(1 - \frac{\lvert x \rvert}{a}\right)$	$\lvert x \rvert < a$	$2\frac{1 - \cos(at)}{a^2t^2}$
Гамма	$\frac{1}{\Gamma(l)} x^l e^{-x}$	$x > 0,\ l > 0$	$\frac{1}{(1 - it)^l}$
Двустороннее показательное	$\frac{1}{2} e^{-\lvert x \rvert}$	$-\infty < x < \infty$	$\frac{1}{1 + t^2}$
Коши	$\frac{1}{\pi} \frac{l}{l^2 + x^2}$	$-\infty < x < \infty$ $l > 0$	$e^{-l(t)}$

Глава 4

СРЕДНИЕ ЗНАЧЕНИЯ В ГЕОХИМИИ И ИХ ОЦЕНКИ

§ 1. Вводные замечания

Понятие «среднего» в геологии и геохимии весьма обширно и столь же неопределенно. Значительная часть острых дискуссий, вспыхивающих периодически на страницах научных журналов, обусловлена именно этой неопределенностью, отсутствием четкой постановки задачи (см. напр., А. Петров, 1962; Карпов, Краснов, 1963 и др.). Поэтому предварительно необходимо уяснить: а) какой смысл вкладывается в термин «среднее»; б) какими свойствами обладает это среднее; в) какие (выборочные) статистики можно применять для оценки различного рода средних; г) какими свойствами должны обладать эти оценки.

Понятие среднего связано с некоторыми *инвариантами*, выбор которых и задает его вид. Так, если инвариантами будут *объем* толщи или рудного тела и *количество* химического элемента в этом объеме, то под средним содержанием надо понимать такое, которое получилось бы при равномерном распределении всего количества элемента во всем указанном объеме, что выразится интегралом

$$\bar{c} = \frac{1}{V} \int c(x, y, z)\, dV, \qquad (4.1.1)$$

где c (x, y, z) — содержание в точке («элементарном объеме») с соответствующими координатами. Располагая элементарные объемы в порядке возрастания содержаний и суммируя их в этом порядке, получим тождественное интегралу (4.1.1) выражение

$$\bar{c} = \frac{1}{V} \int c \cdot V(c)\, dc, \qquad (4.1.2)$$

где V (c) — величина объема с содержанием c. Последний инте-

$c \cdot V(c)$ т.к. $V(c) = \frac{dV}{dc}$ → объём соответствующий единичному

грал по своей структуре аналогичен математическому ожиданию случайной величины c: [1]

$$\bar{c} = \int cp(c)\,dc = \mathbf{M}(c), \qquad (4.1.3)$$

так как $V(c)/V = p(x)$, где $p(x)$ — плотность вероятности содержаний.

Таким образом, наиболее важный в геологии вид среднего, применяемый при подсчете запасов, расчете кларков, составлении геохимического баланса (см. гл. 5 и 6), совпадает с математическим ожиданием. Проблема оценки такого среднего есть проблема оценки математического ожидания. В дальнейшем именно такое среднее в узком смысле слова мы будем иметь в виду, если не оговорено иное. Кроме указанного среднего, в геохимии могут применяться и применяются другие *характеристики положения* распределения концентраций. Например, при диагностике и сопоставлении различных геологических объектов более важной оказывается такая характеристика положения, которая обладала бы наибольшей устойчивостью; при этом не важно, является ли она оценкой математического ожидания или нет. Для островершинных одномодальных распределений этой характеристикой может быть мода, а для плосковершинных с быстрым убыванием встречаемости сильно удаленных от среднего значения содержаний — медиана. При такой постановке задачи единственное существенное требование, предъявляемое к характеристике положения, — возможность получения наиболее эффективной оценки.

§ 2. Оценка среднего в условиях асимметричных распределений

Симметричные распределения, для которых график функции плотности вероятности может быть получен отражением одной из «половинок» относительно некоторой оси, можно охарактеризовать *центром распределения*, совпадающим с абсциссой оси симметрии. Центр такого распределения совпадает с математическим ожиданием, а также с медианой и модой (последнее — не для U-образных или полимодальных распределений). Центр распределения как наглядная и устойчивая характеристика давно привлекает внимание геохимиков. Работая с асимметричными распределениями, во многих случаях можно подобрать такое преобразование изучаемой величины (например, концентрации x) $z = f(x)$, чтобы распределение z оказалось симметричным, и оха-

[1] Как показал Ж. Матерон (1968), функция $\mathbf{P}\{C \leqslant c\}$ и формально, и по свойствам тождественна функции распределения случайной величины, несмотря на их принципиальное отличие по существу. Для различения будем называть первую *функцией распространенности*.

рактеризовать его «центром» как средним арифметическим функции $f(x)$. Именно из этих соображений вытекает целесообразность рассмотрения различного рода средних. Наиболее распространенными видами среди них являются средние арифметическое, гармоническое, геометрическое, квадратическое и кубическое (напр., И. П. Шарапов, 1965, стр. 46—54). Последние два вида средних — разновидности так называемого среднего степенного (табл. 7).

Таблица 7

Некоторые виды средних и способы их вычисления

Среднее	Функция	Расчетная формула
Арифметическое	$z = x$	$\bar{z} = \frac{1}{n}\sum z_i \quad \bar{x}_{ар} = \frac{1}{n}\sum x_i$
Гармоническое	$z = \frac{1}{x}$	$\bar{z} = \frac{1}{n}\sum z_i \quad \bar{x}_{гарм} = \frac{n}{\sum \frac{1}{x}}$
Степенное	$z = x^k$	$\bar{z} = \frac{1}{n}\sum x_i^k \quad \bar{x}_{степ} = \sqrt[k]{\frac{1}{n}\sum x_i^k}$
Квадратическое	$z = x^2$	
Кубическое	$z = x^3$	
Геометрическое	$z = \log_a x$	$\bar{z} = \frac{1}{n}\sum \log_a x \quad \bar{x}_{геом} = a^{\overline{\log x}}$
Экспоненциальное	$z = e^x$, $\ln z = x$	$\bar{z} = \frac{1}{n}\sum e^{x_i} \quad \bar{x}_{эксп} = \frac{1}{n}\sum \ln z_i$

Из таблицы ясно, что различные средние являются оценками центра распределения соответствующей величины $z=f(x)$ и не имеют ничего общего с математическим ожиданием величины x, состоятельную и несмещенную оценку которого дает только среднее арифметическое.

Однако тот факт, что разные виды средних отвечают центру «своего» распределения, причем этот центр в условиях асимметричных распределений никогда не совпадает со средним арифметическим, часто вводил геологов в заблуждение. В печати неоднократно высказывалось мнение о том, что в условиях асимметричных распределений среднее арифметическое «занижает» или «завышает» истинное среднее. В. В. Богацкий, как бы обобщая эти неверные представления, утверждает (1963, стр. 117): «. . . способ оценки среднего определяется характером исследуемого геологического объекта». Ему вторит Ю. Г. Шестаков (1966, стр. 309): «Если для нормальных кривых наиболее достоверное (!) среднее соответствует среднему арифметическому, логнормальных — среднему логарифмическому, и гиперболоподобных —

среднему гармоническому, то для промежуточных кривых оно должно быть промежуточным». *Аналогичных взглядов придерживались* Л. Ф. Залата (1963) и многие другие авторы. Подобные высказывания, с математической стороны неверные (часто именно *вследствие неопределенности понятия* среднего у ряда авторов), были подвергнуты справедливой критике (Суражский, Рощин, 1964; Каждан, Шумилин, 1966), однако публикация ошибочных статей продолжалась (Богацкий, 1968).

Итак, математическая статистика утверждает, что выборочное среднее арифметическое является несмещенной оценкой истинного среднего — математического ожидания в смысле (4.1.3). Это означает, что математическое ожидание отклонения выборочного среднего от истинного равно нулю: $M(\bar{x}-\mu)=0$, другими словами, — отсутствие систематической ошибки. Однако в условиях малых выборок из резко асимметричных распределений распределение выборочных средних арифметических «наследует» эту асимметрию; по мере роста объема выборки эта асимметрия *уменьшается и, начиная с некоторого* момента, их распределение, согласно центральной предельной теореме, становится нормальным. Характер асимметрии распределения *выборочных средних зависит* от исходного распределения признака: чем сильнее асимметрия исходного распределения, тем сильнее уклоняется от нормального распределение выборочных средних. Так, на рис. 11 видно, что при резко асимметричном распределении золота средние из выборок объемом в 5 проб дают распределение, все еще характеризующееся сильной асимметрией; при 25 пробах в выборке асимметрия незначительна, но еще ощутима, а при средних из 100 проб распределение становится практически нормальным.

Предположим теперь, что совокупность, изображенная на рис. 11, опробовалась одной выборкой объемом в 5 проб. Среднее, найденное по этой выборке, скорее всего, окажется меньшим, чем «истинное» содержание, равное 7.6 г/т. Действительно, вероятности получить значения в интервале от 0 до 7.6 г/т намного больше, чем в интервале от 7.6 до $>$ 99 г/т. *Подсчет частот показывает*, что вероятность занизить истинное среднее в данном случае оказалась вдвое большей, чем вероятность *завысить его. Аналогичное явление* можно проиллюстрировать на другом полезном компоненте, имеющем резко асимметричное распределение — горном хрустале. Анализ кривых распределения количества этого полезного ископаемого по единичным пробам, а также кривых распределения среднего из нескольких проб приводит к выводу, что они резко асимметричны: вероятность появления в единичной пробе содержания ниже среднего значительно превышает вероятность встречи пробы с содержанием выше среднего. Следовательно, содержание, оцененное

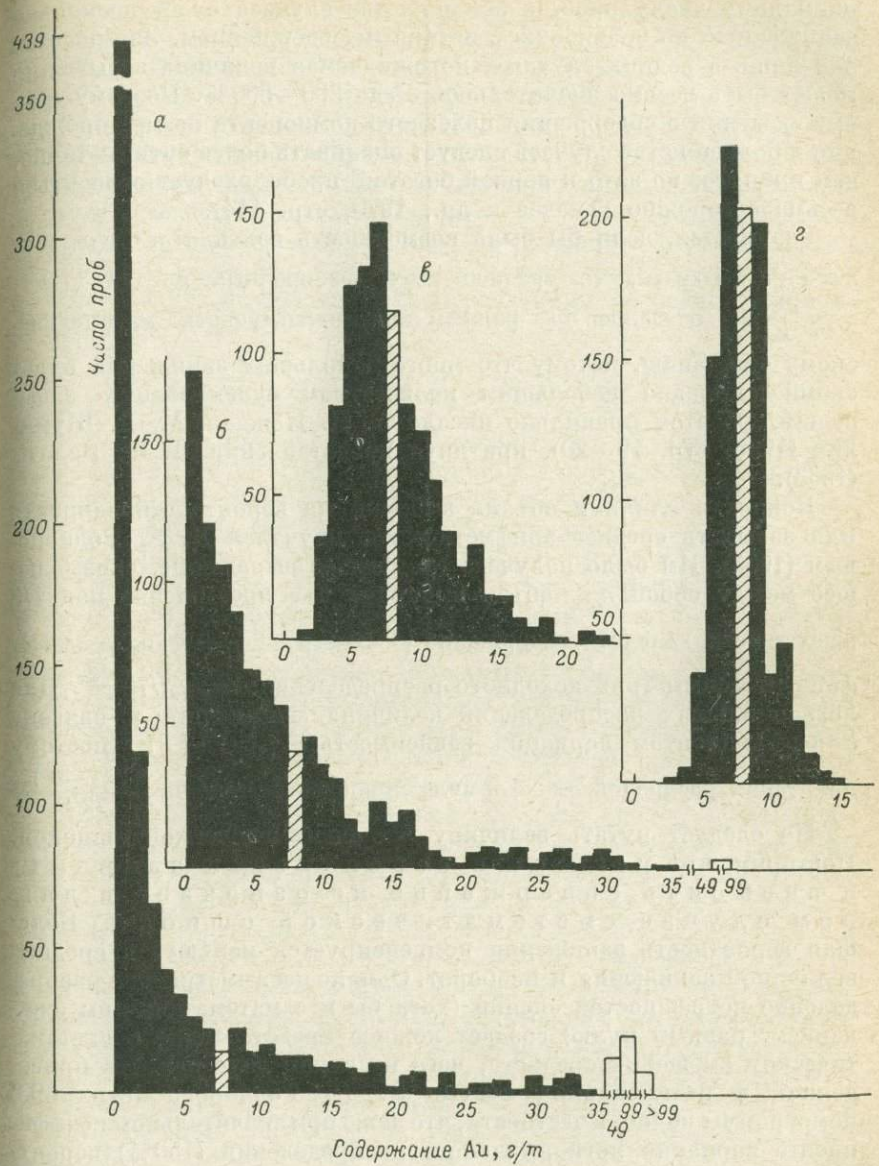


Рис. 11. Иллюстрация центральной предельной теоремы по материалам опробования на золото в шахтном поле Хоумстейк майн, США.

а — первичное распределение анализов; *б*, *в*, *г* — распределение выборочных средних по выборкам объемом в 5, 25 и 100 проб соответственно. Всего случайным образом из первичной совокупности (900 проб, условно приведенных к 1000 пробам) было сделано по 1000 случайных выборок заданного объема. Составлено по данным Дж. Коха и Дж. Линка (Koch, Link, 1970, стр. 72, табл. 3.14).

по малому числу проб, в большинстве случаев будет несколько заниженным по сравнению с истинным содержанием. Завышается оно лишь в редких случаях, но ожидаемая величина завышения может быть весьма значительной — до 300—500%. Поэтому первые сведения о содержании полезного компонента на месторождении в большинстве случаев следует оценивать более оптимистично, чем принято, но зато к первой богатой пробе следует относиться весьма осторожно (Ткачев и др., 1970, стр. 114).

Разумеется, если бы была возможность взять м н о г о м а л ы х в ы б о р о к, то среднее всех выборочных средних $\left(\hat{x}=\frac{1}{n}\sum \bar{x}_i\right)$ оказалось бы равным или почти равным математическому ожиданию, потому что много небольших занижений будет скомпенсировано небольшим количеством существенных завышений. Об этом правильно писали А. Б. Каждан, М. В. Шумилин (1966, стр. 19—20), критикуя представления Л. Ф. Залаты (1963).

Вопрос о влиянии объема выборки на вероятности занизить или завысить среднее арифметическое изучался Д. А. Родионовым (1963). Им было получено следующее выражение, связывающее между собой n — объем выборки, δ — превышение над 1/2 вероятности занизить содержание $\left(\mathbf{P}\{\bar{x}\leqslant\mu\}=\frac{1}{2}+\delta\right)$ и γ_3 — коэффициент асимметрии исходного распределения: $n=\gamma_3^2/72\pi\delta^2$. Для логнормальных распределений величина асимметрии γ_3 связана с коэффициентом вариации зависимостью $\gamma_3=V^3+3V$, поэтому для таких распределений формула приобретает вид $n=\frac{(V^3+3V)^2}{72\pi\delta^2}$.

Не следует путать величину δ с систематической ошибкой. Повторим, что р а з н и ц а в в е р о я т н о с т и з а н и з и т ь и з а в ы с и т ь с о д е р ж а н и е н е о з н а ч а е т в д а н н о м с л у ч а е с и с т е м а т и ч е с к о й о ш и б к и. Бóльшая вероятность занижения компенсируется меньшей в среднем величиной занижения и наоборот. Однако несимметричное распределение погрешностей оценки (хотя бы и с математическим ожиданием, равным нулю) создает ложное представление о систематическом смещении среднего, чего на самом деле нет. Из приведенной в цитированной статье Д. А. Родионова (стр. 693) номограммы видно, в частности, что даже при значительном коэффициенте вариации логнормального распределения (150%) вероятность занизить среднее выборкой объемом в 25 проб всего на 10% больше 1/2.

Методика Д. А. Родионова не нашла широкого применения, ибо значения δ не связаны с понятием *точности выборочного среднего* — вопросом, который так актуален в практике геологоразведки и геохимии.

Более обещающим представляется другой подход: для заданных величин n и γ_3 или V получить набор распределений выборочного среднего. Имея такие распределения, можно оценивать не только вероятность смещения в ту или иную сторону выборочного среднего, но и для заданного доверительного интервала получать доверительную вероятность (или наоборот). Например, если считать распределение на рис. 11, *б* природным (выборка очень большая, поэтому частоты можно отождествлять с вероятностями), то можно сказать, что при $n=5$ вероятность попадания истинного среднего в интервал 5—15 г/т равна 34%, а в интервал 15—20 г/т — всего 4%. Методика получения распределения выборочных средних по исходным распределениям изложена в предыдущей главе (стр. 82).

§ 3. Геохимический фон, медиана и мода

Большинство геологов интуитивно понимает под *геохимическим фоном* наиболее типичные содержания элементов, «рассеянное нахождение элементов в геохимически однородных средах» (Смирнов, 1963, стр. 333), «нормальные значения содержаний химических элементов в горных породах» (Юфа, Гурвич, 1964, стр. 817). Этому понятию противопоставляется понятие *геохимической аномалии*. Следовательно, они составляют единство противоположностей и одно не может быть определено без другого. Если фон — это рассеянные, нормальные, типовые содержания, то аномалия, напротив, характеризует концентрированные формы нахождения элементов, значительно более редкие, не типичные и т. д. Очевидно, что математическое ожидание и мода не всегда могут быть подходящей характеристикой фона.

Неудобство среднего арифметического как оценки математического ожидания состоит в том, что на него сильно влияют отдельные выдающиеся значения признака; кроме того, его трудно вычислить, когда много результатов оказывается за порогом метода анализа. Мода лучше характеризует геохимический фон. Однако выборочная мода может сильно отличаться от моды в генеральной совокупности вследствие влияния ошибок анализа, произвольного задания ширины интервала группирования, неучтенного тренда в расположении проб и т. д. Кроме того, некоторые распределения вообще не имеют моды (например, в «гиперболическом» распределении модальное содержание равно нулю), в других (плосковершинных и полимодальных) нахождение моды связано с неопределенными ошибками.

Недостатков этих статистик, по мнению Б. Я. Юфы и Ю. М. Гурвича (1964), лишена медиана, т. е. срединное значение в ранжированном вариационном ряду. Она мало зависит от крайних «выскакивающих» значений вариационного ряда; ее можно определить, даже если до 25% проб оказалось «за порогом» чувствительности

анализа; она устойчивее средней арифметической в условиях эксцессивных (островершинных) распределений; наконец, она совершенно не зависит от распределения случайной величины, ибо ее положение не изменяется при любых преобразованиях изучаемой величины, например, при логарифмировании. В качестве меры рассеяния вокруг медианы авторы рекомендуют применять величины $\Delta_в$, $\Delta_н$, находимые с помощью двух квартилей (Q) — абсцисс функции распределения, отвечающих 25 и 75% накопленных частот:

$$\left.\begin{aligned} \Delta_в &= Q_3 - \mathrm{Me}, \quad \Delta_н = \mathrm{Me} - Q_1; \\ \sigma_в &= \frac{3}{2}\Delta_в, \qquad \sigma_н = \frac{3}{2}\Delta_н. \end{aligned}\right\} \tag{4.3.1}$$

Опыт показал, что применение медианы весьма удобно в практике геохимических исследований элементов-примесей, содержания которых характеризуются обычно сильнейшей изменчивостью. К тому же природные дисперсии увеличиваются аналитическими, поскольку определение большой группы элементов-примесей производят обычно при помощи грубого полуколичественного спектрального анализа. В этих условиях среднее арифметическое, найденное по небольшим выборкам, неустойчиво, и геохимики предпочитают пользоваться средним геометрическим. Но для логнормальных распределений среднее геометрическое совпадает с медианой. При обработке массовых полуколичественных спектральных анализов в баллах (см. гл. 2) переход от среднего балла b к среднему в весовых процентах также означает нахождение медианы:

$$b = a \cdot \overline{\log x} + b_0, \quad \mathrm{Me} = x_{\text{геом}} = 10^{\frac{b-\bar{b}}{a}}. \tag{4.3.2}$$

Часто, особенно в работах американских геохимиков, публикуются одновременно обе оценки средних: среднее арифметическое и медианное. Первое служит для характеристики кларкового содержания, второе — целям сравнения объектов друг с другом.

Мы коснулись проблемы фона лишь в связи с видами средних и их применением, не затрагивая вопроса выделения геохимических аномалий, которому посвящена обширная литература. Общепринятых методов отделения фона от аномалий по выборкам до сих пор нет, хотя предложено немало эмпирических приемов. Сразу же заметим, что критерий распознавания аномалий может быть построен лишь на основе информации о распределении как фоновых значений, так и аномальных. Когда такой информации нет, то обычный подход к выделению фоновых и аномальных концентраций заключается в произвольном назначении верхнего предела фона A или нижнего предела аномалии в виде неравенства

$$A \geqslant \hat{x} + t_q \sigma, \tag{4.3.3}$$

где $\hat{x}$ — оценка среднего; σ — стандартное отклонение от этого среднего; t_q — коэффициент, зависящий от доверительной вероятности q. В случае, когда фоновые значения характеризуют медианой, нижний предел аномалии рекомендуют (Юфа, Гурвич, 1964) определять из соотношения

$$A = \mathrm{Me} + 1.5t\,(Q_3 - \mathrm{Me}). \qquad (4.3.4)$$

Задавая уровень значимости q, равный 2% (он не должен быть очень малым, чтобы не выделить большое количество «ложных» аномалий, но не должен быть и очень большим, чтобы не пропустить «настоящую» аномалию), авторы указывают, что в условиях любых распределений величина t_q (односторонний критерий) не превысит 4. Таким образом, формула для назначения аномального содержания приобретает вид

$$A = \mathrm{Me} + 6\,(Q_3 - \mathrm{Me}), \qquad (4.3.5)$$

где Q_3 — третий квартиль распределения.

§ 4. К вопросу об оценке средних в условиях известного закона распределения

Рассмотрим следующий простой случай. Требуется найти медиану по выборке из логнормального распределения. Простейший способ заключается в нахождении $(n-1)/2+1$-го члена ранжированного ряда значений. Более эффективный способ состоит в расчете среднего арифметического значения логарифмов (как оценки математического ожидания величины $\log x$). В условиях нормального распределения логарифмов эта величина служит одновременно и оценкой медианы. Известно, что оценка среднего значения при нормальном распределении в 1.571 раза эффективнее оценки медианы, так что полученная этим способом оценка медианы более чем в полтора раза эффективнее обычной ее оценки (значения среднего номера ранжированного ряда).

Повышение эффективности оценки связано здесь с двумя моментами. Во-первых, с выбором статистики, использующей больше информации: расчет среднего арифметического включает всю имеющуюся информацию, медианы — нет. Во-вторых, возможность использовать среднее арифметическое логарифмов как оценки медианы опиралась на знание закона распределения, и тонкость заключается в том, что закон распределения должен быть известен априори. Суждение об этом на основе той же выборки возвращает нас в замкнутый круг, создавая лишь видимость увеличения эффективности. В дальнейшем мы столкнемся с целым рядом случаев принципиально такого же рода.

Для логарифмически нормального распределения между математическим ожиданием μ, медианой γ и логарифмической дисперсией $\sigma^2_{\ln x} = \sigma^2_*$ существует зависимость

$$\mu = \gamma \cdot e^{\frac{1}{2}\sigma_*^2}, \quad \ln \mu = \ln \gamma + \frac{1}{2}\sigma_*^2, \quad \lg \mu = \lg \gamma + 0.217\sigma_*^2. \qquad (4.4.1)$$

Кроме этого, существует «обычная» оценка для математического ожидания — среднее арифметическое $\bar{x}$. Х. Сишелом в 1952 г. было показано (цит. по: Криге, 1968), что оценка, полученная на основе (4.4.1) методом максимального правдоподобия (m), обладает большей эффективностью, чем $\bar{x}$. Для этого необходимо воспользоваться выборочными оценками величин γ и σ_*^2:

$$\gamma = e^{\overline{\ln x}}, \quad \ln \gamma = \overline{\ln x};$$
$$s_*^2 = \frac{1}{n}\sum (\ln x_i - \overline{\ln x})^2. \qquad (4.4.2)$$

Оценка для m дается выражением

$$m = e^{\overline{\ln x}} \cdot \Psi(s_*^2, n). \qquad (4.4.3)$$

Функция Ψ табулирована и часть ее приводится в работе Д. Г. Криге (1968, стр. 254). Если логарифмическая дисперсия σ_*^2 известна, то значение m получается непосредственно из (4.4.1). Такая методика оценки посредством так называемого логнормального среднего стала известна многим советским геохимикам благодаря статье Д. А. Родионова (1962). Таким образом, в условиях выборок из логнормальных совокупностей эффективность оценки математического ожидания может быть повышена; казалось бы, следует в тех случаях, когда логнормальный закон не отвергается, вычислять статистику m вместо статистики $\bar{x}$. К сожалению, эта рекомендация не всегда дает выигрыш в точности. Дело в том, что наше заключение о логнормальном законе распределения есть не более чем результат проверки статистической гипотезы, т. е. этот результат не точен. Выигрыш получается лишь тогда, когда закон распределения известен априори или по крайней мере на основании более широких исследований. «Формула логнормального среднего дает более точную оценку, чем среднее арифметическое в том случае, когда истинное (неизвестное) распределение исходной совокупности является логнормальным, а данные опробования следуют логнормальному закону лишь приближенно» (Карлье, 1966, стр. 126). Поскольку статистический вывод о логнормальности делают, как правило, по той же самой выборке, по которой рассчитывают статистики, то выигрыш в точности обесценивается проигрышем за счет неточности диагноза самого закона распределения! Таким образом, если исследователь не располагает никакой дополнительной информацией о законе распределения, кроме той, которую дает ему выборка, нахождение «логнормального среднего» лишь создает иллюзию преимущества в точности по сравнению с обычным средним арифметическим.

Аналогичная методика существует и для более эффективной оценки дисперсии логарифмически нормального распределения. Д. А. Родионов (1962) приводит следующую формулу для ее расчета:

$$b^2 = e^{2\overline{\ln x}} \cdot \left[\Psi_n(2s_*^2) - \Psi_n\left(\frac{n-2}{n-1}s_*^2\right)\right], \qquad (4.4.4)$$

где s_*^2 — дисперсия натуральных логарифмов; b^2 — дисперсия изучаемой величины; Ψ_n — уже упоминавшаяся функция, аргументами которой является число наблюдений n и выражение в скобках, содержащее логарифмическую дисперсию. Все прежние замечания о фиктивной точности относятся и к этому случаю.

Следует заметить, что логарифмическая дисперсия при не очень больших ее значениях численно равна так называемой относительной дисперсии — квадрату коэффициента вариации (см. прим. Ю. В. Рощина на стр. 22 книги Э. Карлье, 1966).

§ 5. Влияние функции распределения погрешностей анализа на смещение среднего значения

В главе 2 указывалось, в частности, что при спектральном анализе нормальному распределению подчиняются логарифмы содержаний. Если одна проба проанализирована n раз, то несмещенной оценкой среднего будет антилогарифм среднего арифметического логарифма. Среднее арифметическое нелогарифмированных значений будет смещенной оценкой истинного содержания в пробе. Величина смещения в логарифмическом масштабе зависит от дисперсии воспроизводимости (также в логарифмах):

$$\begin{aligned} \ln \bar{x} - \overline{\ln x} &= 0.5\sigma^2_{*\text{воспр}}, \\ \lg \bar{x} - \overline{\lg x} &= 1.15\sigma^2_{*\text{воспр}}, \\ \varepsilon = \lg \Delta &= 1.15\sigma^2_{*\text{воспр}}. \end{aligned} \qquad (4.5.1)$$

Из этого следует важный практический вывод: при расчете среднего арифметического какой-либо совокупности, проанализированной спектральным способом, необходимо учитывать смещение, особенно, если это полуколичественный анализ, характеризуемый низкой воспроизводимостью. Без преувеличения можно сказать, что расчет среднего арифметического (после перевода баллов-логарифмов в содержания) без учета этого смещения является самой распространенной ошибкой, приводящей к завышению содержаний.

Математическое ожидание завышения (в абс. ед.) при анализе каждой пробы ($\varepsilon_i = \lg \Delta_i$), как видно из (4. 5. 1), зависит от содержания, т. е. налицо пропорциональный тип систематической ошибки. Она устраняется вычитанием из полученного результата среднего арифметического $\bar{x}$ или логнормального среднего m (при

логнормальном распределении содержаний изучаемой совокупности), рассчитываемого методом Сишела-Родионова — величины $\bar{x} \cdot 1.15\ \sigma^2_{\text{воспр}}$, где $\sigma^2_{\text{воспр}}$ — логарифмическая дисперсия воспроизводимости анализа. Если ее величина составляет 0.5 (стандартное отклонение равно 0.7 порядка, что соответствует полуколичественному анализу плохого качества), то завышение составит почти 60% среднего значения!

§ 6. Среднее взвешенное и среднее арифметическое. Совместная оценка содержания и мощности

До сих пор мы рассматривали «правильное» опробование, в котором пробы имели равные статистические веса: каждая проба была равноправным представителем генеральной совокупности. В практике геохимических исследований и разведки месторождений это довольно редкий случай. Чаще пробы, например бороздовые, отобранные по падению и простиранию жилы или пласта, представляют большее или меньшее количество материала в зависимости от мощности тела. Их «веса» не остаются постоянными, в связи с чем возникает проблема взвешивания.

Вопрос применимости средних взвешенных и средних арифметических оценок содержаний является предметом длительной дискуссии среди геологоразведчиков. Квалифицированных работ, подводящих итоги этой дискуссии, не появилось, и в них, пожалуй, отпала необходимость. В связи с развитием и применением геостатистики этот вопрос в геологоразведке получил новое освещение, тогда как в геохимии, напротив, интерес к нему только зарождается. Учитывая, что геохимическое опробование в целом обычно ведется по редкой сети, геостатистические методы здесь еще не столь актуальны.

Краткие итоги дискуссии о взвешенных средних заключаются в следующем.

1. Среднее взвешенное дает несмещенную оценку запасов, среднее арифметическое может оказаться смещенным, т. е. давать систематические ошибки.

2. Среднее арифметическое содержание дает смещенную оценку запасов в случае наличия положительной или отрицательной корреляции между мощностью и содержанием. В первом случае запасы будут занижены, во втором — завышены. При отсутствии корреляции между указанными параметрами среднее арифметическое позволяет получить несмещенную оценку.

3. Дисперсия и коэффициент вариации среднего арифметического меньше, чем соответствующие величины среднего взвешенного.

К сожалению, не все из этих выводов правильно понимаются и используются. В частности, сначала мы покажем, в каких условиях среднее арифметическое дает действительно более эф-

фективную оценку, чем среднее взвешенное. Далее будут обсуждены пути увеличения точности средних взвешенных путем включения в расчет точек опробования, в которых измерены не все необходимые параметры, например только мощность или только содержание. Такие данные далее называются *непарными*. *Решение поставленных вопросов дано для* случаев наличия прямой и косвенной информации о корреляционной связи между отдельными параметрами.

Дальнейшее изложение ведется на примере расчета среднего содержания какого-либо химического элемента и его «запаса» в каком-либо пласте или жиле. При этом мы будем применять геологоразведочную терминологию — содержание, мощность, метропроцент, *запас*, *объемный вес*, *сечение* и т. д.

При подсчете запасов среди различного рода средних содержаний применяется такое, каким обладало бы количество руды Q во всех своих частях, чтобы запасы компонента в ней P остались прежними (см., например, В. Петров, 1965): $c=P/Q$.

Аналогично за средний объемный вес принимается такой, каким он должен быть во всех частях руды при условии сохранения ее количества и объема рудного тела V, т. е. $\rho=Q/V$. Наконец, для средней мощности имеем $\bar{m}=V/S$, где S — площадь тела (сечения).

Из этих примеров видно, что такие понятия, как средние значения мощностей, содержаний и объемных весов, являются производными от других понятий: запасов компонента, руды и объема рудного тела. В этом смысле можно сказать, что запасы как инвариант служат основой при определении этих понятий.

В самом деле, если $P=\bar{c}\bar{\rho}\bar{m}S$, а $\bar{c}=\frac{\sum m_i c_i \rho_i}{\sum m_i \rho_i}$, $\bar{\rho}=\frac{\sum m_i \rho_i}{\sum m_i}$, $\bar{m}=\frac{\sum m_i}{n}$, то

$$P=\frac{S}{n}\sum m_i c_i \rho_i=\sum m_i c_i \rho_i \Delta S=\sum p_i. \qquad (4.6.1)$$

Подсчет запасов, как видно из равенства (4. 6. 1), является по существу суммированием запасов в ячейках, охарактеризованных отдельными пробами, а содержание — производным понятием. Эта оценка запасов является состоятельной, так как при неограниченном увеличении числа проб (неограниченном уменьшении *ячейки и в пределе совмещения ее с пробой) совпадает с* истинными запасами. Из физического смысла этой оценки вытекает, что она является также несмещенной, если только замеры мощности и содержания не имеют систематических ошибок.

Пусть мы имеем в n точках замеры мощностей m и содержаний c, т. е., по нашей терминологии, располагаем только парными данными, и не имеем никакой другой информации об интересую-

щих нас параметрах. Из выборочного значения коэффициента корреляции [1]

$$r = \frac{1}{s_m s_c}\left(\frac{1}{n}\sum m_i c_i - \frac{1}{n}\sum m_i \frac{1}{n}\sum c_i\right), \qquad (4.6.2)$$

где s_m, s_c — оценки стандартных отклонений соответствующих величин, имеем

$$\overline{mc} = \bar{m}\cdot\bar{c} + \mathrm{cov}_{mc}. \qquad (4.6.3)$$

Следовательно, при наличии корреляции произведение средних арифметических мощности и содержания, представленное первым слагаемым правой части равенства (4.6.3), является смещенной оценкой среднего взвешенного (среднего метропроцента) на величину ковариации. Оно меньше взвешенного при положительной корреляции и больше его — при отрицательной. При отсутствии корреляционной связи эти оценки тождественно равны, и вопрос о преимуществе какой-либо из них отпадает. Бо́льшая простота расчета среднего арифметического оказывается кажущейся: применимость среднего арифметического необходимо прежде обосновать, а для этого потребуется расчет коэффициента корреляции, включающий в себя по сути расчет обеих названных оценок.

Рассмотрим дисперсии правой и левой частей равенства (4.6.3). Дисперсия среднего значения метропроцента может быть вычислена обычным способом по результатам n выборочных значений этой величины. Для нас будет иметь значение ее выражение через дисперсии мощности и содержания. Если колебания этих признаков малы, то

$$s^2_{\overline{mc}} = \frac{1}{n}s^2_{mc} \approx \frac{1}{n}(s^2_m\bar{c}^2 + s^2_c\bar{m}^2 + 2\,\mathrm{cov}_{mc}\,\bar{m}\bar{c}). \qquad (4.6.4)$$

Более типичным при разведке случаем является как раз противоположный — когда колебания признаков велики, и предыдущее равенство будет давать лишь грубое приближение. Тогда воспользуемся правой частью равенства (4.6.3), состоящей из двух независимых слагаемых (произведение $\bar{m}\cdot\bar{c}$ зависит от набора значений этих величин, но не зависит от их изменчивости и сочетаний; ковариация, напротив, зависит от дисперсий этих признаков и способа их сочетаний, но не зависит от средних значений). Имеем:

$$s^2\{\bar{m}\bar{c} + \mathrm{cov}_{mc}\} \approx s^2_{\bar{m}\bar{c}} + s^2_{\mathrm{cov}_{mc}}. \qquad (4.6.5)$$

Дисперсия первого слагаемого есть выражение в скобках в формуле (4.6.4) и обладает меньшими погрешностями, так как дисперсия средних арифметических по сравнению с дисперсиями признаков уменьшилось в n раз:

$$s^2_{\bar{m}\bar{c}} = s^2_{\bar{m}}\bar{c}^2 + s^2_{\bar{c}}\bar{m}^2 + 2r_{\bar{m}\bar{c}}s_{\bar{m}}s_{\bar{c}}\bar{m}\bar{c}.$$

[1] Подробнее о корреляции и описывающих ее величинах см. в гл. 7.

Дисперсию ковариации находим как дисперсию произведения трех независимых величин: $s^2_{\text{cov}} = s^2_{(r,\, s_m,\, s_c)}$. После небольших преобразований получаем $s^2_{\text{cov}} = 1/n\, s^2_m s^2_c [(1 - r^2)^2 + r^2]$, где n — число пар, по которым оценен коэффициент корреляции. Тогда, пренебрегая ошибками в определении s^2_m и s^2_c, можно записать

$$s^2_{\overline{mc}} = s^2_{\bar{m}} \bar{c}^2 + s^2_{\bar{c}} \bar{m}^2 + 2r_{\bar{m}\bar{c}} s_{\bar{m}} s_{\bar{c}} \bar{m}\bar{c} + \frac{1}{n} s^2_m s^2_c [(1 - r^2)^2 + r^2]. \quad (4.6.6)$$

При $r = 0$ равенство (4.6.5) превращается в точную формулу для дисперсии произведения двух независимых величин:

$$s^2_{\overline{mc}} = \frac{1}{n} s^2_{mc} = \frac{1}{n} (s^2_m \bar{c}^2 + s^2_c \bar{m}^2 + s^2_m s^2_c). \quad (4.6.6а)$$

Сравним теперь дисперсию среднего взвешенного (4.6.6а) и дисперсию среднего арифметического:

$$s^2_{\bar{m}\bar{c}} = s^2_{\bar{m}} \bar{c}^2 + s^2_{\bar{c}} \bar{m}^2 + s^2_{\bar{m}} s^2_{\bar{c}}.$$

Разность между ними равна

$$\frac{s^2_m s^2_c}{n} - \frac{s^2_m s^2_c}{n^2} \approx \frac{s^2_m s^2_c}{n} = s^2_{\text{cov}}, \quad (4.6.6б)$$

т. е. в точности равна дисперсии ковариации. В данном случае при $r = 0$ дисперсия ковариации s^2_{cov}, определенная по n парам значений, равна $s^2_m s^2_c / n$. Впрочем, это непосредственно видно из равенства (4.6.5).

Можно было бы подумать, что в этих условиях, при наличии n пар данных, указывающих на отсутствие значимого отклонения выборочного коэффициента корреляции от нуля и как следствие этого — справедливости равенства

$$\overline{mc} = \bar{m} \cdot \bar{c}, \quad (4.6.7)$$

необходимо применять среднее арифметическое как обладающее меньшей дисперсией и вследствие этого будто бы дающее бо́льшую точность. Именно в таком смысле и был многими понят изложенный в начале этого параграфа один из результатов дискуссии по средним. Следует заметить, что в таком истолковании кроется серьезная методологическая ошибка. Уменьшение дисперсии среднего арифметического здесь только кажущееся, так как нулевое значение ковариации, позволяющее записать равенство (4.6.7), не является точным, а имеет характер статистического равенства и обладает указанной выше (4.6.6б) дисперсией. Прибавление ее приводит к прежнему результату. Практический вывод из этих результатов таков: необходимо использовать среднее взвешенное во всех случаях, в том числе и тогда, когда выборочное значение коэффициента корреляции незначимо отличается от нуля. Исключение составляют весьма редкие ситуации, когда отсутствие корреляции установлено не на основании используемой выборки (n замеров мощностей и содержаний), а с привлече-

нием дополнительной информации: другие выборки, аналогия, геологические соображения. Только в этом случае мы можем быть уверены в получении более эффективной оценки при использовании среднего арифметического. Иллюстрацию данных выводов читатель найдет в статье Ю. А. Ткачева (1974).

Нередко при разведке месторождений одни из изучаемых признаков могут быть определены в большем числе пунктов, чем другие. Например, из-за недостаточного выхода керна содержание компонента в некоторой части скважин оказывается непредставительным, в то время как мощность рудного тела может быть определена с помощью каротажа достаточно точно. В горных выработках, пройденных по простиранию или падению рудных тел и вскрывающих их на полную мощность, число проб для определения содержания ограничивается экономическими соображениями, а мощность может быть замерена по геологической документации практически в неограниченном числе точек. Характер изменения содержания в интервале между пробами в этом случае не известен, а характер изменения мощности можно непосредственно наблюдать. Можно представить себе также случай, когда по большему числу точек известно содержание. Таким образом, при разведке могут быть получены парные данные по некоторому числу точек и непарные, когда известен лишь один из параметров. Вопросы методики уточнения запасов путем использования непарных данных в литературе не обсуждались.

Пусть имеется n сопряженных замеров мощностей и содержаний и k замеров мощностей в других точках, где содержание не определено. В настоящее время такие непарные данные при подсчете запасов могут использоваться лишь в тех редких случаях, когда обоснована применимость средних арифметических. В этих случаях, естественно, $\bar{m}$ и $\bar{c}$ рассчитываются по всем замерам мощностей и содержаний. Среднее значение метропроцента $\overline{mc}$ по n точкам не является теперь лучшей оценкой, так как часть данных (дополнительные замеры мощностей) остается неиспользованной. Действительно, не можем же мы вычислить значение метропроцента в точках, где неизвестны значения одного из параметров! Именно по этой причине в огромном числе случаев вопреки сложившемуся мнению мы при расчете не используем всей имеющейся информации и часто не подозреваем о возможности ее использования. Между тем оценку среднего метропроцента $\overline{mc}$ можно дать косвенным расчетом по формуле (4.6.3) $\overline{mc}=\bar{m}\bar{c}+\mathrm{cov}_{mc}$. Из этого равенства в результате преобразований получаем другое, справедливое для рассматриваемого случая, когда при расчетах используются непарные данные:

$$\overline{mc}_{n+k} \approx \bar{m}_{n+k}\bar{c}_n + \mathrm{cov}_n \frac{s^2_{m,\,n+k}}{s^2_{m,\,n}}. \qquad (4.6.8)$$

Здесь расчет средних арифметических $\bar{m}_{n+k}$ и $\bar{c}_n$, входящих в равенство (4. 6. 8), производится по всем имеющимся для каждого параметра данным:

$$\bar{m}_{n+k} = \frac{1}{n+k} \sum_{i=1}^{n+k} m_i; \quad \bar{c}_n = \frac{1}{n} \sum_{i=1}^{n} c_i.$$

В расчет ковариации между мощностью и содержанием cov_{mc}, естественно, входит только n парных замеров:

$$\mathrm{cov}_n = \frac{1}{n} \sum (m_i - \bar{m})(c_i - \bar{c}). \qquad (4.6.9)$$

Отличие формулы (4. 6. 8) от (4. 6. 3) заключается, кроме того, в поправочном коэффициенте $s^2_{\bar{m},\,n+k}/s^2_{\bar{m},\,n}$, зависящем от значений мощностей в дополнительных точках их замера.

Покажем, что точность оценки среднего значения метропроцента, полученной по предложенным формулам с использованием непарных данных, — выше, чем при использовании только парных данных по формуле $\overline{mc} = \frac{1}{n} \sum_{i=1}^{n} m_i c_i$.

Дисперсия полученной оценки определяется, как и прежде, суммой дисперсий слагаемых в формуле (4. 6. 8). Уменьшение дисперсии такой оценки по сравнению с дисперсией среднего метропроцента из парных замеров происходит за счет двух обстоятельств: а) уменьшения дисперсии величины m ввиду большего числа определений мощности, б) более точного определения коэффициента корреляции. Этот второй случай менее распространен, но и он возможен. Например, если при изучении ряда объектов известно, что на данном типе месторождений не наблюдается корреляция между признаками, т. е. можно считать, что $\mathbf{M}(r) = \rho = 0$, тогда $\overline{mc} = \bar{m} \cdot \bar{c} + 0$ и $s^2_{\overline{mc}} + s^2_{\mathrm{cov}} = s^2_{\bar{m} \cdot \bar{c}}$, так как благодаря знанию ρ дисперсия оценки ковариации равна нулю. При этом следует еще раз заметить, что значение коэффициента корреляции должно быть известно не на основании имеющейся выборки, а по значительно большему косвенному материалу и более точно.

Возможен, наконец, случай, когда наряду с парными данными имеется дополнительная информация как в виде непарных замеров мощностей, так и в форме точного значения коэффициента корреляции. Увеличение эффективности оценки за счет более полного использования имеющейся информации здесь еще более значительно: в «благоприятных» условиях, как показано на нескольких модельных примерах (Ткачев, 1974), оно может достигать 3—4-кратного. Итак:

1. В большинстве практических случаев при оценке средних значений и подсчете запасов наиболее эффективной и несмещен-

ной оценкой является среднее взвешенное. При отсутствии другой информации оно обладает наименьшей дисперсией.

2. Среднее взвешенное необходимо использовать даже тогда, когда выборочный коэффициент корреляции между мощностью и содержанием равен нулю или незначительно отличается от него. Меньшая дисперсия среднего арифметического в этих случаях является фиктивной, следствием методически неверного применения формул.

3. В большинстве практических случаев в наличии имеется дополнительная информация в виде «непарных» замеров одного из признаков по большему количеству точек. Ее использование приводит к уменьшению дисперсии оценки в большей или меньшей степени в зависимости от того, какую долю составляют непарные данные. Аналогичным образом для увеличения эффективности оценки можно использовать и данные, уточняющие коэффициент корреляции, например аналогию, геологические соображения о наличии корреляции или ее отсутствии. Практически дисперсия за счет более полного использования информации может быть понижена в 3—4 раза.

Перечисленные выводы в равной степени относятся и к объемному весу пород и полезных ископаемых, так как операции подсчета запасов и оценки средних симметричны относительно всех трех параметров: мощности, содержания и объемного веса.

Глава 5

ОЦЕНКА ПАРАМЕТРОВ НЕОДНОРОДНЫХ СОСТАВНЫХ СОВОКУПНОСТЕЙ

§ 1. Вводные замечания

Все чаще и чаще исследователи сталкиваются с необходимостью сопоставлять и обобщать данные разных авторов или свои собственные, характеризующие отдельные части сложного неоднородного объекта, т. е. объединять выборки различной представительности с отысканием *сложных средних*. При этом выборки в общем случае могут отличаться численностью (объемом), а значения изучаемых величин могут быть получены в результате опробования и анализа различными методами. Оценки среднего значения в изучаемом объекте, вычисленные по разным выборкам, будут отличаться значением дисперсии σ^2:

$$\sigma^2 = \mathbf{M}(\hat{\Theta} - \Theta)^2, \qquad (5.1.1)$$

где $\hat{\Theta}$ — оценка параметра, Θ — его истинное значение.

Задача состоит в том, чтобы по данным разных выборок получить оценку среднего, свободную от грубых промахов и характеризующуюся минимальной (насколько это возможно) дисперсией.

§ 2. Оценка среднего значения в простом объекте с помощью нескольких серий измерений

Под простым объектом подразумевается такой, в котором значение признака можно получить в результате одного измерения. Это может быть мощность слоя *в данном разрезе*, содержание химического элемента или ингредиента *в единичной пробе*, твердость выбранного зерна минерала и т. д. Среднее содержание в протяженном неоднородном объекте практически нельзя охарактеризовать одной пробой. Такой объект не подходит под определение «простого».

С целью повышения точности можно провести несколько измерений или несколько *серий* измерений, выполненных различными методами. Пусть имеется m серий измерений одной и той же величины различными методами. Это могут быть радиометрические анализы одной и той же пробы, измерения мощности пласта в скважине с помощью различных видов каротажа, определение глубины залегания маркирующего горизонта различными геофизическими методами и т. д. Из этих серий, которые могут состоять из одного или n_i измерений ($i=1, 2\ldots, m$), необходимо получить среднее значение величины с минимально возможной дисперсией. Известно (например, Урбах, 1964, стр. 119), что это достигается взвешиванием оценок средних Θ_i, вычисленных по отдельным сериям измерений, на величины, обратно пропорциональные дисперсиям этих оценок. Значение Θ_i по каждой серии определяется очевидным образом:

$$\Theta_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}. \qquad (5.2.1)$$

Дисперсия оценки среднего [1] по i-й серии $\sigma^2_{\Theta_i}$ определяется точностью i-го метода измерения $h_i^2 = 1/\sigma_i^2$ и числом измерений n_i, откуда

$$\hat{\Theta} = \sum_{i=1}^{m} n_i h_i^2 \Theta_i \left(\sum_{i=1}^{m} n_i h_i^2 \right)^{-1}. \qquad (5.2.2)$$

Если все применяемые методы равноточны, то взвешивание производится по числу измерений в каждой серии, что соответствует вычислению среднего арифметического по совокупной выборке. При неравноточных методах эта среднеарифметическая оценка будет тем больше отличаться от вычисленной по формуле (5.2.2), чем сильнее отличаются по точности методы наблюдений и полученные с их помощью оценки средних. Дисперсия оценки параметра, рассчитанного по (5.2.2), вычисляется по известной формуле

$$\sigma^2_{\hat{\Theta}} = \left(\sum n_i h_i^2 \right)^{-1} = 1 \Big/ \sum \frac{n_i}{\sigma_i^2}, \qquad (5.2.3)$$

знаменатель которой есть сумма весовых коэффициентов формулы (5.2.2). Другими словами, эффективность h^2 оценки среднего одной и той же генеральной совокупности (при условии независимости серий) равна сумме эффективностей оценок каждой из этих серий. Легко убедиться, что дисперсия среднего арифметического из серий $\sigma_x^2 = \sum \sigma^2_{\Theta_i}/m^2$ или среднего арифметического из всех n

[1] Формулы этой главы относятся как к дисперсиям $\sigma^2_{\Theta_i}$, так и к их оценкам $s^2_{\Theta_i}$, поэтому мы будем употреблять первое из этих обозначений.

измерений $\sigma_{\bar{x}}^2 = \sum_{j=1}^{n} \sigma_j^2/n^2$ не меньше, чем $\sigma_{\hat{\theta}}^2$. Этим и оправдывается применение рекомендованной формулы (5. 2. 2) для оценки средних значений. Равенство значений среднеарифметической оценки и оценки, вычисленной по (5.2.2), как и равенство их дисперсий, достигается только в том случае, *когда точность* измерений в сериях одинакова.

Пример. Определить содержание урана в эталонной пробе по данным двух различных методов радиометрического анализа (табл. 8).

Таблица 8

Результаты определения урана двумя методами

Характеристики	Значения характеристик	
	по методу 1 ($i=1$)	по методу 2 ($i=2$)
Число повторений n_i	8	10
Дисперсия воспроизводимости метода σ_i^2	$1 \cdot 10^{-6}$	$4 \cdot 10^{-6}$
Стандартное отклонение σ_i, вес.%	$1 \cdot 10^{-3}$	$2 \cdot 10^{-3}$
Среднее арифметическое Θ_i, вес.%	0.012	0.009
Общее среднее по формуле (5. 2. 2)	$\hat{\theta} = \dfrac{0.012 \cdot 8/1 \cdot 10^{-6} + 0.009 \cdot 10/4 \cdot 10^{-6}}{8/1 \cdot 10^{-6} + 10/4 \cdot 10^{-6}} = 0.011$	
Дисперсия оценки общего среднего по формуле (5. 2. 3)	$\sigma_{\hat{\theta}}^2 = 1/8 \cdot 10^{-6} + 10/4 \cdot 10^{-6} = 1 \cdot 10^{-7}$	
Стандартное отклонение	$\sigma_\Theta = 1 \cdot 10^{-7} = 0.00032 = 3.2 \cdot 10^{-4}$	
	$\hat{\theta} = 0.0113\% \pm 0.0003\%$	

§ 3. Оценка среднего значения в изменчивом объекте при объединении результатов

В противоположность приведенному выше случаю здесь мы рассматриваем оценку сложного, протяженного и изменчивого объекта, например оценку среднего содержания в массиве, средней пористости пласта коллекторов или запаса руды в месторождении. Значение указанных величин нельзя определить одним измерением (по крайней мере, при современном состоянии техники!). Оно получается как результат обработки многих (часто — многих тысяч) измерений или анализов. Методика такой обработки изложена в главе 2. Здесь нас будет интересовать методика усреднения, необходимого при объединении результатов, которые получены по данному объекту различными исследователями, когда последние приводят только средние и дисперсии их оценок.

Пусть имеется несколько *независимых серий* измерений признака сложного изменчивого в пространстве объекта. Каждая серия — результат опробования одним автором — отличается от другой числом проб, их весом и видом, системой расположения и ориентировкой. Независимость серий означает, что сети опробования, примененные различными авторами, расположены независимым друг от друга образом, который неизвестен исследователю, объединяющему данные (в противном случае можно найти лучшую оценку).

Среднее значение и дисперсия его оценки по объединенной выборке определяются, как и в первом случае, по формулам (5. 2. 2) и (5.2.3). Отличительной особенностью расчета является то, что дисперсии оценок средних значений по отдельным сериям (выборкам), вообще говоря, не равны дисперсиям самих значений, деленным на число наблюдений, т. е. σ_i^2/n_i, как это было в предыдущем случае. Дело здесь в том, что теперь это не просто n_i измерений, а система опробования *пространственно упорядоченной случайной величины*. Формулы оценок средних и дисперсий должны быть переписаны здесь в более общем виде:

$$\hat{\Theta} = \sum_{i=1}^{m} \frac{\Theta_i}{\sigma_{\hat{\Theta}_i}^2} \Bigg/ \sum_{i=1}^{m} \frac{1}{\sigma_{\hat{\Theta}_i}^2}; \qquad (5.3.1)$$

$$\sigma_{\hat{\Theta}}^2 = 1 \Bigg/ \sum_{i=1}^{m} \frac{1}{\sigma_{\hat{\Theta}_i}^2}. \qquad (5.3.2)$$

Дисперсия оценки среднего по каждой выборке должна определяться здесь в соответствии с основными положениями *геостатистики*, например с помощью *вариограмм*. Лишь тогда, когда между значениями признака не наблюдается корреляции на удалениях, равных или меньших, чем расстояния между пробами в любой из объединенных систем опробования, дисперсия оценки среднего значения по каждой из них $\sigma_{\hat{\Theta}_i}^2$ станет равной σ_i^2/n_i.

С численными примерами расчета среднего значения параметра и его дисперсии, специально подобранными для этого случая и для всех остальных, описанных ниже (§§ 4—6), читатель может ознакомиться в работе Ю. А. Ткачева (1973).

§ 4. Оценка среднего значения в объекте по средним в частях этого объекта и в группе объектов по средним в отдельных объектах или подгруппах

Часто исследуемые объекты состоят из разнородных частей: пластов различного литологического состава, фациальных разновидностей пород в массиве, блоков разного минерального состава на месторождении и т. д. Части объекта, являясь самостоя-

тельными объектами исследований, могут и не отличаться по существу, но всегда наступает момент, когда исследователю необходимо объединить свои данные или данные различных авторов, чтобы определить среднее значение по всему объекту в целом. Существенно, что исследованные части полностью составляют объект. Например, требуется определить среднее содержание в пробе, составленной из нескольких проб, если известны содержания в них. Здесь вместе с объединением выборок происходит объединение генеральных совокупностей, представленных этими выборками. С другой стороны, эту ситуацию можно рассматривать как разделение генеральной совокупности на части и раздельное опробование частей, так называемое *слоевое опробование*.

Если изучаемым является *абсолютный признак* (количество элемента, запасы руды), который складывается при объединении совокупностей, то

$$\hat{\Theta} = \sum \Theta_i \tag{5.4.1}$$

и

$$\sigma^2_{\hat{\Theta}} = \sum \sigma^2_{\Theta_i}. \tag{5.4.2}$$

Пример — мощность объекта, когда она равна сумме мощностей его частей, его объем, масса, запасы руды и т. д. *Относительный признак* — усредняется при объединении совокупностей (среднее содержание, средняя мощность). В этом случае обладающая минимальной дисперсией оценка среднего значения по объединенной выборке определяется по формуле

$$\hat{\Theta} = \sum_{i=1}^{m} p_i \Theta_i, \tag{5.4.3}$$

где p_i — доли частей, составляющих генеральную совокупность $p_i = v_i/v$; v_i — объем i-й части; v — объем общей генеральной совокупности (объекта).[1] Дисперсия общей оценки равна

$$\sigma^2_{\hat{\Theta}} = \sum_{i=1}^{m} p_i^2 \sigma^2_{\Theta_i}, \tag{5.4.4}$$

$\sigma^2_{\Theta_i}$ — дисперсия оценки среднего значения в i-й части (*слое*) генеральной совокупности. Если объединяются части месторождения, то $\sigma^2_{\Theta_i}$ является функцией плотности сети и изменчивости месторождения, и лишь в случае редких сетей и изменчивых месторождений эта величина может быть определена делением дисперсии значений признака в отдельных пробах на их число. Формула (5.4.4) годится также для расчета относительной дисперсии оценки суммирующегося признака (вес, объем), т. е. величины $\sigma'^2 = \sigma^2_{\hat{\Theta}}/\Theta^2$.

[1] Объем — в статистическом смысле, хотя численно иногда совпадает с объемом в геометрическом смысле.

Тождественная по существу формула для расчета дисперсии относительной погрешности количества металла в объединяемых пробах приводится у Э. Карлье (1966, стр. 37, формула IV. 23).

В специальном случае, когда объемы выборок пропорциональны объемам слоев (*пропорциональное слоевое опробование*), эффективность оценки, определенной по формуле (5. 4. 3), по сравнению с простой среднеарифметической оценкой равна

$$\sigma^2_{\bar{x}}/\sigma^2_{\hat{\Theta}} \approx 1 + s^2_m/\sigma^2_b, \qquad (5.4.5)$$

где σ^2_m — междуслойная дисперсия; $\sigma^2_m = \Sigma p_i (\Theta_i - \hat{\Theta})^2$; σ^2_b — средняя внутрислойная дисперсия; $\sigma^2_b = \Sigma \sigma^2_i (v_i - 1)/(v-1)$. Если средние значения по слоям равны, т. е. $\sigma^2_m = 0$, то выигрыша в эффективности не достигается. Напротив, если между частями совокупности предполагается различие в средних, то предпочтительнее пропорциональное слоевое опробование. Например, при расчете среднего содержания, изменяющегося с глубиной, необходимо отдельно подсчитывать средние содержания по горизонтам и взвешивать их на зоны влияния горизонтов. Если, кроме того, по данным предварительного опробования известны внутрислойные дисперсии σ^2_i, то объемы выборок можно подобрать лучше, чем при пропорциональном опробовании слоев. Определив объемы выборок из слоев n_i по формуле

$$n_i = n \frac{v_i \sigma_i}{v} \Bigg/ \sum_{i=1}^{m} \frac{v_i \sigma_i}{v}, \qquad (5.4.6)$$

где n — общее число проб, получим *оптимальную расслоенную выборку*, в которой число проб из слоя пропорционально произведению объема слоя на внутрислойный стандарт. Дисперсия оценки по оптимальной расслоенной выборке равна

$$\sigma^2_{\hat{\Theta}} = \frac{1}{n} \left(\sum_{i=1}^{m} p_i \sigma_i \right)^2 - \frac{1}{v} \sum_{i=1}^{m} p_i \sigma^2_i \qquad (5.4.7)$$

и меньше, чем при любом другом размещении числа проб. Поэтому опробование рекомендуется проводить по крайней мере в два этапа, оценив на первом этапе относительную изменчивость выделенных слоев. При работах в один этап важен учет хотя бы приближенных сведений об изменчивости частей. Поэтому вполне оправдано более детальное опробование контактовых зон как наиболее изменчивых по сравнению с центральной частью, но при последующем усреднении необходимо правильно взвешивать результаты.

Часто возникает задача определения характеристик в г р у п п е более или менее разнородных объектов, например среднего содержания химического элемента в гранитных массивах определенного района, в формациях, типах месторождений и т. д., по со-

держанию этого элемента в отдельных объектах (Уилкс, 1967).

Если оценка производится в результате обобщения данных по отдельным опробованным объектам, причем опробован (изучен) каждый объект совокупности, и если объекты не объединяются в какие-либо подгруппы, то эта задача принципиально не отличается от предыдущей, хотя здесь имеется одна особенность, а именно вопрос о целесообразном выборе элементов генеральной совокупности. Ими могут быть возможные значения оценок изучаемого параметра в объектах, т. е. о б ъ е к т ы будут выступать в качестве равноценных неделимых элементов совокупности независимо от их размеров или других признаков.

В качестве элементов генеральной совокупности могут быть выбраны возможные значения изучаемого параметра во всей массе вещества, составленной из вещества изучаемых объектов, как например при определении средних содержаний элементов в типах горных пород для изучения баланса элементов в земной коре. Первый вариант более подходит для изучения некоторого процесса, поскольку для его усредненной характеристики важно число реализаций, т. е. число объектов. В соответствии с отмеченными выше особенностями постановки задачи весá p_i, которые входят в последующие формулы, должны пониматься либо как доли ч и с л а о б ъ е к т о в, либо как доли их м а с с в слоях. Для краткости в дальнейшем это оговариваться не будет. Все расчеты проводятся по формулам (5. 4. 1)—(5. 4. 7).

Немногим более сложен случай, когда изученные объекты объединяются в отдельные *подгруппы*, либо существенно отличающиеся, либо выделенные условно, например ввиду их опробования различными авторами. Особенно целесообразно объединять в подгруппы объекты одного типа. Например, при изучении терригенно-карбонатной толщи все пласты известняков могут быть объединены в одну подгруппу, песчаников — в другую, глин — в третью. Усреднение производится в два этапа: 1) по уже известным формулам получают оценки параметров в подгруппах из их значений в отдельных объектах; 2) производят окончательный расчет по оценкам в подгруппах. Пусть, например, имеется k литотипов по m_s пластов каждого литотипа (s=1, 2 . . . k). Тогда

$$\hat{\Theta} = \sum_{s=1}^{k} p_s \Theta_s, \qquad (5.4.8)$$

$$\sigma^2_{\hat{\Theta}} = \sum_{s=1}^{k} p_s^2 \sigma^2_{\hat{\Theta}_s}, \qquad (5.4.9)$$

где p_s — веса «слоев» (в данном случае — не пластов, а литотипов!). Величины $\sigma^2_{\hat{\Theta}_s}$ и $\hat{\Theta}_s$, т. е. дисперсия оценки по слою s и среднее, определяются в свою очередь по такого же типа формулам:

$$\hat{\Theta}_s = \sum_{i=1}^{m_s} p_{si} \Theta_{si}; \qquad (5.4.10)$$

$$\sigma^2_{\bar{\theta}_s} = \sum_{i=1}^{m_s} p^2_{si}\sigma^2_{\bar{\theta}_{si}}, \qquad (5.4.11)$$

где индекс si означает величину, относящуюся к i-му объекту s-го «слоя», т. е. по пласту. Дисперсия $\sigma^2_{\bar{\theta}_{si}}$ представляет собой обычную дисперсию оценки среднего значения в объекте, в данном случае — в пласте. В результате двухступенчатого усреднения будут получены точно такие же оценки средних с такой же точностью, какие мы получили бы сразу, не объединяя объекты в подгруппы. И все-таки такое объединение полезно по следующим причинам. Во-первых, попутно получают сведения по «подгруппе» однородных объектов, например по литотипам. Во-вторых, при обобщении данных различных авторов редко можно располагать такими подробными сведениями, которые позволяют вычислить $\sigma^2_{\bar{\theta}_s}$. Кроме того, приводимые сведения часто относятся не к отдельным объектам, а к их группам, составляющим «слой» данного исследователя. В этих условиях, предполагая, что объекты опробованы небольшим числом проб, отобранных независимо друг от друга, в качестве величины, приближенно характеризующей $\sigma^2_{\bar{\theta}_s}$, можно принять величину

$$\sigma^2_{\bar{\theta}_s} \approx \sigma^2_s/n_s, \qquad (5.4.12)$$

где n_s — число проб по s-му слою. Наконец, такая методика позволяет легко перейти к более сложному случаю, излагаемому ниже.

§ 5. Оценка среднего по частично опробованным группам объектов

Часто ставится задача оценить среднее значение параметра по обширной совокупности объектов данного типа, вплоть до среднего значения по всем объектам такого типа в земной коре. В отличие от предыдущего случая здесь опробованы не все объекты каждой подгруппы. Например, та же терригенно-карбонатная толща может состоять из очень большого числа пластов известняков, песчаников и глин. Из каждого литотипа опробованными могут быть лишь 2—5 пластов.

Важно, что объекты разных подгрупп (подтипов) отличаются друг от друга, и каждый подтип объектов (может быть, исследованных определенным автором) составляет самостоятельный слой опробования. Кроме того, величины p_s, т. е. статистические веса слоев, известны (например, доли литотипов в разрезе). Если взвешивание происходит по числу объектов, то $p_s = M_s/M$, где M_s — число объектов в слое s, M — общее число объектов. Среднее значение по каждому слою опробования, если оно не выведено самим автором, у которого заимствуются данные, определяется по формуле (5.4.10). Если взвешивание ведется по числу объектов, то $p_{si} = 1/m_s$, где m_s — число опробованных объектов в s-м слое.

Среднее значение генеральной совокупности подсчитывается взвешиванием полученных средних по слоям $\hat{\Theta}_s$ на доли слоев p_s в генеральной совокупности, т. е. по формуле (5.4.8). Дисперсия оценки общего среднего $\sigma^2_{\hat{\Theta}}$ определяется по формуле (5.4.9), так как сумма слоев составляет исследуемую совокупность.

Однако сумма объектов в слое не составляет всего слоя, поэтому дисперсия оценки среднего по s-му слою $\sigma^2_{\hat{\Theta}_s}$, на основе закона аддитивности дисперсий, будет состоять из двух слагаемых:

$$\sigma^2_{\hat{\Theta}_s} = \sum_{i=1}^{m_s} p^2_{si}\sigma^2_{\hat{\Theta}_{si}} + (1-p_s)\frac{1}{m_s^2}\sum_{i=1}^{m_s}(\hat{\Theta}_{si}-\hat{\Theta}_s)^2. \qquad (5.5.1)$$

Первое слагаемое нам уже знакомо и представляет обычную дисперсию оценки среднего, относящуюся только к опробованной части слоя. Второе слагаемое есть *дисперсия распространения* этого среднего на весь слой (включая и неопробованную часть). Опробование слоя (подгруппы) происходит по схеме безвозвратной выборки, поэтому дисперсия оценки среднего по слою (дисперсия распространения) равна внутрислойной дисперсии $\frac{1}{m_s}\sum(\hat{\Theta}_{si}-\hat{\Theta}_s)^2$, умноженной на коэффициент $(1/m_s - 1/M_s)$ *степени опробованности* слоя (ср. формулу (1.5.7)):

$$\left(\frac{1}{m_s}-\frac{1}{M_s}\right)\frac{1}{m_s}\sum_{i=1}^{m_s}(\hat{\Theta}_{si}-\hat{\Theta}_s)^2 = \left(1-\frac{m_s}{M_s}\right)\frac{1}{m_s^2}\sum_{i=1}^{m_s}(\hat{\Theta}_{si}-\hat{\Theta}_s)^2, \quad (5.5.2)$$

где $m_s/M_s = p_s$. Если в каждом слое опробованы все объекты, то $m_s = M_s$, и данный случай вырождается в предыдущий. Таким образом, отличительной особенностью этого сложного случая является то, что сумма слоев составляет общую генеральную совокупность, тогда как сумма опробованных объектов в слое не составляет целиком этого слоя.

Сопоставим теперь два способа расчета средних в изложенной ситуации: 1) с подразделением на слои, как изложено выше; 2) как общее средне-взвешенное, не принимая во внимание никакое подразделение на слои. Очевидно, что пласты одного и того же литотипа меньше отличаются по ряду изучаемых признаков, чем пласты различных литотипов. Поэтому, если мы сначала проведем расчет средних содержаний по отдельным типам пород (по группам или «слоям»), а затем усредним групповые средние с учетом распространенности пород, то получим большой выигрыш в точности.

В нашей работе (Ткачев, 1973б) приведены подробные примеры расчета таких оценок. Пример, относящийся к рассматриваемому случаю, показывает, как изменяется точность оценки в зависимости от взгляда на эту оценку, т. е. в зависимости от того, ч т о

подразумевать под объектом оценки. Так, одно и то же значение роста человека мы можем рассматривать: а) как оценку роста данного человека; б) как оценку среднего роста группы людей, из которой выбран этот индивидуум. Во втором случае по своему смыслу оценка будет менее точной за счет изменчивости роста внутри изучаемой группы людей, т. е. за счет второй составляющей в формуле (5. 5. 1). В различных случаях две составляющие дисперсии оценки находятся в различном отношении. В геологической ситуации обычно основной вклад в общую дисперсию вносят большие колебания содержания от пласта к пласту. Дисперсией оценки содержания в отдельных пластах, обусловленной изменчивостью содержания внутри пласта и плотностью сети опробования, нередко можно пренебречь.

§ 6. Оценка среднего по нескольким формально составленным группам объектов

В ряде случаев объекты, исследованные различными авторами, не выделяются какими-нибудь характерными признаками и их объединение в слой вызвано только необходимостью использовать уже обобщенные данные. Если нет других соображений, следует выделить из генеральной совокупности *подсовокупность*, состоящую из опробованных объектов. Тогда будем иметь подсовокупность, разделенную на «слои» (данные различных авторов), причем в каждом слое опробованы все объекты, составляющие данный слой. Оценка среднего и его дисперсии в подсовокупности сведется тогда к случаю из § 4. Расчеты производятся по формулам (5. 4. 8)—(5. 4. 11).

Дисперсия оценки среднего, если рассматривать ее по отношению ко всей генеральной совокупности, по закону аддитивности равна сумме дисперсии оценки среднего в подсовокупности, подсчитываемой по формулам (5. 4. 9) и (5. 4. 11), и дисперсии распространения этого среднего на всю генеральную совокупность. В зависимости от полноты имеющейся информации для приближенной оценки дисперсии распространения существует несколько вариантов. Если известны данные только по слоям, то в соответствии с (1. 5. 7) и (5. 5. 2) получим

$$(1-p)\frac{1}{k^2}\sum_{s=1}^{k}(\hat{\Theta}_s-\hat{\Theta})^2\approx\left(1-\frac{m}{M}\right)\frac{1}{k^2}\sum_{s=1}^{k}(\hat{\Theta}_s-\hat{\Theta})^2, \quad (5.6.1)$$

где p — доля опробованной подсовокупности в генеральной совокупности, m — общее число опробованных объектов; M — общее число объектов в генеральной совокупности. Если известны данные по отдельным объектам, то искусственное объединение в слои не имеет смысла, и общее средневзвешенное среднее будет тождественно равно среднему, полученному двухступенчатым способом через средние по слоям. Дисперсия оценки среднего (по отноше-

нию к среднему по всей генеральной совокупности!) и здесь будет состоять из двух слагаемых. Первое находят обычным способом через две ступени по формулам (5. 4. 9) и (5. 4. 11), или, что то же самое, непосредственно используя данные по отдельным объектам:

$$\sigma^2_{\hat{\Theta}} = \sum_{i=1}^{m} p_i^2 \sigma^2_{\hat{\Theta}_i}, \qquad (5.6.2)$$

где m — общее число опробованных объектов. Второе слагаемое можно оценить по формуле

$$(1-p)\frac{1}{m^2}\sum_{i=1}^{m}(\hat{\Theta}_i - \hat{\Theta})^2, \qquad (5.6.3)$$

незначительно отличающейся от (5. 6. 1). Если выбор объектов в слой случаен (именно такую ситуацию мы здесь и рассматриваем), то $\frac{1}{m^2}\sum_{i=1}^{m}(\hat{\Theta}_i - \hat{\Theta})^2 = \frac{1}{m}\sigma^2_{об} \approx \frac{1}{k^2}\sum_{i=1}^{k}(\hat{\Theta}_s - \hat{\Theta})^2 = \frac{1}{k}\sigma^2_m$, так как междуслойная дисперсия σ^2_M во столько раз меньше дисперсии между объектами $\sigma^2_{об}$, во сколько число объектов больше числа слоев. В пределе, когда $p=1$, получается случай из § 4. В другом предельном случае, когда p очень мало по сравнению с единицей, это второе слагаемое становится обратно пропорциональным объему подсовокупности.

Таким образом, окончательно имеем

$$\sigma^2_{\hat{\Theta}} = \sum_{s=1}^{k} p_s^2 \sigma^2_{\hat{\Theta}_s} + (1-p)\frac{1}{k^2}\sum_{s=1}^{k}(\hat{\Theta}_s - \hat{\Theta})^2, \qquad (5.6.4)$$

или

$$\sigma^2_{\hat{\Theta}} = \sum_{i=1}^{m} p_i^2 \sigma^2_{\hat{\Theta}_i} + (1-p)\frac{1}{m^2}\sum_{i=1}^{m}(\hat{\Theta}_i - \hat{\Theta})^2. \qquad (5.6.4а)$$

Рассмотренный в упомянутой статье (Ткачев, 1973б) относящийся к этой ситуации пример показывает, что из двух составляющих дисперсии — дисперсии оценки среднего в подсовокупности и дисперсии распространения этого среднего на всю совокупность — последняя имеет подавляющее значение, а первой практически можно пренебречь. Кроме того, оказывается, что точность оценки по сравнению с предыдущим вариантом значительно понизилась. Когда опробованы все объекты (пласты) каждого «слоя», а «слои» полностью составляют изучаемую совокупность (пример из § 4), дисперсия оценки равна $0.84 \cdot 10^{-6}$. Если «слои» состоят не только из опробованных пластов, а заключают большее их число, как в случае из § 5, соответствующая дисперсия равна $30 \cdot 10^{-6}$. Нако-

нец, если разными авторами опробовано по нескольку терригенных и карбонатных пластов, но результаты приведены в виде средних по изученным группам, то дисперсия оценки в толще равна $36.9 \cdot 10^{-6}$. Картина была бы еще более показательной, если бы в составе опробованных каждым автором пластов преобладала какая-нибудь одна литологическая разновидность.

Мы видим, что отступление от принципа «слоевого» опробования и формирование выборок без учета литологических (в общем случае — любых) особенностей приводит при том же количестве проб и анализов к значительной потере информации и уменьшению точности оценок.

§ 7. Весовые коэффициенты как случайные величины

В предыдущих разделах главы при оценке сложных средних предполагалось, что весовые коэффициенты, отражающие соотношение опробованных частей объекта или подсовокупностей, являются точными величинами, т. е. определенными без погрешностей. Это значительно упростило изложение. Кроме того, такой случай действительно не редок. При разведке месторождения его разделение на горизонты, на участки или блоки имеет характер назначения, и весовые коэффициенты как бы задаются геологоразведчиком. При четком и однозначном подразделении осадочного разреза на пласты или пачки определить соотношение их мощностей с высокой точностью также не представляет затруднений.

В большинстве других случаев погрешности оценки соотношения типов пород в регионе или объекте — одного порядка с погрешностями оценки содержаний. Поскольку оценка среднего рассчитывается как произведение (сумма произведений) двух случайных величин — значения параметра и весового коэффициента, которые мы будем считать независимыми,

$$\hat{\Theta} = \sum p_i \Theta_i, \qquad (5.7.1)$$

то дисперсия такой оценки определяется формулой

$$s^2_{\hat{\Theta}} = \sum \hat{p}_i^2 s^2_{\hat{\Theta}_i} + \sum \Theta_i^2 s^2_{p_i}. \qquad (5.7.2)$$

Величины p и Θ выступают в ней и в других аналогичных формулах этой главы как симметричные. Так, формулы (5.4.4), (5.4.9), (5.4.11), (5.5.1), (5.6.2), (5.6.4) и (5.6.4а), содержащие сумму (5.6.1), должны быть дополнены так же, как и формула (5.7.2), в случае, если весовой коэффициент — величина с погрешностью.

Любопытно, что весовой коэффициент и уменьшение его погрешности выступают как форма учета геологических данных для повышения надежности средних. К сожалению, методика определения погрешности (дисперсии оценки) весовых коэффициентов

при картировании, составлении разрезов, геохимическом опробовании и петрографических исследованиях разработана очень слабо. Пример принципиальной переоценки точности, связанной с весовыми коэффициентами, приведен в гл. 6.

§ 8. Распространенные ошибки при расчете сложных средних. Выводы

При кажущейся простоте определить, к какому случаю отнести ту или иную практическую ситуацию при расчете сложных средних и кларков, иногда бывает затруднительно. Ошибки и недоразумения чаще всего возникают из-за недостаточно внимательного анализа ситуации. В этом смысле показателен пример из статьи Д. А. Родионова и В. В. Иванова (1967). Требовалось оценить среднее содержание в месторождениях определенного типа, при этом была опробована только небольшая часть их. Исходные данные представлены в виде средних содержаний по небольшим группам месторождений (и отдельным месторождениям), которые в дальнейшем рассматриваются нами как слои, и в виде дисперсии результатов внутри каждого слоя. Таким образом, пример из названной статьи относится к случаям из §§ 5, 6, между тем как авторы при расчетах пользовались схемой случая из § 2! По приведенным у авторов отрывочным данным нельзя точно рассчитать дисперсию оценки среднего значения по каждому слою. Сделав несколько реальных допущений, мы рассчитали оценки параметров (средние и дисперсии), опирающиеся на подробный анализ ситуации (Ткачев, 1973б). При этом мы различали две точки зрения: а) считать полученное среднее оценкой содержания только в группе опробованных месторождений; б) считать его оценкой содержания вообще в месторождениях данного типа. В (а) дисперсия оценки подсчитывалась по формуле (5. 4. 9), в (б) — по (5. 6. 4), где первое слагаемое вычислялось по формуле (5. 4. 9), а второе — из данных по слоям, т. е. по (5. 6. 1). Прежде чем сравнивать полученные данные (табл. 9), необходимо еще раз оговориться, что для иллюстрации возможных отклонений ввиду недостатка исходных данных необходимо было сделать два допущения: а) дисперсия оценки среднего в слоях высчитана не геостатистическим способом; б) взвешивание производилось по числу объектов в слоях (а не по сумме запасов в этих объектах), причем число объектов принималось в некоторых группах условно.

Из табл. 9 видно, что оценки дисперсий, вычисленные для рассмотренного случая по приведенным выше формулам, во много раз превышают оценки, рассчитанные Д. А. Родионовым и В. В. Ивановым. Причина в том, что формулы, использованные ими, годятся лишь для первого и второго случаев объединения выборок, но не подходят для описанной ими ситуации. Это можно проиллюстрировать следующими рассуждениями: если диспер-

сия оценки среднего значения признака (например, мощность пласта в заданной точке) одним из способов равна нулю, то и суммарная дисперсия, как следует из формулы (5. 2. 3), также равна нулю. Однако вполне очевидна ошибочность утверждения, что оценка среднего значения по группе или даже целому типу

Таблица 9

Расхождение оценок точности средних в зависимости от метода вычислений

Месторождения	Оценка среднего значения			Оценка точности среднего значения (дисперсия)		
	средневзвешенная		среднеарифметическая по всей совокупности проб	по данным Д. А. Родионова и В. В. Иванова (1967)	если считать его оценкой в данной совокупности объектов	если считать его оценкой в данном типе месторождений
	весами являются точности оценок (Родионов, Иванов, 1967)	весами являются относительные объемы объектов				
Силикатно-сульфидные оловорудные	3430	3334	3366	15870 (126)	6180 (249)	77150 (278)
Высокотемпературные свинцово-цинковые	1872	2264	2050	6060 (77)	105900 (326)	163300 (404)
Среднетемпературные свинцово-цинковые в изверженных породах, песчаниках и сланцах	3195	3340	3559	4499 (67)	10900 (105)	131400 (362)

Примечание. В скобках — стандартное отклонение.

месторождений будет равна нулю, если среднее по одному или нескольким месторождениям определено абсолютно точно.

Из других возможных ошибок отметим формальное взвешивание, т. е. неправильный выбор весовых коэффициентов, смешивание дисперсии признака с дисперсией оценки среднего значения этого признака, упущение из виду погрешностей анализа и обработки проб как составных частей погрешностей в любом из рассмотренных случаев, наконец, неучет дисперсии распространения.

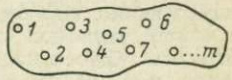
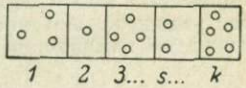
В связи с огромными темпами нарастания различной геологической информации, в частности информации о содержаниях химических элементов, о запасах и т. д., встала как никогда остро проблема обобщения данных, в том числе данных различных ав-

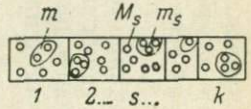
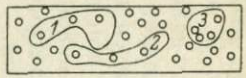
Таблица 10

Типовые ситуации при расчете сложных средних

Характеристика случаев	Оценка параметра	Дисперсия оценки параметра	Весовые коэффициенты	Примечания, примеры
1. m независимых серий по n_i измерений одной и той же величины (в точке), $i=1, 2 \ldots, m$. Среднее по i-й серии $\hat{\theta}_i$, дисперсия измерения в i-й серии σ_i^2	$\hat{\theta}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$, $\hat{\theta} = \sum_{i=1}^{m} \frac{n_i\hat{\theta}_i}{\sigma_i^2} \Big/ \sum_{i=1}^{m} \frac{n_i}{\sigma_i^2}$	$\sigma_{\hat{\theta}}^2 = \frac{1}{\sum_{i=1}^{m} \frac{n_i}{\sigma_i^2}}$	Все измерения одной серии равноточны, их веса одинаковы и равны $1/n_i$	Измеряемая величина не является средним значением пространственно изменчивых величин в пределах *заданного объема* (см. § 2)
2. m независимых серий *по n_i измерений* пространственно изменчивой величины. Сети опробования, составляющие отдельные серии проб, размещены на объекте независимым друг от друга образом. Каждая серия охватывает весь объект	$\hat{\theta} = \sum_{i=1}^{m} \frac{\hat{\theta}_i}{\sigma_{\hat{\theta}_i}^2} \Big/ \sum_{i=1}^{m} \frac{1}{\sigma_{\hat{\theta}_i}^2}$, $\sigma_{\hat{\theta}_i}^2$ — геостатическая оценка дисперсий среднего по i-й *системе опробования*	$\sigma_{\hat{\theta}}^2 = \frac{1}{\sum_{i=1}^{m} \frac{1}{\sigma_{\hat{\theta}_i}^2}}$	Весовыми коэффициентами для среднего, как и в предыдущем случае, являются величины, обратные дисперсиям	Если системы опробования расположены известным автору обобщения образом, то они объединяются в одну, что позволяет получить более эффективную оценку. Опробование месторождений, массивов, толщ различными авторами с выдачей *конечных* цифр. См. § 3
3. Объект разделен на m неперекрывающихся частей, каждая из которых охарактеризована n_i пробами, $i=1, 2, 3 \ldots m$	$\hat{\theta} = \sum_{i=1}^{m} p_i\hat{\theta}_i$, $\hat{\theta}_i$ определяется в зависимости от обстоятельств, как в случае 2	$\sigma_{\hat{\theta}}^2 = \sum_{i=1}^{m} p_i^2\sigma_{\hat{\theta}_i}^2$, $\sigma_{\hat{\theta}_i}^2$ определяется, *как в случае 2*	p_i являются долями частей объекта к целому (*в зависимости от* смысла — по весу, *площади или по* объему)	Опробование различных частей месторождений, слоев толщ, различных *разновидностей пород* в массиве и т. д. См. § 4

Таблица 10 (*продолжение*)

Характеристика случаев	Оценка параметра	Дисперсия оценки параметра	Весовые коэффициенты	Примечания, примеры
4. Оцениваемая совокупность состоит из m объектов, каждый из которых опробован	$\hat{\Theta} = \sum_{i=1}^{m} p_i \hat{\Theta}_i;$ $\hat{\Theta}_i$ рассчитывается в зависимости от обстоятельств, как в случае 1, 2 или 3	$\sigma^2_{\hat{\Theta}} = \sum_{i=1}^{m} p_i^2 \sigma^2_{\hat{\Theta}_i}$	Как в случае 3	Опробование различных массивов одного типа, одного района, различных пластов неоднородных толщ. См. § 4. 
5. Оцениваемая совокупность состоит из k групп объектов, каждая из которых состоит из m_s опробованных объектов; $s = 1, 2, 3 \ldots, k$, si означает i-й объект, s-й группы, n_s — число проб по s-й группе	$\hat{\Theta} = \sum_{s=1}^{k} p_s \hat{\Theta}_s.$ Если известны данные по отдельным объектам, то $\hat{\Theta}_s = \sum_{i=1}^{m_s} p_{si} \hat{\Theta}_{si},$ и этот случай вырождается в случай 4. $\hat{\Theta}_{si}$ определяется, как в случае 1—3	$\sigma^2_{\hat{\Theta}} = \sum_{s=1}^{k} p_s^2 \sigma^2_{\hat{\Theta}_s},$ $\sigma^2_{\hat{\Theta}_s} = \sum_{i=1}^{m_s} p_{si}^2 \sigma^2_{\hat{\Theta}_{si}},$ $\sigma^2_{\hat{\Theta}_{si}}$ определяется, как в случае 2 или 3, либо приближенно как σ_s^2/n_s	p_s — доля (по площади, объему или весу в зависимости от смысла) s-й группы объектов в общей совокупности; p_{si} — доля i-го объекта по отношению к объектам s-й группы	Если известны данные по отдельным объектам, то этот случай вырождается в случай 4. Оценка относится только к опробованным объектам. См. § 4. 

Характеристика случаев	Оценка параметра	Дисперсия оценки параметра	Весовые коэффициенты	Примечания, примеры
6. Оцениваемая совокупность состоит из k отличающихся групп объектов; s-я группа состоит из M_s объектов, из которых опробовано m_s. Общее число объектов M, общее число опробованных — m	$\hat{\theta}=\sum_{s=1}^{k} p_s\hat{\theta}_s$, $\hat{\theta}_s=\sum_{i=1}^{m_s} p_{si}\hat{\theta}_{si}$	$\sigma^2_{\hat{\theta}}=\sum_{s=1}^{k} p_s^2\sigma^2_{\hat{\theta}_s}$, $\sigma^2_{\hat{\theta}_s}=\sum_{i=1}^{m_s} p_{si}^2\sigma^2_{\hat{\theta}_{si}}+\left(1-\frac{m_s}{M_s}\right)\frac{1}{m_s^2}\times\sum_{i=1}^{m_s}(\hat{\theta}_{si}-\hat{\theta}_s)^2$. $\hat{\theta}_{si}$ определяется, как в случае 2 или 3 ($\hat{\theta}_{si}$ есть $\hat{\theta}_i$)	Как в случае 4 (вследствие определенности групп их весá p_s по отношению к генеральной совокупности вполне определены)	Оценка относится не только к опробованным объектам, но ко всем опробованным группам (в которых имеется и неопробованные объекты!). См. § 5 
7. Оцениваемая совокупность состоит из M объектов. Опробовано k групп, которые выделены в подсовокупность из m объектов, $m=\sum n_s$; s-я группа объединяет n_s объектов, каждый из которых опробован	$\hat{\theta}=\sum_{s=1}^{k} p_s\hat{\theta}_s$. Если известны данные по отдельным объектам, то среднее рассчитывается, как в случае 3 или 4: $\hat{\theta}=\sum_{i=1}^{m} p_i\hat{\theta}_i$	$\sigma^2_{\hat{\theta}}=\sum_{s=1}^{k} p_s^2\sigma^2_{\hat{\theta}_s}+\left(1-\frac{m}{M}\right)\times\frac{1}{m^2}\times\sum_{i=1}^{m}(\hat{\theta}_i-\hat{\theta})^2$. Если неизвестны данные по отдельным объектам, то второе слагаемое можно определить как $\left(1-\frac{m}{M}\right)\cdot\frac{1}{k^2}\sum_{s=1}^{k}(\hat{\theta}_s-\hat{\theta})^2$	p_s — весá отдельных групп в опробованной подсовокупности, p_{si} — весá отдельных объектов в ней, m/M — вес опробованной подсовокупности по отношению к генеральной	 $1+2+3$ — опробованная подсовокупность. См. § 6

торов. Наиболее «ходовой» метод обобщения — вычисление средних арифметических по всей совокупности данных — в большинстве случаев не годится, так как не учитывает относительную значимость обобщаемых данных и их взаимное положение в сложном геологическом объекте.

Анализ встречающихся в геохимии, геологии и геологоразведочном деле ситуаций при опробовании и интерпретации данных показывает, что здесь можно выделить по крайней мере семь случаев (табл. 10). Для каждого случая приводится формула расчета средних и их дисперсий.

Мы надеемся, что применение изложенных здесь способов оценки средних значений и их дисперсий по выборкам различной представительности позволит производить сводку данных различных авторов по одному и тому же объекту, ряду объектов группы и в целом по типам геологических образований.

Глава 6

ГЕОЛОГО-СТАТИСТИЧЕСКАЯ ПРОБЛЕМА КЛАРКОВ

По своему содержанию проблема кларков геологическая, а по методам обработки данных — типично статистическая, ибо касается нахождения параметров распределения случайных величин — результатов анализов проб. Она имеет самое непосредственное отношение к проблемам оценивания сложных средних, объединения совокупностей и другим вопросам, излагавшимся в предыдущей главе.

§ 1. Основная литература. Постановка задач

К концу девятнадцатого века накопилось значительное количество данных о химическом составе минералов, горных пород, природных вод и газов. Вначале исследования проводились преимущественно в лабораториях Европы, но затем их важность была осознана и в США. Официальную историю проблемы оценки среднего состава земной коры можно начать с 1889 г., когда Фрэнк Уиглсворт Кларк, занявший за пять лет до этого пост Главного химика Геологической службы США, опубликовал работу «Относительная распространенность химических элементов». Это была первая попытка обобщения имеющихся анализов горных пород для суждения о среднем составе земной коры, затрагивающая также проблему происхождения химических элементов.

В 1908 г. Кларк выпускает первое издание своих знаменитых «Данных геохимии», которые в дополненном виде переиздавались еще четыре раза, вплоть до 1924 г. (пятое издание). В 1924 г. вышло и самое известное сочинение Кларка в соавторстве с Х. С. Вашингтоном — «Состав земной коры». В 1923 г. А. Е. Ферсман предложил ввести в науку термин «кларк» для характеристики среднего содержания химического элемента в космическом теле или в его части.

В настоящее время в связи с огромным потоком информации по геохимии отдельных элементов литература по кларкам стала практически необозримой. Если опираться только на книгу

А. А. Беуса (1972), содержащую огромную библиографию, то даже «избранная» литература по кларкам насчитывает в настоящее время не менее 100 названий. Среди этих работ в свою очередь можно выделить следующие особо важные исследования.

Ф. У. Кларк (Clarke, 1899, 1908, 1904)	Кларки породообразующих элементов в горных породах и в земной коре
В. И. Вернадский, 1909—1911, 1924	Широкое применение спектрального анализа для оценки кларков. Первые кларки для биосферы
У. Дж. Мead (Mead, 1907)	Идея геохимического баланса
Р. А. Дэли (Daly, 1914)	Оценка распространенности горных пород в надрегиональном масштабе
А. Кнопф (Knopf, 1916)	Использование данных Дэли для оценки состава земной коры
П. Н. Чирвинский (Chirvinsky, 1925)	Определение кларка по сборной пробе
Й. Седерхолм (Sederholm, 1925)	Региональные кларки — часть Балтийского щита; взвешивание по площади
И. Фогт (Vogt, 1931)	То же
В. М. Гольдшмидт (Goldschmidt, 1933, и др.)	Определение кларков многих редких элементов
А. Полдерваарт (Poldervaart, 1955)	Глобальные кларки со взвешиванием по объемам блоков земной коры
А. Б. Ронов, А. А. Ярошевский, 1967	То же, более современные оценки
А. Б. Ронов, 1952, 1965 и др.	Серия работ по среднему составу осадочных толщ на региональном и надрегиональном уровнях
Э. Д. Голдберг (Goldberg, 1963)	Кларки для океана
А. П. Виноградов, 1967	То же
А. П. Виноградов, 1949, 1956, 1962	Серия полных таблиц глобальных кларков и кларков для основных типов пород
К. Турекьян, К. Ведеполь (Turekian, Wedepohl, 1961)	Оценки кларков редких элементов в осадочных породах
М. М. Ермолаев, 1967	Использование кларков Турекьяна и Ведеполя со взвешиванием, с учетом строения земной коры
В. А. Кутолин, 1972	Лучшая из сводок о петрогенных элементах в базальтах
А. А. Беус, 1972	Лучшая работа по кларкам петрогенных элементов надрегионального и глобального уровней
В. В. Ляхович, 1967, 1968, 1972	Оценки кларков редких элементов в породообразующих минералах; оценки распространения акцессорных минералов
Н. М. Страхов, 1973 и др.	Проблема точности геологического взвешивания при расчете кларков
Р. А. Паркер (Parker, 1967)	Полезный общий обзор проблемы кларков

На наш взгляд, в проблеме кларков целесообразно выделять следующие аспекты: а) *уровень*, на котором выполняется расчет: земная кора в целом, отдельные блоки коры, регионы и т. д.; б) *объект*, для которого рассчитывается кларк: литосфера, изверженные горные породы, отдельные типы горных пород, минералы и т. д.; в) *способ расчета* и характер использованных данных в статистическом смысле; г) способ расчета и характер используемых данных в геологическом смысле (сюда же относятся и все вопросы косвенных способов расчета кларков).

§ 2. Разновидности кларков

Проблема кларков есть часть проблемы распределения химических элементов. Можно рассматривать кларки нескольких видов. Их типизация должна включать в себя по крайней мере два признака: характеристику объекта, для которого вычисляется кларк, и указание уровня, на котором оценивается кларк для данного объекта. Следуя А. А. Беусу (1972), мы будем рассматривать локальные, региональные и глобальные параметры (кларки), дополнив эту типизацию надрегиональными кларками (рис. 12).

Локальные параметры по А. А. Беусу характеризуют распределение элементов в «различных ограниченных по площади геологических образованиях — интрузивных массивах или комплексах, эффузивных сериях, метаморфических или осадочных свитах пород и т. д. В породах, не затронутых деятельностью каких-либо рудообразующих процессов, оценки локальных параметров распределения характеризуют так называемый геохимический фон — термин, широко применяемый в практике геохимических поисков» (стр. 21).

Региональные параметры характеризуют распределение химических элементов на уровне геохимических провинций. Развивая учение А. Е. Ферсмана, А. А. Беус геохимической провинцией называет «крупные структурные элементы земной коры, характеризующиеся едиными чертами геохимической эволюции, отраженными в химическом составе слагающих их геологических комплексов, а также в составе эндогенных и экзогенных концентраций рудных и нерудных химических элементов» (стр. 20—21). А. А. Беус считает понятие геохимической провинции более широким и рациональным, чем понятия металлогенической или петрографической провинции. Хороший пример влияния геохимической провинции на оценку кларка приводит Д. Шоу (1969). По данным Штрока (1936 г.), среднее содержание лития в гранитах было оценено в 179 г/т. Однако позднейшие исследования Хорстмана на более обширном материале дали среднее содержание лития не более 40 г/т. По этому поводу Д. Шоу замечает, что граниты, проанализированные Штроком, расположены в горах

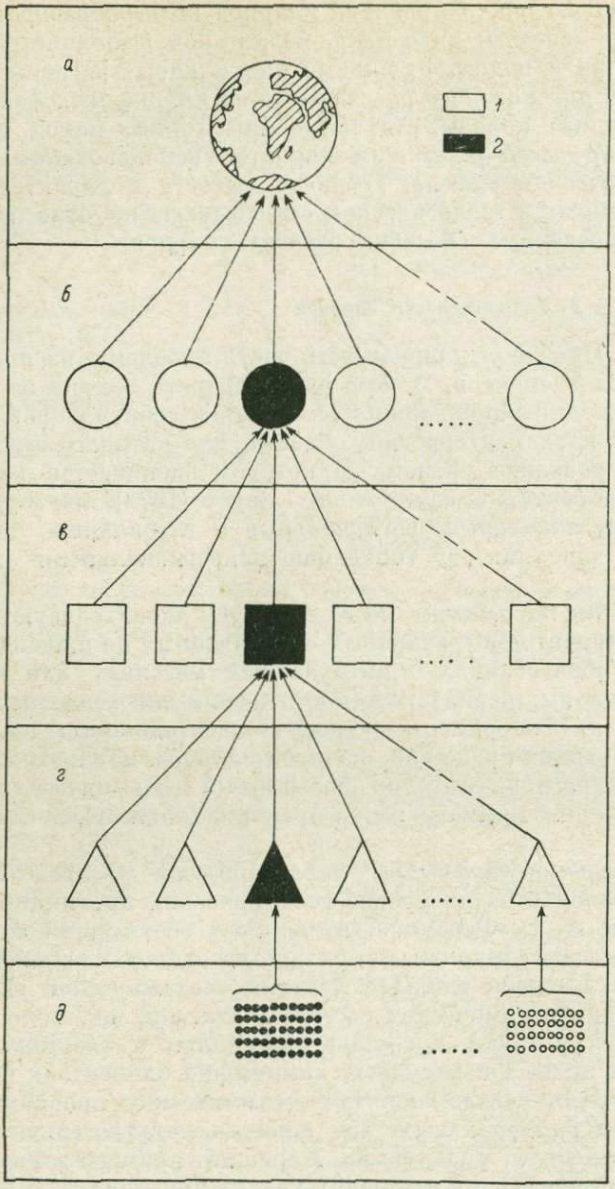


Рис. 12. Схема расчета кларков разных уровней.

Уровни: *а* — глобальный, *б* — надрегиональный, *в* — региональный, *г* — локальный; *д* — отдельные пробы. Фигуры различной формы изображают кларки, составляющие разные базисные совокупности. *1* — любой объект данного уровня с номером *i*; *2* — конкретный объект данного уровня, для которого изображены составляющие его объекты низшего уровня.

Гарца, где широко распространены признаки литиевой минерализации.

Надрегиональные параметры характеризуют распределение химических элементов на уровне основных типов (сегментов) земной коры, например, по А. Б. Ронову и А. А. Ярошевскому (1967), — на уровне континентального (материкового), переходного и океанического типов.

Таблица 11

Типизация кларков

Уровень изучения	Минерал F	Горная порода G	Формация горных пород R	Сообщества родственных формаций D	Все формации C
Глобальный (вся литосфера)	F_4	G_4	R_4	D_4	C_4
Надрегиональный (типы земной коры)	F_3	G_3	R_3	D_3	—
Региональный (крупные структурные элементы)	F_2	G_2	R_2	D_2	—
Локальный (массивы, свиты, месторождения и др.)	F_1	G_1	(R_1)	—	—

Примечание. Символы кларков даны в честь ученых, внесших значительный вклад в проблему: А. Е. Ферсмана (F), В. М. Гольдшмидта (G), А. Б. Ронова (R), Р. А. Дэли (D), Ф. У. Кларка (C).

Глобальные параметры характеризуют распределение химических элементов на уровне планеты (точнее — доступной ее части) — например, состав всех изверженных или осадочных пород земного шара, либо средний состав всей литосферы, либо всего океана. Оценка состава вод имеет свои особенности и с трудом поддается типизации в принятых рамках (одна река может пересекать платформенную и орогенную области и т. д.). Попытка типизации разновидностей кларков в пределах литосферы сделана в табл. 11. Условными значками здесь показаны *виды кларков* и *уровни*, на котором они рассчитываются.

Символы кларков мы сопровождаем цифровыми индексами в порядке повышения уровня. Будем считать, что кларк, расположенный в таблице выше и правее, имеет более высокую *ступень*, т. е. более высокий уровень, или характеризует более сложный геологический объект, или и то и другое вместе. Особенностью такой типизации является невозможность существования некоторых видов кларков. Так, если локальная изучаемая территория меньше, чем территория, занимаемая формацией, то локальные кларки для формации R_1 или их группы D_1 не имеют смысла. Кларки D_1, C_1, C_2, C_3

не имеют смысла ни при каких условиях. Может оказаться, что кларк определенного уровня тождественно равен кларкам более высоких уровней. Например, известно, что щелочные граниты — платформенные образования. Поэтому кларк циркония в щелочных гранитах литосферы G_4 или континентов G_3 — то же самое, что и в щелочных гранитах платформ G_2.

§ 3. Особенности кларков как случайных величин

В большинстве опубликованных таблиц кларков отсутствуют оценки их точности. Как правило, в этих таблицах кларки даются в форме, нарушающей основное правило метро-

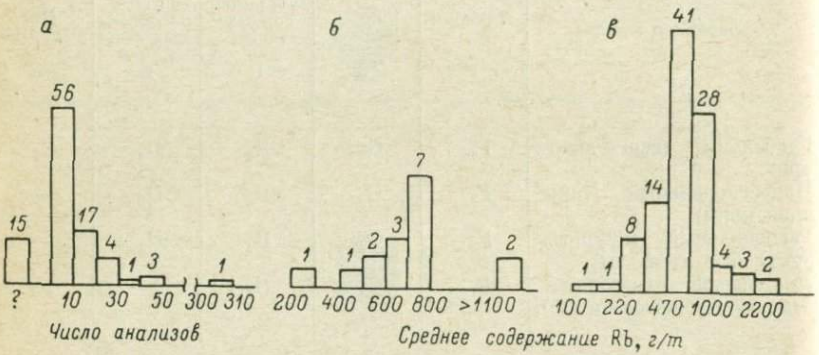


Рис. 13. Характеристика совокупности опубликованных данных, по которым рассчитан кларк рубидия в биотите.

Распределение: *а* — публикаций по числу использованных анализов; *б* — средних содержаний по публикациям; *в* — средних содержаний по регионам. Ось ординат — частоты.

логии — с числом значащих цифр, превышающем то, которое соответствует реальной точности. Например, в таблице кларков В. И. Попова (1963) для «средней материковой осадочной породы» мы находим содержание цезия, равное 0.00041%, бора — 0.0086, скандия — 0.0011, гафния — 0.00028, тантала — 0.000665% и т. д. Для всех перечисленных элементов можно поручиться всего лишь за порядок величины. Относительная ошибка приведенных кларков заведомо не менее 50%, а для таких слабо изученных элементов, как гафний или тантал, — не менее 100—200%. Правильная запись кларков этих элементов — $3 \cdot 10^{-4}$, $7 \cdot 10^{-4}$ соответственно.

Следует помнить, что всякая оценка кларка — это, по необходимости, продукт компиляции с большой долей произвола компилятора. Обычно авторы таблиц кларков не входят в детали самой техники расчетов. Например, из монографий В. В. Ляховича (1967, 1968, 1972) весьма трудно уяснить, каким образом он получал средние цифры: необходимые пояснения либо отсут-

ствуют, либо даны настолько неясно, что могут трактоваться неоднозначно. Приведем пример, показывающий реальную точность кларков, вычисляемых на основе литературных данных: рассчитаем кларк рубидия в биотитах из пород гранитоидного

Таблица 12

Оценка кларка рубидия в биотитах из гранитоидов, полученная различными способами

Вид среднего	Базовая совокупность	Объем базовой совокупности	Среднее, г/т
Среднее арифметическое:			
без взвешивания по числу анализов	Средние и единичные данные из опубликованных работ	102	693.15
со взвешиванием по числу анализов	То же	102	785.67
из средних по региональным группам	Средние по регионам	16	748.78
из средних по региональным группам со взвешиванием по числу анализов в регионе	То же	16	664.53
Среднее геометрическое:			
без взвешивания по числу анализов	Средние и единичные данные из опубликованных работ	102	617.00
со взвешиванием по числу анализов	То же	102	543.10
Среднее, определенное из частотного графика в логарифмическом масштабе	Анализы, сгруппированные в логарифмические интервалы	9(см. рис. 13, *в*)	600

Примечание. Для последнего случая оценки точности кларка: $s_{\lg} = 0.2393$, $s_x \approx 444$.

состава. Необходимые источники (102 публикации по 16 регионам) были нам любезно предоставлены М. П. Кетрис. При этом число проб по регионам и «объемы регионов» оказались самыми различными, например, «Канада», «Кольский полуостров», «Урал», «Судеты» и т. д. (рис. 13). Средние были рассчитаны М. П. Кетрис семью различными способами (табл. 12). Из приведенных данных можно заключить: 1) оценка кларка рубидия в биотитах, только в зависимости от выбранного способа расчета, изменяется на 32% (максимальное расхождение: 600 и 786 г/т); 2) наиболее

устойчиво к «помехам расчетов» среднее геометрическое; 3) численные методы расчета (первые 6 способов) не имеют никаких преимуществ перед грубым и простым графическим способом (способ 7), описанным в статье Я. Э. Юдовича и сотр. (1972); 4) оценка точности полученного кларка также достаточно впечатляюща; принимая, что распределение рубидия в биотите аппроксимируется логнормальным законом, получаем пределы колебания среднего логарифма: $\overline{\log x} = -1.224 \pm 0.2394$, что дает для среднего геометрического ($\tilde{x}$) громадный доверительный (68%) интервал: от 334 до 1036 г/т! Этот пример можно считать вполне типовым; более того, полученный кларк (600 г/т), в котором метрологически «верной» является только первая значащая цифра, значительно точнее, чем многие кларки у В. В. Ляховича, найденные на более ограниченном материале. (Это не помешало, однако, В. В. Ляховичу записывать их с пятью значащими цифрами). Иногда непонимание особенностей кларков как случайных величин принимает странные формы. Так, И. Я. Фурман (1968) подверг резкой критике таблицу кларков А. П. Виноградова 1949 г. (цит. по: Сауков, 1950) на том основании, что сумма первых 25 элементов в этой таблице превысила 100% (и составила 100.96%). Но от суммы независимых случайных величин, каковыми являются кларки отдельных элементов, н е л ь з я т р е б о в а т ь т о ч н о г о р а в е н с т в а 100%, хотя бы они и были выражены в процентах. Эта сумма неизбежно отклоняется от 100% тем сильнее, чем менее точно определены кларки. Сумма содержаний элементов в силикатном анализе из одной навески и сумма кларков — не одно и то же. Действительно, величина суммы средних значений силикатных анализов может выйти за пределы допуска ($\pm 0.5\%$), хотя бы каждый анализ в отдельности и не выходил за эти пределы, поэтому точность оценки кларка, как правило, значительно ниже точности отдельных химических анализов.

В гл. 4 обсуждались различного рода средние как оценки характеристики положения распределения на оси концентраций. Здесь для определенности отметим, что под кларком мы будем понимать оценку интеграла концентрации по всему изучаемому объему $\frac{1}{V}\int x \cdot dV$, т. е. не что иное, как оценку математического ожидания концентраций в пробах из этого объема.

Случайная величина кларка как оценки состава соответствующего объекта есть функция других случайных величин двоякого рода — *весовых коэффициентов* и *базисных содержаний*. Базисные случайные величины могут быть непосредственным результатом анализов единичных проб, либо оценками средних значений более простых объектов, из которых состоит оцениваемый. В первом случае кларк — оценка математического ожидания результатов

анализов, во втором — оценка математического ожидания кларков низших ступеней. Например, кларк типа D_3 — содержание марганца в базальтовых формациях континентов — может быть вычислен двояко: а) усреднением отдельных анализов (базисные величины — результаты анализов всех континентальных базальтов); б) усреднением кларков D_2 — средних по различным формациям базальтов с учетом распространения этих пород в пределах континентов. Кларки, рассчитанные первым способом назовем *первичными*, вторым — *вторичными*.

В целях экономии анализов кларки часто определяют не путем усреднения анализов частных проб, а анализом *сборной пробы* (метод П. Н. Чирвинского). Если при этом не были известны результаты анализов частных проб (как правило, это так), то кларк по одному анализу сборной пробы имеет неизвестную дисперсию оценки. Такую случайную величину мы назовем кларком *с неопределенным приближением*.

Если сборные пробы составлены по отдельным формациям базальтов, то они выступают как отдельные «пробы формаций». «Подъем» с уровня D_2 на D_3, например от платформ к континентам, будет означать просто увеличение числа сборных проб, в данном случае — охват ими не только платформ, но и геосинклиналей и т. д. При этом на основании дисперсии базисных величин (разброса значений средних проб вокруг генерального среднего, т. е. дисперсии кларков D_2) можно вычислить только часть дисперсии оценки D_3, ибо дисперсия оценки каждой сборной пробой «своей» формации остается неизвестной. По сути дела, мы найдем нижний предел дисперсии оценки. Такую случайную величину, вычисленную по серии сборных проб, будем называть кларком *с неполно определенным приближением*.

Каждый кларк по необходимости иногда заменяет соответствующие кларки более высокого уровня (при недостатке информации о более высоких уровнях). Например, *локальный кларк* F_1 — содержание титана в биотите массива — есть оценка математического ожидания содержаний титана в совокупности биотитов данного массива. Но допустим, что исследователь не располагает более никакими сведениями — ни о биотитах из других массивов данной геохимической провинции (региональный уровень), ни о биотитах в пределах платформ, геосинклиналей или целых континентов (надрегиональный уровень). В этом случае оценка кларка для биотита данного массива по необходимости служит оценкой кларков более высоких уровней, в том числе и *генерального кларка* (F_4). В чем разница оценки F_1, используемой на своем уровне и вместо оценок F_2, F_3, F_4? Оценка F_1 на своем уровне имеет о п р е д е л е н н у ю дисперсию, а на более высоких уровнях ее точность (хотя и заведомо меньшая) неизвестна, неопределенна — она превращается в кларк с неопределенным или неполно определенным приближением. Кроме того, что осо-

бенно неблагоприятно, такая оценка является кларком с неопределенной репрезентативностью (табл. 13).

На примере определения кларка лития в гранитах (Шоу, 1969) мы видели, как распространение в сущности регионального кларка лития на глобальный уровень привело к грубо ошибочному определению глобального кларка. Другим показательным примером является первоначально завышенный кларк германия

Таблица 13

Интерпретация кларков как случайных величин

Кларки как случайные величины	Способ получения кларков	
	первичные кларки: по первичным базисным совокупностям — индивидуальным анализам проб	вторичные кларки: по вторичным базисным совокупностям — кларкам низших ступеней
С известным приближением	Кларки F, G, R, C по результатам анализов проб	Кларки на основе кларков более низких ступеней (кроме F_1)
С неполно определенным приближением	Кларки по совокупности сборных проб	То же, с неполно определенным приближением
С неопределенным приближением	Кларки по одной сборной пробе	То же, с неопределенным приближением
С неопределенной репрезентативностью	Кларки, если они используются на более высоких уровнях без расширения опробуемой совокупности	То же, но используемые на более высоких уровнях

в каменных углях. Случайно оказалось так, что первооткрыватель германия в углях В. М. Гольдшмидт анализировал угли Рура и Англии — очень богатые германием и рудными элементами в сульфидной форме. Последующие исследователи, составлявшие таблицы кларков, механически переносили данные Гольдшмидта на угли всего мира. Например, К. Краускопф (Krauskopf, 1955), который вместо обобщения большого количества новых данных привел в своей сводке старые данные Гольдшмидта, сильно завысил кларк германия для ископаемых углей.

§ 4. О точности первичных кларков

Методика оценки первичных кларков по одной выборке проб, т. е. по одной совокупности базисных величин, элементарна и заключается в расчете среднего арифметического. Точность первичных кларков $1/s^2_{\Theta}$ определяется дисперсией величин первичной базисной совокупности $s^2_{б_{аз}}$, т. е. дисперсией результатов анализов отдельных проб, и их числом n: $s^2_{\Theta}=s^2_{б_{аз}}/n$. Оценка первичных кларков по результатам нескольких независимых сово-

купностей базисных величин в точности соответствует случаю, изложенному в § 3 гл. 5 с применением формул (5. 3. 1) и (5. 3. 2).

Рассмотрим вопрос о соотношении точности первичных кларков различных уровней. Она может увеличиваться, уменьшаться или оставаться без изменения в зависимости от соотношения двух факторов — увеличения числа проб при переходе на более высокий уровень и увеличения степени неоднородности геологического объекта (табл. 14).

Таблица 14

Изменение точности первичных кларков различных уровней

Уровень	Число проб	Дисперсия базисной совокупности	Кларк	Дисперсия оценки кларка
Локальный	10 (10)	30 (30)	$F_1=50$	3 (3)
Региональный	100 (30)	60 (120)	$F_2=20$	0.6 (4)
Надрегиональный	200 (90)	100 (480)	$F_3=60$	0.3 (5.3)

Примечание. Перед скобкой — первый пример, в скобках — второй.

В первом примере точность кларка более высокой ступени возросла, во втором — понизилась. Практически, если при расчете кларка средняя «опробованность» объектов более низкого уровня не уменьшается, то этот кларк будет более точен.

§ 5. О точности вторичных кларков

Вторичными кларками мы назвали такие, базисной совокупностью для которых служили кларки более низких ступеней, т. е. или более низкого уровня (для D_2 — кларки D_1), или более простого вида того же уровня (для D_2 — кларки F_2 всех входящих в данную породу минералов).

Точность вторичного кларка зависит от числа базисных кларков более низких ступеней, их точности, разброса значений вокруг среднего и отношения объема выборки базисных кларков к объему опробуемой совокупности. Оценка вторичных кларков полностью соответствует оценке сложных средних по схемам, описанным в §§ 5.4, 5.5. Действительно, среднее по подгруппе является ни чем иным, как базисным кларком; опробованная подгруппа объектов — выборка из объектов данного уровня, причем она практически никогда не охватывает всех его объектов (см. табл. 10, случай 6). Рассмотрим методику расчета вторичных кларков двумя принципиально различными путями: с помощью кларков более простого вида (например, расчет R_1 по G_1, табл. 11) и с помощью кларков более низкого уровня (G_2 по G_1).

Базисными величинами являются кларки более простого вида. Характерной особенностью этого пути расчета является к о н е ч н ы й набор базисных кларков. Например, при расчете кларков формации используются кларки составляющих ее 3—5 типов горных пород: при расчете кларка горной породы необходимы кларки составляющих ее всего нескольких минералов и т. д. Набор базисных кларков всегда можно (и нужно) сделать п о л н ы м.

Таблица 15

Расчет кларка типа R_1

Тип породы	Среднее Θ_i, вес. %	Дисперсия оценки s_i^2	Доля пласта в исследованной выборке p_i
Терригенные	0.035	$4 \cdot 10^{-6}$	0.14
	0.041	$5.7 \cdot 10^{-6}$	0.25
	0.029	$2.9 \cdot 10^{-6}$	0.28
	0.032	$4.4 \cdot 10^{-6}$	0.33
Карбонатные	0.056	$12 \cdot 10^{-6}$	0.23
	0.074	$9 \cdot 10^{-6}$	0.16
	0.048	$6.8 \cdot 10^{-6}$	0.61

1. Пусть требуется определить кларк титана в перидотитах данного массива (D_1) по содержанию титана в составляющих эту породу минералах (по G_1). Содержание TiO_2 в оливине, пироксене и магнетите составляет 0.01, 0.025 и 3.7% соответственно. Точность (дисперсия) оценки TiO_2 в этих минералах равна соответственно $1 \cdot 10^{-6}$, $1 \cdot 10^{-6}$, $4 \cdot 10^{-2}$, а содержание минералов в породе (статистические p_i веса для кларков F_1) — 65.3, 35.3 и 0.4%. Для расчета кларка G_1 необходимо воспользоваться формулой (5.4.3): $\Theta_{G_1} = \sum_{i=1}^{3} p_i \Theta_i = 0.653 \cdot 10^{-2} + 0.353 \cdot 0.025 + 4 \cdot 10^{-3} \cdot 3.7 = 0.030\%$. Точность оценки будет характеризоваться следующей дисперсией: $s_{G_1}^2 = \sum_{i=1}^{3} p_i^2 \sigma_i^2 = 0.425 \cdot 10^{-6} + 0.124 \cdot 10^{-6} + 4 \cdot 10^{-2} = 1.2 \cdot 10^{-6}$. Расчет соответствует схеме из § 5.4.

2. Определить кларк стронция во флишоидной формации данного района (R_1) по кларкам G_1, т. е. по содержанию его в карбонатных и терригенных породах. Попутно проведем расчет кларков G_1 (табл. 15). Опробован каждый пятый пласт карбонатов (всего 3) и каждый десятый пласт терригенных пород (всего 4). По каждому пласту известны средние содержания и дисперсии их оценок, вычисленные как дисперсии содержаний, деленные на число проб. Расчет кларка G_1 производится по фор-

муле (5.4.3). Для терригенной части разреза $\Theta_{G_1} = \sum_{i=1}^{4} p_i \cdot \Theta_i =$
$= 0.035 \cdot 0.14 + 0.041 \cdot 0.25 + 0.029 \cdot 0.28 + 0.032 \cdot 0.33 = 0.0341.$

Дисперсия оценки среднего в данных четырех пластах рассчитывается по формуле (5.4.4): $s_т^2 = \sum_{i=1}^{4} p_i^2 \cdot s_i^2 = 0.14^2 \cdot 4.10^{-6} +$
$+ 0.25^2 \cdot 5.7 \cdot 10^{-6} + 0.28^2 \cdot 2.9 \cdot 10^{-6} + 0.33^2 \cdot 4.4 \cdot 10^{-6} = 1.1 \cdot 10^{-6}.$

Аналогичным образом вычисляются средние и их дисперсии для карбонатов: $\Theta_{G_i} = 0.0540$, $s_k^2 = 3.4 \cdot 10^{-6}$. Кларк по формации R_1:

$$\Theta_{R_1} = \sum_{i=1}^{2} p_i \Theta_i = 0.4 \cdot 0.0341 + 0.6 \cdot 0.0540 = 0.046, \quad (6.5.1)$$

где 0.4 и 0.6 — доли терригенной и карбонатной частей в разрезе формации. Если бы опробованные пласты исчерпывали разрез изучаемой формации, дисперсию оценки кларка R_1 мы нашли бы по формуле

$$s_{R_1}^2 = \sum_{i=1}^{2} p_i^2 \cdot s_{G_1, i}^2 = 0.4^2 \cdot 1.1 \cdot 10^{-6} +$$
$$+ 0.6^2 \cdot 3.4 \cdot 10^{-6} = 1.4 \cdot 10^{-6}. \quad (6.5.2)$$

Как видим, точность этого кларка была бы в несколько раз выше, чем точность среднего по отдельным пластам. Но опробованные пласты составляют только небольшую часть разреза. Вычисленные дисперсии $s_т^2$ и s_k^2 характеризуют точность средних, о т н о с я щ и х с я т о л ь к о к о п р о б о в а н н о й ч а с т и. Это лишь одна составная часть дисперсии оценки кларка G_1, тем меньшая, чем больше пласты отличаются друг от друга. Дисперсия кларка G_1 определяется здесь по схеме § 5.5:

$$s_{G_1, т}^2 = \sum_{i=1}^{n} p_i^2 \cdot s_i^2 + \left(1 - \frac{n}{N}\right) \frac{1}{n^2} \sum_{i=1}^{n} (\Theta_i - \hat{\Theta})^2 =$$
$$= \sum_{i=1}^{n} p_i^2 \cdot s_i^2 + \left(1 - \frac{n}{N}\right) \frac{1}{n} \sum_{i=1}^{n} p_i (\Theta_i - \hat{\Theta})^2. \quad (6.5.3)$$

Второе слагаемое представляет уже известную нам *дисперсию распространения*. Подставляя в (6.5.3) численные значения величин, получим

$$s_{G_1, т}^2 = 1.1 \cdot 10^{-6} + \left(1 - \frac{1}{5}\right) \frac{1}{4} 20.7 \cdot 10^{-6} =$$
$$= 1.1 \cdot 10^{-6} + 4.1 \cdot 10^{-6} = 5.2 \cdot 10^{-6}.$$

Как видим, фактическая дисперсия кларка почти в пять раз превышает дисперсию оценки содержания только в опробованных

четырех пластах. Аналогично для карбонатной части имеем: $s^2_{G_{1,k}} = 3.4 \cdot 10^{-6} + \left(1 - \frac{1}{10}\right)\frac{1}{3}\, 87 \cdot 10^{-6} = 3.4 \cdot 10^{-6} + 25.1 \cdot 10^{-6} = 28.5 \cdot 10^{-6}$. Теперь мы вправе применить формулу (6. 5. 2) для расчета дисперсии оценки кларка R_1: $s^2_{R_1} = 0.4^2 \cdot 5.2 \cdot 10^{-6} + 0.6^2 \times 28.5 \cdot 10^{-6} = 0.6 \cdot 10^{-6} + 10.3 \cdot 10^{-6} \approx 11 \cdot 10^{-6}$.

Сравнение с (6. 5. 2) показывает, что решающий вклад в общую точность кларков вносят именно дисперсии распространения. Они зависят от «контрастности» объектов и отношения опробованного их числа к общему. При подсчете первичных кларков объем проб несоизмерим с объемом опробуемой совокупности: общее число «проб», из которых «состоит» объект, можно считать бесконечным. Например, если опробуется сравнительно небольшой гранитный массив объемом 1 куб. км, а объем пробы составляет 1 куб. дм (достаточно солидная проба весом порядка 2.6 *кг*!), то в массиве «содержится» триллион таких проб! Однако ситуация в корне меняется при расчете вторичных кларков. Здесь число опробованных объектов (объем выборки) может быть одного порядка с объемом всей опробуемой совокупности. В нашем случае опробовался каждый пятый и каждый десятый пласт, и коэффициенты при дисперсиях распространения были равны соответственно (1—1/5)=4/5 и (1—1/10)=9/10. Если бы были опробованы все пласты формации, вторые слагаемые в кларках $G_{1,k}$ и $G_{1,\tau}$ исчезли бы, так как коэффициент $(1-n/N)$ стал бы равен нулю. Точность кларка в этом случае станет равной средневзвешенной дисперсии оценок, составляющих этот кларк.

Можно вообразить и такую редкую ситуацию, когда вторичный кларк найден вообще без ошибки — с нулевой дисперсией. Это произойдет в случае, если, во-первых, на этом уровне исчерпана опробуемая совокупность и, во-вторых, если базисные кларки также определены без ошибки. Например, изучается минерализация озер данного региона. Всего озер 20. В каждом озере разброс значений минерализации по отдельным пробам не превосходит погрешности анализа, и его можно считать нулевым: средняя взвешенная дисперсия оценки по всем опробованным озерам нулевая. При этом опробованы все 20 озер. Тогда погрешность оценки средней минерализации всех 20 озер не будет превышать погрешности анализа.

Базисными величинами являются кларки более низкого уровня. «Восхождение» к кларкам более высокого уровня производится путем усреднения кларков нижележащего уровня. Например, для расчета G_2 необходимы кларки G_1 многих массивов, входящих в данный регион. Характерной особенностью движения «по вертикали» (табл. 11) является то, что лишь в редких случаях мы будем располагать кларками по всем массивам региона. В большинстве случаев опробованные объекты будут представлять мень-

шинство, составляющее опробованную подсовокупность, т. е. рассчитывать необходимо по схеме § 6 гл. 5. Значение G_2 рассчитывается как средневзвешенное: $\Theta_{G_2} = \sum_{i=1}^{n} p_i \cdot \Theta_{G_{1,i}}$, где p_i — статистические веса базисных кларков, т. е. массивов, которые они представляют. Расчет дисперсии оценки кларка производится по формуле (5. 6. 4а), в которой p — вес опробованной подсовокупности, Θ_i — значения базисных кларков. Опять мы видим, что дисперсия оценки состоит из двух слагаемых — средневзвешенной дисперсии базисных кларков и дисперсии распространения, зависящей от числа опробованых объектов (числа базисных кларков) и контрастности значений, т. е. «неоднородности» региона.

Таким образом, один и тот же вторичный кларк, например R_3, можно вычислить двумя способами — «по горизонтали», т. е. взвешивая кларки по распространенности типов пород и «по вертикали», т. е. расширяя выборку формаций, охватывая ею более высокий уровень. Ясно, что в обоих случаях при использовании одной и той же информации и правильном взвешивании мы должны получить одну и ту же величину кларка R_3; одинаковой будет и его точность, так что один из расчетов может служить контролем другого. Однако здесь имеется тонкость, на которую следует обратить внимание. Пусть, например, терригенно-карбонатная толща опробована на стронций в нескольких обнажениях региона. Для простоты примем, что содержание стронция в каждом литотипе постоянно и определено с незначительной погрешностью, которой можно пренебречь. Тогда при расчете кларка стронция в толще «по горизонтали», т. е. из кларков литотипов, по формуле (6. 5. 3) получим следующую дисперсию оценки:

$$s^2_{G,\,т} = \sum_{i=1}^{n} p_i^2 \cdot 0 + \left(1 - \frac{n}{N}\right)\frac{1}{n}\sum_{i=1}^{n} p_i\,(0)^2 = 0,$$

$$s^2_{G,\,k} = \sum p_i^2 \cdot 0 + \left(1 - \frac{n}{N}\right)\frac{1}{n}\sum p_i\,(0)^2 = 0,$$

$$s^2_{R_1} = \sum_{i=1}^{2} p_i^2 s^2_{G_i} = \sum_{2} p_i^2 \cdot 0 = 0.$$

Получается, что кларк стронция в толще определен с погрешностью, не превышающей погрешность анализа. Так было бы лишь в том случае, если бы весовые коэффициенты терригенной и карбонатной частей в регионе были определены без погрешностей. До сих пор мы в своих рассуждениях исходили из этой предпосылки. Практически весовой коэффициент — такая же случайная величина, как и содержание. Учет этого обстоятельства приведет к дисперсии, которую мы получили бы, проводя расчет кларка «по вертикали», т. е. определяя среднее содержание по разрезам и взвешивая их

на зоны влияния разрезов в районе. Поскольку содержание в каждом литотипе постоянно, колебания содержаний от разреза к разрезу обусловлены изменением соотношения литотипов, т. е. именно той величиной, которую при расчете «по горизонтали» мы считали достоверной. Определение и уточнение весовых коэффициентов производится при картировании, составлении разрезов и т. д. В большинстве случаев это менее трудоемкая и более творческая задача, чем анализ дополнительного количества проб. Поэтому при расчете вторичного кларка предпочтительнее тот путь, который ведет через базисную совокупность с меньшей дисперсией (при условии надежного определения весовых коэффициентов).

То же самое относится и к сопоставлению эффективности расчета вторичных кларков по сравнению с первичными. В рассмотренном случае базисной совокупностью для расчета первичных кларков были бы анализы проб всех литотипов из всех разрезов. Затем все анализы были бы усреднены. Здесь не нужны сведения о соотношении литотипов или каких-либо других весовых коэффициентах, но именно по этой причине при том же количестве проб кларк будет определен менее точно. Вообще следует заметить, что расчет вторичных кларков по сравнению с расчетами первичных кларков (путем расширения совокупности анализов) имеет преимущество с чисто эвристической стороны. Можно рассчитать кларк бириллия по 1000 пробам гранитов какого-то региона (кларк типа G_2). Но для геолога далеко не безразличны сведения о том, что эти 1000 проб представляют 20 массивов, каждый из которых характеризуется своей дисперсией содержаний и своим средним содержанием.

§ 6. О косвенных методах определения кларков

Дж. Мид (Mead, 1907) был, по-видимому, первым, кто высказал идею геохимического баланса элементов в земной коре. Предполагалось, что вследствие вторичного характера осадочных пород по отношению к изверженным средние составы тех и других должны точно соответствовать друг другу. Если записать уравнение, связывающее составы этих пород, то из него можно было бы предсказать количественные соотношения между типами горных пород и тем самым получить независимый критерий для оценки этих типов. Так, в работе 1914 г. Дж. Мид (Mead, 1914) дал графическое решение уравнения

$$x \cdot \text{гранит} + y \cdot \text{базальт} = a \cdot \text{сланец} + b \cdot \text{песчаник} + c \cdot \text{известняк}.$$

При этом соотношение $x : y$ задавалось равным 63 : 35. Решение этого уравнения относительно a, b, c позволило грубо оценить теоретические соотношения между осадочными породами. Позже идея геохимического баланса была развита Ф. У. Кларком (Clarke,

1924) в отношении натрия. Приняв, что весь натрий, освобождаемый при эрозии изверженных пород, скапливается в океане, и зная содержание натрия в океанической воде и изверженных породах, Кларк рассчитал тот объем изверженных пород, который должен был быть эродирован для создания наблюдаемого запаса натрия в океане, и получил 54.8 млн. куб. миль, или $^1/_{30}$ всех изверженных пород коры мощностью 10 миль. Б. Мейсон (1971, стр. 192) уточнил этот расчет, учтя, что не весь натрий попадает в океан, а 35% его, освобожденого при выветривании, остается в составе осадочных пород. Известен целый ряд работ в этом направлении, принадлежащих перу П. Кюнена (Kuenen, 1941), П. Барта (Barth, 1961), М. Хорна и Дж. Адамса (Horn, Adams, 1966), А. Б. Ронова и А. А. Ярошевского (1967), А. А. Беуса (1972) и многих других исследователей. Такой подход позволил выявить две важные геохимические закономерности: 1) количество натрия в осадочных породах и океане в сумме меньше, чем в изверженных породах; 2) содержание летучих элементов (хлора, йода, брома, фтора) в осадочных породах и океане в сумме больше, чем в изверженных породах. Эти факты заставили, во-первых, не пренебрегать гипотезой гранитизации осадков, в процессе которой фоссилизированный натрий переходит в состав метаморфических пород, и, во-вторых, уделить серьезное внимание процессам поступления вещества в океан непосредственно из мантии. Замечательно, что эти неочевидные и смелые построения были сделаны с помощью простого сравнения кларков!

Методом геохимического баланса пользуются также для оценки кларков осадочных пород в том случае, если не стремятся к большой точности, а прямых определений недостаточно. Так, Дж. Адамс и Ч. Уивер (Adams, Weaver, 1958) применили этот метод для оценки кларков урана и тория в песчаных породах, для которых имелось очень мало анализов. Например, для тория было составлено уравнение

$$\underset{(13.5 \pm 1.5\ \text{г/т Th})}{\text{изверженные породы}} = \underset{(12 \pm 1\ \text{г/т Th})}{\text{сланцы}\ (46\%)} + \underset{(x\ \text{Th})}{\text{песчаники}\ (32\%)} + \underset{(1.7 \pm 0.7\ \text{г/т Th})}{+\ \text{известняки}\ (22\%)}$$

и по нему высчитано содержание тория в песчаниках: $x = 24 \pm 7$ г/т. Аналогично для урана нашли 4.1 ± 1.5 г/т. Заметим, что в отличие от натрия содержания тория в морской воде настолько ничтожны, что ими в расчете баланса пренебрегают.

В. М. Гольдшмидту (Goldschmidt, 1933) принадлежит также остроумная идея косвенной оценки глобальных кларков по составу осадочных пород. Рассматривая материковое оледенение четвертичного времени как бы в качестве громадного бульдозера, механически срезавшего с поверхности Балтийского щита представительную среднюю пробу, практически не затронутую химическими про-

цессами, Гольдшмидт считал такой «пробой» ледниковые глины Норвегии. Действительно, среднее из 77 анализов таких глин показало большое сходство со средними данными Кларка и Вашингтона для литосферы. К сожалению, уязвимость исходной посылки (отсутствие химических процессов при образовании ледниковых глин), не способствовала развитию этого метода.

И. Фогт (Vogt, 1931) использовал для оценки кларков малых элементов выдержанность отношений их к породообразующим элементам, кларки которых были уже известны. Так, с учетом отношений марганца, никеля и кобальта к железу для разных пород по содержанию последнего он вычислил средние содержания этих элементов. Позже эта идея была развита С. Р. Тэйлором (Taylor, 1964, 1965), который использовал ее в комбинации с идеей геохимического баланса. Основываясь на установленных соотношениях редких земель в основных типах пород, он рассчитал, в каких соотношениях должны быть взяты кислые и основные породы для получения наблюдаемого распределения лантанидов в осадочных породах при допущении, что последние суть продукты эрозии первых. Оказалось, что соотношение мафического и фельзического компонентов в земной коре должно составлять 1 : 5. Недостатком «метода отношений» является неустойчивость исходной посылки, поскольку соотношения редких земель в процессах седиментации могут нарушаться. Однако для приближенных оценок этот метод вполне удовлетворителен.

Недавно Ф. П. Кренделев (1972) предложил новый способ определения локальных кларков путем выбора для анализа п р е д с т а в и т е л ь н о й п р о б ы.[1] Суть методики заключается в следующем. «Представительная проба должна иметь средний состав по любому из параметров: химическому и/или минералогическому составу (плотность, объемный вес, электропроводность, магнитная восприимчивость и т. д.). Следовательно, можно воспользоваться одним из экспрессных методов измерения одного или нескольких физических свойств (или содержаний) прямо в поле, с тем, чтобы оценить характер распределения этого свойства в пределах изучаемого объекта и выбрать точку, в которой это свойство точно отвечает среднему. Ясно, что выбор точки осуществляется квалифицированным геологом, обеспечивающим отбор представительной в геологическом смысле пробы» (стр. 4). В частности, для отбора представительной пробы при определении локального кларка радиоактивных элементов Ф. П. Кренделев прибегает к предварительным массовым замерам радиоактивности для построения гистограммы распределения этого признака. «В случае, если гистограмма имеет вид правильной Гауссовой кривой, для определения кларков отбирается одна представительная проба», отвечаю-

[1] Правильнее называть ее т и п и ч е с к о й — для отличия от представительной в статистическом смысле.

щая моде (*Ю. Т., Я. Ю.*). В случае «многогорбой» кривой количество проб должно быть больше и отражать все разновидности пород. . . При вычислении кларков должен быть учтен процент участия каждой разновидности. . .» (там же).

Косвенные способы определения кларков позволяют получить важную геохимическую информацию с минимальными затратами труда или получить такую информацию, которую трудно добыть иным способом (например, прямые наблюдения по дегазации мантии в современных вулканах дают пока еще слишком неопределенные оценки поступления из мантии летучих компонентов). Однако в большинстве этих способов обычно отсутствует статистический аспект расчетов. Например, неизвестна точность коэффициентов, используемых в уравнении геохимического баланса, поэтому трудно судить и о точности конечного результата. Оригинальный метод Ф. П. Кренделева в известной мере также страдает этим недостатком, но здесь можно все же оценивать дисперсию найденной величины по дисперсии того физического свойства (например, общая радиоактивность), с которым данная величина коррелирует.

§ 7. О достоверности кларков в связи с геологической проблемой взвешивания

Выше при расчетах кларков мы оговорили одно важное условие: объем выборки по каждому объекту был пропорционален объему самого объекта. Наглядной моделью такого опробования явилась бы сеть с квадратными ячеями, наброшенная на группу объектов. Практически плотность сети неправильно изменяется, а некоторые объекты остаются вообще неопробованными (рис. 14).

Как правило, одни объекты опробовались гораздо чаще других, значительно больших по объему. Например, среди тысяч химических анализов на уран значительная часть принадлежит аномальным ураноносным толщам, тогда как подавляющие по объему толщи с фоновыми содержаниями охарактеризованы значительно слабее. Угольные бассейны промышленных районов подробно изучались на содержание в углях элементов-примесей, тогда как гигантские бассейны, Ленский и Тунгусский, заключающие львиную долю запасов угля СССР, охарактеризованы всего лишь несколькими сотнями проб. Если при подсчете кларков элементов-примесей в углях допустить, что объемы выборок пропорциональны объемам объектов (в данном случае — запасам угля), то будет сделана грубая ошибка, обесценивающая расчеты (Юдович и др., 1972).

Таким образом, и в наши дни проблема геологического взвешивания стоит столь же остро, как во времена дискуссии Кларка и Вашингтона с Дэли, Фогтом и другими исследователями. Кларк и Вашингтон фактически исходили как бы из модели равномерной сети опробования. Их оппоненты указывали, что редкие типы пород анализировались чаще, чем широко распространенные. Прак-

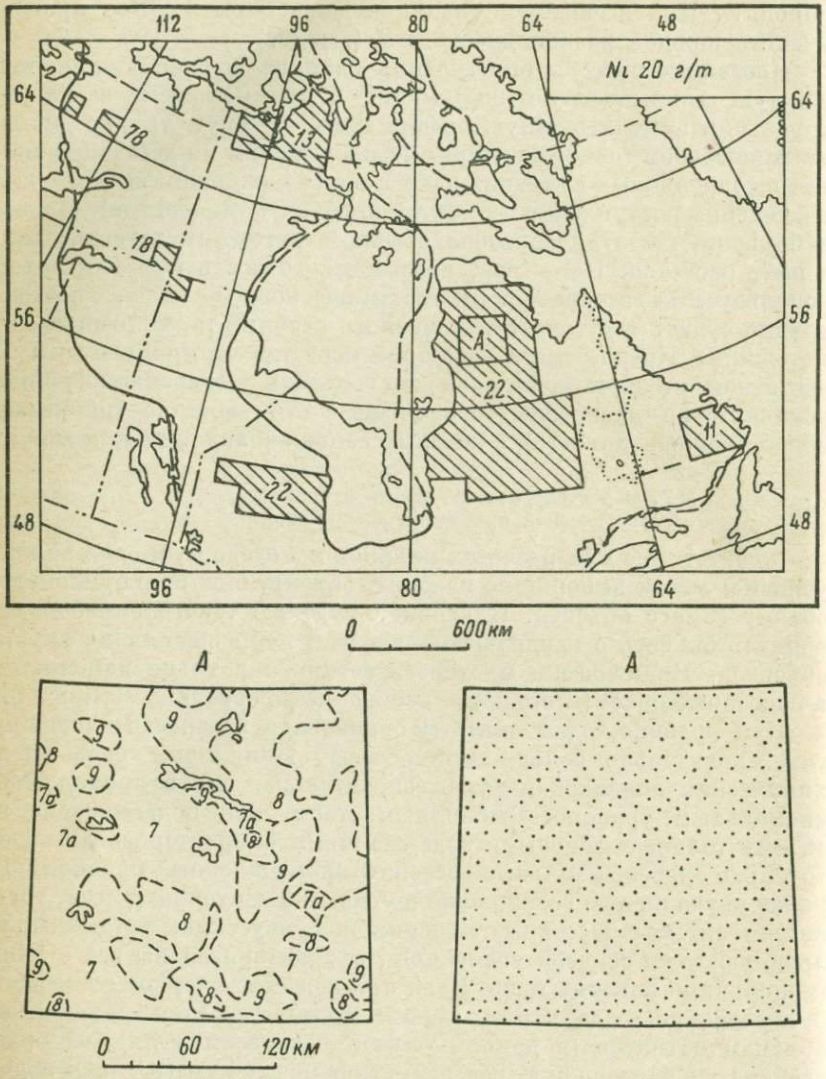


Рис. 14. Пример расчета кларка никеля для Канадского щита (изученные площади заштрихованы).

Вверху — шесть локальных кларков и выведенный на основе взвешивания по площадям выходов пород региональный кларк для Канадского щита. *А* — геологическая карта части провинции Нью-Квебек, где цифры отвечают отдельным литологическим комплексам, справа — схема опробования. Взято из работы К. Ида, У. Фарига (Eade, Fahrig, 1971) с некоторыми упрощениями.

тика последующих геохимических исследований показала, что взвешивание содержаний совершенно необходимо, если хотят, чтобы кларки действительно выполняли роль своего рода геологических констант, с которыми можно сравнивать содержания в реальных объектах. Уже первые взвешенные кларки Й. Седерхолма (Sederholm, 1925) и И. Фогта (Vogt, 1931) по территории Балтийского щита дали значительно более кислый состав «земной коры», чем по Кларку и Вашингтону. Седерхолм и Фогт взвешивали содержания по площадям выходов горных пород на дневную поверхность. Гораздо труднее произвести взвешивание по объемам горных пород в литосфере. Эта проблема обсуждается начиная с этапной работы А. Полдерваарта (Poldervaart, 1955), который впервые рассчитал надрегиональные и глобальные кларки с учетом строения земной коры по геофизическим данным. Поэтому приходится согласиться с акад. Н. М. Страховым (1973) в том, что время количественной геохимии даже осадочных толщ в рамках континентов и крупных структурных единиц (например, платформ) еще не настало, и что существующие методы дают возможность получать пока лишь прикидочные оценки с неизвестной погрешностью, дающие только порядок величины.

В настоящее время проблема правильности геологического взвешивания более остро стоит для петрогенных элементов, чем для редких элементов, оценки кларков которых содержат громадные погрешности. Для редких элементов проблема геологического взвешивания в масштабе осадочных пород платформ и континентов попросту еще «неактуальна».

Пример. По одним оценкам доля известняков в карбонатных формациях некоего региона составляет 40% (доломиты — 60%), а при уточнении оказалось, что эта доля равна 30%. При этом содержание стронция в известняках и доломитах первоначально было оценено в 0.1 и 0.01%, а затем с помощью более совершенного анализа существенно уточнено, и составило 0.05 и 0.008% соответственно. В первом случае кларк стронция для формации изменился от 0.023 до 0.018%, т. е. на 22%, а во втором случае — от 0.023 до 0.0064%, т. е. на 72%.

Для уточнения вторичного кларка стронция в карбонатной формации в данном примере важнее более точно определить кларки стронция в типах пород, чем соотношения типов. Однако ясно, что прогресс геохимических исследований (увеличение числа и точности анализов) приводит ко все более достоверным оценкам локальных кларков для горных пород и минералов. Отсюда вытекает правило относительной роли статистического и геологического взвешивания: по мере накопления геохимической информации о химическом составе горных пород и минералов возрастает роль геологического взвешивания при расчете вторичных кларков и убывает роль статистического взвешивания, связанного с объемом выборок.

Глава 7

ИССЛЕДОВАНИЕ ЗАВИСИМОСТЕЙ. ПАРНАЯ КОРРЕЛЯЦИЯ И РЕГРЕССИЯ

Из всех методов математической статистики наибольшее применение в геохимии нашел *корреляционный анализ*. Под этим термином мы будем понимать установление *меры* (силы, тесноты) *связи* между случайными переменными. *Регрессионный анализ* состоит в изучении *формы* этой связи. Оба понятия тесно связаны, что и позволяет рассматривать их как единый метод статистического исследования зависимостей. Ему посвящено огромное количество литературы, в том числе десятки (!) превосходных монографий. Наша задача заключается в том, чтобы: а) дать читателю необходимый минимум математических основ; б) обратить особое внимание на те «тонкости», которые или недостаточно освещены в доступной геологам литературе, или трактуются ошибочно; в) рекомендовать для практической работы наиболее удобные формулы, ясно обнажающие предметный смысл выражаемых соотношений; г) затронуть ряд новых или слабо освещенных в литературе проблем.

§ 1. Мера линейной связи. Коэффициент корреляции

Рассмотрим два ряда попарно сопряженных случайных величин, например, содержаний золота в пробах и мощность жилы в местах взятия этих проб. Просмотр таких данных может подсказать, что бóльшим значениям мощности чаще отвечают и бóльшие содержания. Это беглое впечатление можно проверить, построив *точечную диаграмму* в координатах «x (мощность)—y (содержание золота)». Очевидно, что чем более вытянут на диаграмме *эллипс рассеяния*, тем теснее связь переменных. Можно применить еще два приема: ранжировать одну из переменных (т. е. расположить, например, значения мощностей по возрастанию) и посмотреть, как будут вести себя значения другой переменной,

или составить *корреляционную таблицу*. Составление последней требует предварительной группировки данных; в клетках такой таблицы записывают частоты совместного появления значений переменных в заданных интервалах. Чем больше частот располагается вдоль диагонали корреляционной таблицы, тем более тесной является связь. Все перечисленные приемы дают наглядное представление о возможном наличии связи и ее тесноте. Однако отсутствие количественной меры тесноты сильно обесценивает эти, в общем полезные (и, заметим, — выработанные веками!),[1] практические приемы.

В математической статистике такой мерой является ковариация. По аналогии с дисперсией $D(x)$, являющейся математическим ожиданием квадрата отклонений о д н о й случайной величины от ее среднего значения: $\mathbf{D}(x)=\mathbf{M}(x-\mu)^2=\mathbf{M}(x-\mu)\cdot(x-\mu)$, ковариация представляет собой математическое ожидание произведения отклонений д в у х случайных величин x и y:

$$\operatorname{cov}(x,\ y)=\mathbf{M}(x-\mu_x)(y-\mu_y). \qquad (7.1.1)$$

Значение ковариации зависит от масштабов измерения изучаемых величин, и потому она неудобна как показатель силы связи: выбор масштаба произволен и не должен изменять характера зависимости! Образовав нормированные величины $z_x=\frac{x-\mu_x}{\sigma_x}$ и $z_y=\frac{y-\mu_y}{\sigma_y}$, получим безразмерный показатель

$$\rho_{xy}=\mathbf{M}(z_x\cdot z_y)=\frac{\operatorname{cov}_{xy}}{\sigma_x\sigma_y}. \qquad (7.1.2)$$

Величина ρ_{xy} — нормированная ковариация — и есть введенный английским биологом и статистиком К. Пирсоном в упомянутом выше журнале *коэффициент корреляции*. Коэффициент корреляции обладает следующими удобными свойствами.

1. При наличии предельно тесной (функциональной) зависимости его величина равна ± 1 (знак плюс отвечает положительному, минус — отрицательному приращению функции при положительном приращении аргумента). Действительно, если $y=bx+a$, то $\sigma_y=b\sigma_x$, $\mu_y=b\mu_x+a$, откуда

$$\rho_{xy}=\mathbf{M}\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{bx-b\mu_x}{\sigma_x b}\right)=\mathbf{M}\frac{(x-\mu_x)(x-\mu_x)}{\sigma_x\sigma_x}=1.$$

2. При отсутствии связи его величина равна нулю. Действительно, если x и y — независимые величины, то

$$\operatorname{cov}(x,\ y)=\mathbf{M}[(x-\mu_x)(y-\mu_y)]=\mathbf{M}(x-\mu_x)\,\mathbf{M}(y-\mu_y)=0\cdot 0=0.$$

---

[1] Например, журнал Biometrica, т. II, ноябрь 1902—ноябрь 1903 гг. (главный журнал по применению статистических методов того времени) почти наполовину состоит из корреляционных таблиц!

3. Его величина не изменяется при линейных преобразованиях переменных. Прежде всего это означает, что изменение масштаба измерения одной или обеих переменных или перенос точки отсчета не изменяет значения коэффициента корреляции.

4. Значение ρ тем больше, чем ближе зависимость между x и y к прямой линии, и тем меньше, чем сильнее искривлена эта зависимость (даже если эта зависимость очень тесная!).

Выражение (7.1.2) — это «истинное» значение коэффициента корреляции ρ_{xy}, другими словами— значение его в генеральной совокупности изучаемых величин. Его выборочной оценкой служит величина r_{xy}. Она вычисляется по аналогичной формуле, в которой и ковариация, и стандарты также являются выборочными оценками соответствующих «истинных» величин:

$$r_{xy} = \frac{\operatorname{kov}(x,\ y)}{s_x s_y}. \qquad (7.1.3)^2$$

Существует целый ряд расчетных формул для нахождения r_{xy}; почти все они потеряли практическое значение в связи с распространением ЭВМ. Отметим только две из них. Поскольку

$$\overline{xy} = \bar{x} \cdot \bar{y} + \operatorname{kov}(x,\ y), \quad \text{то} \quad r_{xy} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{s_x s_y}. \qquad (7.1.4)$$

Как видим, в числителе стоит величина «среднее произведение минус произведение средних», которую легко запомнить, если обратить внимание на сходство с одним из выражений для расчета дисперсии: $\overline{xx} - \bar{x}\bar{x}$. Среднее значение произведения двух величин $\overline{xy}$ определяется по формуле

$$\overline{xy} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i.$$

Другая формула получается из (7.1.2) заменой параметров их оценками:

$$r_{xy} = \frac{1}{n} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} \cdot \frac{y_i - \bar{y}}{s_y} = \frac{1}{n} \sum t_x t_y. \qquad (7.1.5)$$

В ней коэффициент корреляции предстает как среднее значение произведения центрированных и нормированных значений переменных. Строго говоря, величина r, подсчитанная по любой из этих формул, является смещенной оценкой ρ, а именно $\mathbf{M}(r) = \rho - \rho(1-\rho^2)/2n$.

---

[2] Напомним, что два ряда обозначений — σ и s, cov и kov и т. д. — отвечают генеральным и выборочным величинам соответственно.

§ 2. Проверка гипотез относительно коэффициентов корреляции

Вычисленное значение коэффициента корреляции, или *выборочный коэффициент корреляции* (r), является всего лишь оценкой истинного значения коэффициента корреляции, и в этом смысле сам является случайной величиной. Функция распределения r сложна и имеет следующие особенности: она симметрична только при $\rho=0$, и асимметричность увеличивается с увеличением значения ρ; дисперсия r при нормальных распределениях x и y зависит как от объема выборки n, так и от значения ρ:

$$\sigma_r^2=\frac{1-\rho^2}{n-2}. \qquad (7.2.1)$$

Поскольку значение ρ неизвестно, то, заменяя его на r, вычислим лишь оценку дисперсии:

$$s_r^2=\frac{1-r^2}{n-2}; \quad s_r=\frac{\sqrt{1-r^2}}{\sqrt{n-2}}. \qquad (7.2.2)$$

Геохимика и геолога для обоснования своих выводов интресуют прежде всего следующие вопросы: а) можно ли считать зависимость существенной, т. е. исключает ли данный экспериментальный материал гипотезу независимости; б) можно ли считать, что данная зависимость существенно отличается от заданной — ρ_0; в) можно ли считать, что две экспериментальные зависимости, охарактеризованные выборочными коэффициентами корреляции r_1 и r_2, существенно отличаются (или не отличаются) по тесноте.

Перечисленные вопросы решаются статистической проверкой двух гипотез. 1) $\mathbf{H}_0$: $\rho=a$; проверкой этой гипотезы при $a=0$ решается вопрос (a), при $a=\rho_0$ — вопрос (б).

2) $\mathbf{H}_0$: $\rho_1=\rho_2$; здесь при заданных оценках r_1 и r_2 решается вопрос (в).

Испытание $\mathbf{H}_0$: $\rho=0$ называют *проверкой значимости* коэффициента корреляции. Простейшим методом проверки является расчет статистики t:

$$t=\frac{|r-0|}{s_r}=\frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}}. \qquad (7.2.3)$$

Если бы величина r была распределена нормально, то статистика t была бы распределена по Стьюденту с $f=n-2$ степенями свободы.[1] Считают, что при не очень близких к единице значениях r и большом числе наблюдений это условие выполняется. Практически объем выборки должен составлять 200—300 наблюдений. Если вычисленное значение статистики t превышает табличное значение при заданном уровне значимости ($t>t_{\alpha,f}$), то нулевая гипотеза отвергается. Это означает, что н е з а в и с и м ы е величины

[1] Предварительно было оценено по два параметра каждой переменной: средние значения и дисперсии.

не могут дать такого (имеется в виду — такого большого) значения r и что с вероятностью $1-\alpha$ они должны считаться коррелированными. Если нулевая гипотеза не отвергается, определенного вывода сделать нельзя и «вопрос остается открытым»: мы не можем определенно доказать ни отсутствие, ни наличие зависимости. При отсутствии таблиц распределения t для проверки значимости r можно применить простое неравенство: $r\sqrt{n-1} \geqslant 3$. Если оно выполняется, зависимость считается значимой.

Более строго гипотеза значимости коэффициента корреляции проверяется с помощью замечательного преобразования Р. А. Фишера:

$$z = \frac{1}{2} \ln \frac{1+r}{1-r}. \qquad (7.2.4)$$

Величина z нормально распределена даже при очень малых выборках. Важно, что распределение z не зависит от значения ρ. Его дисперсия зависит только от n и равна

$$\sigma_z^2 = \frac{1}{n-3}; \quad \sigma_z = \frac{1}{\sqrt{n-3}}. \qquad (7.2.5)$$

Таким образом, проверку гипотез относительно величины r можно заменить значительно более простой проверкой уже известной нам гипотезы относительно нормально распределенной величины z с заданной дисперсией σ_z^2. Например, значимость коэффициента корреляции (читай: значимость отличия коэффициента корреляции от нуля) проверяется с помощью статистики

$$u_z = \frac{z-0}{\sigma_z} = z \cdot \sqrt{n-3}. \qquad (7.2.6)$$

Гипотезу $z=0$ отвергают, если $u_z > u_\alpha$, где u_α (значение аргумента нормального распределения для заданного уровня значимости) находят по таблицам нормального распределения. Аналогичным образом проверяется гипотеза $\rho = \rho_0$: обе величины r и ρ_0 подвергают преобразованию Фишера и образуют статистику $u = \frac{z - z_0}{\sigma_z} = (z - z_0)\sqrt{n-3}$.

Таким же образом при необходимости производят расчет доверительных интервалов для ρ по заданному r. Он отличается от соответствующих расчетов для любой другой нормально распределенной величины лишь предварительным преобразованием r в z и обратным преобразованием результатов.

Все переходы от r к z и обратно табулированы, так что на практике даже не приходится вычислять величин u_z, ибо таблицы составлены сразу в форме *критических значений* коэффициента корреляции $r_{кр}$ для заданного объема выборки n и заданного уровня значимости α. Если коэффициент корреляции меньше «критического», то связь считается незначимой. К сожалению, во многих

учебниках эти «критические величины» приведены только для одного (0.05) или двух (0.05 и 0.01) уровней значимости, тогда как на практике этого не всегда достаточно. Поэтому умение пользоваться преобразованием Фишера необходимо: ведь таблицы функции u_α (таблицы нормального распределения) имеются в большинстве руководств.

Гипотеза о значимости различия двух истинных коэффициентов корреляции ρ_1 и ρ_2 по их выборочным значениям r_1 и r_2 проверяется с помощью статистики

$$u = \frac{|z_1 - z_2|}{\sigma_{z_1 - z_2}} = \frac{|z_1 - z_2|}{\sqrt{\sigma_{z_1}^2 + \sigma_{z_2}^2}} = \frac{|z_1 - z_2|}{\sqrt{1/(n_1 - 3) + 1/(n_2 - 3)}}, \quad (7.2.7)$$

где n_1 и n_2 — объемы выборок, по которым рассчитаны значения r_1 и r_2. В случае $u > u_\alpha$ делается вывод (с заданным уровнем значимости α), что $\rho_1 \neq \rho_2$. Еще раз заметим, что если нулевая гипотеза (в данном случае $H_0 : \rho_1 = \rho_2$) не отвергнута, это еще не служит доказательством того, что $\rho_1 = \rho_2$. Статистические методы в данном случае, как и в большинстве других, могут исключить какое-либо утверждение, какую-либо модель (например, утверждение $\rho = 0$ или $\rho = a$, или $\rho_1 = \rho_2$), но если нулевая гипотеза не отвергается, то это еще не может служить доказательством справедливости таких равенств. В переводе на естественно-научную терминологию это означает, что статистический материал может исключать какую-либо модель, одновременно подразумевая возможность согласования данного экспериментального материала с другой или даже с целым рядом других моделей.

Приведем пример, имеющий немаловажное значение. При отсутствии корреляции между мощностью и содержанием в качестве оценки среднего содержания рекомендуют среднее арифметическое. Однако доказательство отсутствия корреляции ($\rho = 0$) подменяют проверкой гипотезы $H_0 : \rho = 0$. Если эта гипотеза не отклонена, то отсутствие корреляции не исключено, но отнюдь не доказано.

§ 3. Уравнение прямой регрессии. Оценка параметров уравнений регрессии

Вычислением r заканчивается первая часть статистического исследования зависимостей. К сожалению, во многих геохимических работах исследование связи этим и ограничивается. Вторая часть задачи заключается в установлении параметров уравнения, связывающего x и y, в предположении, что эта связь прямолинейна. Поясним, что́ будет представлять из себя искомая линия. Поскольку изучается зависимость между величинами x и y, то имеет смысл говорить об условном математическом ожидании одной величины Y при заданном значении другой $M\{Y/X = x\}$.

Предположение о прямолинейной зависимости означает, что это условное математическое ожидание изменяется прямо пропорционально величине X. Линия условных математических ожиданий и называется уравнением регрессии «y на x». Аналогично линию условных математических ожиданий $M\{X/Y=y\}$ называют уравнением регрессии «x на y». Прямую регрессии «y на x» можно провести приближенно по точкам *частных средних*. Разбив совокупность на несколько интервалов по оси x, подсчитаем частное среднее $\bar{y}_{x_i}$ для каждого i-го интервала. Через полученные точки «на глаз» проводится прямая так, чтобы отклонение ее от точек частных средних было минимальным.

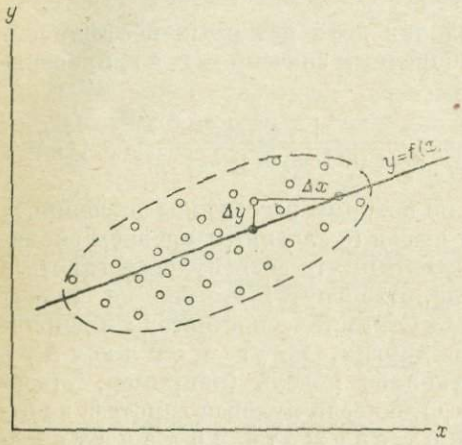


Рис. 15. Эллипс рассеивания и уравнение прямой регрессии y на x.
Минимизируется сумма квадратов отрезков Δy.

Последняя операция в значительной степени субъективна, и каждый может оценивать степень отклонения точек от прямой по-своему. Поэтому необходимо строго определить, что понимается под минимальным отклонением. Из метода наименьших квадратов вытекает (рис. 15), что минимизировать необходимо сумму квадратов отклонений значений y от искомой прямой, т. е. сумму квадратов отрезка Δy (по вертикали!): $\sum_{i=1}^{n} (\Delta y_i)^2$. При проведении прямой регрессии «x на y», естественно, необходимо минимизировать сумму отклонений точек от аппроксимирующей прямой по горизонтали: $\sum_{i=1}^{n} (\Delta x_i)^2$. Можно показать, что обе прямые регрессии проходят через точку с координатами $(\bar{x}, \bar{y})$.

Итак, в общей форме обе прямых регрессии могут быть записаны в виде

$$y = \beta_{yx} \cdot x + \alpha_y,$$
$$x = \beta_{xy} \cdot y + \alpha_x,$$

где β_{yx}, β_{xy} — угловые коэффициенты этих прямых, а α_x, α_y — свободные члены (отрезки, отсекаемые соответственно по осям x и y). Величины β и α называются параметрами уравнения регрессии; достаточно найти их оценки по выборке, чтобы построить выборочные прямые регрессии: $y=b_{yx} \cdot x + a_y$, $x=b_{xy} \cdot y + a_x$, где b и a —

выборочные оценки «истинных» параметров уравнения регрессии. Метод наименьших квадратов дает следующие формулы для этих величин: $b_{yx} = r\frac{s_y}{s_x}$, $b_{xy} = r\frac{s_x}{s_y}$, откуда (7.3.1)

$$r^2 = b_{xy} \cdot b_{yx}.$$

Заметим, что истинные значения параметров уравнений регрессии связаны с истинными значениями коэффициента корреляции и стандартами-равенствами, аналогичными предыдущим. Расчет оценок параметров по (7. 3. 1) — не что иное, как компактная и удобно приспособленная для данного случая форма метода наименьших квадратов.

Угловые коэффициенты прямых регрессии прежде всего зависят от соотношения стандартных отклонений переменных: увеличение дисперсии величины x растягивает корреляционный эллипс по горизонтали, а дисперсии y — по вертикали. При фиксированной дисперсии y уменьшение дисперсии x «задирает» корреляционный эллипс кверху: величина b_{yx} увеличится, наклон прямой регрессии станет более крутым. Он зависит, следовательно, от масштабов измерения переменных, что следует иметь в виду при построениях. Распространенный казус состоит в том, что для компактности чертежей масштабные отрезки по осям x и y принимаются разными, и тогда неожиданно выясняется, что тангенс угла наклона прямой на чертеже не совпадает с аналитическим значением b. Угловые коэффициенты прямых регрессии зависят также от коэффициента корреляции: при заданном отношении стандартов σ_y/σ_x наклон прямой регрессии «y на x», как видно из (7. 3. 1), будет уменьшаться с уменьшением r. Подробная интерпретация этого явления будет дана в следующей главе.

Квадрат коэффициента корреляции, значение которого определяется формулой (7. 3. 1), называется в статистике коэффициентом детерминации B. Он указывает, какая доля изменчивости одной величины обязана влиянию другой величины. Положим, что между содержанием золота и мощностью жил найден коэффициент корреляции 0.8. Тогда величина $(0.8)^2=0.64$ означает, что изменчивость содержаний золота на 64% контролируется мощностями жил и на 36% — иными факторами.

Поскольку прямые регрессии обязательно проходят через точку $(\bar{x}, \bar{y})$, то для их построения достаточно знания только одного параметра — углового коэффициента b. Параметр a (свободный член) не имеет самостоятельного значения. Он выражается через параметр b и положение срединной точки

$$a_y = \bar{y} - b_{yx}\bar{x}, \qquad (7.3.2)$$
$$a_x = \bar{x} - b_{xy}\bar{y}.$$

Зная выражение для параметров уравнений регрессии, мы можем,

наконец, написать расчетную формулу уравнений регрессии:

$$y = r_{xy}\frac{s_y}{s_x}\cdot x + \left(\bar{y} - r_{xy}\frac{s_y}{s_x}\cdot \bar{x}\right),$$
$$x = r_{yx}\frac{s_x}{s_y}\cdot y + \left(\bar{x} - r_{yx}\frac{s_x}{s_y}\cdot \bar{y}\right). \qquad (7.3.3)$$

Перенеся переменные, получим уравнения, которые чаще всего приводятся в литературе:

$$y - \bar{y} = b_{yx}(x - \bar{x}),$$
$$x - \bar{x} = b_{xy}(y - \bar{y}). \qquad (7.3.4)$$

Подчеркнем, что уравнения (7. 3. 4) — разные, описывающие разные, несовпадающие линии. Иначе было бы $b_{yx}=1/b_{xy}$ и $a_x=a_y/b_{yx}$. Из равенства (7. 3. 1) видно, что это возможно лишь при условии $r=1$. Тогда $b_{yx}=s_y/s_x$, $b_{xy}=s_x/s_y$. Чем меньше абсолютное значение r, тем больше угол между прямыми регрессии Θ:

$$\operatorname{tg}\Theta = \frac{1 - b_{xy}\cdot b_{yx}}{b_{xy} + b_{yx}} = \frac{1 - r^2}{r}\cdot\frac{1}{s_x/s_y + s_y/s_x}. \qquad (7.3.5)$$

При $r=1$ получается $\Theta=0$, при $r=0$ — неопределенность, так как при отсутствии корреляции прямых регрессии не существует: они теряют смысл.

§ 4. Проверка гипотез относительно параметров уравнения регрессии

Вычисленные параметры уравнения регрессии, будучи выборочными оценками соответствующих истинных значений, являются случайными величинами. Из математической статистики известно, что величина тангенса угла наклона прямой регрессии b распределена по нормальному закону с дисперсией:

$$s^2_{b_{yx}} = \frac{s^2_y}{s^2_x}\cdot s^2_r = \frac{s^2_y}{s^2_x}\cdot\frac{1 - r^2}{n - 2}. \qquad (7.4.1)$$

Интуитивно кажется очевидным, что погрешность в определении угла наклона, действительно, прямо пропорциональна погрешности в определении коэффициента корреляции. Кроме того, ясно, что при фиксированных r и объеме выборки n погрешность в тангенсе угла наклона должна быть пропорциональна самому тангенсу угла, который при фиксированном r пропорционален отношению стандартов переменных — s_y/s_x. Любопытно, что относительная погрешность тангенса угла наклона не зависит от наклона. Изложенного достаточно, чтобы построить критерий для проверки гипотез относительно b. Действительно, величина

$$t = \frac{b_{yx} - \beta}{s_{b_{yx}}} \qquad (7.4.2)$$

должна быть распределена по закону Стьюдента с $n-2$ степенями свободы. Нулевая гипотеза заключается в том, что найденное по выборке уравнение регрессии имеет наклон, одинаковый с заданным $H_0: \beta=\beta_0$.

Особыми значениями β_0, на которых следует отстановить внимание читателя, являются $\beta_0=0$ и $\beta_0=1$. Первое возможно только при отсутствии коррелированности переменных; проверяя гипотезу $H_0: \beta=0$, мы одновременно проверяем гипотезу $\rho=0$, что ясно из равенства $b_{yx}=rs_y/s_x$. Значение $\beta_0=1$ часто (но не совсем правильно — см. гл. 8) используют при проверке гипотезы о том, что в уравнении предполагаемой функциональной зависимости между переменными $y=kx+a$ значение коэффициента пропорциональности $k=1$. Как и всегда, нулевая гипотеза отвергается при $t>t_{\alpha, f}$, где $t_{\alpha, f}$ — табличное значение аргумента распределения Стьюдента при $f=n-2$ степенях свободы. Как видим, проверка гипотез относительно параметров уравнения регрессии может понадобиться при построении статистических моделей геохимических процессов. Обычно в основе такой модели лежит предполагаемая функциональная зависимость. Внося в нее ряд случайных факторов, определяют положение «теоретической» (модельной) прямой регрессии и затем с помощью изложенной выше процедуры сравнивают его с эмпирическим уравнением регрессии.

Может быть, более важное практическое значение имеет сравнение двух эмпирических уравнений регрессии. К сожалению, методика такого сравнения разработана лишь для случая равенства остаточных дисперсий $\sigma^2_{\text{ост. 1}}=\sigma^2_{\text{ост. 2}}$. Поэтому сначала необходимо проверить гипотезу о равенстве остаточных дисперсий [1] по критерию Фишера $F=s^2_{\text{ост. 1}}/s^2_{\text{ост. 2}}$, где $s^2_{\text{ост. 1}}$ — оценка большей из остаточных дисперсий. Если эта гипотеза не отклоняется, рассчитывают статистику

$$t=\frac{|b_1-b_2|}{s_{b_1-b_2}}=\frac{|b_1-b_2|}{\sqrt{s^2_{b_1}+s^2_{b_2}}}. \qquad (7.4.3)$$

Величина $s^2_{b_1}+s^2_{b_2}$ определяется следующим образом:

$$s^2_{b_1}+s^2_{b_2}=\frac{s^2_{y_1}(1-r^2_1)}{s^2_{x_1}(n_1-2)}+\frac{s^2_{y_2}(1-r^2_2)}{s^2_{x_2}(n_2-2)}=\frac{s^2_{\text{ост.1}}}{s^2_{x_1}(n_1-2)}+\frac{s^2_{\text{ост.2}}}{s^2_{x_2}(n_2-2)}.$$

Так как остаточные дисперсии $\sigma^2_{\text{ост}}=(1-\rho^2)\,\sigma^2_y$ считаются равными, а $s^2_{\text{ост. 1}}$ и $s^2_{\text{ост. 1}}$ — их независимые оценки, то, взвешивая их на число степеней свободы, получим оценку $\hat{\sigma}^2_{\text{ост}}$ по двум выборкам:

$$\hat{\sigma}^2_{\text{ост}}\frac{(n_1-2)\,s^2_{\text{ост. 1}}+(n_2-2)\,s^2_{\text{ост. 2}}}{n_1+n_2-4}. \qquad (7.4.4)$$

[1] Об остаточных дисперсиях подробнее см. § 7.

Таким образом,

$$s_{b_1}^2 + s_{b_2}^2 = \left(\frac{1}{s_{x_1}^2 (n_1 - 2)} + \frac{1}{s_{x_2}^2 (n_2 - 2)}\right) \hat{\sigma}_{\text{ост}}^2.$$

Сравнивая значение t из (7. 4. 3) с табличным значением для заданного уровня значимости, отклоняем или принимаем гипотезу $\mathbf{H}_0 : \beta_1 = \beta_2$. В последнем случае линии регрессии можно считать параллельными. Важно подчеркнуть, что отвергнуть в данном случае нулевую гипотезу — значит доказать неравенство $\beta_1 > \beta_2$ или $\beta_1 < \beta_2$; не отвергнуть нулевую гипотезу — не значит доказать $\beta_1 = \beta_2$. Позитивный смысл здесь имеет только факт отклонения нулевой гипотезы.

Наконец, в условиях параллельности сравниваемых прямых регрессии остается проверить равенство свободных членов, что будет означать полное совпадение прямых. В этих условиях значение b должно быть равно

$$b = \frac{\bar{y}_1 - \bar{y}_2}{\bar{x}_1 - \bar{x}_2} = b^*. \qquad (7.4.5)$$

В итоге имеем две независимые оценки углового коэффициента: по (7. 4. 5) и как средневзвешенное из двух выборок (коэффициенты сравниваемых уравнений регрессии) — $\hat{b}$. При определении этой средневзвешенной оценки весовыми коэффициентами являются их точности, т. е. величины, обратные дисперсиям:

$$\hat{b} = \frac{b_1/s_{b_1}^2 + b_2/s_{b_2}^2}{1/s_{b_1}^2 + 1/s_{b_2}^2} = \frac{b_1 s_{x_1}^2 (n_1 - 2)/\hat{\sigma}_{\text{ост}}^2 + b_2 s_{x_2}^2 (n_2 - 2)/\hat{\sigma}_{\text{ост}}^2}{s_{x_1}^2 (n_1 - 2)/\hat{\sigma}_{\text{ост}}^2 + s_{x_2}^2 (n_2 - 2)/\hat{\sigma}_{\text{ост}}^2}.$$

Постоянный множитель $\sigma_{\text{ост}}^2$ можно опустить. Разность $b^* - \hat{b}$ в условиях параллельности имеет нулевое математическое ожидание, что позволяет образовать распределенную по Стьюденту статистику $t = \dfrac{b^* - \hat{b}}{s_{b^* - \hat{b}}}$. Для расчета дисперсии этой разности $s_{b^* - \hat{b}}^2 = s_{b^*}^2 + s_{\hat{b}}^2$ дадим формулу без пояснений:

$$s_{b^*}^2 + s_{\hat{b}}^2 = \hat{\sigma}_{\text{ост}}^2 \left[\frac{1}{(\bar{x}_1 - \bar{x}_2)^2}\left(\frac{1}{n_1} - \frac{1}{n_2}\right) + \frac{1}{s_{x_1}^2 (n_1 - 1) + s_{x_2}^2 (n_2 - 1)}\right], \qquad (7.4.6)$$

где $\hat{\sigma}_{\text{ост}}^2$ берется из (7. 4. 4).

Совсем просто проверяется гипотеза $\mathbf{H}_0 : a = a_0$ о значении свободного члена уравнения регрессии вне связи со сравнением двух эмпирических уравнений регрессии, где a_0 — любое наперед заданное число, например $a_0 = 0$:

$$t = \frac{|a - a_0|}{s_a}, \quad s_a^2 = s_{\text{ост}}^2 \left(\frac{1}{n} + \frac{(\bar{x})^2}{(n-1)\, s_x^2}\right), \qquad (7.4.7)$$

где s_a^2 — дисперсия оценки свободного члена.

Это значение дисперсии становится понятным, если a выразить следующим образом: $a=\bar{y}-b_{yx}\bar{x}$, откуда

$$s_a^2 = s_y^2 + (\bar{x})^2 s_b^2 = \frac{s_{\text{ост. }y}^2}{n} + (\bar{x})^2 \frac{s_{\text{ост. }y}^2}{(n-1)\, s_x^2}. \qquad (7.4.8)$$

Практически при изучении регрессии начало координат (изменением начала отсчета значений x и y) необходимо перенести в точку $(\bar{x}, \bar{y})$. Тогда $a=0$. Кроме того, масштабы измерения желательно подобрать так, чтобы $b=1$.

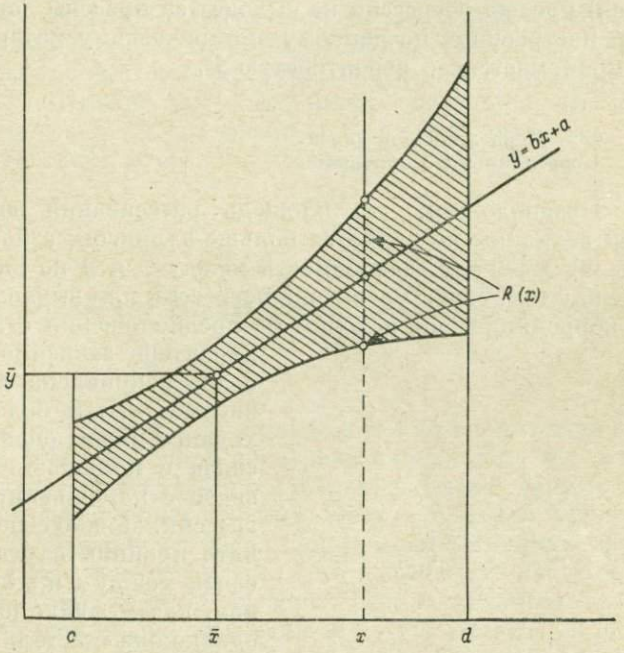


Рис. 16. Доверительная зона положения истинного уравнения регрессии на отрезке $[c, d]$, построенная по таблицам с использованием величины R (x).

Очень наглядное представление о надежности уравнения регрессии дает доверительная зона для него. Перед ее построением необходимо определить диапазон значений независимой переменной $[c, d]$, в котором будут проводиться построения. Затем определяют величины $C=\frac{c-\bar{x}}{s_x}$, $D=\frac{d-\bar{x}}{s_x}$, $\lambda=\sqrt{\frac{1}{2}\left[1-\frac{1-nCD}{\sqrt{(1-nC^2)(1-nD^2)}}\right]}$.

Задавшись доверительной вероятностью P, по специальным таблицам (например, Большев, Смирнов, 1965, табл. 4.6а, 4.6б) для значений λ и числа степеней свободы $f=n-2$ определяют значение

$u_{f,\lambda,P}$. Доверительную зону строят следующим образом (рис. 16). В выбранном интервале $[c, d]$ проводят линию регрессии и для нескольких значений x (число их должно быть достаточным для построения плавных кривых) по вертикали вверх и вниз откладывают величину $R(x)$, равную

$$R(x) = \pm u_{f,\lambda,P} \cdot s_{\text{ост}} \sqrt{\frac{(x-\bar{x})^2}{s_x^2} + \frac{1}{n}}, \qquad (7.4.9)$$

где $s^2_{\text{ост}}$ — остаточная дисперсия величины y. Если теоретическая (модельная) прямая регрессии не выходит за пределы доверительной зоны, построенной по данному эмпирическому материалу, то модель «принимается» с вероятностью P.

§ 5. Мера линейной связи. Корреляционное отношение

Предположим, что площадь рассеивания двумерной случайной величины не является больше эллипсом, а напоминает серпообразную фигуру: зависимость между x и y не вполне линейна. Между тем, строя уравнение регрессии или вычисляя коэффициент корреляции, мы исходим из предположения о линейном характере зависимости. Чем более искривлена линия зависимости, тем более неподходящи наши оценки силы связи с помощью r и формы связи с помощью прямой регрессии. Можно себе представить крайний случай, когда связь между x и y очень тесная, почти функциональная; однако она «настолько криволинейна», что коэффициент корреляции равен нулю (рис. 17). Проблема состоит в том, что мы априори не знаем, какова форма связи, и не располагаем никакой другой информацией, кроме той, которую дает нам выборка парных значений x и y. «При любом характере усредненной зависимости y от x прямая регрессии y на x дает наилучшее линейное приближение по методу наименьших квадратов. Но из этого вовсе еще не следует, что средние значения $\bar{y}_x$, подсчитанные для каждого значения x, действительно лежат на прямой линии (точнее, что их отклонения от прямой обусловлены только колебаниями этих средних)» (Румшиский, 1971, стр. 120).

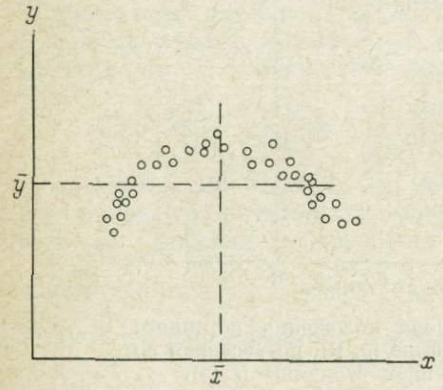


Рис. 17. Пример зависимых переменных с коэффициентом корреляции между ними, равным нулю.

В качестве наиболее общей меры связи используют величину корреляционного отношения η_{xy}. Для наглядного и логически естественного перехода от коэффициента корреляции к корреляционному отношению рассмотрим первую из этих величин с полезной для этой цели точки зрения. Уравнение регрессии y на x, как известно, является линией (в данном случае — прямой) условных математических ожиданий Y при заданных значениях X, т. е. $\mathbf{M}(Y/X=x)$. Переменную y можно представить себе как сумму двух независимых составляющих — упомянутого математического ожидания, однозначно определяемого значением x, и случайной составляющей — остаточной величины Δy. В соответствии с этим ее дисперсия может быть представлена также в виде суммы двух дисперсий. Первая из них носит название *обусловленной* $s^2_{\bar{y}/x}$ (она определяется колебаниями значений x и углом наклона прямой регрессии), вторая — *остаточной* $s^2_{\text{ост. }y}$:

$$s^2_y = s^2_{\bar{y}/x} + s^2_{\text{ост. }y}. \qquad (7.5.1)$$

Известно, что $r^2 = \frac{s^2_{\bar{y}/x}}{s^2_y}$ или $r^2 = 1 - \frac{s^2_{\text{ост. }y}}{s^2_y}$. По аналогии с величиной r^2 (коэффициента детерминации) можно ввести величину

$$\eta^2_{yx} = 1 - \frac{s^2_{\text{ост. }y}}{s^2_y}. \qquad (7.5.2)$$

Ее отличие от r^2 заключается только в том, что остаточная величина Δy для расчета остаточной дисперсии измеряется не от прямой линии, а от кривой произвольной формы, проходящей через точки частных средних. Практически совокупность разбивается по значениям x на ряд классов, подсчитываются средние значения y для каждого класса (так называемые частные средние $\bar{y}_x$) и дисперсия частных средних. Она и является выборочной оценкой обусловленной дисперсии $s^2_{\bar{y}/x}$ в равенстве (7.5.1). Из него определяется и другая компонента дисперсии, а из равенства (7.5.2) — корреляционное отношение.

Желательно, чтобы число интервалов при расчете частных средних и их дисперсии было не меньше 8—10. Расчет производят по формуле

$$s^2_{\bar{y}/x} = \sum_{i=1}^{k} (\bar{y}|_{x_i} - \bar{y})^2 \cdot n_i,$$

где i — номер интервала, k — общее число интервалов, n_i — доля значений, приходящихся на этот интервал. Идея корреляционного отношения станет наглядной из рис. 18. На рис. 18, *а* связь между переменными x и y отсутствует: любому x_i отвечает одно и то же частное среднее значение $\bar{y}/_x$. Аппроксимирующая прямая линия регрессии параллельна оси абсцисс. На рис. 18, *б* между x и y существует положительная корреляция: линия регрессии

приобрела наклон. Легко заметить, что в этой ситуации возросла общая дисперсия s_y^2 (ибо появились большие отклонения от $\bar{y}$), тогда как дисперсия, характеризующая отклонения y_i от линии регрессии, осталась той же самой.

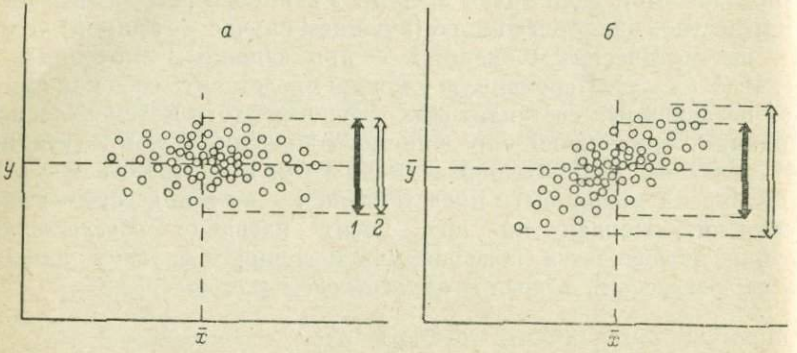


Рис. 18. График, иллюстрирующий идею корреляционного отношения.

а — при отсутствии связи остаточная дисперсия (*1*) равна общей дисперсии (*2*); *б* — при наличии связи остаточная дисперсия меньше общей дисперсии.

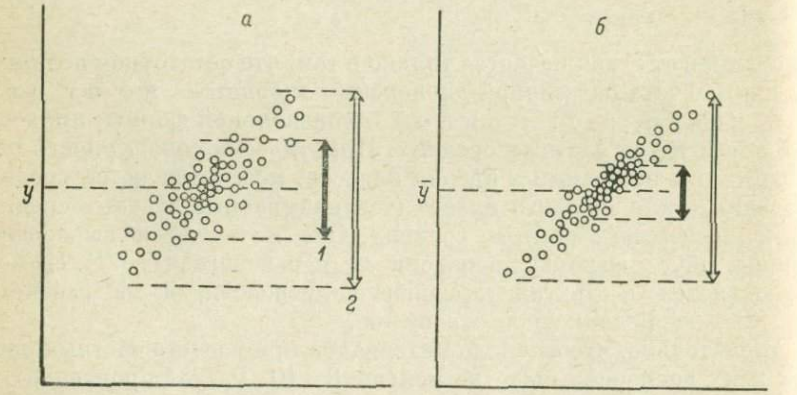


Рис. 19. Зависимость величины корреляционного отношения от тесноты связи.

а — связь не очень тесная, остаточная (*1*) и общая (*2*) дисперсии различаются не очень сильно; *б* — связь тесная, остаточная дисперсия намного меньше общей дисперсии.

Как видим, в случае отсутствия связи обе дисперсии s_y^2 и $s_{ост}^2$ равны — отклонения отсчитываются от одной и той же величины, равной $\bar{y}$. Как только появляется связь, ситуация изменяется: остаточная дисперсия остается прежней, тогда как общая дисперсия обязательно возрастает (или, что то же самое, общая диспер-

сия остается постоянной, а остаточная уменьшается). Поэтому справедливо неравенство $s_y^2 \geqslant s_{\text{ост}}^2$. Совершенно очевидно, что связь будет тем теснее, чем больше превышение общей дисперсии над остаточной (ср. рис. 19, *а* и 19, *б*).

Можно убедиться, что отношение $s_{\text{ост}}^2/s_y^2$ является универсальным показателем тесноты связи: оно сохраняет свою «чувствительность» при любой, самой сложной форме связи.

Величина η — *корреляционное отношение*, как и коэффициент корреляции, введена в статистику К. Пирсоном. Она также является величиной безразмерной, изменяющейся от 0 до 1. В отличие от r эта величина по сути своей всегда положительна (одна и та же криволинейная зависимость на одних участках может быть положительной, на других — отрицательной). В общем случае величины $\eta_{yx}^2 = 1 - s_{\text{ост. }y}^2/s_y^2$ и $\eta_{xy}^2 = 1 - s_{\text{ост. }x}^2/s_x^2$ не равны друг другу, при этом всегда $\eta_{xy} \geqslant r \leqslant \eta_{yx}$. Знак равенства имеет место в случае линейной зависимости $\eta_{xy} = \eta_{yx} = r$.

§ 6. Критерии линейности связи

В общем, чем сильнее корреляционное отношение η отличается от коэффициента корреляции r, тем существеннее отклонение зависимости от линейной. Но для уверенного суждения необходим строгий количественный критерий. Следует отметить, что вычисление критерия линейности и рекомендации о его применении — недостаточно разработанная область регрессионного анализа. Известен, например, критерий Блекмана: $\xi = \eta^2 - r^2$. Однако, как указывает А. Б. Вистелиус (1963а, стр. 119), величина стандарта этого критерия может быть оценена только приблизительно, и пользоваться таким критерием затруднительно. В книге А. М. Длина (1958, стр. 254) приведено громоздкое выражение для вычисления критерия линейности Θ_ξ, причем автор вообще не раскрывает связи его с величинами r и η.

В. Ю. Урбах (1964, стр. 298, формула 8. 30) приводит критерий линейности в форме

$$F = \frac{\eta^2 - r^2}{1 - \eta^2} \cdot \frac{n - k}{k - 2}, \qquad (7.6.1)$$

где n — объем выборки, k — число интервалов группировки при расчете η. Эта же формула приведена у Л. З. Румшиского (1971, стр. 121). Гипотеза линейности отвергается, если $F > F_\alpha$, где F_α — табличное значение критерия Фишера с уровнем значимости α и степенями свободы $n-k$ и $k-2$.

На наш взгляд, критерий линейности можно построить одним из трех следующих способов.

Первый способ (вычисление критерия $u_{\text{лин}}$). Основан на эмпирически установленном положении (которое, по-видимому, будет

доказано), что преобразование Фишера $z_\eta = \frac{1}{2}\ln\frac{1+\eta}{1-\eta}$ приводит к величине, которая $\left(\text{как и величина } z_r = \frac{1}{2}\ln\frac{1+r}{1-r}\right)$ имеет нормальное распределение с дисперсией $\sigma^2_{z_\eta} = 1/f$, где f — число степеней свободы; $f = n-k-1$, где n — объем выборки, k — число интервалов группирования.

Статистика

$$u_{\text{лин}} = \frac{|z_\eta - z_r|}{\sigma_{z_\eta - z_r}} = \frac{|z_\eta - z_r|}{\sqrt{\frac{1}{n-3} + \frac{1}{n-k-1}}} \qquad (7.6.2)$$

распределена по нормальному закону, а гипотеза линейности отвергается, если $u_{\text{лин}} > u_\alpha$, где u_α — табличное значение аргумента нормального распределения для уровня значимости α.

Второй способ (вычисление критерия $F_{\text{лин}}$). Заключается в сравнении оценок двух остаточных дисперсий: вычисленной по *остаткам* Δy от прямой регрессии $s^2_{\text{ост.}\,y}$ и по остаткам от частных средних (т. е. по существу — по остаткам от кривой регрессии $s'^2_{\text{ост.}\,y}$). Поскольку эти оценки независимы и первая всегда больше второй, то статистика

$$F_{\text{лин}} = \frac{s^2_{\text{ост.}\,y}}{s'^2_{\text{ост.}\,y}} = \frac{1-r^2}{1-\eta^2} \qquad (7.6.3)$$

распределена по Фишеру со степенями свободы $f_1 = n-2$ для числителя и $f_2 = n-k$ для знаменателя. В условиях справедливости нулевой гипотезы (т. е. линейности связи) $\mathbf{M}(F) = 1$. Если $F_{\text{лин}} > F_\alpha$, где F_α — табличное значение критерия Фишера для уровня значимости α, то гипотеза линейности отклоняется.

В превосходном руководстве В. В. Налимова (1960, стр. 269) в качестве критерия проверки линейности градуировочных графиков (построенных по k эталонам, которые проанализированы n раз каждый) используется критерий, очень сходный с описанным. Если средние значения по эталонам $\bar{y}_x$ (с известными содержаниями x) лежат на прямой линии (прямая регрессии), то для дисперсии этих частных средних относительно прямой регрессии можно получить две оценки. Во-первых,

$$s^2_{\bar{y}_x} = \frac{1}{k-2}\sum (y_{x_i} - Y)^2. \qquad (7.6.4)$$

Во-вторых, еще одна, независимая от этой, оценка, которая равна дисперсии воспроизводимости, деленной на число анализов каждого эталона n, т. е. $s^2_{\text{воспр}}/n$. В случае криволинейной зависимости дисперсия частных средних $s^2_{\bar{y}_x}$ будет превышать величину $s^2_{\text{воспр}}/n$ (рис. 20). Таким образом,

$$F = \frac{s^2_{\bar{y}_x}}{s^2_{\text{воспр}}/n} = \frac{n s^2_{\bar{y}_x}}{s^2_{\text{воспр}}}. \qquad (7.6.5)$$

Легко видеть, что $s^2_{\text{воспр}}$ есть остаточная криволинейная дисперсия, а $s^2_{\bar{y}_x} \cdot n$ — то же, но прямолинейная. Следовательно, критерий (7. 6. 5) тождествен критерию (7. 6. 3). Число степеней свободы для числителя $k-2$, для знаменателя $k\,(n-1)=N-k$. Несходство числа степеней свободы для числителя и знаменателя объясняется особенностями нахождения остаточных дисперсий (прямолинейных) в том и другом случаях.

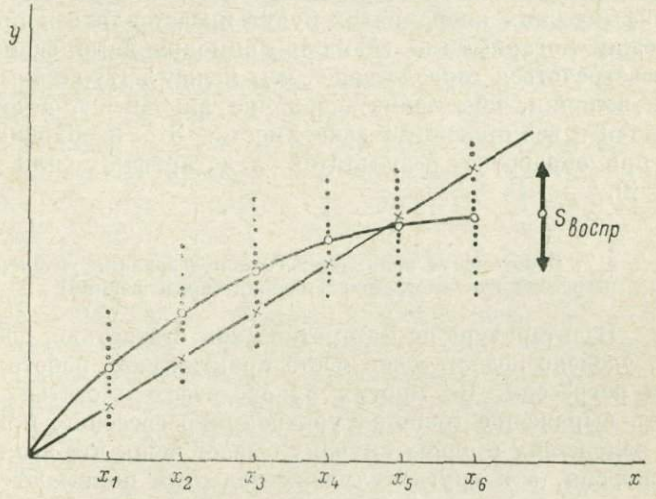


Рис. 20. К проверке линейности градуировочных графиков. Соотношение частных средних (*кружки*) и теоретических значений y, высчитанных по линейному уравнению регрессии (*крестики*).

$s_{\text{воспр}}$ — дисперсия воспроизводимости данного эталона (на чертеже эталона № 6), проанализированного n раз.

Третий способ (вычисление критерия $t_{\text{лин}}$). Всякая криволинейная зависимость может быть аппроксимирована степенным многочленом с любым нужным приближением. Начало отсчета при измерении переменных всегда можно выбрать так, чтобы $\bar{x}=0$, $\bar{y}=0$. Тогда прямолинейная зависимость выразится уравнением $y=ax$, а криволинейная, если ограничиться первым членом полинома, — $y \approx ax^b$. Логарифмируя эти функции, получим

$$\log y = \log x + \log a, \qquad (7.6.6а)$$

$$\log y = b \log x + \log a. \qquad (7.6.6б)$$

В уравнении (7.6.6а) тангенс угла наклона прямой (в логарифмических координатах) равен единице, в уравнении (7. 6. 6б) — b. Таким образом, все линейные зависимости при изучении их в логарифмических коор-

динатах дадут прямые линии, угол наклона которых будет равен 45°. Для проверки линейности, следовательно, достаточно прологарифмировать обе переменные, рассчитать уравнение регрессии для логарифмов и проверить гипотезу о равенстве единице углового коэффициента. Последняя процедура уже была описана.

В связи с изложенным остановимся на важном в геохимии и геологии обстоятельстве. Все линейные и степенные зависимости в логарифмических координатах будут представляться прямыми, т. е. между логарифмами величин прямолинейные зависимости должны встречаться гораздо чаще, чем между натуральными значениями величин, ибо кроме «первично линейных», в линейные обращаются и все степенные зависимости. Это необходимо учитывать при обработке результатов в логарифмах или баллах (см. гл. 2).

§ 7. Особенности практического использования уравнений регрессии как инструмента оценки (предсказания)

В литературе по математической статистике, доступной геологу, уделено недостаточно места практической работе с уравнениями регрессии. Во многих руководствах и статьях можно встретить выражение «ошибка уравнения регрессии», причем по смыслу изложения в одних случаях бывает ясно, что это остаточная дисперсия, а в других случаях под этим понимают ошибки параметров уравнения регрессии b и a (см. напр., Длин, 1958; Раевский, Шурубор, 1958, и др.). Опыт практического использования уравнений регрессии показывает, что здесь имеется немало тонкостей в понимании характера ошибок. Находя значение y по некоторому значению x с помощью регрессии, мы сталкиваемся с ошибкой предсказания. Она тем больше, чем слабее корреляция между x и y, или чем сильнее зависимость между ними отклоняется от прямой линии. Мерой, измеряющей ошибку предсказания по уравнению регрессии, является *остаточная дисперсия*

$$\sigma^2_{\text{ост}.y} = \frac{1}{n}\sum_{i=1}^{n}(\Delta y_i)^2. \qquad (7.7.1)$$

Остаточная дисперсия — это средняя сумма квадратов отклонений значений y от линии регрессии. Ее можно определить как средний квадрат отклонений индивидуальных значений Δy_i (по вертикали) от истинного уравнения регрессии (см. формулу (7. 7. 1)) либо как функцию параметров корреляционной зависимости, что более удобно:

$$\sigma^2_{\text{ост}.y} = (1 - \rho^2_{xy})\,\sigma^2_y. \qquad (7.7.2)$$

В литературе более распространен первый способ (см. напр., Бондаренко, 1970, стр. 97). Находя остаточную дисперсию $\sigma^2_{ост}$, мы подразумеваем, что уравнение регрессии определено точно. На практике, однако, линию регрессии проводят по выборочным данным; при повторении выборки всякий раз получаются несколько иные уравнения регрессии, не совпадающие друг с другом. Ситуация вполне аналогична процедуре оценки параметра генеральной совокупности по выборке: как, допустим, по выборочному среднему оценивают величину генерального среднего, так и всякое найденное по выборке уравнение регрессии является *оценкой регрессии в генеральной совокупности*. Если регрессионная зависимость в генеральной совокупности выражается «истинным уравнением регрессии», то эту прямую можно представить себе как ось симметрии некоторого множества (пучка) выборочных уравнений регрессии, параметры которых b и a в той или иной мере отличаются от параметров генеральной регрессии β и α. Следовательно, с выборочным уравнением регрессии связана не только ошибка предсказания, вызванная разбросом точек около линии регрессии, измеряемая остаточной дисперсией, но и ошибка, обусловленная отклонением выборочной линии регрессии от линии «истинной регрессии». Это отклонение может быть тем большим, чем меньше объем выборки и меньше коэффициент корреляции (рис. 21). Таким образом, ошибка предсказания по выборочной линии регрессии есть сумма двух ошибок: собственно ошибки предсказания и «ошибки в положении линии регрессии». Обе они независимы друг от друга.

При работе с уравнением регрессии может возникнуть несколько типовых ситуаций, для каждой из них ошибка предсказания различна. Рассмотрим эти ситуации на примере, заимствованном из книги И. П. Шарапова (1965, стр. 165), где приведены следующие параметры корреляционной зависимости между содержаниями ZrO_2 и Nb_2O_5 на одном из коренных месторождений колумбита:

$$x = ZrO_2; \quad \bar{x} = 0.115\%; \quad s_x^2 = 0.0026; \quad s_x = 0.0510; \quad r = 0.65;$$
$$y = Nb_2O_5; \quad \bar{y} = 0.039\%; \quad s_y^2 = 0.00015; \quad s_y = 0.0124; \quad n = 834.$$
$$y = 0.159x + 2.1.$$

Для удобства расчетов переведем значения в десятитысячные доли, что соответствует сотым долям процента: $\bar{x} = 11.5$; $s_x^2 = 26$; $s_x = 5.1$; $\bar{y} = 3.9$; $s_y^2 = 1.54$; $s_y = 1.24$; $s_{\bar{y}}^2 = s_y^2/n = 1.8 \cdot 10^{-2}$; $r^2 = 0.422$; $1 - r^2 = 0.578$. Рассмотрим точность предсказания содержаний ниобия по содержаниям циркония в тех ситуациях, которые могут возникнуть на практике.

1. **Средняя ошибка предсказания y по значениям x.** В этом случае нас интересует точность, которой будет характеризоваться

предсказанное значение y по любому, заранее не оговоренному значению x. Единственное ограничение: значение x должно относиться к данной совокупности, в нашем случае — проба должна быть взята из данного месторождения. Будем для простоты считать, что 834 пробы исчерпывают генеральную совокупность, т. е. приведенные выше параметры характеризуют «истинное»

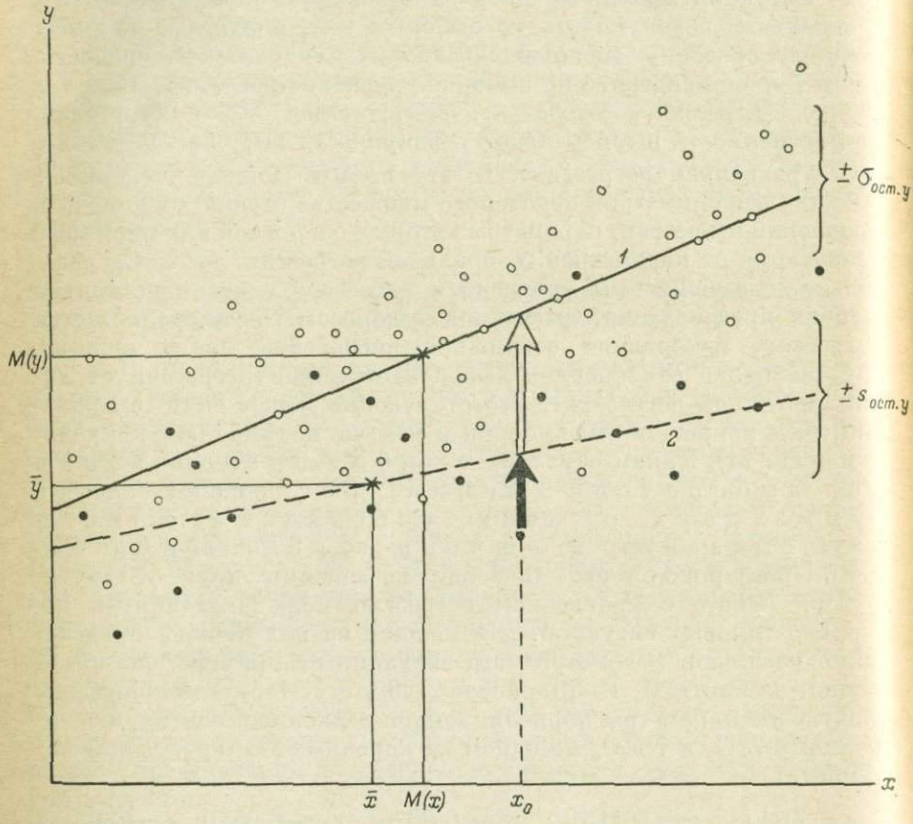


Рис. 21. Схема, показывающая «истинную линию регрессии» (*1*) и одну из ее выборочных линий (*2*).

Светлые кружки — элементы генеральной совокупности; темные кружки — то же, попавшие в выборку.

уравнение регрессии. Тогда ошибка предсказания будет в точности равна остаточной дисперсии: $\sigma^2_{y_x} = \sigma^2_{ост.\,y} = (1 - \rho^2) \cdot \sigma^2_y$. Она, как видно из формулы, прямо пропорциональна коэффициенту недетерминированности (так иногда называют величину $1 - \rho^2$) и дисперсии величины y в исследованной совокупности; эта дисперсия выступает здесь в роли масштабного множителя (ибо ее абсолютное значение, как и значение всей ошибки предсказания,

зависит от масштаба измерения величины y). Для нашего примера имеем: $\sigma^2_{y_x} = 0.578 \cdot 1.54 = 0.89$; $\sigma_{y_x} = \sqrt{0.89} = 0.94$. Таким образом, в среднем предсказанное по уравнению регрессии индивидуальное значение y ляжет в интервал ± 0.94 с вероятностью 68%. Если вообразить, что процедура нахождения содержаний ниобия по содержаниям циркония есть своего рода «анализ» (в котором роль прибора играет уравнение регрессии), то ошибка ± 0.94 будет «ошибкой метода». Пусть теперь истинное уравнение регрессии неизвестно (что чаще всего и бывает), а приведенные выше параметры характеризуют уравнение регрессии, найденное по выборке из 20 проб. В этом случае ошибка предсказания увеличится за счет неопределенности в положении выборочной линии регрессии и представится в следующем виде:

$$\tilde{s}^2_{y_x} = s^2_{\text{ост. } y} + \frac{s^2_{\text{ост. } y}}{n} + \frac{s^2_{\text{ост. } y}}{n} = s^2_{\text{ост. } y}\left(\frac{2}{n} + 1\right). \qquad (7.7.3)$$

Остаточную дисперсию $s^2_{\text{ост. } y}$ можно определить по формуле

$$s^2_{\text{ост. } y} = \frac{1}{n-2} \sum_{i=1}^{n} (\Delta y_i)^2, \qquad (7.7.4)$$

где Δy_i — отклонения значений y от эмпирической линии регрессии. Величина $s^2_{\text{ост. } y}$ служит несмещенной оценкой $\sigma^2_{\text{ост. } y}$, так как деление производится на число степеней свободы $n-2$ (при оценке положения прямой регрессии было использовано два параметра). Как и в случае (7. 7. 2), остаточная дисперсия может быть рассчитана из выборочных характеристик корреляционной зависимости:

$$s^2_{\text{ост. } y} = (1 - r^2)\, s^2_y. \qquad (7.7.5)$$

В нашем примере $s^2_{\text{ост. } y} = (1 - 0.65^2)\, 1.54 = 0.89$. Подставляя в уравнение (7.7.3) значения $s^2_{\text{ост. } y} = 0.89$ и $n = 20$, получаем: $s^2_{y_x} = 0.89 \cdot \left(\frac{2}{20} + 1\right) = 0.98$; $s_{y_x} = \sqrt{0.98} \approx 1$. Расчет показал, что ошибка предсказания по выборочной линии регрессии больше, чем по истинному уравнению регрессии, примерно на 10%.

2. ***Ошибка предсказания частного среднего.*** Ситуация отличается от предыдущей тем, что предсказывают не индивидуальное значение y, а среднее арифметическое по группе проб с известными значениями x. Очевидно, что ошибка предсказания в этом случае равна дисперсии индивидуального предсказания, деленной на число проб m, по которым оценивают частное среднее:

$$\sigma^2_{\bar{y}_x} = \sigma^2_{\text{ост. } y}/m. \qquad (7.7.6)$$

Так, ошибка предсказания среднего содержания ниобия по выборке из 10 проб (не оговаривая заранее, каково содержание циркония в этих пробах!) равна $0.89 : 10 = 0.089$, $\sigma_{\bar{y}_x} = \sqrt{0.089} = 0.3$.

Таким образом, точность предсказания частного среднего (измеряемая стандартом $\sigma_{\bar{y}_x}$) оказалась в 3 раза выше, чем точность индивидуального предсказания.

Если истинное уравнение регрессии неизвестно, то используются те же формулы, только вместо $\sigma^2_{\text{ост. }y}$ подставляем s^2_{yx}: $s^2_{\bar{y}_x} = 0.98/10 = 0.098$; $s_{yx} = \sqrt{0.098} = 0.34$.

3. **Ошибка предсказания индивидуального значения y по назначенному (наперед известному) значению x.** Здесь речь идет о погрешности предсказания y не по всему диапазону значений x, и не в среднем, а для конкретного, выбранного значения x_0. Таким образом, здесь нас интересует ошибка конкретного анализа, которая, как известно, может быть и больше и меньше, чем усредненная «ошибка метода». Оказывается, такая постановка задачи имеет смысл только для выборочного уравнения регрессии. Действительно, для истинного уравнения регрессии средняя ошибка индивидуального предсказания в точности равна ошибке предсказания по любому заданному значению x_0, потому что величина $\pm\sigma_{\text{ост. }y}$ образует зону, параллельную линии регрессии, и в любом месте по оси x величина остаточной дисперсии одна и та же. Иная ситуация в случае выборочного уравнения регрессии. Здесь ошибка предсказания будет тем больше, чем дальше назначенное значение x_0 отстоит от его среднего значения $\bar{x}$. Она находится по формуле

$$s^2_{y_{x_0}} = \frac{s^2_{\text{ост. }y}}{n} + s^2_{b_{yx}}(x - \bar{x})^2. \qquad (7.7.7)$$

В это выражение входит величина $s^2_{b_{yx}}$, т. е. ошибка углового коэффициента выборочного уравнения регрессии. Подставляя значение $s^2_{b_{yx}}$ из (7. 4. 1.), получим расчетную формулу

$$s^2_{y_{x_0}} = \frac{s^2_{\text{ост. }y}}{n} + \frac{s^2_{\text{ост. }y}(x - \bar{x})^2}{s^2_x \cdot n} + s^2_{\text{ост. }y} =$$

$$= s^2_{\text{ост. }y}\left[\frac{(x - \bar{x})^2}{n \cdot s^2_x} + \frac{1}{n} + 1\right]. \qquad (7.7.8)$$

Положим, $x_0 = 0.055\%$. Подставляя это значение, получим

$$s^2_{y_{x_0}} = 0.89\left(\frac{5.5 - 11.5^2}{20 \cdot 26} + \frac{1}{20} + 1\right) \approx 1.$$

Как видим, ошибка конкретного предсказания оказалась чуть больше, чем ошибка «предсказания вообще», или «средняя ошибка предсказания». Назначим теперь значение x_0, сильно отличающееся от среднего значения, например, $x_0 = 0.355\%$. В этом случае получим:

$$s^2_{y_{x_0}} = 0.89\left(\frac{(35.5 - 11.5)^2}{20 \cdot 26} + \frac{1}{20} + 1\right) = 1.92.$$

Ошибка предсказания получилась значительно больше средней ошибки.

4. **Ошибка предсказания единичного значения y по назначенному интервалу значений $x_i \div x_k$.** В этом случае исследователь знает не конкретное значение, а лишь интервал, в котором оно лежит. Например, для обогатительной фабрики может представить интерес руда, содержащая заранее известное количество нерудных примесей. Так, на Кемпирсайском месторождении хромитовых руд найдено (Кузнецов, Панов, 1971), что между содержаниями Cr_2O_3 и SiO_2 существует жесткая корреляционная связь по 2290 пробам (месторождение «Миллионное»): $r = -0.852$. Уравнение регрессии кремнезема на хром: $SiO_2 = 31.05 - 0.5\ Cr_2O_3$. Это позволяет, зная содержание рудного компонента Cr_2O_3, предсказывать содержание вредного компонента.

Вернемся к нашему примеру и оценим ошибку предсказания содержаний Nb_2O_5 для интервала содержаний ZrO_2 от 0.14 до 0.26%. Очевидно, что средняя квадратичная ошибка для всего интервала получится как средняя взвешенная, где весами будут частоты значений x внутри интервала или, учитывая, что стандартное отклонение пропорционально отклонению $(x - \bar{x})$, как средняя арифметическая из квадратичных погрешностей для концов указанного интервала:

$$s^2_{y_{x_1}} = 0.89\left(\frac{(14 - 11.5)^2}{20 \cdot 26} + \frac{1}{20} + 1\right) = 0.94, \quad s_1 = 0.96;$$

$$s^2_{y_{x_2}} = 0.89\left(\frac{(26 - 11.5)^2}{20 \cdot 26} + \frac{1}{20} + 1\right) = 1.34, \quad s_2 = 1.16;$$

$$s_{y_{x_1 \div x_2}} = \frac{0.96 + 1.16}{2} = 1.06.$$

§ 8. Ошибка предсказания индивидуального значения y по уточненному значению x.

При расчете ошибок оценки y с помощью уравнения регрессии предполагалось, что значение x (например, содержание компонента в пробе) определено без погрешностей. В большинстве случаев это не так, и к вычисленным погрешностям оценок y как составную часть необходимо прибавлять погрешность в определении x. Простейшим выражением дисперсии итоговой погрешности предсказания является следующая очевидная формула:

$$s^2_{y^*_x} = s^2_{y_x} + b^2_{yx} s^2_{\text{анал.}x}, \qquad (7.8.1)$$

где $s^2_{y_x}$ — дисперсия оценки, вычисленная по формулам § 7; b_{yx} — угловой коэффициент уравнения регрессии y на x, $s^2_{\text{анал.}x}$ — дисперсия определения x (например, погрешности воспроизводимости).

Интуитивно ясно, что если бы удалось определить значение x с большей точностью, то и ошибка предсказания значения y должна снизиться. При уменьшении аналитической дисперсии, например в связи с переходом к более точному анализу или к использованию средних значений k-кратных анализов пробы, соответственно уменьшается второе слагаемое суммы (7. 8. 1):

$$s^2_{y_x^*} = s^2_{y_x} + b^2_{yx}\frac{s^2_{\text{анал.}\,x}}{k}. \qquad (7.8.2)$$

Однако такой способ дает лишь самое грубое приближенное решение. Действительно, при уменьшении $s^2_{\text{анал.}\,x}$ изменяются параметры корреляционной зависимости: уменьшается общая (наблюдаемая) дисперсия величины x, так как $s^2_x = s^2_{\text{прир}} + s^2_{\text{анал}}$, увеличивается коэффициент корреляции, уменьшается остаточная дисперсия и изменяется коэффициент регрессии b. Существо этих явлений будет рассмотрено в гл. 8 и 9.

Конечно, можно было бы построить новые уравнения регрессии, взяв за основу средние значения x из k определений каждой пробы и вычислив новый коэффициент корреляции, после чего использовать вновь полученную зависимость для оценки. Однако такой подход не продвинул бы нас ни в практическом, ни в теоретическом отношении. Действительно, для этой цели нам были бы необходимы k-кратные анализы каждой пробы, использованной при построении новых уравнений регрессии. А нам желательно получить оценку y по k-кратно (или вообще более точно) проанализированным пробам в оперативном порядке, когда k может меняться от одной пробы к другой.

Статистические методы, изложенные в следующей главе, позволяют внести простые коррективы в исходные уравнения регрессии. Рассмотрим их на том же примере. Положим, что содержания двуокиси циркония определены с аналитической дисперсией, составляющей 30% общей дисперсии: $s^2_{\text{анал}} = 0.3 \cdot 26 = 7.8$.

Пусть для улучшения точности предсказания содержаний ниобия по уравнению регрессии одну пробу проанализировали на Zr четыре раза. Тогда дисперсия x уменьшится, поскольку снизится аналитическая дисперсия. Если прежняя аналитическая дисперсия составляла $s^2_{\text{анал.}\,x}$, то новая будет $s^2_{\text{анал.}\,x}/k$, и величина уменьшения будет равна

$$\Delta s^2_{\text{анал.}\,x} = s^2_{\text{анал.}\,x} - \frac{s^2_{\text{анал.}\,x}}{k} = s^2_{\text{анал.}\,x}\left(1 - \frac{1}{k}\right). \qquad (7.8.3)$$

Исправленная дисперсия $s^2_{x\,\text{испр.}} = s^2_x - \Delta s^2_{\text{анал.}\,x}$. Исправленный коэффициент регрессии определится из формул гл. 8:

$$b_{yx\,\text{испр}} = b_{yx}\left(1 - \frac{\Delta s^2_{\text{анал.}}}{s^2_x}\right). \qquad (7.8.4)$$

Для рассматриваемого примера $\Delta s^2_{\text{анал}} = 7.8\left(1-\frac{1}{4}\right)=5.9$, $b_{yx\,\text{испр.}} = 1.59 : \left(1-\frac{5.9}{26}\right)=2.01$.

Как видим, исправленная линия регрессии имеет более крутой наклон, чем первичная. Для нахождения исправленного коэффициента корреляции воспользуемся выражением $r=\sqrt{b_{xy}\cdot b_{yx}}$ и найдем предварительно величину b_{xy} из равенства $r^2 = b_{xy}\cdot b_{yx}$: $b_{xy} = r^2 : b_{yx} = (0.65)^2 : 1.59 = 0.265$. Теперь из равенства $r^2_{\text{испр}} = b_{yx\,\text{испр}}\cdot b_{xy}$ определяем $r_{\text{испр}} = \sqrt{b_{yx\,\text{испр}}\cdot b_{xy}} = \sqrt{0.265\cdot 2.01} = \sqrt{0.534} = 0.73$.

Повышение точности определения циркония (путем четырехкратного повторения анализа) позволило заметно увеличить коэффициент корреляции цирконий—ниобий (0.74 вместо 0.65). Дисперсии оценок всех перечисленных видов были пропорциональны остаточной дисперсии. Ее исправленное значение найдем из равенства

$$s^2_{\text{ост.}\,y\,\text{испр}} = (1-r^2_{\text{испр}})\,s^2_y. \qquad (7.8.5)$$

Насколько точнее удастся теперь предсказать содержания ниобия? $s^2_{\Delta y\,\text{испр}} = (1-r^2_{\text{испр}})\,s^2_y = [1-(0.74)^2]\cdot 1.5\cdot 10^{-8} = 6.8\cdot 10^{-9}$, $s_{\Delta_{\text{испр}}} = 8.2\cdot 10^{-5}$.

Итак, с помощью повторных (четырехкратных) анализов проб на цирконий нам удалось снизить дисперсию на 22% ($6.8\cdot 10^{-9}$ против $8.7\cdot 10^{-9}$) и улучшить среднюю точность предсказания на 11.8% ($8.2\cdot 10^{-5}$ против $9.3\cdot 10^{-5}$).

Глава 8

КОРРЕЛЯЦИОННАЯ ЗАВИСИМОСТЬ И ЕЕ СООТНОШЕНИЕ С ЛИНЕЙНОЙ ФУНКЦИОНАЛЬНОЙ МОДЕЛЬЮ

§ 1. Постановка вопроса и его современное состояние

Проблема соотношения функциональной и корреляционной зависимостей издавна привлекает внимание статистиков и естествоиспытателей, применяющих статистические методы. Предметом их обсуждения является главным образом интерпретация существования двух уравнений регрессии и объяснение этого явления с профессиональных, математических и даже философских позиций. Особенно оживленные дискуссии происходят среди биологов и геологов, применяющих математические методы.

Существо обсуждаемого здесь явления заключается в том, что в результате корреляционного и регрессионного анализов получается *два уравнения*, которые называют уравнениями *регрессии*. Этот термин был введен в литературу К. Пирсоном (Pearson, Lee, 1903) по следующему поводу. Изучалась связь между ростом отцов и ростом сыновей, и было замечено, что средний рост отцов, имеющих сыновей заданного роста A, оказался *ниже*, чем та величина роста отцов, сыновья которых имеют средний рост A. С этого времени проблема неоднократно обсуждалась статистиками английской школы. Следует упомянуть, что это явление исследовал также акад. Б. И. Срезневский в 1888 г. и в более поздних малоизвестных зарубежным статистикам работах. Обсуждая «нормы бурь», т. е. низшие пределы ветра, опасного для судоходства, он пишет: «. . . зная норму бури для одной станции A, можно было бы определить норму бури для всякой соседней станции B, как среднее из отметок силы ветра, сделанных на B в то время, когда на A замечается буря (т. е. по уравнению регрессии B на A, — *Ю. Т., Я. Ю.*). Но оказывается, что такой способ приводит к слишком низкой оценке нормы бури для B» (цит. по: А. А. Конюс, 1950, стр. 222).

Описываемое явление очень наглядно изображается графически (рис. 22). Оно оказалось всеобщим, зависящим не от природы исследуемых величин, а от специфических свойств корреляционной зависимости, порожденным именно наличием двух уравнений регрессии. Большое число выразительных примеров можно привести из геологии и геохимии. При подсчете запасов сопутствующих элементов часто пользуются уравнениями регрессии «содержание сопутствующего элемента—содержание основного элемента», например, «содержание Zn—содержание Cu». Оно позволяет в пробе с известным содержанием меди оценить содержание цинка. Если бы теперь мы захотели оценить содержание меди по содержанию цинка, то нам следовало бы воспользоваться другим уравнением регрессии, в результате чего мы получили бы иное соотношение концентраций (рис. 22). Действительно, если содержанию меди y_0 соответствует содержание цинка x_0, то x_0 в пробе соответствует y_1, но не y_0, как подсказывает «здравый смысл».

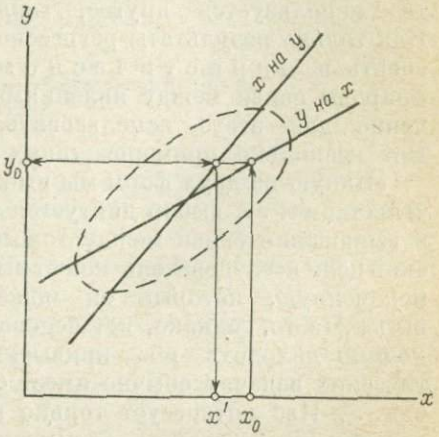


Рис. 22. Иллюстрация закона регрессии.

y_0 — оценка по x_0 с помощью уравнения регрессии y на x; x' — оценка по y_0 с помощью уравнения регрессии x на y.

С момента открытия этого явления оно не перестает настораживать естествоиспытателей, которые справедливо считают, что статистическая (корреляционная) связь является лишь оболочкой, скрывающей под собой реально существующую функциональную зависимость.

Поиски такой функциональной зависимости или «единой линии корреляционной связи» предпринимали в дальнейшем и другие известные статистики. Классик русской статистической науки А. А. Чупров (1960) так писал в 1924 г. об идее единой линии корреляционной связи: «. . . своеобразие необратимых взаимоотношений между уравнениями регрессии воспринимается как нечто несуразное. Он (исследователь, — *Ю. Т., Я. Ю.*) не может преодолеть чувства, что дело здесь в несовершенстве обычных способов статистического подхода к изучению стохастической связи, — в недодуманности, которую надо преодолеть путем вычисления единого уравнения регрессии. Поиски единого уравнения, выражающего закон функциональной зависимости между y и x, представляются во всех подобных случаях вполне законными. . .

уравнение это не имеет ничего общего с уравнениями регрессии» (стр. 44—45).

Вопрос для исследователей остается ясным лишь до тех пор, пока уравнения регрессии используются для оценочных целей: если требуется оценивать (предсказать) значения y по заданным значениям x, то используется уравнение регрессии y на x, если требуется оценивать значения x по заданным значениям y, то используется другое, *сопряженное* с первым уравнение. Как только результаты регрессионного анализа пытаются использовать в эвристическом плане для характеристики природы связи между явлениями, то рекомендации, какое уравнение для этого использовать, становятся неопределенными. Вот несколько примеров таких рекомендаций.

«Какую из двух форм мы выберем, будет определяться обстоятельствами. . ., выбор диктуется целью исследования. Например, в выражении связи между током и сопротивлением в электрической цепи исследователь может выбрать тот фактор за независимую переменную, который он может непосредственно контролировать. Часто, однако, нет оснований для такого выбора, и необходимо находить обе кривые» (Юл, Кендэл, 1960). «В практических задачах обычно имеет смысл лишь одна прямая регрессии. . . Нас интересует только вопрос о том, зависит ли и как зависит выход стали. . . от угара кремния, а не наоборот: угар кремния от выхода стали» (Шторм, 1970, стр. 206). Существуют и другие мнения на этот счет: «. . . практические ситуации, в которых такая постановка задачи (т. е. вопрос выбора требуемого уравнения регрессии и т. д., — *Ю. Т., Я. Ю.*) оказывается содержательной, крайне редки. . .» (Айвазян, 1968, стр. 127).

Двойственность уравнений регрессии Я. И. Лукомский (1958) называет «законом регрессии», который выражается в том, что изменение признака замедляется, когда он выступает в корреляционной зависимости на правах функции по сравнению с изменением его на правах аргумента. Причину существования закона регрессии Я. И. Лукомский усматривает в том, что для проведения линии регрессии y по x данные группируются по значениям x, а для сопряженной линии регрессии — по значениям y. При перемене признака группировки соотношение между сгруппированными средними, по мнению Я. И. Лукомского, видоизменяется. Эти, вообще говоря, правильные рассуждения можно назвать, однако, лишь введением к объяснению. В них остается неясным главный вопрос: почему указанное соотношение видоизменяется и притом в сторону уменьшения приращения *оцениваемой* величины.

Заметим, наконец, что многие авторы, в особенности «чистые» математики, не придают значения «необратимости» уравнений регрессии, считая, что каждое из них является решением самостоятельной задачи: либо оценки y по x по принципу наименьших

квадратов, либо аналогичной оценки значений x по y. В ответ на желание практиков найти «истинное» соотношение между y и x математики указывают, что такая постановка вопроса бессодержательна, и статистическое исследование не может и не должно решать такой вопрос. Мы не можем полностью согласиться с таким мнением, хотя действительно во многих случаях статистический материал не позволяет дать однозначной оценки положения графика функциональной зависимости.

§ 2. Анализ простейшей математической модели корреляционной зависимости

Добиться ясной интерпретации двойственности уравнений регрессии можно анализом исходной математической модели, приводящей к линейной корреляции между двумя переменными. Обозначим величины, связанные функциональной зависимостью, через x' и y', а соответствующие им измеряемые (наблюдаемые) величины — x и y.

Пусть величина x' имеет распределение со средним значением $\mathbf{M}(x')$ и дисперсией $\sigma^2_{x'}$; а величина y' связана с ней функциональной зависимостью $y'=bx'+a$. Положим также, что $y=y'+\xi$, $x=x'$; тогда

$$y = bx + a \pm \xi, \tag{8.2.1}$$

где ξ — независимая от x и y случайная величина, имеющая распределение с дисперсией σ^2_ξ и средним значением $\mathbf{M}(\xi)=0$. Последнее условие несущественно и принято ради удобства. После замены $\xi=\mathbf{M}(\xi)\pm\Delta\xi$ всегда можно прийти к аналогичной модели: $y=bx+[a+\mathbf{M}(\xi)]\pm\Delta\xi=bx+a'\pm\Delta\xi$, в которой $\mathbf{M}(\Delta\xi)=0$.

Величина ξ в этой модели отражает действие некоторого случайного фактора, «мешающего» проявлению функциональной зависимости в «чистом» виде. В частном случае это может быть погрешность наблюдений величины y. Дисперсию величины ξ назовем для краткости *дисперсией мешающего фактора*. Дисперсия величины y в модели (8. 2. 1) будет равна

$$\sigma^2_y = b^2\sigma^2_x + \sigma^2_\xi = b^2\sigma^2_{x'} + \sigma^2_\xi. \tag{8.2.2}$$

Первое слагаемое отражает изменчивость величины x и является следствием функциональной зависимости между x' и y'. Величина σ^2_ξ является уже известной нам остаточной дисперсией, связанной с коэффициентом корреляции r_{xy}, откуда получаем

$$1-r^2 = \sigma^2_\xi/\sigma^2_y = \sigma^2_s/(b^2\sigma^2_x+\sigma^2_\xi), \tag{8.2.3}$$

или

$$r^2 = \frac{b^2\sigma^2_x}{b^2\sigma^2_x+\sigma^2_\xi} = \frac{\sigma^2_{y'}}{\sigma^2_y}. \tag{8.2.4}$$

Очевидное тождество σ^2_ξ с остаточной дисперсией является доказа-

тельством того, что уравнение регрессии y на x совпадает с уравнением функциональной зависимости между y' и x', т. е. $b_{yx}=b$. Действительно, из постулированной модели (8. 2. 1) непосредственно вытекает, что $M(Y/X=x)=bx+a$, где левая часть означает линию регрессии y на x.

Каким образом будет располагаться сопряженная линия регрессии x на y, т. е. линия математических ожиданий $M(X/Y=y)$? Для определения ее углового коэффициента b_{yx} можно воспользоваться равенством $b_{xy}b_{yx}=r^2_{xy}$, откуда

$$b_{xy}=r^2/b_{yx}=r^2/b=b\cdot\sigma^2_x/\sigma^2_y, \qquad (8.2.5)$$

или

$$b_{xy}=\frac{b\sigma^2_x}{b^2\sigma^2_\xi+\sigma^2_\varepsilon}, \text{ или } \frac{b_{xy}}{b_{yx}}=\frac{\sigma^2_x}{\sigma^2_y}. \qquad (8.2.5а)$$

Прямая регрессии x на y тем сильнее будет отклоняться от прямой функциональной зависимости с угловым коэффициентом $1/b$, чем меньше r_{xy} и чем больше, следовательно, σ^2_ξ. Таким образом, чем больше погрешность в определении величины y (больше дисперсия мешающего фактора!), тем на больший угол отклоняется уравнение регрессии на эту переменную при условии, что другая переменная определена без погрешностей. Дисперсия мешающего фактора не обязательно порождается погрешностями наблюдений или анализов. Пусть, например, предполагается, что содержание германия в угольном пласте c при прочих равных условиях обратно пропорционально его мощности m, т. е. величины $1/c$ и m связаны прямой пропорциональной зависимостью. Однако в процессе внутриформационных размывов и неоднородного уплотнения первоначальная мощность пласта искажена. В этом случае уравнение регрессии $1/c$ на m будет отклонено на некоторый угол, который можно вычислить и по которому можно оценить степень искажающего влияния размывов. Силу связи x и y (читай: $1/c$ и m) можно, как известно, охарактеризовать с помощью дисперсий остаточных величин: $1-r^2=\sigma^2_{ост.x}/\sigma^2_x$; $1-r^2=\sigma^2_{ост.y}/\sigma^2_y=\sigma^2_\xi/\sigma^2_y$. Из этих равенств можно получить величину остаточной дисперсии оцениваемой переменной $\sigma^2_{ост.x}$, которая и будет характеризовать мешающее действие размывов:

$$\sigma^2_{ост.x}/\sigma^2_x=\sigma^2_{ост.y}/\sigma^2_y; \quad \sigma^2_{ост.x}=\sigma^2_\varepsilon\sigma^2_x/\sigma^2_x, \qquad (8.2.6)$$

или

$$\sigma^2_{ост.x}/\sigma^2_{ост.y}=\sigma^2_x/\sigma^2_y=b_{xy}/b_{yx}. \qquad (8.2.7)$$

Последние равенства любопытны тем, что отношение остаточных дисперсий в общем случае не равно отношению величин x' и y', как это могло бы показаться из анализа модели (8. 2. 1), а равно отношению дисперсий наблюденных коррелируемых величин: x и y.

Поскольку остаточные дисперсии характеризуют точность оценки одной из коррелируемых величин по значениям другой, то из (8. 2. 7) следует важный в практическом отношении вывод: абсолютные точности оценки одной коррелируемой величины по значениям другой относятся как их дисперсии. Из этого же равенства следует, что «нормированная погрешность» оценки одинакова для обеих переменных и равна $\sqrt{1-r^2}$. «Нормированная погрешность» не синоним относительной погрешности. Последняя означает, какую долю составляет погрешность по отношению к среднему значению этой величины, а «нормированная погрешность» является долей погрешности оценки величины в общей ее изменчивости. Она получается при выражении величин и их погрешностей в долях стандартных отклонений. Изложенное является статистической формой выражения того свойства информации, что ее количество, которое несет первое из взаимосвязанных явлений о втором, равно количеству информации второго явления о первом.

Рассмотрим теперь модель, в которой мешающие факторы влияют только на величину x: $y'=bx'+a$, $x=x'\pm\eta$, $y=y'$, отсюда $y=y'=b\,(x\pm\eta)+a$. Решая это равенство относительно x и обозначая $b'=1/b$, $a'=-b'a$, получим модель, аналогичную модели (8. 2. 1) с переменой осей координат:

$$x=b'y+a'\pm\eta. \qquad (8.2.8)$$

Теперь прямая функциональной зависимости совпадает с прямой регрессии x на y, а сопряженная прямая регрессии отклоняется от нее на угол Θ, определяемый коэффициентом корреляции. Все относящиеся к модели (8. 2. 1) рассуждения, естественно, относятся и к модели (8. 2. 8) с заменой ξ на η.

Из проведенного анализа моделей вытекает, что прямая функциональной зависимости совпадает с прямой регрессии на ту из переменных, которая не подвергалась действию «мешающих» факторов.

В общем случае, когда мешающие факторы действуют на обе переменные, прямая графика функциональной зависимости не совпадает ни с одним из уравнений регрессии, располагаясь между ними. Положение ее определяется соотношением дисперсий мешающих факторов на изучаемые величины (рис. 23, 24): она располагается ближе к прямой регрессии на ту переменную, которая в меньшей степени подвергалась «искажению».

К одному и тому же статистическому результату, а именно к одному и тому же коэффициенту корреляции и одним и тем же двум уравнениям регрессии приводит бесчисленное множество функциональных зависимостей, графики которых заключены между прямыми регрессии. На основании упомянутого эмпири-

ческого материала принципиально невозможно однозначно найти положение линии функциональной зависимости.

Коэффициент корреляции до сих пор рассматривался и рассматривается как мера тесноты связи между величинами, однако не менее важно (а для нас — более важно) видеть в нем мера неопределенности формы линейной связи между

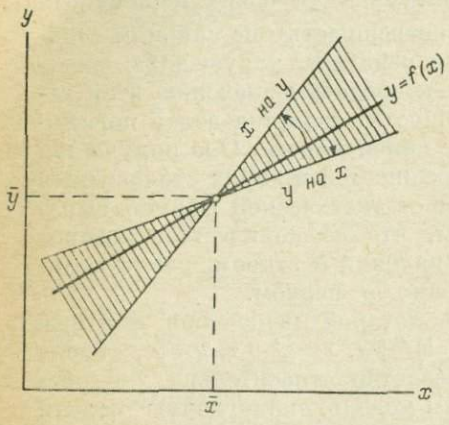


Рис. 23. Зона положения прямой функциональной зависимости между прямыми регрессии.

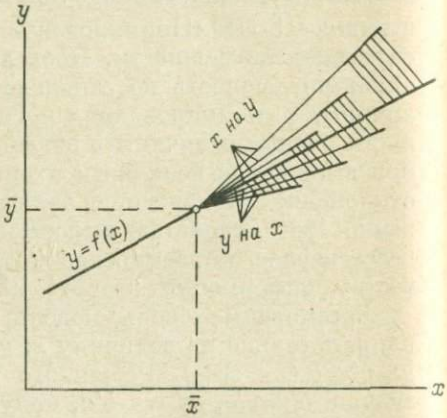


Рис. 24. Возможные положения уравнений регрессии по отношению к уравнению функциональной зависимости при заданном коэффициенте корреляции.

ними. Говорить о «единой линии связи», о линии функциональной зависимости, если между рассматриваемыми величинами коэффициент корреляции не равен единице, — такой же абсурд, как и говорить о точном значении величины, если дан только интервал ее возможных значений. Тем не менее в каждом конкретном случае такая функциональная зависимость существует. Затруднение состоит только в том, что в уравнениях регрессии и коэффициенте корреляции нет достаточной информации для ее установления.

Таким образом, вопрос естествоиспытателей о том, какое из уравнений регрессии следует брать в качестве оценки линии функциональной зависимости, нуждается в следующем уточнении: какую из возможных прямых, заключенных между прямыми регресии (включая и их самих), следует считать оценкой графика функциональной зависимости? Ответить на этот вопрос без дополнительной информации об изучаемых переменных невозможно. Такая информация должна заключаться в значениях дисперсий мешающих факторов, отнесенных к каждой переменной, и если она имеется, то положение графика функциональной зависимости может быть либо значительно уточнено, либо определено однозначно.

§ 3. О так называемой единой линии связи

Отыскание «единой линии» связи как оценки функциональной зависимости по наблюдениям, отягощенным погрешностями, является продолжением проблемы двух линий регрессии. Одна из первых попыток принадлежит Б. И. Срезневскому (цит. по: А. Конюс, 1950), предложившему в качестве такой линии прямую с угловым коэффициентом $b = \sigma_y^2 / \sigma_x^2$. Далее будет показано, что это решение является частным случаем, подходящим для рассмотренного им примера нормы бури: очевидно, станции А и В находились «в равном положении» при измерении силы ветра.

К. Пирсон (Pearson, 1901) в качестве единой линии предложил использовать так называемую главную ось корреляционной поверхности, т. е. главную (длинную) ось эллипса рассеивания (корреляции). Он считал, что она более подходит изучению такой характеристики, как скорость роста (приращение функции при данном приращении аргумента). Легко понять стремление к «единой линии», особенно среди геологов, геохимиков и палеонтологов, имеющих дело с малыми значениями коэффициентов корреляции. В этих условиях угол наклона регрессионной линии будет зависеть от того, какой из двух признаков выбран в качестве аргумента. В палеобиометрических исследованиях (Миллер, Кан, 1965), где функцию и аргумент нередко выбирают без реального обоснования, усовершенствованию методики построения единой линии уделяется серьезное внимание.

Так, Г. Тиссьер (цит. по: Миллер, Кан, 1965, стр. 211) предложил единую линию, которую он назвал сокращенной главной осью, являющуюся в современной терминологии прямой ортогональной регрессии. Как и уравнения регрессии, она строится по методу наименьших квадратов, но минимизируется сумма квадратов отклонений не по горизонтали или вертикали, а по перпендикуляру к искомой линии. Еще ранее А. Вальд (Wald, 1940) предложил следующий метод проведения линии функциональной зависимости: а) для «локализации» прямой используется точка $(\bar{x}, \bar{y})$; б) для определения угла наклона используются две точки с координатами $(\bar{x}_1, \bar{y}_1)$ и $(\bar{x}_2, \bar{y}_2)$, где $x_{1,\,2}$, $y_{1,\,2}$ — значения средних в двух половинах выборки, разбитой по возрастанию x. Совершенствуя метод Вальда, М. С. Бартлетт (Bartlett, 1949) для определения угла наклона использует также две точки с координатами $(\bar{x}_1, \bar{y}_1)$ и $(\bar{x}_3, \bar{y}_3)$, определенными по средним значениям переменных в первой и третьей равночисленных группах, на которые разбита совокупность по оси x. Встречаются также работы с использованием в качестве единой линии биссектрисы угла между прямыми регрессии.

Авторы перечисленных методов подразумевали, что параметры предлагаемых линий ближе к параметрам функциональной зависимости, чем параметры любого из двух уравнений регрессии.

Теперь ясно, что: а) никакие математические приемы не устраняют неопределенности в положении линии функциональной зависимости, поскольку в распоряжении исследователя нет для этого объективных данных; б) любая прямая, располагающаяся между прямыми регрессии, одинаково хороша в качестве единой линии; в) информация, необходимая для нахождения параметров линии функциональной зависимости, заключается в значениях дисперсий мешающих факторов; г) правильный путь лежит не в поисках новых формальных методов построения какой-либо «единой» линии, а в конкретном исследовании дисперсий мешающих факторов.

§ 4. Корреляция величин, отягощенных аналитическими погрешностями. Задача Берксона. Зависимость между суммами коррелирующих величин

Выше мы рассмотрели модели корреляции, в основе которых лежит прямая пропорциональная зависимость, осложненная случайными «мешающими» факторами, или случайными слагаемыми. Напрашивается вывод, что эти слагаемые целесообразно разделить на две группы. К одной группе отнесем те из них, для которых известны: а) к какой из переменных они относятся; б) их величина и закон распределения. К другой группе отнесем остальные, для которых известно, что они существуют и то зачастую только потому, что исключение слагаемых ξ и η не приводит к «чистой» функциональной зависимости. Классический пример «мешающих» факторов первой группы — погрешности наблюдений или анализов, влияние которых мы подробно рассмотрим. Регистрация переменных, например содержаний химических элементов, всегда производится с известной погрешностью (погрешность воспроизводимости), которая и будет выступать в роли мешающего фактора для каждого определяемого элемента.

Интерес к влиянию погрешностей наблюдений на положение уравнений регрессии отражен в очень интересных работах М. Бартлетта (Bartlett, 1949), Дж. Берксона (Berkson, 1950), А. Вальда (Wald, 1940) и Д. Линдли (Lindley, 1953). Однако результаты этих работ остались мало известными широкому кругу естественников. М. Бартлетт выделяет два уже известных нам типа задач в регрессионном анализе: получение уравнения регрессии, которое затем используется для предсказания, и оценку линейной функциональной зависимости, если она существует. Он подчеркивает, что вторая задача не имеет единственного решения без допущений об относительной величине ошибок (ошибок наблюдений, — *Ю. Т., Я. Ю.*) обеих переменных.

Рассмотрим результаты фундаментальной статьи Дж. Берксона (Berkson, 1950), относящиеся к исследованию регрессии признаков, которые измерены с заданными погрешностями на объектах

данной совокупности. Путем анализа математической модели он пришел к следующим простым формулам для оценки угловых коэффициентов прямых регрессии. Пусть b_{xy} и b_{uv} — угловые коэффициенты прямых регрессии y на x и v на u, где y и x — измеренные значения изучаемых признаков, отягощенные ошибками наблюдений; v и u — истинные их значения (не путать с y' и x', между которыми постулируется прямолинейная функциональная зависимость!). Тогда

$$b_{vu} = b_{yx} \frac{\sigma_u^2 + \sigma_\eta^2}{\sigma_u^2}, \qquad (8.4.1)$$

где σ_η^2 — дисперсия погрешностей измерения величины x, т. е. $x = u \pm \eta$, $M(\eta) = 0$. Из равенства (8.4.1) следует, что «истинная» линия регрессии тем сильнее отклоняется от наблюденной линии регрессии, чем больше погрешность наблюдений той переменной, на которую проведена линия регрессии (рис. 25). В частном случае, когда погрешность наблюдений x равна нулю ($\sigma_\eta^2 = 0$), линия регрессии на эту переменную совпадает с прямой функциональной зависимости: $b_{uv} = b_{yx} \frac{\sigma_u^2 + 0}{\sigma_u^2} = b_{yx}$. Этот вывод как частный случай совпадает с рассмотренным в предыдущем разделе. Аналогично можно написать для линии регрессии x на y:

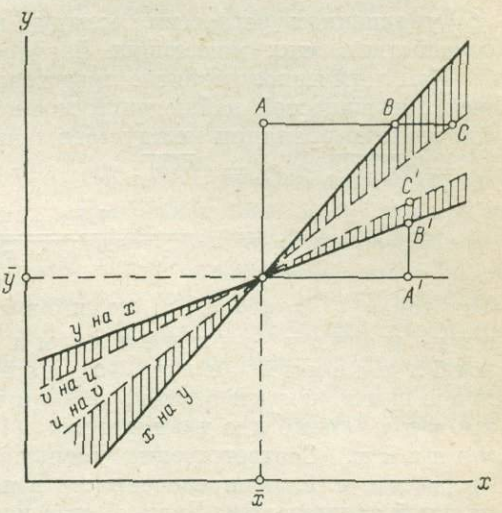


Рис. 25. Соотношение уравнений регрессии «истинных» переменных и переменных, отягощенных аналитическими погрешностями.

$$b_{uv} = b_{xy} \frac{\sigma_v^2 + \sigma_\xi^2}{\sigma_v^2}. \qquad (8.4.2)$$

Так как нам непосредственно известны наблюденные дисперсии величин, то равенства (8.4.1) и (8.4.2) удобнее переписать в следующем виде:

$$b_{vu} = b_{yx} \frac{\sigma_x^2}{\sigma_x^2 - \sigma_\eta^2} = b_{yx} \frac{\sigma_x^2}{\sigma_u^2}; \qquad (8.4.3)$$

$$b_{uv} = b_{xy} \frac{\sigma_y^2}{\sigma_y^2 - \sigma_\xi^2}. \qquad (8.4.4)$$

Последние уравнения легко связать с выводами о разделении мешающих факторов на две группы. Действительно, если влияние части мешающих факторов изучено настолько, что имеется возможность выразить его дисперсиями σ_η^2 и σ_ξ^2, то эти факторы мы относим к первой группе, после чего правомерно поставить вопрос об уравнениях регрессии между величинами, «очищенными» от действия этих мешающих факторов.

Как логическое развитие представлений об искажающем влиянии погрешностей наблюдений можно поставить вопрос о коэффициенте корреляции между «очищенными» величинами. Так как $r_{xy} = \sqrt{b_{xy} b_{yx}}$, а $r_{uv} = \sqrt{b_{uv} b_{vu}}$, то

$$r_{uv} = r_{xy} \frac{\sigma_x \sigma_y}{\sqrt{(\sigma_x^2 - \sigma_\eta^2)(\sigma_y^2 - \sigma_\xi^2)}} = r_{xy} \frac{\sigma_x \sigma_y}{\sigma_u \sigma_v}. \qquad (8.4.5)$$

Так как $\sigma_u \leqslant \sigma_x$ и $\sigma_v \leqslant \sigma_y$, то коэффициент корреляции между очищенными от погрешностей наблюдений (или от действия других мешающих факторов) величинами выше, чем между наблюденными значениями переменных. Соответственно уменьшается угол между прямыми регрессии, т. е. неопределенность в положении прямой функциональной зависимости. Если σ_η^2 и σ_ξ^2 исчерпывают действие мешающих факторов, то в этом, и только в этом случае, прямая функциональной зависимости может быть локализована однозначно, и *расчетный коэффициент корреляции* между «очищенными» величинами становится равным единице.

Какой из коэффициентов корреляции и какие из уравнений регрессии должны использоваться в практике: между «очищенными» значениями переменных или между наблюденными? Ответ на этот вопрос зависит от постановки задачи. Следует лишь различать, какие переменные при этом подразумеваются. Например, в геохимической практике всегда вычисляются и обсуждаются коэффициенты корреляции, которые называют коэффициентами корреляции между содержаниями элементов, хотя на самом деле они являются коэффициентами корреляции между измеренными значениями содержаний. Очевидно, что они занижены по сравнению с «истинными», «геохимическими» коэффициентами в пропорции, определяемой формулой (8. 4. 5). Значения их непостоянны и зависят от точности анализов: с уменьшением последней оценки коэффициентов будут уменьшаться. Пользуясь малоточными методами, например полуколичественными спектральными анализами, мы улавливаем лишь малую долю геохимических корреляций, о чем не следует забывать при интерпретации. Такая постановка вопроса кажется настолько очевидной, что поневоле

удивляет отсутствие в статистической литературе упоминаний об этом.[1] Имеет ли право на существование коэффициент корреляции между измеренными значениями концентраций? Разумеется, имеет, особенно если соответствующие ему уравнения регрессии используются в оценочных целях (прогноз содержания одного элемента по другому). Но для обоснования геохимических выводов больше подходил бы «геохимический» коэффициент корреляции.

Мы рассмотрели случай с единственным источником мешающих факторов — погрешностями анализа. Однако природные зависимости осложнены целым рядом мешающих факторов, лишь часть из которых поддается расшифровке; их после такой расшифровки мы относим к факторам первого рода. Полученные результаты легко обобщаются для случая, когда исследуемые величины x и y образуются из n слагаемых (по числу действующих факторов). Прежде мы рассмотрели случай, когда погрешности анализа ξ и η были некоррелированы как между собой ($r_{\xi\eta}=0$), так и с другими составляющими ($r_{\xi v}=0$, $r_{\eta u}=0$). Довольно легко учесть и эти корреляции, если они наблюдаются. В итоге получаем следующие формулы для коэффициента корреляции между суммами (коррелированных и некоррелированных) случайных величин:

$$r_{xy}=\sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij}k_{x_i}k_{y_j}, \qquad (8.4.6)$$

где r_{ij} — коэффициент корреляции между i-м слагаемым из суммы, составляющей переменную x (всего n слагаемых), и j-м слагаемым из суммы, составляющей переменную y (всего m слагаемых); k_{x_i}, k_{y_j} — отношения стандартного отклонения данного слагаемого к стандарту суммы:

$$k_{x_i}=\frac{\sigma_{x_i}}{\sigma_x};\quad k_{y_i}=\frac{\sigma_{y_i}}{\sigma_y}. \qquad (8.4.7)$$

Этот любопытный результат означает, что коэффициент корреляции между двумя величинами равен взвешенной сумме коэффициентов корреляции между слагаемыми этой суммы, т. е. так называемых *парциальных коэффициентов корреляции* (не путать с частными коэффициентами корреляции!).

В частном случае при $n=2$ $m=2, r_{1,2}=0, r_{2,2}=0, r_{2,1}=0$ получаем формулу (8.4.5). Интерпретируя этот результат в терминах погрешностей анализов, можно заключить, что коэффициент корреляции между результатами анализов равен взвешенной сумме коэффициентов корреляции между истинными содержаниями

[1] Формула, аналогичная (8.4.5), приведена без вывода в недавней работе А. И. Гавришина и Г. А. Востроknутова (1971).

$r_{x'y'}\frac{\sigma_{x'}\sigma_{y'}}{\sigma_x\sigma_y}$, погрешностями анализа $r_{\xi\eta}\frac{\sigma_\eta\sigma_\xi}{\sigma_x\sigma_y}$ и погрешностями анализа одного элемента и содержанием в пробе другого элемента: $r_{x\xi}\frac{\sigma_x\sigma_\xi}{\sigma_x\sigma_y}$, $r_{y\eta}\frac{\sigma_y\sigma_\eta}{\sigma_x\sigma_y}$. Отсюда следует важный вывод о необходимости изучения трех последних корреляций «аналитического» происхождения, с тем, чтобы освободиться от них при восстановлении геохимической корреляции.

§ 5. Универсальная линейная модель корреляции

Полученный выше результат может служить основой универсальной линейной модели корреляции. Представим, что каждый из изучаемых признаков x и y зависит от нескольких факторов (общим числом n). Некоторые из факторов влияют на обе переменные, остальные — на одну из них. В этом случае пара значений (x, y), например содержаний двух элементов в пробе, определяется набором значений факторов. Это можно выразить следующим образом:

$$\left.\begin{aligned} x &= x_1 + x_2 + \ldots + x_i + \ldots + x_n \\ y &= y_1 + y_2 + \ldots + y_i + \ldots + y_n \end{aligned}\right\}. \qquad (8.5.1)$$

Например, если всего факторов пять (F_1, F_2, . . ., F_5) и признак x не зависит от F_5, а признак y — от F_4, то $x=x_1+x_2+x_3+x_4+0$, $y=y_1+y_2+y_3+0+y_5$. Слагаемые x_i, y_i являются линейными функциями соответствующих факторов F_i:

$$\left.\begin{aligned} x_i &= a_{ix}F_i \\ y_i &= a_{iy}F_i \end{aligned}\right\} i = 1,\ 2,\ \ldots n. \qquad (8.5.2)$$

Система линейных уравнений (8. 5. 2) исчерпывает информацию о зависимости между x и y. В частности, она определяет парциальные коэффициенты корреляции между слагаемыми наших переменных. Между слагаемыми с одинаковыми номерами (теми, которые определяются одноименными факторами) парциальные коэффициенты корреляции будут равны единицам, так как $\frac{x_i}{y_i}=\frac{a_{ix}F_i}{a_{iy}F_i}=\frac{a_{ix}}{a_{iy}}=\text{const}$. Все остальные парциальные коэффициенты в случае взаимной независимости факторов будут равны нулю, в противном случае, очевидно $r_{x_i, y_j}=r_{F_iF_j}$. Таким образом (табл. 16), $r_{xy}=\sum_{i=1}^{n}\sum_{j=1}^{n} r_{F_iF_j}k_{x_i}k_{y_j}$, или, уплотняя обозначения,

$$r_{xy}=\sum_{i=1}^{n}\sum_{j=1}^{n} r_{ij}k_{ij}. \qquad (8.5.3)$$

Сумма квадратов весовых коэффициентов k_{ij}, являющихся долями

общей дисперсии, равна единице: $\sum_{i=1}^{n}\sum_{j=1}^{n} k_{x_i}^2 k_{y_j}^2 = \sum\sum \frac{\sigma_{x_i}^2 \sigma_{y_j}^2}{\sigma_x^2 \sigma_y^2} =$

$= \frac{1}{\sigma_x^2 \sigma_y^2} \sum \sigma_{x_i}^2 \sum \sigma_{y_j}^2 = 1.$

Рассмотрим пример. Некоторая горная порода содержит n ингредиентов, несущих по крайней мере один из элементов x и y, например содержит минералы A, B, C. Все остальные минералы, не несущие x и y, в рассмотрение не берутся. Пусть минерал B состоит целиком из элемента x, минерал C — из элемента y, а в минерале A имеются оба элемента в постоянных количествах. В этом примере изучаемые переменные — содержания элементов, действующие факторы — содержания минералов в породе. Коэффициент корреляции между x и y определится по формуле (8. 5. 3), а значения парциальных коэффициентов корреляции и весовых коэффициентов сведены в табл. 16.

Таблица 16

Структура коэффициента корреляции между компонентами x и y

Содержание минерала в породе	Содержание элемента в минерале	Элементы в породе	c_A	c_B	c_C
			x_A	x_B	x_C
			$x_1 = x_A \cdot c_A$	$x_2 = x_B \cdot c_B$	$x_3 = x_C \cdot c_C$
c_A	y_A	$y_1 = y_A c_A$	$\frac{r_{AA}=1}{x_A \sigma_A y_A \sigma_A}$	$\frac{r_{AB}}{x_B \sigma_B y_A \sigma_A}$	$\frac{r_{AC}}{x_C \sigma_C y_A \sigma_A}$
c_B	y_B	$y_2 = y_B c_B$	$\frac{r_{AB}}{x_A \sigma_A y_B \sigma_B}$	$\frac{r_{BB}=1}{x_B \sigma_B y_B \sigma_B}$	$\frac{r_{CB}}{x_C \sigma_C y_B \sigma_B}$
c_C	y_C	$y_3 = y_C c_C$	$\frac{r_{AC}}{x_A \sigma_A y_C \sigma_C}$	$\frac{r_{CB}}{x_B \sigma_B y_C \sigma_C}$	$\frac{r_{CC}=1}{x_C \sigma_C y_C \sigma_C}$

Примечание. Над чертой — парциальный коэффициент корреляции, под чертой — часть весового коэффициента (ее надо разделить на $\sigma_x \sigma_y$).

В простейшем случае, когда содержания минералов-носителей некоррелированы, «общий» коэффициент корреляции (между x и y) будет равен взвешенной сумме коэффициентов из главной диагонали, которые равны единице (остальные парциальные коэффициенты равны нулю). В итоге имеем просто сумму весовых коэффициентов:

$$r_{xy} = 1 \cdot k_{x_A} k_{y_A} + 1 \cdot k_{x_B} k_{y_B} + 1 \cdot k_{x_C} k_{y_C}. \qquad (8.5.4)$$

Общий коэффициент корреляции r_{xy} в этом случае будет опреде-

ляться только изменчивостью содержаний минералов-носителей, так как только они определяют значения весовых коэффициентов. Например,

$$k_{x_A} k_{y_A} = \frac{\sigma_{x_1}\sigma_{y_1}}{\sigma_x \sigma_y} = \frac{\sigma_A x_A \sigma_A y_A}{\sigma_x \sigma_y} = \frac{x_A y_A \sigma_A^2}{\sigma_x \sigma_y}. \qquad (8.5.5)$$

Таким образом, $r_{xy} = \frac{1}{\sigma_x \sigma_y}(x_A y_A \sigma_A^2 + x_B y_B \sigma_B^2 + x_C y_C \sigma_C^2)$, но так как у нас, по условию $x_B = 1$, $y_B = 0$, $x_C = 0$, $y_C = 1$, то

$$r_{xy} = \frac{x_A y_A \sigma_A^2}{\sigma_x \sigma_y} = \frac{x_A y_A \sigma_A^2}{\sqrt{(x_A^2 \sigma_A^2 + \sigma_B^2)(y_A^2 \sigma_A^2 + \sigma_C^2)}}. \qquad (8.5.6)$$

Этот любопытный результат показывает, что коэффициент корреляции в данном случае зависит от диапазона изменчивости того минерала, который является носителем обоих элементов.

Зная средние содержания элементов в минералах и дисперсии содержаний этих минералов, зависящие от величины пробы и крупности слагающих зерен, можно для основных типов горных пород рассчитать своего рода «кларки коэффициентов корреляции». Сравнение их с фактическими корреляциями будет способствовать решению геохимических задач. Таким образом, можно констатировать многозначительную эволюцию «геохимических констант», или «норм», которые использует геохимик в своей работе. Вначале были только средние содержания — кларки ($\bar{x}$), затем в поле зрения геохимика была включена другая важная характеристика природного распределения — дисперсия (s_x^2). Теперь мы стоим перед назревшей задачей нахождения «кларковых корреляций» для различных типов пород.

«Расчленение» связи между двумя характеристиками на отдельные независимые или коррелирующие факторы (например, минералы, содержащие изучаемые элементы) представляет собой необходимый и в рассмотренном аспекте законченный анализ линейной зависимости двух случайных величин. Вообще, исследование зависимости не может считаться законченным до тех пор, пока не будут учтены все действующие (мешающие) факторы, т. е. пока мы не расчленим корреляционную зависимость на ряд функциональных зависимостей между факторами и изучаемыми величинами.

Остается заметить, что одним из действующих факторов на величину y во многих случаях может быть величина x, как в самой простой модели с двумя действующими факторами ($y = ax + \xi$), являющейся частным случаем рассмотренной. Здесь один из факторов — сама величина x, другой — случайная погрешность $x = x_1 = F_1$, $y = y_1 + y_2 = ax_1 + \xi = aF_1 + \xi$. Только один парциальный

коэффициент корреляции здесь не равен нулю, а именно $r_{x_1y_1} = 1$. Весовой коэффициент при нем равен

$$k_{x_1y_1} = \frac{\sigma_{F_1} a\sigma_{F_1}}{\sigma_x \sigma_y} = \frac{\sigma_x a\sigma_x}{\sigma_x \sigma_y} = \frac{a\sigma_x}{\sqrt{a^2\sigma_x^2 + \sigma_\xi^2}}. \qquad (8.5.7)$$

В заключение отметим, что изложенный материал, как нам кажется, подготовит читателя к восприятию идей факторного анализа и восполнит тот разрыв, который существует между литературой по корреляционному и регрессионному анализу, с одной стороны, и факторному анализу — с другой.

Глава 9

ВОПРОСЫ ИНТЕРПРЕТАЦИИ И ИСПОЛЬЗОВАНИЯ РЕЗУЛЬТАТОВ ИССЛЕДОВАНИЯ ЗАВИСИМОСТЕЙ

§ 1. Интерпретация уравнений регрессии как источника информации

В предыдущих разделах было рассмотрено соотношение прямых регрессии с уравнением функциональной зависимости, уменьшение угла между прямыми регрессии в результате устранения влияния некоторых мешающих факторов, например погрешностей наблюдений, а также объяснение «закона регрессии» как результата неопределенности положения прямой функциональной зависимости при наблюденном коэффициенте корреляции, отличном от нуля. Рассмотрим теперь механизм «закона регрессии» и дадим ему подробную интерпретацию с двух позиций: «что это означает» и «как это получается».

В основу дальнейшего положено коренное изменение взгляда на существо оценки y по известному значению x с помощью уравнения регрессии. Эта оценка основывается на информации двоякого рода. Во-первых, она опирается на полученное в данном измерении значение x, т. е. на зависимости между переменными. Когда говорят об оценке по уравнению регрессии, то обычно подразумевают использование именно этой информации. Во-вторых (что менее очевидно!), эта оценка основывается на параметрах распределения оцениваемой величины y в изучаемой совокупности — на его среднем значении $\bar{y}$ и дисперсии s_y^2. Действительно, если бы у нас не имелось никаких других сведений о значении признака y_i в i-м элементе, например пробе, случайным образом выбранной из месторождения, мы вынуждены были бы положить $y_i = \bar{y}$, причем дисперсия этой оценки была бы равна s_y^2.

Как поступить, если, кроме этой (скажем, весьма неопределенной) информации, имеется другая, почти столь же неопределен-

ная? Известно, что наиболее эффективная оценка с использованием нескольких источников информации есть средневзвешенное значение. Весовыми коэффициентами служат точности соответствующих источников информации, в данном случае — величины, обратные дисперсиям. Таким образом, можно записать

$$\hat{y}_x = \frac{\bar{y}/\sigma_y^2 + f(x)/\sigma_{yx}^2}{1/\sigma_y^2 + 1/\sigma_{yx}^2}, \qquad (9.1.1)$$

где $\hat{y}_x$ — оценка значения y по данному x; $f(x)$ — функция зависимости y от x; σ_{yx}^2 — дисперсия оценки y по x с помощью $f(x)$. Обратимся к рис. 26, иллюстрирующему это равенство. Пусть прямая OA есть линия функциональной зависимости $y=bx+a$, проходящая через точку O с координатами $(\bar{x}, \bar{y})$. Тогда две рассматриваемые оценки y при данном значении x_0, т. е. $\bar{y}$ и $f(x_0)$, будут соответствовать ординатам точек A и B, а средневзвешенное значение по уравнению (9.1.1) — ординате точки C, которая делит отрезок AB в отношении

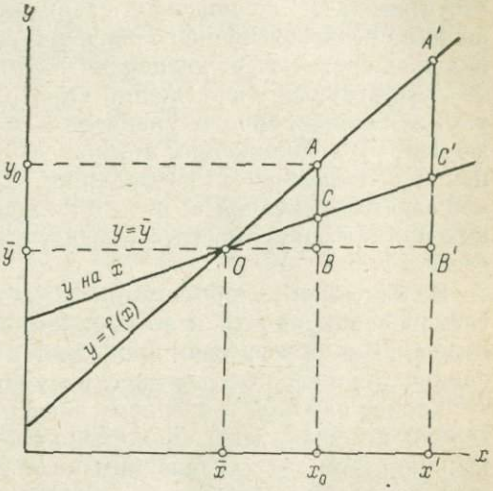


Рис. 26. График, иллюстрирующий происхождение уравнения регрессии из двух источников информации. Пояснения в тексте.

$$AC/CB = \sigma_{yx}^2/\sigma_y^2. \qquad (9.1.2)$$

Отношение AC/CB остается постоянным для любого значения x, так как σ_y^2 и σ_{yx}^2 — величины постоянные для данной совокупности и данной зависимости. Поэтому точка C' делит отрезок $A'B'$ в том же отношении, следовательно, лежит на прямой OC. Таким образом доказано, что линия средневзвешенных оценок представляет собой п р я м у ю.

Рассмотрим природу весовых коэффициентов. Один из них — дисперсия наблюденных значений σ_y^2, которая является одновременно дисперсией оценки среднего по единичному значению из совокупности. Существо другого весового коэффициента — дисперсии σ_{yx}^2 — выясняется из следующих рассуждений. Представим себе, что величины y и x связаны функциональной зависимостью. Тогда, если бы нам было известно точное значение одной из переменных — x, то значение другой было бы оценено путем

вычислений также без погрешностей. Если x определены с погрешностью σ_η, точность оценки y целиком и полностью определялась бы этой погрешностью:

$$\sigma^2_{\bar{y}_x} = \sigma^2_\eta b^2, \qquad (9.1.3)$$

т. е. погрешность определения значения функции по значению аргумента есть выражение влияния мешающего фактора на величину x. Используя это, можно доказать,[1] что прямая OC совпадает с уравнением регрессии, а дисперсия оценки средневзвешенного значения есть не что иное, как остаточная дисперсия, т. е. дисперсия оценки по уравнению регрессии.

Аналогичные рассуждения справедливы и для оценки x по y, т. е. для сопряженного уравнения регрессии. И здесь мы убеждаемся, что положение прямой функциональной зависимости может быть выбрано произвольно, если не определены дисперсии мешающих факторов σ^2_η и σ^2_ξ, а в случае равенства нулю одной из них прямая зависимости совпадает с соответствующей прямой регрессии.

Изложенная интерпретация наглядно показывает влияние изучаемой совокупности и ее параметров на оценки значений переменных. Закон регрессии проявляется в том, что средневзвешенная оценка (являющаяся оценкой по уравнению регрессии) оказывается более близкой к среднему значению, чем оценка, полученная только по уравнению функциональной зависимости. Она всегда располагается между средним значением и значением, полученным по этому уравнению. Объяснение заключается в том, что мы придаем определенный вес сведениям о среднем значении оцениваемой величины

---

[1] **Доказательство.** Дисперсия оценки средневзвешенного значения

$$\sigma^2 = \frac{1}{1/\sigma^2_y + 1/\sigma^2_\eta b^2} = \frac{\sigma^2_y \sigma^2_\eta b^2}{\sigma^2_\eta b^2 + \sigma^2_y} = \frac{\sigma^2_\eta \sigma^2_y}{\sigma^2_\eta + \sigma^2_y / b^2}. \qquad (9.1.4)$$

Из равенства $\sigma^2_{\text{ост}.y}/\sigma^2_\eta = \sigma^2_y/\sigma^2_x$, откуда $\sigma^2_{\text{ост}.y} = \dfrac{\sigma^2_\eta \sigma^2_y}{\sigma^2_x}$, а также из равенства $\sigma^2_y = b^2(\sigma^2_x + \sigma^2_\eta)$, откуда $\sigma^2_x = \dfrac{1}{b^2}\sigma^2_y + \sigma^2_\eta$, получаем

$$\sigma^2_{\text{ост}.y} = \frac{\sigma^2_\eta \sigma^2_y}{\sigma^2_x} = \frac{\sigma^2_\eta \sigma^2_y}{\sigma^2_\eta + \sigma^2_y / b^2}, \qquad (9.1.5)$$

т. е. дисперсия средневзвешенного (9.1.4) действительно равна остаточной дисперсии величины y. Для доказательства, что OC является прямой регрессии, определим ее угловой коэффициент $b = \dfrac{AC + CB}{OB}$, $b_{yx} = \dfrac{CB}{OB}$, $\dfrac{b}{b_{yx}} = \dfrac{AC + CB}{CB} = \dfrac{\sigma^2_\eta b^2 + \sigma^2_y}{\sigma^2_y}$, или $b_{yx} = \dfrac{b\sigma^2_y}{\sigma^2_\eta b^2 + \sigma^2_y}$. Он оказался равным угловому коэффициенту прямой регрессии, что и требовалось доказать.

в изучаемой совокупности. Среднее значение «подтягивает» к себе оценку, данную только на основании сведений о зависимости. Уравнения регрессии оказываются конструкциями, как бы специально приспособленными для наиболее эффективной «комплексной» оценки одной величины по всей имеющейся информации: о характере и тесноте связи исследуемых величин, о среднем значении и дисперсии оцениваемой величины в изучаемой совокупности. О последней мы судим по выборке, использованной для расчета уравнения регрессии. Оценки значений параметров изучаемой совокупности выступают перед получением значения в роли априорной информации.

Рассмотрим закон регрессии с позиций «как это получается». Разобьем изучаемую совокупность на узкие интервалы по значению x и будем для простоты считать, что все x, входящие в i-й интервал, равны x_i. Если имеет место прямая пропорциональная зависимость между x и y, то каждому x_i соответствует y_i. При наличии мешающего фактора, действующего на переменную x с дисперсией σ^2_η и $M(\eta)=0$, p-я часть значений попадает в соседние интервалы, т. е. будет иметь значения x_{i-1} и x_{i+1}. Аналогично в i-й интервал попадет p-я часть значений из $i-1$ и $i+1$-го интервалов. Среднее значение y для объектов, имеющих теперь значения $x=x_i$, не будет равно y_i. Действительно, в этом интервале теперь будет $(1-p)n_i$ объектов со значением $y=y_i$, $\frac{p}{2}n_{i-1}$ объектов со значением $y=y_{i-1}=y_i-\Delta y$ и столько же объектов со значениями $y=y_{i+1}=y_i+\Delta y$ (объектами, попавшими в i-й интервал из $i\pm2$, $i\pm3$ и т. д., пренебрегаем ввиду малой вероятности этого при достаточно больших интервалах). Тогда среднее значение для объектов с $x=x_i$ будет равно $y_i(1-p)n_i+(y_i-\Delta y)\frac{p}{2}n_{i-1}+(y_i+\Delta y)\times$ $\times\frac{p}{2}n_{i+1}=y_i+p\Delta y\Delta n$, где Δn — разность частот в соседних интервалах. Таким образом, среднее значение y для заданного интервала значений x в результате действия мешающего фактора на x изменится. А именно там, где плотность вероятности с увеличением номера интервала уменьшается (т. е. там, где $\Delta n<0$), среднее значение y уменьшится. В подавляющем большинстве практических случаев плотность вероятности уменьшается от средней части к краям распределения, так как почти всегда модальное значение Мо располагается в средней части. Поэтому для значений $x_i>x_{Mo}$ значение y_i будет уменьшаться, а для $x_i<x_{Mo}$ — увеличиваться по сравнению с оценкой по функциональной зависимости. Таким образом, закон регрессии y на x по сути дела является следствием совокупного действия мешающего фактора на x и уменьшения плотности вероятности от x_{Mo} в сторону бо́льших и меньших значений. Приведем несколько приближенных качественных схем для иллюстрации линий (не обязательно прямых) «регрессии» в тех случаях, когда плотность

распределения x изменяется необычным образом (рис. 27). На рис. 27, *а* показаны функциональная зависимость $y=f(x)$ — прямая OO' и кривая оптимальных оценок (кривая регрессии) $M'OM$ для совокупности, характеризующейся тем, что плотность распределения x является усеченным нормальным распределением,

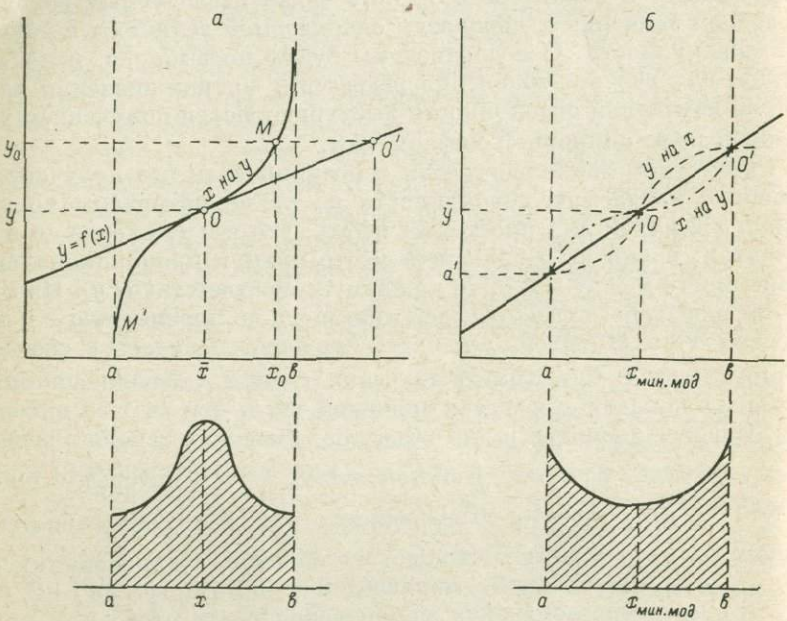


Рис. 27. Уравнения регрессии при необычных распределениях переменной.

Вверху — приближенная форма уравнения регрессии, внизу (*заштриховано*) — плотность вероятности переменной.

ограниченным крайними значениями a и b. Из рисунка видно, что как бы ни было велико значение y_0, x_0 не может превышать значение b. Кривая регрессии $M'OM$ асимптотически приближается к вертикалям $x=b$ и $x=a$.

Такая кривая регрессии не лишена практического смысла. При анализе минералов содержание компонентов не может превысить стехиометрического значения. Между тем вследствие погрешностей анализа содержание компонента, найденное по градуировочному графику, иногда соответствует содержанию $x > b$, что невозможно. Например, в нашей практике, при спектральном анализе богатых кремнеземом пород, фотометрическое сравнение интенсивности спектральных линий кремния с эталонными иногда приводило к оценке содержания SiO_2 в пробах, равному 110—115% и более! Единственный выход из этого положения — построение кривых регрессии по типу, изображенному на рис. 27, *а*.

На рис. 27, *б* изображена линия регрессии гипотетической совокупности, распределенной на интервале $a < x < b$ с заданными вероятностями. Представим себе руду, где в числе прочих есть два минерала — А и В, несущих элемент x: в первом $x=a$, во втором $x=b$. В процессе дробления и высвобождения минералов распределение содержаний элемента x в отдельных зернах будет приближаться к указанному. Каждое зерно представлено либо минералом А, либо минералом В, либо их сростками между собой и с прочими минералами, не содержащими x. При уменьшении количества сростков кривая распределения будет все более приобретать указанный на рисунке вид. Своей формой, особенно ветвями, асимптотически приближающимися к $x=a$ и $x=b$, кривая регрессии x на y будет напоминать соответственную линию регрессии на рис. 27, *а*. Существенное и принципиальное ее отличие от последней заключается в том, что на интервале OO' линия регрессии лежит н и ж е линии функциональной зависимости, т. е. оценка по ней подтягивается не к среднему, а к значениям a и b от точки с минимодальным значением x. Появляется как бы «регрессия наоборот».

Особенности рассмотренной интерпретации уравнений регрессии связаны с тем, что признаки x и y измеряются на элементах о п р е д е л е н н о й с о в о к у п н о с т и в том смысле, что существование этой совокупности необходимо. Часто можно видеть случаи применения корреляционного и регрессионного анализов в ситуации, когда такой совокупности нет и она не может даже подразумеваться. Например, в книге Дж. Юла и М. Кендэла (1960) говорится о линиях регрессии для связи тока и сопротивления в электрической цепи, т. е. в условиях такого физического эксперимента, в котором нет совокупности, а есть величина, назначаемая исследователем (например сопротивление), и величина измеряемая. В этих условиях до опыта априори нет совокупности электрических цепей с некоторым средним значением силы тока и некоторым средним сопротивлением. Роль априорной информации ($y=\bar{y}$) сводится к нулю. Естественно, что в этих условиях методы корреляционного и регрессионного анализов[2] для исследования зависимости не годятся. Если в корреляционном и регрессионном анализах мерой связи является коэффициент корреляции, а мерой неопределенности — нормированная остаточная дисперсия $s^2_{\text{ост}.x}/s^2_x$, то в условиях *контролируемого эксперимента* величины s^2_y и s^2_x зависят от решений экспериментатора, и коэффициент корреляции как мера связи теряет предметный смысл. Вместо него следует использовать ненормированный показатель — остаточную дисперсию оцениваемой переменной.

---

[2] В нашей терминологии — анализа формы связи в неконтролируемом эксперименте. В литературе эта ситуация иногда называется *конфлюэнтным* анализом.

§ 2. Соотношение регрессионного и тренд-анализа. Статистика градуировочных графиков

В связи с обсуждаемыми вопросами, в частности с различной интерпретацией результатов регрессионного анализа, самостоятельной проблемой встает выяснение роли изучаемой генеральной совокупности, или, шире — роли характера источника величин, между которыми исследуется зависимость. Изложенное выше убеждает нас в существенном влиянии оценок параметров изучаемой совокупности при регрессионной оценке — в том, что в уравнениях регрессии эти оценки уже используются.

В корреляционном и регрессионном анализах различают два типа ситуации, или два типа эксперимента, называемые Дж. Берксоном (Berkson, 1950) *неконтролируемым* и *контролируемым*. Неконтролируемый эксперимент заключается в изучении случайной выборки: для каждого ее элемента заранее неизвестно значение аргумента x, оно измеряется, как и значение y, после попадания элемента в выборку, и в этом смысле изучаемые характеристики «симметричны». Применение результатов регрессионного анализа также начинается с выбора элемента, например образца горной породы, из той же совокупности. Значение аргумента x в образце до его отбора неизвестно, и оно никак не влияет на выбор образца. Равным образом значение функции y также не оказывает влияния на вероятность взятия данного образца. Именно поэтому одна из двух оценок y равна среднему значению этой характеристики в совокупности. Все предыдущие результаты относились именно к такому (неконтролируемому) эксперименту.

В контролируемом эксперименте изучаемой совокупности элементов в статистическом смысле нет, а во многих случаях нет ее и в смысле физического существования. Экспериментатор по своей воле устанавливает значение аргумента, например напряжение в электрической цепи, и измеряет значение функции, например силу тока. Если элементы, на которых производятся измерения, физически существуют, то контролируемый эксперимент заключается в выборе элементов с заданным значением аргумента с последующим измерением на этом объекте значения функции. Так поступают при построении эталонировочных графиков, приготавливая предварительно образцы с заданными содержаниями x и измеряя аналитический сигнал от полученных эталонов. Совокупность изученных элементов (или совокупность пар измеренных значений x и y, если отдельные элементы физически не существуют) образуют апостериорную совокупность, которую нельзя назвать статистической выборкой. Значения аргумента в этой выборке встречаются с частотами, регулируемыми самим исследователем в пределах его возможностей. Здесь имеет смысл только эмпирическая функция распределения x, которая, однако, не является оценкой функции распределения какой-либо генеральной сово-

купности, и это приводит к совершенно иной интерпретации результатов. В геологии и геохимии проф. А. Б. Вистелиус называет такой эксперимент изучением поверхности отклика. В США (а затем и в других странах) распространилось название «анализ поверхности тренда» (trend surface analysis). Отличительная особенность тренд-анализа состоит в том, что аргументом является координата в пространстве и (или) времени, например глубина, расстояние от начала профиля, возраст. В экономике и социальных исследованиях такие совокупности называются динамическими рядами.

Основной принцип построения линий или поверхностей отклика тот же, что и в регрессионном анализе — метод наименьших квадратов. Если минимизируют отклонения функции y от аппроксимирующей линии, то она будет полной формальной аналогией прямой регрессии y на x. Рассмотрим отличия в использовании и интерпретации обычной линии регрессии (неконтролируемый эксперимент) и линейного тренда (контролируемый эксперимент). Значение аргумента — контролируемая переменная — устанавливается волей исследователя, и это делается обычно с точностью, значительно превышающей точность определения функции. Например, при изготовлении эталонных проб для построения эталонировочного графика в них вводится очень точная навеска исследуемого вещества. При построении поверхности тренда аргументом является координата, отмечаемая на графике с высокой точностью. Это значительно уменьшало бы отклонение линии регрессии на контролируемую переменную от функциональной прямой. Однако указанного отклонения все равно не возникнет по причине отсутствия генеральной совокупности и изменения плотности распределения величины x. Построенная прямая — уравнение регрессии на контролируемую переменную — совпадает с функциональной зависимостью, является ее оценкой. Если мы попытаемся построить линию регрессии контролируемой переменной на неконтролируемую (x на y), то полученное уравнение не будет оценкой функциональной зависимости. Итак, информация о том, что одна из переменных — контролируемая, достаточна для того, чтобы однозначно отождествить линию функциональной зависимости с одной из прямых регрессии.

Поверхность тренда используется как в эвристических целях, так и в оценочных (например, в горном деле для оценки содержаний и запасов в участках с заданными координатами), т. е. поверхность тренда является «оценочным графиком». Таковыми и являются градуировочные графики в химико-аналитической практике. Содержание в эталонах является контролируемой переменной, график строится путем минимизации отклонений аналитического сигнала (рис. 28), т. е. как уравнение регрессии аналитического сигнала на содержание, а используется противоположным

образом — для оценки содержания по аналитическому сигналу. В этом состоит принципиальное отличие градуировочных (эталонировочных) графиков и уравнений регрессии. Можно коротко сказать, что «уравнения регрессии используют так же, как и строят», т. е. если построено уравнение регрессии y на x, то его используют для оценки значений y по значениям переменной x, а не наоборот. Поэтому в регрессионном анализе описанное выше «перевернутое» использование градуировочного графика было бы грубо ошибочным.

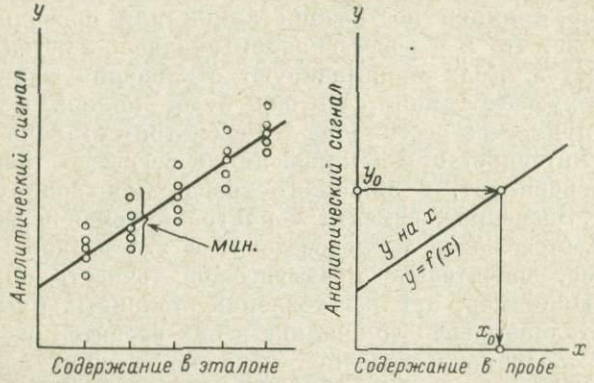


Рис. 28. Построение (*слева*) и использование градуировочного графика (*справа*).

Применимость градуировочного графика в указанном смысле объясняется только тем, что регрессия «аналитический сигнал — содержание» является одновременно оценкой функциональной зависимости, так что мы фактически используем в качестве графика линию функциональной зависимости. Только поэтому такие графики годятся для изучения различных совокупностей (отличающихся средними значениями содержаний x и их распределением: ведь данная, изучаемая, совокупность никак не участвовала в построении графика).

Оценка по уравнению функциональной зависимости (например, по эталонировочному графику) не всегда является «наилучшей». Она эффективнее других только тогда, когда о параметрах изучаемой совокупности, например о среднем содержании и дисперсии в изучаемом типе пород, ничего не известно. Однако, если нам известны эти параметры, мы можем их использовать, конструируя «искусственное» уравнение регрессии. Возможность для этого имеется, так как известны прямая функциональной зависимости (эталонировочный график) и параметры изучаемой совокупности. Проведем прямую $x=x$ (рис. 29) и разделим отрезок AB в отношении $AC/CB=\sigma^2_{x_y}/\sigma^2_x$, где $\bar{x}$ — среднее содержание в изучаемой совокупности, σ^2_x — дисперсия содержания, $\sigma^2_{x_y}$ — дисперсия оценки по эта-

лонировочному графику; $\sigma^2_{x_y} = b^2\sigma^2_{\Delta y}$, где b — угловой коэффициент эталонировочного графика, $\sigma^2_{\Delta y}$ — остаточная дисперсия аналитического сигнала. Прямая OC есть созданное нами (таким необычным способом!) уравнение регрессии x на y. По одному и тому же аналитическому сигналу мы можем получить, таким образом, две оценки: оценку по эталонировочной прямой (x_0), которую аналитики всегда и выдают геохимику, и оценку, исправленную с учетом априорных сведений об изучаемом материале (x'_0). При переходе к другой генеральной совокупности уравнение регрессии изменится и будет проходить, например, через точку O'. Создается впечатление, что мы как бы «подгоняем» результат. Однако это является лишь разумным использованием дополнительной информации. Мы сознательно конструируем такую средневзвешенную оценку, какая сама собой получилась бы при регрессионном анализе. Оправданность таких действий можно проиллюстрировать двумя способами.

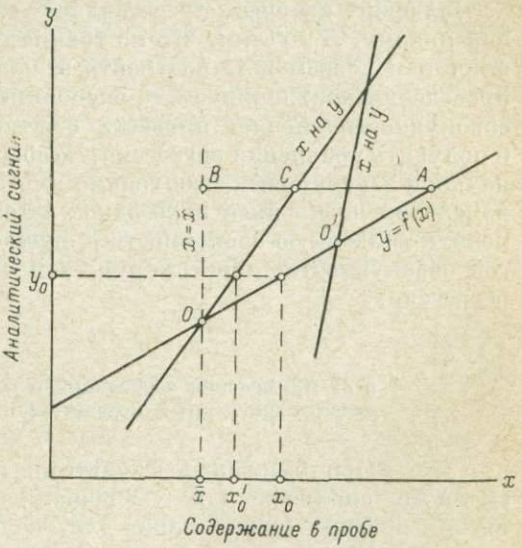


Рис. 29. Использование градуировочного графика и уравнений регрессии при переходе к анализам проб другой совокупности. Пояснения в тексте.

Если бы эталонировочный график был построен не на эталонных образцах, а по способу неконтролируемого эксперимента на материалах данной совокупности, то мы пришли бы именно к этому уравнению регрессии. Допустим, при анализе однородной совокупности с очень устойчивым содержанием (например, сплавов железа с его содержанием 90 %) получен результат (по градуировочному графику) 80 %. Переоценка содержания по уравнению регрессии, вероятно, значительно приблизило бы ее к 90 %.

Априорная информация меняет взгляд на значение наиболее эффективной оценки. Величина этой «корректировки» зависит от соотношения точности градуировочного графика и параметров распределения изучаемой совокупности: чем точнее график, тем меньше корректировка, чем менее изменчив изучаемый материал, тем ценнее сведения о среднем содержании этой совокупности.

Нам не известны случаи, когда бы описанная процедура использовалась в аналитической практике. Инициатива в этом должна принадлежать геологу и геохимику, которые всегда имеют какие-то предварительные сведения о направленном на анализ материале.

Большой интерес представляет случай, противоположный описанному. Допустим, что по техническим или принципиальным причинам невозможно построить эталонировочный график путем проведения контролируемого эксперимента. Тогда на подходящей совокупности следует провести обычный регрессионный анализ и получить уравнения регрессии. Чтобы перейти к оценке функциональной зависимости, необходимо иметь сведения о погрешностях измерения по крайней мере одной из величин. Затем можно применить найденную зависимость в оценочных целях на любой другой совокупности, конструируя, как описано, другие уравнения регрессии.

§ 3. Применение корреляций и регрессий при оценке средних значений и подсчете запасов

Использование результатов корреляционного и регрессионного анализов для увеличения эффективности (точности) оценок средних представляет для геологии и геохимии большой интерес. Среди опубликованной литературы каких-либо разработок по этому вопросу авторами не отмечалось. Здесь он будет рассмотрен на примере подсчета запасов, но полученные результаты аналогичным образом можно использовать при оценке средних содержаний количества элементов (компонентов) в массивах и толщах горных пород.

Имеющуюся информацию для подсчета запасов можно условно подразделить на прямую (данные о мощностях, содержаниях, объемных весах, границах рудного тела и других параметрах, входящих в подсчет запасов) и косвенную. Косвенная информация представляет собой закономерности, связывающие значения перечисленных параметров между собой и с другими наблюдаемыми при изучении месторождений признаками в виде корреляционных зависимостей. До настоящего времени этот последний вид информации не используется. Основные пути ее использования, приводящие к существенному повышению точности оценки запасов, будут рассмотрены на примере подсчета запасов на месторождениях, представленных пластами, линзообразными и пластообразными залежами, жилами и жильными зонами, штокверками и полями продуктивных жил. Значения признаков в соседних точках будем считать независимыми.

Простейшим случаем, на котором можно проанализировать общий принцип использования косвенной информации, является

подсчет запасов на месторождении, представленном пластом, пластообразной залежью или жилой, разведанной горными выработками и скважинами (Ткачев, 1973а). Предположим, что содержание в керновых пробах из скважин c^* определяется с большой погрешностью, характеризуемой дисперсией s_1^2. Этот случай довольно обычен из-за небольшого выхода керна, мало представительного веса получающейся из него пробы, избирательного истирания керна, приводящего к появлению систематических погрешностей, учет которых с помощью поправочных коэффициентов не приводит к уточнению, так как они сами определены с большими погрешностями.

Опробование по горной выработке, напротив, характеризуется погрешностями, которыми можно пренебречь по сравнению с предыдущими. Пусть при обработке данных опробования горных выработок установлена корреляционная зависимость между мощностью m и содержанием c:

$$c = f(m) \pm \xi_2, \tag{9.3.1}$$

где ξ_2 — случайная погрешность оценки содержания по мощности с помощью уравнения регрессии, характеризуемая стандартным отклонением так называемых остаточных величин s_2^2. По каждой разведочной скважине для оценки содержания имеется, таким образом, два источника информации: содержание, определяемое непосредственно по керну, и содержание, оцененное по формуле (9. 3. 1) на основании корреляционной зависимости между m и c. Оптимальная оценка содержания должна определяться в этом случае по формуле средневзвешенного

$$\tilde{c} = \frac{c^*/s_1^2 + f(m)/s_2^2}{1/s_1^2 + 1/s_2^2}, \tag{9.3.2}$$

где весовыми коэффициентами являются величины, обратные дисперсиям: $1/s_1^2$ и $1/s_2^2$. В частном случае, когда содержание по керну не устанавливается, $\tilde{c} = f(m)$. В другом частном случае, когда корреляция между m и c отсутствует или не обнаружена, $\tilde{c} = c^*$. Подсчет запасов на рассматриваемых месторождениях производится по формуле

$$Q = \overline{mc}\rho S, \tag{9.3.3}$$

где ρ — объемный вес руды; S — площадь рудного тела. Величина метропроцента $\overline{mc}$ должна определяться в этом случае по формуле

$$\overline{mc} = \frac{1}{n}\left(\sum_{i=1}^{k} m_i c_i + \sum_{i=k+1}^{n} m_i \tilde{c}_i\right), \tag{9.3.4}$$

где k равно числу точек, в которых m и c определены в горных выработках, т. е. с незначительной погрешностью, а n — общее

число точек; $\tilde{c}_i$ — оптимальная оценка содержания в i-й точке (скважине), определяемая по формуле (9. 3. 2).

Учитывая, что реально в большинстве случаев изучение зависимостей ограничивается линейными уравнениями, рассмотрим частный случай, когда содержание в точках $k+1$, $k+2$, . . . , $k+i$, . . . , n оценивается только по линейной корреляционной зависимости с мощностью. Тогда

$$\tilde{c}_i = a_i m_i + b, \quad a = r\frac{s_c}{s_m}, \quad b = \bar{c} - a\bar{m}, \qquad (9.3.5)$$

где a и b — параметры корреляционной зависимости, высчитанные по материалам опробования горных выработок. Подставляя $\tilde{c}_i$ в формулу (9. 3. 4) и проведя ряд преобразований, получим

$$(\overline{mc})_{nk} = \bar{m}_n \bar{c}_k + \mathrm{cov}_k \frac{s^2_{m,\,n\cdot}}{s^2_{m,\,k}}, \qquad (9.3.6)$$

где $(\overline{mc})_{nk}$ означает искомое нами среднее значение метропроцента, в подсчет которого входит n значений мощностей и k значений содержаний ($n > k$); m_n — среднее значение мощности, подсчитанное по всем данным; c_k — среднее значение содержания, полученное по k данным (из горных выработок); cov_k — ковариация мощности и содержания, подсчитанная по такому числу точек, для которых имеются значения обоих параметров, т. е. по данным опробования горных выработок (в дальнейшем такие данные будем называть парными). Множитель $s^2_{m,\,n\cdot}/s^2_{m,\,k}$ приближенно характеризует отношение дисперсий мощности, подсчитанных по всем n данным и по k данным из горных выработок. Точнее, $s^2_{m,\,n\cdot}$ представляет собой некоторую аналогию дисперсии — дисперсию по смешанному количеству данных

$$s^2_{m,\,n\cdot} = \overline{m^2_n} - \bar{m}_n \cdot \bar{m}_k, \qquad (9.3.7)$$

тогда как $s^2_{m,\,n} = \overline{m^2_n} - m^2_n = \overline{m^2_n} - \bar{m}_n\bar{m}_n$ (значения нижних индексов те же). Дисперсию, высчитанную по формуле (9. 3. 7), будем называть смешанной.

Расчеты по формуле (9. 3. 6) значительно проще соответствующих им расчетов среднего метропроцента по формуле (9. 3. 4). Они позволяют сразу находить среднее значение, не обращаясь к уравнению регрессии для оценки каждого значения содержания $\tilde{c}_i$ по значению мощности m_i, что особенно удобно, если таких значений много. Кроме того, и это не менее важно, равенство (9. 3. 6) позволяет наглядно видеть влияние на рассчитываемую величину разных источников информации.

Рассмотрим также общий случай, когда кроме «парных» данных, имеется дополнительно b значений мощностей и h значений содержаний. Тогда получится $b+k=l$ значений мощностей и $h+k=p$ значений содержаний и $(\overline{mc})_{lp} = \bar{m}_l \bar{c}_p + \mathrm{cov}_k \frac{s^2_{m,\,l\cdot} s^2_{c,\,p\cdot}}{s^2_{m,\,k} s^2_{c,\,k}}$,

где $s^2_{m,k}$ и $s^2_{c,k}$ — дисперсии соответственно мощности и содержания по k данным; $s^2_{m,l}$, $s^2_{c,p}$ — смешанные дисперсии тех же величин, вычисляемые по формулам типа $s^2_{m,l} = \overline{m_l^2} - \bar{m}_l \cdot \bar{m}_k$. Естественно, что мощность и содержание по смыслу могут меняться местами.

Изложенные выше принципы можно использовать также и на месторождениях, тело полезного ископаемого которых представляет собой не сплошной пласт или жилу, а серию прожилков (жильную зону) или штокверк. Среднее значение метропроцента, необходимое для подсчета запасов по формуле (9. 3. 3), выражается следующим образом:

$$mc = \frac{1}{n}\sum_{l=1}^{n} m_l c_l, \qquad (9.3.8)$$

где m_l — мощность l-го сечения жильной зоны или штокверка, c_l — содержание полезного компонента в нем. При наличии n сечений с известным содержанием (по горным выработкам) из общего количества n формулу (9. 3. 8) можно привести к следующему виду:

$$(\overline{mc})_{nk} = \bar{m}_n \bar{c}_k + \mathrm{cov}_k \frac{s^2_{m,n}}{s^2_{m,k}}. \qquad (9.3.9)$$

В этой формуле cov_k представляет собой ковариацию «спрессованной» мощности (m) прожилков в данном пересечении и среднего содержания c в этих прожилках, высчитанную по k парным данным. $s^2_{m,n}$ является смешанной дисперсией спрессованной мощности прожилков по n и k сечениям, $s^2_{m,k}$ — дисперсией той же величины по k сечениям. При таком способе расчета среднего значения метропроцента используется информация по спрессованной мощности рудного тела как в сечениях из выработок, так и из скважин, а также по содержанию в сечениях из горных выработок и корреляционной зависимости между этими величинами. Информация о мощностях отдельных прожилков в сечениях из скважин не используется. Не используется, следовательно, и информация, заключенная в корреляционной зависимости между мощностью отдельных прожилков и содержанием в них полезного компонента. Если таковая изучена, то для оценки содержания (и метропроцента) по сечениям из скважин имеются уже три источника информации: определения содержаний по керновым пробам, оценка его по первому уравнению регрессии (спрессованная мощность прожилков—содержание) и оценка по второму уравнению регрессии (мощность отдельных прожилков—содержание). Основой суммарной оценки может служить формула (9. 3. 2) соответственно с тремя слагаемыми. Поскольку нашей целью является разработка способов использования косвенной информации, мы будем рассматривать лишь два последних ее источника. В резуль-

тате использования первого источника информации оценка метропроцента дается по формуле (9. 3. 9). Оценка на основании второго источника информации производится аналогичным образом с той разницей, что величины, обозначенные звездочкой, и их нижние индексы относятся не к числу сечений, а к числу подсеченных прожилков (n^* — общее число подсеченных скважинами и опробованных в горных выработках прожилков, k^* — их число, опробованное в горных выработках). Пусть в результате получена оценка

$$(\overline{mc})^*_{n^*k^*} = \bar{m}^*_{n^*}\bar{c}^*_{k^*} + \mathrm{cov}^*_{k^*}\frac{s^2_{m^*n^*}}{s^2_{m^*k^*}}. \qquad (9.3.9а)$$

Поскольку эта оценка относится к отдельным прожилкам, а не к сечениям, она не сопоставима с оценкой среднего метропроцента по сечениям. Для определения запасов обе эти оценки необходимо (не учитывая объемного веса) умножить на площади S. Площадь для второй оценки S_2 является площадью рудного тела (ее горизонтальной проекцией, если используется вертикальная мощность). Площадь для первой оценки S_1 является площадью всех прожилков. Она во столько раз больше рудного тела, во сколько раз число пересечений прожилков больше числа сечений: $S_1/S_2 = n^*/n$. Таким образом, если величина $(\overline{mc})_{nk}$ употребляется с коэффициентом 1, то $(\overline{mc})^*_{n^*k^*}$ должна складываться с ней после умножения на коэффициент n^*/n. Кроме того, при нахождении результирующего среднего значения метропроцента эти величины должны взвешиваться на дисперсии оценок. При их определении будем исходить из того, что, во-первых, оценки метропроцентов по формулам (9. 3. 9) и (9. 3. 9а) эквивалентны оценкам, вычисленным по формуле (9. 3. 4); во-вторых, дисперсией оценки мощности можно пренебречь; в-третьих, содержание по горным выработкам определено с ошибкой, которой также можно пренебречь. Тогда

$$s^2_{(\overline{mc})_{nk}} = \frac{n-k}{n^2}\bar{m}^2 s^2_{\bar{c}}. \qquad (9.3.10)$$

Аналогично

$$s^2_{(\overline{mc})^*_{n^*k^*}} = \frac{n^*-k^*}{n^{2*}}\bar{m}^{2*}s^2_{\bar{c}^*}. \qquad (9.3.11)$$

Учитывая, что отношение числа опробованных в горных выработках прожилков к их общему числу в среднем равно отношению числа сечений по горным выработкам к их общему числу, $\frac{n-k}{n} \approx \approx \frac{n^*-k^*}{n^*}$, в качестве весовых коэффициентов при расчете общего среднего значения метропроцента можно использовать величины $\frac{m^2}{n}s^2_{\bar{c}}$ и $\frac{m^{2*}}{n^*}s^2_{\bar{c}^*}$, где $s^2_{\bar{c}}$ и $s^2_{\bar{c}^*}$ — средние дисперсии оценки содержа-

ния по мощности соответственно в сечении и прожилке (остаточные дисперсии); $\bar{m}$ и $\bar{m}^*$ — средние мощности соответственно спрессованных прожилков в сечении и отдельных прожилков. Поскольку в среднем $\bar{m}n \approx \bar{m}^* n^*$ (и равно общей подсеченной мощности рудного тела) и учитывая, что оценка $(\overline{mc})^*$ при подсчете общего среднего значения метропроцента средневзвешенным способом должна входить в расчет с множителем n^*/n, окончательно можно написать

$$\widetilde{\overline{mc}} = \frac{\overline{mc} \cdot n/\bar{m}^2 s_{\bar{c}}^2 + \overline{mc}^* n^*/\bar{m}^{2*} s_{\bar{c}^*}^2}{n/\bar{m}^2 s_{\bar{c}}^2 + n^*/\bar{m}^{2*} s_{\bar{c}^*}^2}. \qquad (9.3.12)$$

Из этой формулы видно, что влияние отдельных оценок метропроцента (по корреляции со спрессованной мощностью жильной зоны и с мощностью отдельных прожилков) на суммарную оценку $\widetilde{\overline{mc}}$ определяется дисперсиями оценок содержаний: а) по спрессованной мощности прожилков в сечении, б) по мощности отдельных прожилков. Большее влияние оказывает та оценка, которая основывается на более тесной корреляционной зависимости.

Косвенную информацию двух источников при подсчете запасов на штокверковых месторождениях большой мощности (большой протяженности на глубину) удобнее использовать после разделения рудного тела на горизонтальные слои определенной мощности. При этом источником информации будет служить корреляция между мощностью отдельных прожилков и содержанием в них полезного компонента и между спрессованной мощностью прожилков в пределах выделенного слоя и средним содержанием компонента в прожилках этого слоя.

Изложенные результаты применимы также к месторождениям, представленным полями продуктивных жил. Роль прожилков играют здесь отдельные жилы. В отличие от прожилков последние представляют собой более или менее изолированные тела, поэтому, как показал анализ, тождественные результаты можно получить путем использования корреляции между объемом тел и содержанием (или количеством полезного ископаемого) в них с учетом вероятностей подсечения жил.

Широкое использование косвенной информации при подсчете запасов позволит существенно повысить его точность при тех же объемах геологоразведочных работ. В геологии и геохимии использование ее позволит получить более объективные выводы при том же объеме каменного и аналитического материала.

§ 4. Оценочный и эвристический аспекты корреляционного анализа. Особенности геохимической интерпретации корреляций и регрессий

При интерпретации результатов корреляционного анализа мы различаем две ситуации, такие же как и при аппроксимации эмпирических распределений (см. гл. 3).

В первом случае корреляционный анализ используется как инструмент оценки — для количественного выражения силы линейной связи между x и y (собственно корреляционный анализ) или для предсказания значений y по известным x (регрессионный анализ). Природа связей между признаками здесь несущественна. Переменные x и y могут не стоять между собой в непосредственных причинно-следственных отношениях. Предполагается, что исследуемая зависимость действительно линейна. Если это не так, то оценка тесноты связи становится слишком грубой, а уравнения регрессии дают большую погрешность оценки. К сожалению, линеаризация путем замены переменных (например, $y=x^2$, $y=\log x$) далеко не всегда оказывается достаточной. Кроме того, «преобразование эмпирической формулы нарушает принцип наименьших квадратов» (Румшиский, 1971, стр. 82). Короче, в этих случаях лучше перейти к использованию криволинейных зависимостей.

Желательно, чтобы обе коррелируемые переменные имели нормальное или не сильно от него уклоняющееся распределение. Именно по этой причине мы рекомендуем пользоваться «баллами» (см. гл. 2) при расчете корреляций элементов-примесей: распределение баллов как линейной функции логарифмов обычно не противоречит нормальному закону.

Пусть, например, исследователь выявил отрицательную связь между проницаемостью известняков и содержанием в них титана. Ни одна из переменных в данном случае не относится к другой как причина или следствие, но обе зависят от третьей характеристики: содержания нерастворимого остатка (н. о). Чем выше в породе содержание глинистого н. о., тем выше содержание титана и тем ниже проницаемость. Это отнюдь не означает, что исследователь не имеет права пользоваться содержанием титана для предсказания проницаемости. Заметим, что подобные связи обычно характеризуются — именно вследствие своей опосредствованности — низкими значениями коэффициентов корреляции.

Во втором случае исследователь не ограничивается «потреблением» численного результата, а связывает с ним определенную модель. Эта модель может быть детерминированной или стохастической. В первой случайный характер отклонений эмпирических результатов от следствий модели объясняют влиянием неучтенных факторов, относимых к разряду случайных. Во второй случайный характер отклонений является свойством модели. Такие модели более содержательно отражают существо природных явлений, но хороших примеров их построения пока не так много. Детерминированные модели просты и более распространены.

Хороший пример представляет изучение корреляционных связей стронция в фосфоритах. Как показал В. З. Блисковский (1969), в ряде случаев содержание стронция в фосфоритах жестко коррелирует с содержанием фосфата кальция; коэффициенты кор-

реляции по выборкам 20—40 проб в координатах «P_2O_5, % —Sr, г/т» достигают +0.99. В других случаях связи нет или она отрицательна. Для интерпретации связи автор предложил модель изоморфного замещения кальция на стронций в структуре фосфата. Это замещение происходит при формировании фосфоритов, т. е. еще в седиментогенезе, и зависит, по мнению автора, только от одного параметра — отношения Ca/Sr в воде. Следовательно, если среда седиментации содержала мало стронция, то образующийся фосфат будет резко обеднен этим элементом; в этом случае нефосфатные компоненты фосфоритов могут содержать стронция больше, чем сам фосфат, и поэтому в тех же координатах связь может измениться на обратную (увеличение содержания фосфата приводит к разбавлению валового стронция в фосфорите «пустым» фосфатом). Если же в одну выборку попадут образцы фосфоритов, сформированные при разных исходных значениях Ca/Sr в воде, то связи в названных координатах может вообще не оказаться. Как выяснилось, обе последние модели точно отвечают вторичным фосфоритам, сформированным при переотложении первичных фосфоритов в коре выветривания. При этом отсутствие связи стронция и фосфата означает, что переотложению подверглась только часть первичных фосфоритов (смешанная выборка), а в случае обратной связи фосфориты были, по-видимому, полностью переотложены: фосфатное вещество вновь осаждалось из пресных вод, крайне бедных стронцием. Таким образом, исчезновение положительной корреляции P_2O_5—Sr в фосфоритах или даже изменение знака этой корреляции на обратный может служить диагностическим признаком вторичных фосфоритов.

Широко известным примером, в котором модель корреляции построена заранее, является применение так называемого битумоидного коэффициента β $\left(\text{обычно в форме } \frac{\text{ХБ}}{C_{орг}} \cdot 100\%\right)$, где ХБ — содержание в породе хлороформенного битумоида, $C_{орг}$ — содержание органического углерода. Физический смысл β — степень битуминозности органического вещества, т. е. доля битумных компонентов в органическом веществе породы. Многочисленными исследованиями установлена зависимость между величиной β и валовым содержанием органического вещества, так называемая закономерность Успенского—Вассоевича: по мере убывания содержания органического вещества в породе битуминозность его возрастает. В координатах $C_{орг}$—β наблюдается зависимость, хорошо аппроксимируемая уравнением гиперболы $\beta = \frac{a}{(C_{орг})^n} + b$, где a, b, n — константы. Хотя закономерность Успенского—Вассоевича и не получила однозначного толкования, все исследователи сходятся на том, что она — продукт генетической связи битума с органическим веществом. Напрашивается мысль (Вассоевич, 1958) использовать отсутствие данной закономер-

ности в целях диагностики вторичных битумов, содержание которых не контролируется содержанием $C_{орг}$ в породе. Успехи геохимии органического вещества последних лет, приведшие к разработке осадочно-миграционной теории происхождения нефти, в значительной мере обязаны использованию величины β в указанном аспекте.

Эффективность «модельного» подхода к использованию корреляционного анализа хорошо иллюстрируется исследованиями В. Ф. Мягкова (1969), построившего несложные теоретические модели поведения рудных компонентов с учетом механизма процесса. Так, если в рудном теле имеется по две генерации коррелируемых компонентов, то последние должны состоять в криволинейной зависимости (в идеальном случае — парабола второго порядка). Если два компонента образуют общий парагенезис только на первой стадии, а на второй стадии в структуру рудоотложения поступал только один из них, метасоматически замещая продукты первой генерации, то зависимость описывается выпуклой параболой, а в случае наступления общего парагенезиса лишь на второй стадии (на первой имелся только один компонент) — вогнутой параболой. По мнению В. Ф. Мягкова, в пределах одного парагенезиса (генерации) зависимости, по-видимому, всегда линейны.

Любопытна интерпретация результатов корреляционного анализа В. А. Кутолиным (1972). Изучая связи между породообразующими окислами в различных формациях базальтов, он установил наличие отрицательной корреляции между FeO и Fe_2O_3 в эффузивных базальтах и ее отсутствие в габброидах, что «. . . наводит на мысль о связи между характером корреляции между закисью и окисью железа и фациальной принадлежностью пород» (стр. 118—119). Дальнейшие исследования подтвердили, что при переходе от эффузивных базальтов к гипабиссальным долеритам и далее к мезоабиссальным габброидам наблюдается постепенное ослабление отрицательной корреляционной связи между закисью и окисью железа. Изучение экспериментальных плавок показало, что в быстро остывающих стеклах, в которых не успевает установиться равновесие между FeO и Fe_2O_3, эти окислы находятся в отрицательной корреляции, в выдержанных расплавах корреляция Fe_2O_3—FeO отсутствует. Автор приходит к окончательному генетическому заключению: «В пределах глубинных магматических очагов в расплаве устанавливается равновесие между двух- и трехвалентным железом, статистическим отражением которого является отсутствие корреляции между ними, характерное для габброидов. В процессе движения магмы в верхние структурные ярусы и ее излияния на поверхность установившееся равновесие нарушается, происходит переход части закисного железа в окисное, в результате чего между обеими формами железа возникает сильная отрицательная корреляция, характерная для базальтов» (стр. 119).

Глава 10

НЕКОТОРЫЕ ВОПРОСЫ МНОГОМЕРНОГО АНАЛИЗА

Здесь мы рассмотрим вопросы, хотя и различные по содержанию, но объединенные «многомерностью», означающей, что статистические выводы делаются на основании совместной обработки данных по многим признакам. Методы многомерного анализа уже широко применяются в геохимии, но неиспользованных возможностей и нерешенных проблем здесь больше, чем в других методах.

§ 1. О некоторых критериях сравнения многомерных средних

Р. Миллер и Дж. Кан (1965, стр. 250—278) описывают ряд критериев сравнения многомерных средних: критерий T^2 Хотелинга, так называемое обобщенное расстояние D^2 Махалонобиса, коэффициент расового сходства Пирсона CRL. Недавно Д. А. Родионов (1968) предложил критерий $V(r^2)$, а Д. Шоу (1968) — дистанционный коэффициент d. Все эти критерии как обязательный элемент включают разность средних по каждому признаку $\bar{x}_j - \bar{x}_k$, или

$$\sum_{i=1}^{m} (\bar{x}_{ij} - \bar{x}_{ik})^2, \qquad (10.1.1)$$

где m — число признаков, i — номер признака, j, k — номера объектов.

Рассмотрим более подробно предложенный Д. А. Родионовым и получивший у нас довольно широкое применение критерий $V(r^2)$, который, как легко убедиться, идентичен критерию расового сходства Пирсона (Ткачев, 1972). Согласно Д. А. Родионову (1968, стр. 39), критерием для проверки гипотезы об однородности

геологических объектов по m признакам служит функция

$$V(r^2)=\sum_{j=1}^{m}\frac{(\bar{x}_j^{(1)}-\bar{x}_j^{(2)})^2\, n_1 n_2}{s_j^2\,(n_1+n_2)}, \qquad (10.1.2)$$

которая, по замыслу автора, в случае справедливости нулевой гипотезы является суммой квадратов m нормально распределенных величин с нулевым математическим ожиданием и единичной дисперсией. В силу этого значения функции для каждого элемента r^2 разбиения геологического объекта $V(r^2)$ должны быть распределены как χ^2 с m степенями свободы. Необходимо обратить внимание читателей на те предпосылки, выполнение которых необходимо при использовании разработанной Д. А. Родионовым методики.

1. Значения всех m признаков в изучаемой совокупности должны быть *независимыми* друг от друга. Д. А. Родионов указывает, что «очень часто гипотеза о равенстве нулю коэффициента корреляции не отвергается для „локальных" совокупностей, несмотря на то, что в совокупности, объединяющей „локальные" участки, корреляционная зависимость бывает сильной. Это обстоятельство и позволяет ввести предположение о независимости случайных величин ξ_{t_1}, ξ_{t_2}, ... ξ_{t_j} ... ξ_{t_m}» (стр. 35). Под «локальной» совокупностью понимается участок, охарактеризованный t-й пробой. Это может быть отдельный прослой, разновидность породы или просто незначительный объем изучаемого участка. Однако из существа критерия ясно, что независимость должна выполняться для следующих пар величин: $\xi_{1,i}$ и $\xi_{1,j}$, ξ_{2i} и ξ_{2j}, ... ξ_{ti} и ξ_{tj} ... ξ_{ni} и ξ_{nj} при $i,\ j=1,\ 2\ldots,\ m\ (i\neq j)$, т. е. для любых пар признаков во всей изучаемой совокупности. Только в этом случае будут независимыми слагаемые в формуле (10.1.2). Эта предпосылка ограничивает применимость указанного критерия, так как в значительном числе случаев между изучаемыми признаками наблюдаются корреляционные зависимости. Это относится к содержанию химических элементов, распространенности палеонтологических форм, минералогическому составу пород и т. д. Таким образом, процедура проверки гипотезы однородности должна начинаться с проверки гипотезы об отсутствии корреляции и исключении одного из коррелирующих друг с другом попарно признаков. В пределе такая процедура может привести к исключению из рассмотрения $(m-1)$ признаков, что является, конечно, внушительной потерей информации.

2. Нормированная разность двух средних значений, квадрат которой является отдельным слагаемым в формуле (10.1.2), должна иметь нормальное распределение. Это возможно в двух случаях: а) значения признаков в совокупности распределены нормально, б) условие (а) не выполняется, но рассматриваются только такие

разбиения, число элементов в меньшем классе которых достаточно велико для того, чтобы распределение среднего значения по этому классу $\bar{x}^{(1)}$ приближалось к нормальному. В книге Д. А. Родионова условие нормальности распределения в совокупности заменяется условием нормальности распределения в пределах небольшого участка t изучаемого объекта. Это условие также ограничивает применимость критерия. Особенно это относится к разбиениям, в которых один из классов представлен одним или несколькими элементами. Наша попытка разделить линейно упорядоченное множество, охарактеризованное величинами, которые не подчиняются нормальному закону, привела к тому, что значения $V(r^2)$ закономерно увеличивались с уменьшением числа элементов в меньшем из классов. Кривая значений $V(r^2)$ в зависимости от номера элемента, по которому производится разбиение линейно упорядоченного множества, оказалась почти симметричной с минимумом при разбиении на два равных класса. Причина этого явления заключается в том, что распределение разности средних, одно из которых подсчитано по единичному числу значений, не подчиняющихся нормальному закону, асимметрично, и вероятности больших значений значительно выше, чем при нормальном распределении.

3. Строго говоря, распределение рассматриваемой разности двух средних величин никогда не будет нормальным, так как эти средние величины характеризуют две безвозвратные выборки, составляющие одну ограниченную генеральную совокупность, т. е. не являются величинами независимыми. С увеличением среднего значения по одному классу среднее значение по другому классу уменьшается. Эта зависимость оказывает влияние не только на характер распределения, но и на величину дисперсии разности. Она определяется следующим образом:

$$\mathbf{M}(\bar{x}_j^{(1)} - \bar{x}_j^{(2)})^2 = \mathbf{M}\left(\bar{x}_j^{(1)} - \frac{n\bar{x}_j - n_1\bar{x}_j^{(1)}}{n^2}\right)^2,$$

так как $n_i \cdot \bar{x}_j^{(1)} + n_2\bar{x}_j^{(2)} = n\bar{x}_j$, где $n = n_1 + n_2$. После небольших преобразований получаем $\mathbf{M}(\bar{x}_j^{(1)} - \bar{x}_j^{(2)}) = n^2/n_2^2\mathbf{M}(\bar{x}_j^{(1)} - \bar{x}_j)^2$, откуда с учетом того, что выборка является безвозвратной, получим

$$\frac{n^2}{n_2^2}\mathbf{M}(\bar{x}_j^{(1)} - \bar{x}_j)^2 = \frac{n^2}{n_2^2}s_j^2\left(\frac{1}{n_1} - \frac{1}{n}\right) = \frac{s_j^2 n}{n_1 n_2}.$$

Данное выражение в точности совпадает с дисперсией суммы (разности) двух независимых величин, однако об этом важном обстоятельстве в книге не упомянуто. Наличие упомянутой выше обратной зависимости между средними приводит к тому, что вероятности больших отклонений превышают таковые для нормального распределения с той же дисперсией (за счет соответствующего увеличения числа малых отклонений). По-

этому нулевая гипотеза, если она верна, будет отвергаться чаще, чем в $q\%$ случаев.

4. Применение данного критерия для проверки гипотезы об однородности неупорядоченного пространства нельзя считать обоснованным. Действительно, в этом случае происходит перебор всех возможных разбиений и среди них неизбежно найдется $q\%$ таких, для которых значение $V(r^2)$ будет превышать $\chi^2_{m,q}$.

Если бы мы рассматривали одно *случайное* разбиение или разбиение, продиктованное информацией, которая выходит за рамки рассматриваемого набора значений признаков, то статистическую значимость такого разбиения можно было бы проверить с помощью предложенной методики. Когда речь идет о переборе всех возможных разбиений, эта методика неприменима по своему смыслу. Гипотеза об однородности совокупности, если последняя достаточно велика, при одном или нескольких разбиениях не будет принята, следовательно, она будет отвергнута для данной совокупности вообще. Надо сказать, что если разбиение при полном переборе производится на основании значимой разницы в средних значениях одного или нескольких признаков, то статистическая значимость такой операции уже не может быть проверена с помощью критерия, основанного на значениях этих же признаков. Если для разбиения используются m_1 признаков из числа m, то для проверки уже могут быть использованы только $m-m_1$ оставшихся признаков.

Дело в том что, производя перебор комбинаций признаков, мы тем самым накладываем на систему априорные ограничения, ибо комбинация уже заранее задана, предопределена экспериментатором. Число таких ограничений системы равно числу признаков. Следовательно, рассчитав критерий многомерного различия $V(r^2)$ после перебора всех комбинаций признаков, мы не имеем права проверять значимость этого критерия с помощью χ^2, ибо полученное значение его имеет число степеней свободы, равное нулю!

Таким образом, оценка значимости $V(r^2)$ вообще не может проводиться в терминах проверки статистических гипотез.

Перечисленные замечания в равной степени касаются и процедуры разбиения, основанной на последовательном объединении заведомо однородных участков, хотя чем крупнее каждый однородный участок и чем их меньше, тем обоснованней получится результат.

Таким образом, мы приходим к выводу, что предложенная Д. А. Родионовым методика может быть использована лишь для поисков каких-то границ в геологическом объекте по максимальному (или просто большому) значению $V(r^2)$. Понятие «большое значение» $V(r^2)$ устанавливается сравнением с соответствующими табличными значениями $\chi^2_{m,q}$. Однако это не оз-

начает, что разделение произведено на уровне q. Статистическую значимость разделения можно было приближенно проверить, если бы оно было произведено до расчета $V(r^2)$ на основании другой информации. Правомерна лишь такая постановка задачи: значимо ли отличаются части объекта по комплексу m_1 признаков, если объект разделен по признаку, не входящему в предыдущий комплекс. В частном случае выбранным для разделения критерием может быть максимальное или просто большое значение $V(r^2)$, высчитанное по комплексу m_2 других признаков. Если игнорировать совершенно необходимую здесь зависимость информации, по которой мы разделяем, от информации, по которой мы проверяем значимость разделения, мы придем к такой неправомерной постановке задачи: значимо ли отличаются части объекта по комплексу m_1 признаков, если объект значимо разделен (на уровне q) по комплексу этих же признаков! При разбиении линейно упорядоченного пространства условия уже приближаются к тем, которые необходимы для проверки статистической значимости разделения. В этой ситуации упорядоченность является той информацией, на которой основан выбор класса допустимых разбиений, содержащих (и это важно подчеркнуть) неизмеримо меньше разбиений, чем их общее число.

Итак, для практического использования критерия $V(r^2) \geqslant \chi^2_{m,q}$ следует иметь в виду, что правильная формулировка нулевой гипотезы следующая: геологический объект однороден и характеризуется нормальным распределением взаимно независимых признаков. Альтернативное заключение можно сформулировать в виде утверждений: 1) геологический объект неоднороден при условии, что его части, на которые он может быть разделен, характеризуются нормальным распределением признаков; кроме того, признаки взаимно независимы в пределах всего объекта; 2) распределение признаков в пределах разделяемых частей не подчиняется нормальному закону (точнее: распределение средних значений признаков хотя бы по одной из разделяемых частей не подчиняется нормальному закону); 3) признаки не являются независимыми в пределах всей совокупности.

Еще раз подчеркнем, что даже при справедливости нулевой гипотезы фактический уровень значимости будет отличаться от q вследствие взаимной зависимости средних, причем степень этого отличия оценить трудно.

§ 2. О математическом содержании метода многократной корреляции и условиях его применимости

Метод так называемой многократной корреляции (МК) был введен в геологическую литературу в 1967—1968 гг. Ю. К. Бурковым (1969, 1971, 1972 и др.). За прошедшее время

этот метод получил заметное распространение в научных учреждениях и производственных геологических организациях, оснащенных ЭВМ. Однако ни сам автор, ни его последователи не раскрыли математического содержания метода, не определили сферы его применения; остаются недостаточно ясными и принципы предметной интерпретации результатов. Таким образом, МК — характерное «дитя» века машинной обработки информации, когда стало легче получить численный результат, чем понять, что он означает. Здесь мы попытаемся дать анализ процедуры МК, ее математическую основу, свойства получаемых решений и их геохимическую интерпретацию.

В качестве меры связи между случайными переменными (например, содержаниями химических элементов X и Y) Ю. К. Бурков предлагает использовать наряду с обычными коэффициентами корреляции (КК) вида r_{xy} так называемые коэффициенты корреляции второго, третьего и т. д. порядков. Переменными для расчета КК второго порядка между X и Y служат значения парных КК каждого из этих элементов с остальными. Между этими двумя рядами величин, которые сами являются КК, обычным способом рассчитывается новый КК, который и будет коэффициентом корреляции второго порядка; КК третьего порядка получают из КК второго порядка аналогичным образом. Расчеты, как правило, производятся на ЭВМ по циклической программе с заменой матрицы коэффициентов корреляции i-го порядка коэффициентами $i+1$-го порядка.

Очевидно, что КК второго порядка отражают «косвенное» сходство между переменными, т. е. сходство, основанное на сравнении этих элементов не между собой, а с группой других элементов. Что эти два вида сходства не тождественны, видно хотя бы из зависимости косвенного сходства от состава остальной группы элементов — *базы сравнения*. Сходство, основанное на КК третьего и более высокого порядка, будет «еще более косвенным», хотя смысл этого сходства становится уже не совсем понятным.

Метод МК используется прежде всего для выделения ассоциаций, под которыми автор метода понимает «группу элементов, сходных по своему поведению». Роль и значение МК определяются Ю. К. Бурковым следующим образом (Бурков, 1968, 1971 и др.). Непосредственное применение корреляционного анализа для выделения ассоциаций элементов далеко не во всех случаях дает удовлетворительные результаты. На каждом последующем этапе многократной корреляции будут получаться статистические оценки, отражающие все более высокие уровни взаимоотношений элементов. В целом применение метода многократной корреляции позволяет выделять частные ассоциации, обусловленные связями первого порядка, объединяющие их более общие ассоциации 2-го порядка, еще более общие ассоциации вплоть до установления наиболее общих ведущих ассоциаций.

Такого рода качественные, не строгие представления, разумеется, не могут быть положены в основу точных методов, каковыми является корреляционный анализ, и нуждаются в более четких определениях.

Наиболее замечательным свойством КК порядков выше первого является то, что с увеличением порядка они преимущественно увеличиваются по абсолютной величине, пока не становятся равными $+1$ или -1 и далее остаются постоянными. Обычно это достигается уже на 6—7-м этапах (циклах) корреляции, реже — на 9—11-м. Большие коэффициенты исходной матрицы как бы задают тон и стремятся к ± 1 монотонно и быстро, малые (0.1—0.2) во многих случаях то увеличиваются, то уменьшаются, иногда несколько раз меняют знаки, но к 8—11-му циклам их значения тоже быстро приближаются к ± 1. Эти эмпирически установленные факты проявляются независимо от порядка и особенностей исходной корреляционной матрицы. Эмпирически установлено также, что матрица итоговых КК, состоящая из положительных и отрицательных единиц, не может быть устроена произвольным образом, и расположение знаков в ней строго закономерно. Если какой-либо элемент связан с заданными элементами x и y коэффициентами с одинаковыми знаками, то и все остальные элементы связаны с этими двумя элементами также КК одинаковых знаков. Из этого свойства вытекает также следующее: если какие-либо два элемента связаны с третьим зависимостью одинакового знака, то с любым из остальных они связаны зависимостью также одинаковых знаков. Поэтому в корреляционной матрице порядка $n \cdot n$ число независимых итоговых КК составляет лишь $n-1$. Достаточно знать одну строку (столбец) знаков, чтобы проставить в матрице остальные, пользуясь указанными ее свойствами. Путем простой перестановки столбцов и строк итоговую матрицу всегда можно привести к данному виду →

	Элементы									
	1	2	3	4	5	6	7	8	9	10
1	+	+	+	+	−	−	−	−	−	−
2	+	+	+	+	−	−	−	−	−	−
3	+	+	+	+	−	−	−	−	−	−
4	+	+	+	+	−	−	−	−	−	−
5	−	−	−	−	+	+	+	+	+	+
6	−	−	−	−	+	+	+	+	+	+
7	−	−	−	−	+	+	+	+	+	+
8	−	−	−	−	+	+	+	+	+	+
9	−	−	−	−	+	+	+	+	+	+
10	−	−	−	−	+	+	+	+	+	+

Как видно, в результате МК совокупность элементов распадается на две, и только на две, группы — ассоциации. Они составлены так, что внутри каждой группы элементы связаны положительно, элементы из разных групп — отрицательно. Провести такое подразделение непосредственно по матрице исходных коэффициентов в общем случае невозможно, так как значительная часть исходных КК имеют знаки, противоположные итоговым. Элементы, попавшие в одну ассоциацию по методу МК, могут иметь между собой отрицательные исходные (парные)

КК. Тем не менее основные черты такого подразделения намечаются и по исходной матрице: пары с самыми высокими положительными КК оказываются в одной группе, с отрицательными — в разных.

Истоки идеи МК можно усмотреть в методе сравнения корреляционных профилей, предложенном Р. Трайоном (цит. по: Миллер, Кан, 1965, стр. 304). Он заключается в том, что за меру сходства двух признаков принимают не парный КК между ними, а согласованность их корреляционных профилей. Под корреляционными профилями понимают графики, на которых по оси ординат откладывают парные КК, по оси абсцисс (в произвольном порядке) — наименования признаков. Рис. 30 наглядно показывает, что косвенное сходство в отличие от прямого определяется исключительно о д и н а к о в ы м н а п р а в л е н и е м в изменении КК сравниваемых элементов с третьими элементами, т. е. «симбатностью» линий, а не абсолютной близостью значений КК. Так, признаки А и С имеют меньшее прямое сходство, чем признаки А и Б, но значительно большее косвенное сходство.

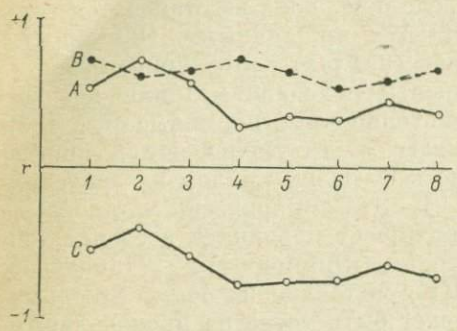


Рис. 30. Сравнение корреляционных профилей трех признаков.

Пусть значения изучаемых переменных (признаков) определяются состояниями (значениями) n *факторов* f_1, f_2, f_3, . . . , f_i, . . . , f_n. В линейном случае, которым мы ограничимся, будем иметь

$$\begin{aligned} x_1 &= a_{11}f_1 + a_{12}f_2 + a_{13}f_3 + \ldots + a_{1n}f_n, \\ x_2 &= a_{21}f_1 + a_{22}f_2 + a_{23}f_3 + \ldots + a_{2n}f_n, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ x_m &= a_{m1}f_1 + a_{m2}f_2 + a_{m3}f_3 + \ldots + a_{mn}f_n. \end{aligned} \qquad (10.2.1)$$

Масштаб измерения значений факторов всегда можно выбрать так, чтобы дисперсия его значений была равна единице: $\sigma^2_{f_i} = 1$. Кроме того, без уменьшения общности можно выбрать масштабы измерения переменных из условия $\sigma^2_{x_i} = 1$. Если значения факторов попарно независимы, то легко показать, что КК между переменными выражается через коэффициенты отображения факторов на переменные a_{ij} следующим образом:

$$r_{it} = \sum_{j=1}^{n} r_{i,t,j} \frac{\sigma_{x_ij}\sigma_{x_tj}}{\sigma_{x_i} \cdot \sigma_{x_t}} = \sum_{j=1}^{n} 1 \cdot \frac{a_{ij} \cdot a_{tj}}{\sqrt{\sum a^2_{ij} \sum a^2_{tj}}} = \sum_{j=1}^{n} a_{ij}a_{tj}, \qquad (10.2.2)$$

где дисперсии переменных равны $\sigma^2_{x_i} = \sum a^2_{ij} = 1$.

Рассмотрим условия, при которых КК второго порядка между переменными x_i и x_t будет равен единице. Для этого необходимо выполнение равенства

$$r_{ki} = br_{kt} + a, \qquad (10.2.3)$$

где b и a — постоянные, а индекс k ($k \neq i \neq t$) охватывает признаки, являющиеся базой для сравнения признаков x_i и x_t. Сумму (10. 2. 2), составляющую КК между признаками, можно представить в виде двух слагаемых:

$$r_{ki} = \sum_{j \in A} a_{kj} a_{ij} + \sum_{j \in \bar{A}} a_{kj} a_{ij}, \qquad (10.2.4)$$

где A — множество индексов, обозначающих факторы, которые одинаково влияют на косвенно сравниваемые переменные x_i и x_t; $\bar{A}$ — множество признаков, объединяющих остальные факторы. Под факторами, одинаково действующими на переменные x_i и x_t, будем понимать такие, коэффициенты отображения у которых для этих переменных либо равны, либо пропорциональны, т. е. или $a_{ij} = a_{tj}$, или $a_{ij}/a_{tj} = \text{const}$ ($i \in A$).

Аналогично равенству (10. 2. 4) можно написать

$$r_{kt} = \sum_{j \in A} a_{kj} a_{tj} + \sum_{j \in \bar{A}} a_{kj} a_{tj}. \qquad (10.2.5)$$

Так как $a_{ij}/a_{tj} = \text{const} = b$, ($j \ni A$), то

$$\sum a_{kj} a_{ij} = b \sum a_{kj} a_{tj} \quad (j \in A), \qquad (10.2.6)$$

откуда

$$r_{ki} = br_{kt} = \sum_{j \in A} a_{kj} a_{ij} - b \sum_{j \in A} a_{kj} a_{tj},$$

$$r_{ki} = br_{kt} + \sum_{j \in A} a_{kj} (a_{ij} - ba_{tj}). \qquad (10.2.7)$$

Сравнивая последнее равенство с (10. 2. 3), заключаем, что КК второго порядка будет равен единице в двух случаях: 1) если второе слагаемое в равенстве (10. 2. 7) постоянно — случай $a_{kj} = \text{const}$ для всех j, т. е. когда коэффициенты отображения a_{kj} для всех j не зависят от k (равны друг другу при всех j), а также в том случае, когда $\Sigma (a_{ij} - ba_{tj}) = a/\text{const}$; 2) если второе слагаемое равно нулю — случай $a_{ij} = ba_{tj}$, означающий равенство парных КК для сравниваемых признаков и очевидный.

На практике трудно ожидать, чтобы множество факторов, определяющих поведение изучаемых признаков, очевидным образом распадалось именно на два подмножества, каждое из которых одинаковым образом действует на «свою» группу признаков. Скорее всего, действие разных факторов может быть приблизительно одинаковым. В соответствии с этим КК второго порядка

между родственными признаками не станет равным единице. Однако можно показать, что максимальный КК (парный) будет обязательно увеличивать свое значение с увеличением порядка, пока не достигнет значения ± 1. Далее, если один из КК n-го порядка примет значение ± 1, то будет увеличиваться абсолютное значение следующего по величине КК. Практически после каждого цикла вычислений увеличивается бóльшая часть КК.

С изложенных позиций попытаемся оценить результаты применения МК в геологии и геохимии. Конечный результат процедуры — разделение множества признаков (содержаний химических элементов) на две антагонистические группы — содержателен только тогда, когда реальная система действительно характеризуется двумя и только двумя группами факторов с одинаковым действием на «свои» группы признаков. Если априори известно или предполагается, что это так, то применение МК целесообразно. В противном случае разделение признаков на две группы будет формальным, не имеющим реальной основы. Применение МК целесообразно также в том случае, когда задача заключается в подразделении признаков или объектов — носителей этих признаков на две и только на две группы по принципу максимального антагонизма. Скорее, это может иметь значение в социологии, например при формировании конкурирующих коллективов, чем в геологии и геохимии. В геологии, палеонтологии, рудной геологии и геохимии зачастую требуется отнести охарактеризованные рядом признаков объекты, например окаменелости, к одному из двух классов, априори не охарактеризованных (только по предположению, что они — представители двух разных совокупностей: вид А—вид Б, перспективный—бесперспективный, русловой—пойменный и т. д.). В этих случаях представляется весьма перспективным применение *обращенной процедуры МК*. От обычной она отличается тем, что в качестве «признаков» рассматривают носители признаков — пробы, массивы, окаменелости, районы и т. д. Парные КК, из которых должна быть построена исходная матрица, рассчитываются необычным образом по формулам

$$r_{ik}=\frac{\mathrm{cov}_{ik}}{s_i s_k};\quad \mathrm{cov}_{ik}=\frac{1}{m}\sum_{j=1}^{m}(x_{ij}-\bar{x}_i)(x_{kj}-\bar{x}_k);$$

$$s_i^2=\frac{1}{m}\sum_{j=1}^{m}(x_{ij}-\bar{x}_i)^2, \qquad (10.2.8)$$

где x_i — значения различных признаков в i-м объекте; $\bar{x}_i$ — среднее значение различных признаков (всех изученных!) в этом же объекте; m — число признаков. Матрица «объекты—признаки» используется здесь способом, противоположным обычному. На-

сколько известно авторам, обращенный метод МК еще не применялся. С известными ограничениями он может служить простой заменой чрезвычайно сложного в вычислительном отношении метода факторного анализа.

Таким образом, из расшифровки математической сущности МК вытекает ограниченность содержательного применения этого метода.

1. Если реальная ситуация характеризуется двумя и только двумя группами факторов, приблизительно одинаковым образом действующих на «свою» группу признаков, метод МК позволяет выделить эти группы признаков. Если факторы в действительности не подразделяются описанным образом или подразделяются на большее число групп, результаты подразделения на две группы с помощью МК имеют формальный характер. Геохимическая интерпретация такого результата едва ли обоснована.

2. Промежуточные результаты МК, которые сторонниками метода подвергаются интерпретации наравне с конечными, не имеют самостоятельного значения.

3. Более широкое применение МК может получить в «обращенном» варианте для подразделения объектов-носителей признаков. Таким образом, мы получим новый метод разбиения совокупности объектов, охарактеризованных m признаками, на две группы. В этом методе используется не сравнение многомерных средних, а характер корреляционных связей между признаками.

§ 3. О ложной корреляции в закрытых системах

Здесь мы подведем итоги длительной дискуссии по вопросу о так называемой ложной корреляции в связи с процентными пересчетами (Chayes, 1948, 1960, 1962; Сарманов, Вистелиус, 1959; Алешин, 1967; Шестаков, 1969; Груза, 1969; Вистелиус, 1970). Нами было показано (Ткачев, Юдович 1972), что математические зависимости, составляющие содержание проблемы, крайне просты.

Будем называть *числовой системой* совокупность чисел таких, для которых имеет предметный смысл сумма $\sum_{i=1}^{n} x_i$, где x_i — число с номером i; n — общее количество чисел. Если сумма чисел является величиной неограниченной, то такую числовую систему назовем открытой; если $\sum_{i=1}^{n} x_i \leqslant a$, назовем ее ограниченной; наконец, если эта сумма постоянна, то такую числовую систему назовем закрытой. Системы процентных величин как раз и являются закрытыми числовыми системами: сумма чисел в них постоянна и равна 100%.

Любую открытую или ограниченную систему можно превратить в закрытую, поделив каждое число на их сумму и получив таким путем новые относительные величины вида $\xi = x_i / \sum x_i$.

Однако обратный переход от закрытой системы к открытой или ограниченной возможен не всегда — для этого необходимо знать величину суммы.

Закрытые (процентные) числовые системы используются геологами постоянно. Особенно велика роль процентных систем в геохимии. Поэтому важно установить, какими специфичными свойствами по сравнению с исходными открытыми и ограниченными системами они обладают (кроме разницы, заложенной в самом определении закрытой системы, — постоянства суммы). Для нас важны следующие свойства.

1. Между компонентами закрытой системы существуют статистические связи, которые, по словам Ф. Чейза (Chayes, 1960), могут быть свойствами этих систем, а не образующих их исходных данных. Другими словами: в закрытой или ограниченной системе имеются связи, которых вообще не было в соответствующей открытой системе.

2. Если в исходной открытой или ограниченной системе уже имелись статистические связи, то при переходе к закрытой системе они могут в различной степени искажаться, причем практически более важно знать свойства перехода от закрытой системы к открытой (или ограниченной).

Первое свойство закрытых систем изучал Ф. Чейз (Chayes, 1958, 1960, 1962). Он показал, что при небольшом числе компонентов в системе процентных величин значения парных коэффициентов корреляции между компонентами полностью контролируются их выборочными дисперсиями. В частности, для i, j, k коэффициент корреляции первых двух запишется в виде $r(ij) = \frac{1}{2} \frac{s_k^2 - (s_i^2 - s_j^2)}{s_i \cdot s_j}$, где s^2 и s — соответственно дисперсии и стандартные отклонения значений компонентов. Очевидно также, что для системы из двух компонентов коэффициент корреляции между ними равен минус единице; при увеличении их числа по меньшей мере два коэффициента корреляции остаются всегда отрицательными. Добавим, что многокомпонентную систему всегда можно рассматривать как трехкомпонентную, приняв за третий сумму всех остальных $(n-2)$ компонентов. Значительный интерес представляет исследование данного свойства для ограниченных систем.

Второе свойство закрытых (процентных) систем — возможность искажений уже имевшихся связей — широко обсуждалось в геохимической литературе как «проблема ложной корреляции». Предметом обсуждения были обычно величины, выражающие состав вещества — концентрации. Поэтому, прежде чем рассмотреть суть проблемы, разберем, что из себя представляют концентрации.

В химии используют два существенно различных вида концентраций, которые можно назвать «абсолютной» и «относитель-

ной» концентрациями.[1] Абсолютная концентрация — это дробь, в которой числитель и знаменатель имеют различную размерность. Например: г/см³, кг/м³, см³/г, моль/л, мг/мл, % — экв./л и т. п. В частности, абсолютной концентрацией является число атомов N_i^V элемента, заключенное в единице объема горной породы, которая равна $10\,000A^3$: $N_i^V = 6.024 \cdot 10^{-2} \cdot d_V A_i$, где d_V — объемный вес породы, A_i — атомный вес данного элемента (Рудник, 1966, стр. 17).

Ограничимся для простоты наиболее распространенными концентрациями вида «масса—объем», которые можно назвать объемными концентрациями. Если массу i-го компонента обозначим m_i, а объем пробы — v, то получим выражение для объемной концентрации x_i данного компонента: $x_i = m_i/v$.

Для суммы всех концентраций в данной пробе получим

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \frac{m_i}{v} = \frac{1}{v} \sum_{i=1}^{n} m_i. \qquad (10.3.1)$$

Какой системой — открытой или закрытой — является набор абсолютных концентраций всех n компонентов в пробе? Как видим, выражение (10. 3. 1) — объемный вес. Теоретически верхний предел для значения объемного веса колоссально высок — это «объемный вес» нейтронной звезды, равный, по расчетам физиков, величине порядка $1 \cdot 10^{15}$ г/см³. Однако для реальных горных пород величины объемных весов не превышают 3.5 г/см³, и лишь для руд тяжелых металлов значения объемных весов могут быть в несколько раз бóльшими.

Таким образом, абсолютные концентрации являются ограниченными системами, поскольку в любой анализируемой партии проб имеется верхний предел объемного веса. С этим обстоятельством связана известная специфичность концентраций по сравнению с другими переменными величинами, используемыми в геохимии.

Относительная концентрация, или процентная доля, — это дробь, в которой числитель и знаменатель имеют одинаковую размерность: мл/л, см³/м³, г/г, γ/г, атомов/1000 атомов, моль/г и т. д. Ограничимся наиболее распространенными в геохимии процентными долями вида «масса—масса», которые называют весовыми концентрациями. Если M_i — масса i-го компонента в пробе, а M — масса самой пробы, то весовая концентрация элемента ξ_i выразится так:

$$\xi_i = M_i/M. \qquad (10.3.2)$$

Таким образом, выражение (10. 3. 2) описывает закрытую систему, и в этом состоит принципиальная разница между абсо-

[1] Кавычки означают условность этих определений, поскольку любая концентрация является относительной величиной.

лютными и относительными концентрациями. Заметим еще раз, что исходная система, соответствующая закрытой системе процентных долей («весовых процентов»), обладает важной особенностью — незначительными колебаниями суммы, являющейся объемным весом проб, т. е., строго говоря, является не открытой, а ограниченной.

Применительно к концентрациям проблема «ложной корреляции» формулируется следующим образом: а) в какой мере коэффициенты корреляции, рассчитанные между весовыми процентами, являются состоятельными оценками коэффициентов корреляции, рассчитанных между абсолютными (например, объемными) концентрациями? б) признавая, что корреляция между весовыми процентами неизбежно искажена по сравнению с корреляцией между исходными абсолютными концентрациями, найти способ расчета неискаженных, «истинных», связей.

Весовая (относительня) концентрация ξ выражается через объемную (абсолютную) концентрацию x с помощью величины плотности пробы δ: $\xi = x/\delta$, если по смыслу объемная концентрация означает вес в единице объема жидкости или твердого тела без учета его пор и трещин; либо с помощью величины объемного веса d: $\xi = x/d$, если по смыслу объемная концентрация означает вес в единице общего объема, занимаемого данной порцией вещества, что бывает необходимо при переходе к объемам толщ, формаций и месторождений.

Известно, что коэффициенты корреляции — величины, инвариантные к линейным преобразованиям переменных. Это означает, что парный коэффициент корреляции для объемных концентраций $\rho(x_1, x_2)$ будет в точности равен коэффициенту корреляции для весовых процентов $\rho'(\xi_1, \xi_2)$ в том (и только в том) случае, если плотность δ или объемный вес d — постоянные величины, не изменяющиеся от пробы к пробе. Действительно, при таком преобразовании концентраций в пробах с постоянной плотностью не происходит перехода к открытой (ограниченной) системе! Если же плотность или объемные веса от пробы к пробе колеблются, то коэффициенты корреляции для весовых процентов получатся несколько иными, чем для объемных концентраций.

В чем формально-математический смысл корреляции между весовыми концентрациями? Чтобы понять это, будем рассматривать d или δ наряду с процентными концентрациями каждого компонента ξ_i как еще одну переменную от пробы к пробе величину — d_i или δ_i. Тогда можно сказать, что коэффициенты корреляции между объемными концентрациями $\rho(x, x_2)$ численно равны и по существу эквивалентны частному коэффициенту корреляции $\rho(\xi_1, \xi_2)_\delta$ между весовыми процентами этих же величин. Действительно, частный коэффициент корреляции вычисляют при условии, что все переменные, кроме двух коррелируемых, постоянны. Следовательно, зная плотности или объемные веса

каждой пробы, можно рассчитать «истинную» корреляцию между переменными в форме весовых процентов по известной формуле для частного коэффициента корреляции:

$$\rho(x_1x_2) = \rho_{1,\,2(\delta)} = \frac{\rho_{1,2} - \rho_{1,\delta}\cdot\rho_{2\delta}}{\sqrt{(1-\rho^2_{1,\delta})(1-\rho^2_{2,\delta})}}. \qquad (10.3.3)$$

Расчеты по формуле (10. 3. 3) могут в каком-то случае оказаться полезными, хотя, располагая данными о плотностях или объемных весах проб, удобнее весовые концентрации предварительно перевести в объемные и далее рассчитывать обычные парные коэффициенты корреляции. Значение выражения (10. 3. 3) состоит в том, что оно обнажает математическую сущность «ложной» корреляции как такой зависимости между двумя переменными концентрациями, которая найдена без учета другого фактора (плотности), влияющего на обе переменные.

Предположим теперь, что среди объемных концентраций одна (например, первого компонента) постоянна: $x_1 = c$. Подставим это значение в соответствующую процентную величину $\xi_1 = c/\delta_1$, а в общем случае для данного компонента $\xi_i = c/\delta_i$, где i — номер пробы. Как видим, постоянная объемная концентрация становится переменной весовой концентрацией вследствие того, что плотности δ_i проб различны.

Что произойдет, если все весовые концентрации ξ_i поделить на весовую концентрацию постоянного компонента ξ_c? Получим некоторые величины z_i, прямо пропорциональные плотности проб:

$$z_i = \frac{\xi_i}{\xi_c} = \frac{\xi_i}{c}\delta_i = \xi_i \cdot c' \cdot \delta_i. \qquad (10.3.4)$$

Теперь сравним выражение (10. 3. 4) и выражение для объемных концентраций:

$$\left.\begin{aligned} z_i &= \xi_i \cdot c' \cdot \delta_i \\ x_i &= \xi_i \cdot \delta_i \end{aligned}\right\} \; z_i = c' x_i. \qquad (10.3.5)$$

Видно, что величины z_i прямо пропорциональны объемным концентрациям, а это значит, что использование величин z_i вместо ξ_i исключает ложную корреляцию.

Таким образом, привлекая к рассмотрению проблемы плотности проб, мы получили результат, широко известный в литературе как «первая теорема Вистелиуса—Сарманова». В 1959 г. О. В. Сарманов и А. Б. Вистелиус доказали две теоремы, позволяющие избавляться от ложной корреляции, оперируя процентными величинами (без использования информации о плотностях проб) в двух случаях: а) при «конкреционной» схеме (среди пере-

менных имеется хотя бы одна постоянная) — случай, разобранный нами выше; б) при «метасоматической» схеме (среди переменных имеется одна или две независимых друг от друга и от всех остальных).

Вторая теорема пока, к сожалению, не нашла большого применения, тогда как первая использовалась исключительно широко в практике петрохимических пересчетов. Согласно этой теореме, для устранения ложной корреляции следует рассчитывать коэффициент корреляции не между процентными величинами (окислы, элементы, грамм-атомы, грамм-молекулы), а между частными от их деления на процентную величину постоянного компонента. В петрохимической практике применение этой теоремы сводится к делению весовых процентов элементов на весовой процент кислорода (иногда — кремния) в анализе. Согласно представлениям Т. Ф. Барта (Barth, 1962), в «стандартной ячейке» изверженных и метаморфических горных пород содержится в среднем 160 атомов кислорода на 100 атомов катионов, т. е. примерно постоянна объемная концентрация кислорода.

Как было показано выражением (10. 3. 5), информация о постоянстве абсолютной объемной концентрации одного из компонентов эквивалента информации о плотностях проб. Это обстоятельство позволяет проверить справедливость гипотезы Т. Ф. Барта. Для этого нужно лишь сравнить коэффициенты корреляции между процентными величинами, предварительно преобразованными, путем: а) деления весовых процентов на процент кислорода; б) пересчета весовых процентов в объемные, используя данные о плотностях проб.

В течение ряда лет процедура деления на кислород при корреляционном анализе в петрохимии считалась обязательной: в противном случае геохимические выводы о связях породообразующих компонентов признавались сомнительными, отягощенными ошибками ложной корреляции. Однако впоследствии подметили, что коэффициенты корреляции, полученные после предварительного деления на кислород и без такого деления, различаются незначительно. С. М. Алешин (1967) пришел к заключению, что деление на кислород во многих случаях излишне. Однако А. Б. Вистелиус (1970, стр. 1390) возразил, что если процедура устранения ложной корреляции даже незначительно изменяет значение коэффициентов корреляции, «то все равно ее (теорему, — *Ю. Т., Я. Ю.*) применять надо, если в дальнейшем предполагается сравнительно полное изучение связей», поскольку в дальнейших вычислениях (например, при расчетах частной корреляции) влияние начальных ошибок быстро увеличивается.

В литературе имеются и другие суждения относительно применения теоремы Вистелиуса—Сарманова, вызванные ее недостаточным пониманием. Так, в 1969 г. Ю. Г. Шестаков, В. П. Романова, Э. П. Володина выступили со статьей, в которой поставили

под сомнение применимость этой теоремы на том основании, что в ней рассматривались процентные величины вида $M_i/M_1+M_2+ +\ldots+M_n$, где n якобы не является числом всех компонентов породы и, следовательно, А. Б. Вистелиус и О. В. Сарманов использовали «не те проценты», которые обычно используются геологами. Ошибочность такого мнения легко устанавливается при внимательном чтении статьи О. В. Сарманова, А. Б. Вистелиуса (1959). На это указали В. В. Груза (1969), а также А. Б. Вистелиус (1970).

К сожалению, долю неясности в обсуждаемую проблему внес и В. В. Груза. Касаясь теоремы Вистелиуса-Сарманова, он писал: «Метод пересчетов, предложенный А. Б. Вистелиусом и О. В. Сармановым, позволяет оценивать корреляцию между изообъемно сопоставимыми величинами, которая. . . может быть сведена к корреляции объемных процентов. Тем самым подобный пересчет. . . не способствует избавлению от „ложной корреляции" в связи с процентными пересчетами», но помогает на основе использования процентов одного рода величин оценивать корреляцию между процентами другого рода величин» (стр. 146). Однако что же понимается под выражением «проценты другого рода величин» и «объемные проценты»? Из текста статьи ясно, что В. В. Груза имеет в виду объемные концентрации, которые, как мы видели, не являются процентами!

Мы уже упоминали о специфичности понятия «концентрация»: даже абсолютные концентрации, строго говоря, не являются полностью открытыми системами. В большинстве прикладных задач эта специфика неощутима. Однако имеются ситуации, где специфичность даже абсолютных концентраций может порождать ложную корреляцию — в тех случаях, когда за «исходные» переменные принимаются какие-то другие величины. Приведем пример.

Пусть переменными величинами являются абсолютные массы, или, что то же самое, — запасы двух компонентов — титана и стронция, накопившихся на морском дне за некоторый промежуток времени на каких-то площадках, например, на 1 кв. м площади дна. При этом мы не имеем возможности подсчитать запасы по площадкам, а вынуждены оперировать только с величинами абсолютных концентраций в пробах с площади 1 кв. м, выраженными, например, в г/см^3 осадка. Будут ли коэффициенты корреляции, которые рассчитаны для абсолютных концентраций титана и стронция в осадках, состоятельными оценками «истинных» коэффициентов корреляции между их запасами, замеренными на отдельных площадках дна бассейна седиментации?

Поскольку стронций входит в карбонатную часть осадка, а титан в терригенную, то в какой бы части бассейна мы не замеряли концентрации, выборочные коэффициенты корреляции между концентрациями будут о т р и ц а т е л ь н ы м и, если порода представлена только двумя частями — терригенной и

карбонатной. Действительно, в фиксированном объеме, например в 1 $см^3$, увеличение карбонатной составляющей неизбежно повлечет разбавление ею терригенной составляющей и наоборот. Совершенно иначе будут вести себя запасы этих элементов: они будут зависеть только от количества осадка, приходящегося на данную площадку. Запасы будут связаны с расстоянием площадки от береговой линии: *чем ближе к берегу, тем бо́льшими* будут массы и терригенные и карбонатные; чем дальше от берега, тем меньшими будут масштабы седиментации и того и другого материала. Это означает, что запасы титана и стронция будут находиться в п о л о ж и т е л ь н о й зависимости.

Наконец, возможны более сложные связи в том случае, если осадок представлен, например существенно кремнистым материалом, а терригенный и карбонатный компоненты играют роль небольшой примеси. В этом варианте корреляционные связи стронция и титана, найденные по их концентрациям, могут быть вполне адекватными корреляционным связям между запасами, т. е. окажутся положительными.

Приведенный пример лишний раз подчеркивает, что в проблеме ложной корреляции, как и во всех геохимических проблемах, решающую роль играет предметное осмысливание цифрового материала, грамотная постановка задачи, учитывающая специфику изучаемых величин.

Итак, рассмотрение особенностей числовых систем процентных величин, проблемы ложной корреляции, а также обзор основной литературы по этим вопросам позволяют заключить следующее.

1. Представление о двух видах концентрации — абсолютной и относительной — оказывается весьма полезным в геохимической практике. Абсолютная концентрация не является процентной величиной, а относительная является.

2. Если состав вещества выражен в абсолютных концентрациях, а постановка задачи не предусматривает использования других величин (например, запасов), то ложная корреляция не возникает — для *корреляционного анализа* нет никаких ограничений.

3. Если состав вещества выражен в относительных концентрациях, возможна ложная корреляция. Для исключения ложной корреляции можно либо воспользоваться рекомендациями О. В. Сарманова, А. Б. Вистелиуса (1959), если среди исходных переменных имеется одна постоянная (теорема № 1), одна или две *независимых* (теорема № 2) величины, либо перейти к абсолютным концентрациям, используя дополнительную информацию о величинах плотностей проанализированных проб. В каких пределах колебания плотностей можно считать несущественными для точности расчетов? Приведем мнение В. А. Рудника (1966, стр. 16): «Процентное выражение состава горных пород,

руд и минералов, не приемлемое в общем случае для сравнительной характеристики указанных объектов, в отдельных случаях может быть с успехом использовано. Такими частными случаями являются: а) сравнение пород близкого минерального состава одной и той же классификационной группы; б) сопоставление микросодержаний элементов в породах близкого минерального состава или в породах, различие удельных весов которых не превышает 20—30%; в) сравнение составов минералов одной и той же группы, выраженного в атомных процентах».

4. Два способа освобождения от ложной корреляции в петрохимии (переход к абсолютным концентрациям с привлечением данных о плотностях проб и применение теоремы № 1 Сарманова—Вистелиуса) математически совершенно эквивалентны. Одновременно первый способ эквивалентен нахождению частной корреляции между процентными величинами, когда плотность рассматривается как исключенная переменная величина.

5. Доказательство эквивалентности информации о плотностях проб — информации о наличии постоянного компонента — позволяет легко проверить правильность гипотезы Т. Ф. Барта (Barth, 1961) о постоянстве объемной концентрации кислорода в изверженных и метаморфических горных породах. Для этого достаточно привлечь данные о плотностях представительных выборок проб и рассчитать абсолютные концентрации всех компонентов. Если коэффициенты корреляции, рассчитанные между абсолютными концентрациями, окажутся такими же, как между весовыми процентами, предварительно поделенными на кислород, гипотеза Барта верна; в противном случае ее, может быть, придется оставить.

6. Ложная корреляция может возникнуть не только при использовании процентных величин; специфика концентраций, даже если это не процентные величины, может в некоторых типах задач также порождать ложную корреляцию.

ЛИТЕРАТУРА

Алешин С. М. К вопросу о ложной корреляции. — Геохимия, 1967, № 4, с. 486—490.

Айвазян С. А. Статистическое исследование зависимостей. М., 1968, 227 с.

Барышев Н. В. Контроль опробования. — В кн.: Материалы ВКЗ по методике разведки и подсчета запасов. М., вып. 2, 1948, с. 88.

Бачурин Э. Ф., Шарапов И. П., Жеребцов А. В. Статистическое исследование мощности, зольности и сернистости углей Кизеловского бассейна. — В кн.: Сб. научн. тр. Пермск. политехн. инст., 1970, № 67, с. 162—170.

Белов Б. И. О геохимическом смысле законов распределения концентраций вещества. — В кн.: Математические методы геохимических исследований. М., 1966, с. 71—81.

Береншейн Л. Е., Фалькова О. Б. Применение математической статистики к исследованию аналитических методик. — Тр. Центр. научн.-исслед. горноразвед. инст., 1967, вып. 77, с. 248—252.

Беус А. А. Геохимия литосферы (породообразующие элементы). М., 1972, 295 с.

Беус А. А. и др. Руководство по предварительной математической обработке геохимической информации при поисковых работах. М., 1965, 120 с.

Блисковский В. З. Геохимия и особенности концентрации элементов-примесей в фосфоритах. — Автореф. канд. дисс. М., 1969, 25 с.

Богацкий В. В. Математический анализ разведочной сети. М., 1963, 212 с.

Богацкий В. В. Оценка средних содержаний малыми выборками при асимметричных распределениях как общегеологическая проблема. — Геол. и геофиз., 1968, № 5, с. 74—80.

Большев Л. Н., Смирнов Н. В. Таблицы математической статистики. М., 1965, 464 с.

Бондаренко В. Н. Статистические методы изучения вулканогенных комплексов. М., 1967, 135 с.

Бондаренко В. Н. Статистические решения некоторых задач геологии. М., 1970, 246 с.

Боровко Н. Н. Обобщенный логнормальный закон распределения содержаний химических элементов в породах и рудах. — Геохимия, 1964, № 3, с. 282—288.

Бостанджиян В. А. Определение плотности вероятности. Необходимый объем выборки. М., 1971, 160 с.

Бурков Ю. К. Линейные парагенезисы малых элементов в осадочных толщах как индикаторы условий седиментогенеза. — В кн.: Физ. и хим. процессы и фации. М., 1968, с. 22—26.
Бурков Ю. К. Изучение условий формирования осадочных толщ методами статистической обработки геохимических данных. — Тр. Всесоюзн. научн.-исследов. геол. инст., 1971, вып. 158, с. 346—365.
Бурков Ю. К. Ассоциации химических элементов и ряды их подвижности. — Тр. Межведомств. стратиграф. ком. АН СССР, 1972, вып. 5, с. 199—210.
Бурков Ю. К., Певзнер В. С. Корреляционный анализ поведения химических компонентов при корообразовании. — В кн.: Материалы семинара по геохимии гипергенеза и коры выветривания. Минск, 1969, с. 51—57.
Вассоевич Н. Б. Образование нефти в терригенных отложениях (на примере чокракско-караганских отложений Терского передового прогиба). — В кн.: Вопросы образования нефти. Л., 1958 (Тр. ВНИГРИ, вып. 128, с. 3—220).
Ван дер Варден Б. Л. Математическая статистика. М., 1960, 434 с.
Вернадский В. М. Заметки о распространении химических элементов в земной коре. Наблюдения 1909—1910 г. (Совместно с Е. А. Ревуцкой и А. А. Твалчрелидзе). — Изв. АН, 1910, с. 1129—1148.
Вернадский В. И. Заметки о распространении химических элементов в земной коре. Наблюдения 1910 г. (Совместно с Б. А. Линденером и Е. Д. Ревуцкой). — Изв. АН, 1911, с. 1007—1018.
Виноградов А. П. Закономерности распределения химических элементов в земной коре. — Геохимия, 1956, № 1, с. 6—52.
Виноградов А. П. Среднее содержание химических элементов в главных типах изверженных горных пород земной коры. — Геохимия, 1962, № 7, с. 555—571.
Виноградов А. П. Введение в геохимию океана. М., 1967, 213 с.
Вистелиус А. Б. Фазовая дифференциация палеозойских отложений Среднего Поволжья и Заволжья. М.—Л., 1963а, 203 с.
Вистелиус А. Б. Проблемы математической геологии. Случайный процесс. — Геол. и геоф., 1963б, № 12, с. 3—10.
Вистелиус А. Б. Теоретические предпосылки стохастических моделей и их проверка в конкретных геологических условиях. — В кн: Математические методы в геологии. М., 1968, с. 11—14.
Вистелиус А. Б. О некоторых ошибках в применении математических методов при анализе геохимических данных. — Геохимия, 1970, № 11, с. 1390—1393.
Вистелиус А. Б., Сарманов О. В. Стохастическое обоснование одного геологически важного распределения вероятностей. — ДАН СССР, 1947, т. 58, № 4, с. 631—634.
Воробьев В. Я. Статистические методы в геохимии. Саратов, 1970, 256 с.
Вострокнутов Г. А. Один из методов статистического исследования усеченных выборок в геохимии. — Изв. вузов, Геол. и разв., 1969, № 9, с. 76—81.
Гавришин А. И., Вострокнутов Г. А. Исследование качества приближенно-количественного спектрального анализа по данным междулабораторного контроля. В кн.: Геохимические методы при поисках и разведках рудных месторождений. Вып. 2, М., 1971, с. 88—122.
Гриффитс Дж. Научные методы исследования осадочных пород. Перев. с англ. М., 1971, 420 с.
Груза В. В. Ложная корреляция и процентные пересчеты в геохимии. — Сов. геол., 1969, № 11, с. 145—147.
Длин А. М. Математическая статистика в технике. М., 1958, 466 с.
Доерфель К. Статистика в аналитической химии. Перев. с нем. М., 1969, 247 с.

Дубов Р. И. Влияние дискретности записей на точность результатов измерений. — В кн.: Спектральный анализ элементов-примесей в горных породах. М., 1972, с. 44—57.
Ермолаев М. М. Основные геометрические и геохимические параметры земной коры. — Изв. Всесоюзн. геогр. общ., 1967, № 5, с. 420—434.
Зайдель А. Н. Элементарные оценки ошибок измерений. Изд. третье, испр. и доп. Л., 1968, 97 с.
Залата Л. Ф. Об определении средних содержаний полезных компонентов в руде. — Разв. и охрана недр., 1963, № 8, с. 22—26.
Зильберштейн Х. И. и др. Спектральный анализ чистых веществ. Л., 1971, 415 с.
Иванов В. В. Генеральные оценки средних содержаний элементов-примесей в главных рудных минералах. — ДАН СССР, 1969, т. 186, № 1, с. 185—186.
Инструкция по внутрилабораторному контролю точности (воспроизводимости) результатов количественных анализов рядовых проб полезных ископаемых, выполняемых в лабораториях Министерства геологии СССР. М., 1968, 15 с.
Каждан А. Б., Шумилин М. В. По поводу статьи Л. Ф. Залаты «Об определении средних содержаний полезных компонентов в руде». Перев. с франц. — Разв. и охрана недр, 1966, № 2, с. 19—20.
Канцель А. Ф. Функция распределения величин концентраций металла в рудах как генетическая характеристика процесса рудообразования.— Изв. АН СССР, 1966, сер. геол., № 10, с. 18—30.
Карлье Э. Методика количественной оценки месторождений урана. — М., 1966, 351 с.
Карпов А. В., Краснов Е. Г. О методе взвешивания при расчете средних содержаний полезных компонентов. — Разв. и охрана недр, 1963, № 12, с. 22—26.
Кельин А. М., Михайлович М. П. Способ вычисления систематической ошибки химических анализов. — Разв. и охрана недр, 1969, № 7, с. 23—25.
Кибисов Г. И., Антропов Н. П., Кубасова И. Б., Резвова М. И. Опыт разработки и применения универсального метода количественного спектрального анализа. Л., 1961, 54 с.
Колмогоров А. Н. О логарифмически нормальном законе распределения частиц при дроблении. — ДАН СССР, 1941, т. 31, № 2, с. 99—101.
Конюс А. А. Исследование корреляций, приводимых к нормальному типу, и построение корреляционных уравнений. — В кн.: Вопросы статистического измерения связей между явлениями (корреляционный анализ). М., 1950, с. 95—130.
Королев В. П. Об обработке результатов анализов внешнего контроля разведочных проб. — Разв. и охрана недр, 1969, № 7, с. 22—23.
Крамбейн У., Грейбилл Ф. Статистические модели в геологии. Перев. с англ. М., 1969, 396 с.
Крамбейн У., Кауфмен М., Мак-Кеммон Р. Модели геологических процессов. Введение в математическую геологию. Перев. с англ. М., 1973, 150 с.
Кренделев Ф. П. Методика отбора представительных проб пород для определения кларков радиоактивных элементов. М., ВИНИТИ АН СССР, 1972, № 5117-72 деп., 13 с.
Крейтер В. М. Поиски и разведка месторождений полезных ископаемых. Ч. II. М., 1960, 390 с.
Криге Д. Г. Роль математической статистики в методах уточненной оценки промышленного оруденения на рудниках Южной Африки. — В кн.: Вопросы математической геологии. Л., 1968, с. 252—272.
Кузнецов В. И., Панов Ю. К. Корреляционная связь содержаний кремнезема и трехокиси хрома в хромитовой руде Кемпирсайских

месторождений и ее практическое использование. — В кн.: Матер. III Уральск. конф. молодых геол. и геофиз. (тезисы докл.). Свердловск, 1971, с. 173—175.

Кутолин В. А. Проблемы петрохимии и петрологии базальтов. — Тр. Инст. геол. и геофиз. СОАН СССР, вып. 189, 1972, 208 с.

Лонцих С. В., Мешалкин Л. Д. Об оценке результатов полуколичественного спектрального анализа. — Заводск. лабор., 1964, № 7, с. 831—837.

Лонцих С. В., Райхбаум Я. Д., Смоляк З. М. Точность визуального фотометрирования спектрограмм. — В кн.: Геология и вещественный состав. М., 1967.

Лонцих С. В. и др. Спектральный анализ при поисках рудных месторождений. Л., 1969, 294 с.

Лукомский Я. И. Теория корреляции и ее применение к анализу производства. М., 1958, 388 с.

Ляхович В. В. Акцессорные минералы в гранитоидах Советского Союза. М., 1967, 447 с.

Ляхович В. В. Акцессорные минералы. М., 1968, 275 с.

Ляхович В. В. Редкие элементы в породообразующих минералах гранитоидов. М., 1972, 199 с.

Маликов С. Ф., Тюрин Н. И. Введение в метрологию. М., 1966, 247 с.

Матерон Ж. Основы прикладной геостатистики. Перев. с франц. М., 1968, 408 с.

Мейсон Б. Основы геохимии. Перев. с англ. М., 1971, 311 с.

Мешалкин Л. Д. Выбор шкалы для представления результатов полуколичественного спектрального анализа. — Заводск. лабор., 1964, № 7, с. 857—860.

Миддлтон Г. В. Возникновение логнормального распределения частот в осадках. — В кн.: Вопросы математической геологии. Л., 1968, с. 37—45.

Миллер Р. Л., Кан Д. С. Статистический анализ в геологических науках. Перев. с англ. М., 1965, 481 с.

Мищенко В. С. К распределению ошибок приближенных методов спектрального анализа. — Заводск. лабор., 1965, № 3, с. 338—341.

Мягков В. Ф. О геологической природе криволинейных зависимостей между содержаниями компонентов в эндогенных рудных телах. — ДАН СССР, 1969, т. 187, № 2, с. 424—427.

Налимов В. В. Применение математической статистики при анализе вещества. Л., 1960, 430 с.

Огнева Э. Я., Огнев В. Р., Райхбаум Я. Д. Использование метода добавок для контроля правильности спектрального анализа горных пород. — Заводск. лабор., 1968, т. 34, № 12, с. 1450—1454.

Орлова К. Б., Кивисилла Я. Я. О применении естественных стандартных образцов при полуколичественном спектральном анализе. — В кн.: Спектральный анализ в геологии. М., 1971, с. 167—168.

Остафийчук И. М., Толстой М. И. Статистические закономерности распределения химических элементов в гранитоидах (на примере Сев. Казахстана). М., 1972, 224 с.

Петров А. А. О неприменимости метода взвешивания для расчета средних содержаний. — Разв. и охрана недр, 1962, № 3, с. 11—14.

Петров В. А. О применении способов среднего арифметического и среднего взвешенного. — Сов. геол., 1965, № 2, с. 112—125.

Петров В. П. Петрология и геохимия метаморфических пород Ладожской серии. — Автореф. канд. дисс. Апатиты, 1970, 25 с.

Попов В. И. Использование кларков для оценки баланса химических элементов и вероятности нахождения осадочных месторождений. —

В кн.: Геохимия, петрография и минералогия осадочных образований. М., 1963, с. 22—67.
Прокофьев А. П. Приемы обработки контрольных химических анализов. — Разв. и охрана недр, 1955, № 3, с. 25—29.
Прокофьев А. П. О внутреннем и внешнем контроле анализов разведочных проб. — Разв. и охрана недр, 1962, № 6, с. 6—10.
Раевский В. И., Шурубор Ю. В. Обработка данных внешнего контроля химических анализов геологических проб. — Изв. вузов, 1958, Геол. и разв., № 11, с. 63—69.
Разумовский Н. К. Характер распределения содержаний металлов в рудных месторождениях. — ДАН СССР, 1940, т. 28, № 9, с. 815—817.
Разумовский Н. К. Логнормальный закон распределения и его особенности. — Зап. ЛГИ, 1948, т. 20.
Разумовский Н. К. К вопросу о выделении аномалий на фоне обычных содержаний элементов в породе при поисковых работах. — В кн.: Вопросы разведочной геофизики, вып. 1. 1962, с. 28—45.
Рац М. В. Статистический анализ масштабных эффектов при изучении свойств горных пород. — В кн.: Математические методы в геологии. М., 1968, с. 117—123.
Родионов Д. А. Об оценивании среднего содержания и дисперсии логнормального распределения компонентов в породах и рудах. — Геохимия, 1962, № 7, с. 629—732.
Родионов Д. А. Особенности распределения среднего арифметического в условиях асимметричных распределений содержаний. — Геохимия, 1963а, № 7, с. 689—693.
Родионов Д. А. Трехпараметрические распределения содержаний элементов в породах. — Геохимия, 1963б, № 2, с. 179—184.
Родионов Д. А. Функции распределения содержаний элементов и минералов в изверженных горных породах. М., 1964а, 102 с.
Родионов Д. А. О внешнем контроле результатов анализов при подсчете запасов рудных месторождений. — Разв. и охрана недр, 1964б, № 5, с. 14—17.
Родионов Д. А. Статистические методы разграничения геологических объектов по комплексу признаков. М., 1968, 158 с.
Родионов Д. А., Иванов В. В. Статистические оценки средних содержаний по совокупности наблюдений разной представительности. — Геохимия, 1967, № 1, с. 109—117.
Ронов А. Б., Ратынский В. М. Метод установленных средних проб. — ДАН СССР, 1952, т. 86, № 4, с. 779—782.
Ронов А. Б., Гирин Ю. П., Казаков Г. А., Илюхин М. Н. Сравнительная геохимия геосинклинальных и платформенных осадочных толщ. — Геохимия, 1965, № 8, с. 961—976.
Ронов А. Б., Ярошевский А. А. Химическое строение земной коры. — Геохимия, 1967, № 11, с. 1285—1309.
Рудник В. А. Атомно-объемный метод в применении к метасоматическому минерало- и породообразованию. М., 1966, 118 с.
Румшиский Л. З. Математическая обработка результатов эксперимента. М., 1971, 192 с.
Рыжов П. А., Гудков В. М. Применение математической статистики при разведке недр. М., 1966, 235 с.
Рябчиков И. Д. Влияние степени неравновесности процесса кристаллизации на поведение элемента-примеси. — Геохимия, 1960, № 4, с. 345—354.
Сарманов О. В., Вистелиус А. Б. О корреляции между процентными величинами. — ДАН СССР, 1959, т. 126, № 1, с. 22—25.
Сауков А. А. Геохимия. М., 1950, 348 с.
Смирнов В. И. и др. Подсчет запасов месторождений полезных ископаемых. М., 1960, 672 с.

Смирнов С. И. Вероятностно-статистическая оценка геохимического фона при поисках месторождений полезных ископаемых. — Геохимия, 1963, № 3, с. 333—343.
Соколов М. Б. Вопросы интерпретации древнейших радиогеологических дат. Сыктывкар, 1968, 45 с.
Страхов Н. М. Об объеме осадочного чехла Русской платформы и глобальной количественной геохимии. — Изв. АН СССР, 1973, сер. геол., № 5, с. 3—12.
Суражский Д. Я., Рощин Ю. В. О книге В. В. Богацкого «Математический анализ разведочной сети». — Геол. рудн. месторожд., 1964, № 4, с. 102—105.
Тепляков В. Г. Способ попеременного фотографирования спектров исследуемого вещества и эталона в процессе одной экспозиции. (Докл. на семинаре по количеств. спектр. анализу минеральн. сырья 15—30 июня 1964 г., Ленинград). — М.—Л., 1964.
Ткачев Ю. А. Несколько замечаний к книге Д. А. Родионова «Статистические методы разграничения геологических объектов по комплексу признаков». — Геол. и геофиз., 1972, № 4, с. 141—143.
Ткачев Ю. А. Повышение надежности оценок средних в геологии без увеличения объема выборок. — Ежегодник-1972, Инст. геол. Коми фил. АН СССР, 1973а, с. 177—183.
Ткачев Ю. А. Оценки параметров неоднородных совокупностей в геологии, геохимии и геологоразведочном деле при объединении выборок разной представительности. М., ВИНИТИ АН СССР, 1973б, № 7500-73 деп.
Ткачев Ю. А. Анализ сравнительной эффективности оценок средних при подсчете запасов нефти. — В кн.: Геология и нефтегазоносность Тимано-Печорской провинции. Сыктывкар, 1975.
Ткачев Ю. А., Брезгунова Г. Д., Брезгунов Н. И. К вопросу о классификации гидротермальных месторождений пьезооптического кварца по степени изменчивости хрусталеносной минерализации. — Тр. Всесоюзн. научн.-исслед. инст. синтеза минер. сырья, 1970, т. 12, с. 45—53.
Ткачев Ю. А., Юдович Я. Э. К вопросу о процентных величинах, ложной корреляции и гипотезе Т. Ф. Барта. — В кн.: Научные основы геохимических методов поисков месторождений полезных ископаемых и оценки потенциальной рудоносности магматических и метаморфических комплексов докембрия. Апатиты, 1972, с. 108—113.
Трошин Ю. П., Белов Б. И., Трошина Г. М. Законы распределения концентраций элементов-примесей в минералах гидротермальных рудных тел (регулярный процесс). — В кн.: Математические методы геохимических исследований. М., 1966, с. 31—38.
Уилкс С. Математическая статистика. М., 1967, 632 с.
Урбах В. Ю. Биометрические методы. М., 1964, 415 с.
Фалькова О. Б., Лифшиц Д. М., Большакова Н. А. О точности полуколичественного спектрального анализа. — В кн.: Спектральный анализ в геологии. М., 1971, с. 238—241.
Феллер В. Введение в теорию вероятностей и ее приложения. Т. 2. М., 1967, 752 с.
Ферсман А. Е. Химические элементы Земли и Космоса. Пгр., 1923.
Францкий И. В. Функция распределения показателей месторождений. — В кн.: Математические методы в поисково-разведочной практике. Иркутск, 1970, с. 230—235.
Фридлендер Н. Г., Юшкова Г. Е. Метод просыпки-вдувания в спектральном анализе геологических проб. — Тр. Инст. геол. Коми фил. АН СССР, вып. 12, 1972, с. 81—90.
Фурман И. Я. К расчетам кларков земной коры. — Тр. Воронежск. гос. унив., 1968, т. 66 (геол. сб.), с. 262—265.

Хитров В. Г., Кортман Р. В. Рекомендуемые содержания породообразующих и малых элементов в стандартных породах ИГЕМ по данным межлабораторного анализа (обзор результатов). М., Ротапринт ИГЕМ АН СССР, 1969, 68 с.

Четвериков Л. И. Закон распределения частот содержания минерального компонента в теле полезного ископаемого. — Сов. геол., 1964, № 7, с. 92—102.

Чупров А. А. Основные проблемы теории корреляции. М., 1960, 174 с.

Шарапов И. П. О контрольных анализах геологических проб. — Разв. и охрана недр, 1954, № 1, с. 16—27.

Шарапов И. П. Применение математической статистики в геологии. М., 1965, 260 с.

Шашкин В. Л. О контрольном опробовании и контрольных анализах геологических проб. — Тр. Инст. геол. АН Киргизск. ССР, 1956, вып. 7, с. 111—124.

Шестаков Ю. Г. К вопросу определения среднего при обработке результатов анализа. — В кн.: Материалы геологической конференции Красноярского геологического управления, 1964 г. Красноярск, 1966, с. 306—309.

Шестаков Ю. Г., Романова В. П., Володина Е. Н. К вопросу корреляции процентных величин химического анализа. — Сов. геол., 1969, № 2, с. 138—141.

Шиманский А. А., Базанов Г. А. О возможности использования распределения Вейбулла при решении геохимических задач. — В кн.: Математические методы геохимических исследований. М., 1966, с. 61—70.

Шиманский А. А., Учакин Ю. М., Базанов Г. А. и др. Использование функции распределения при решении некоторых поисково-разведочных задач. — В кн.: Математические методы в поисково-разведочной практике. Иркутск, 1970, с. 197—229.

Шоу Д. М. О делении данных в аналитической геохимии на две группы с помощью дистанционного коэффициента. — В кн.: Вопросы математической геологии. Л., 1968, с. 98—110.

Шоу Д. М. Геохимия микроэлементов кристаллических пород. Перев. с франц. Л., 1969, 207 с.

Шторм Р. Теория вероятностей. Математическая статистика. Статистический контроль качества. Перев. с нем. М., 1970, 368 с.

Юдович Я. Э. О распределении зольности в углах. — Вестн. МГУ, 1964, сер. 4, № 3, с. 101—104.

Юдович Я. Э., Шасткевич Ю. Г. Зольность углей и содержание в них редких элементов. Изв. вузов, Геол. и разведка, 1966, № 9, с. 68—76.

Юдович Я. Э., Гольдберг Ю. И., Юшкова Г. Е., Иванова Т. И., Соколов М. Б. Полуколичественный спектральный анализ в геологических целях. — Литол. и полезн. ископ., 1970, № 5, с. 131—142.

Юдович Я. Э., Корычева А. А., Обручников А. С., Степанов Ю. В. Средние содержания элементов-примесей в ископаемых углях. — Геохимия, 1972, № 8, с. 1023—1031.

Юл Д. Э., Кендэл М. Д. Теория статистики. М., 1960, 779 с.

Юфа Б. Я. Учет систематических ошибок при подсчете запасов. — Разведка недр, 1951, № 6, с. 23—28.

Юфа Б. Я. Замечания по статье Шарапова И. П. «О контрольных анализах геологических проб (Разв. и охрана недр, 1954, № 1)». — Разв. и охрана недр, 1954, № 5, с. 60—63.

Юфа Б. Я. Об ошибках В. Л. Шашкина в решении некоторых вопросов контроля результатов опробования. — Тр. Инст. геол. АН Киргизск. ССР, 1958, вып. 10, с. 193—200.

Юфа Б. Я., Гурвич Ю. М. Применение медианы и квартилей для оценки нормальных и аномальных значений геохимического поля. — Геохимия, 1964, № 8, с. 817—824.

Яцко Я. Математико-статистические таблицы. М., 1961, 243 с.

Adams J. A. S., Weaver C. E. Thorium-to-uranium ratios as indicators of sedimentary processes: example of concept of geochemical facies. — Bull. Amer. Ass. Petrol. Geol., 1958, vol. 42, N 2, p. 387—430.

Ahrens L. H. A fundamental law of geochemistry. — Nature, 1953, vol. 172, N 4390.

Ahrens L. H. The lognormal distribution of elements. (1—2). — Geochim. Cosmochim. Acta, 1954, vol. 5, p. 49—73; vol. 6, p. 121—131.

Ahrens L. H. Lognormal type distribution. (3) — Geochim. Cosmochim. Acta, 1957, vol. 11, № 4, p. 205—212.

Aitchison J., Brown I. A. The lognormal distribution. Cambridge Univ. Press., 1957, 176 p.

Aubrey K. V. Frequency-distribution of elements in igneous rocks. — Geochim. Cosmoshim. Acta, 1956, vol. 9, № 1, p. 83—89.

Barth T. F. W. Abundances of elements, areal averages and geochemical cycles. — Geochim. Cosmochim. Acta, 1961, vol. 23, № 1, p. 1—8.

Barth T. F. W. Theoretical petrology. — N. Y., 1962, 416 p.

Bartlett M. S. Fitting a strait line when both variables are subjekt to error. — Biometrics, 1949, vol. 5, № 9, p. 207—212.

Berkson J. Are there two regressions? — Journ Am. Stat. Ass., 1950, vol. 45, № 25, p. 164—180.

Cameron E. M. Studies of the frequency distribution of the ore elements in certain rock units the Canadian shield. Projekt 690050. — Geol. Surv. Canada, 1971, Paper, pt. A, p. 68—69.

Chayes F. The lognormal distribution of elements: a discussion. — Geochim. Cosmochim. Acta, 1954, vol. 6, № 2—3, p. 119—120.

Chayes F. A petrographic criterion for the possible replacemant of rocks. — Amer. Journ. Sci., 1948, vol. 246, № 7.

Chayes F. On correlation between variables of constant sum. — Journ. Geophys. Res., 1960, vol. 65, p. 4185—4193.

Chayes F. Numerical correlation and pterographic variation. — Journ. Geol., 1962, vol. 70, № 4, p. 440—4528.

Chirvinsky P. N. Petrographische Värhältnisse der Karbonsedimentgesteine des Donetzbeckens in Russland. — Zeitschr. Deutsch. Gev. Ges., 1925, Bd. 77, Hf. 1.

Clarke F. W. The relative abundance of the chemical elements. Phil. Soc. Washington Bull., 1889, vol. 11, p. 135.

Clarke F. W. The data of geochemistry. — U. S. Geol. Survey Bull., 1908, N 330, 716 p.

Clarke F. W. The data of geochemistry. — U. S. Geol. Survey Bull., 1924, № 770, 841 p.

Clarke F. W., Washington H. S. The composition of the earth's crust. — U. S. Geol. Survey Prof. Paper, 1924, № 127, 117 p.

Daly R. A. Igneous rocks and their origin. N. Y., 1914, 563 p.

Eade K. E., Fahrig W. F. Geochemical evolutionary trends of continental plates — a preliminary study of the Canadian Shield. — Geol. Surv. Canada Bull., 1971, № 179, 51 p.

Fairbairn H. W., Schlecht W. G. et al. A cooperative investigation of precision and accuracy in chemical spectrochemical and modal analysis of silicate rocks. — U. S. Geol. Survey Bull., 1951, № 980, 71 p.

Goldberg E. D. The oceans as a chemical system. — In: The Sea, vol 2, N. Y.—London, 1963, p. 3—25.

Goldschmidt V. M. Grundlagender quantitativen Geochemie. —

Fortschr. Mineralogie, Kristallographie, Petrographie, 1933, Bd. 17, S. 112—156.
Hazen S. W. Jr., Meyer W. L. Using probability models as a basis for making decisions during minerals deposit exploration. — U. S. Bur. Mines Rept. Inv., 1966, № 6778, 83 p.
Horn M. K., Adams J. A. S. Computer-derived geochemical balances and element abundances. — Geochim. Cosmochim. Acta, 1966, vol. 30, p. 279—297.
Knopf A. The composition of the average igneous rock. — Journ. Geology, 1916, vol. 24, p. 620—622.
Koch J. S., Link J. F. Statistical analysis of geological data. New York—London—Sydney—Toronto, 1970, 375 p.
Krauskopf K. B. Sedimentary deposits of rare metals. — Econ. Geol., 1955, 50-th Ann. Vol. p. 413—463.
Kuenen P. H. Ceochemical calculation concerning the total mass of sediments in the Earth. — Amer. Journ. Sci., 1941, vol. 239, p. 161.
Lindley D. V. Estimation of a functional relationship. — Biometrica, 1953, vol. 40, p. 1—2.
Mead W. J. Redistribution of elements in the formation of sedimantary rocks. — Journ. geol., 1907, vol. 15.
Mead W. J. The average igneous rock. — Journ. Geology, 1914, vol. 22, p. 772—781.
Mercury in the environment. — U. S. Geol. Survey Profess. Paper, 1970, № 713, p. 53.
Miller R. L., Goldberg E. D. The normal distribution in geochemistry. — Geochim. Cosmochim. Acta, 1955, vol. 8, № 1/2, p. 53—62.
Parker R. L. Composition of the Earth's crust. — U. S. Geol. Survey Profess. Paper, 1967, № 440-D, 17 p.
Pearson K. On lines and planes of closest fit to systems of points in space. — Phil. Mag., 1901, Ser. 6, vol. 2, N 11, p. 559—572.
Pearson K., Lee A. On the laws of inheritance in man. — Biometrica, 1903, vol. 2, p. 4.
Poldervaart A. Chemistry of the Earth's crust. — Geol. Soc. America, Special paper, 1955, vol. 62, p. 119—144.
Rodgers J. J. W., Adams J. A. S. Lognormality of thorium concentrations in the Conway granite. — Geochim. Cosmochim. Acta, 1963, vol. 27, № 7, p. 775—783.
Sederholm J. J. The average composition of the earth's crust in Finland. — Finlande Comm. Geol. Bull., 1925, № 70, 20 p.
Sichel H. S. — New methods in the statistical evaluation of mine sampling data. — Trans. Inst. Min. Metall., London, 1951, vol. 61, p. 261—288; Discussion: p. 391, 463, 501.
Taylor S. R. Abundance of chemical elements in the continental crust: a new table. — Geochim. Cosmochim. Acta, 1964, vol. 28, N 8, p. 1273—1275.
Taylor S. R. Abundance of chemical elements in the continental crust—amended basaltic rare-earth patterns. — Geochim. Cosmochim. Acta, 1961, vol. 29, p. 145—146.
Turekian K. K., Wedepohl K. H. Distribution of the elements in some major units of the earth crust. — Geol. Soc. America Bull., 1961, vol. 71, № 2, p. 175—192.
Vernadsky W. I. Geochemic. Paris, 1924.
Vistelius A. B. The skew frequency distribution and the fundamental low of the geochemical processes. — Journ. Geol., 1960, vol. 68, N 1, p. 1—22.
Vogt J. H. L. On the average composition of the earth's crust with particular reference to the contents of phosphoric and titanic acid. — Norsk.-Akad. Skr., Mat.-Naturw. Kl., 1931, N 7, p. 1—48.
Wald A. The fitting of strait lines if both variable are subject to error. — Annals of Mathem. Statistics, 1940, vol. 11, p. 284—300.

ОГЛАВЛЕНИЕ

Юрий Андреевич Ткачев,
Яков Эльевич Юдович

СТАТИСТИЧЕСКАЯ ОБРАБОТКА
ГЕОХИМИЧЕСКИХ ДАННЫХ
Методы и проблемы

Утверждено к печати
Президиумом Коми филиала АН СССР

Редактор издательства *Т. П. Жукова*
Художник *Я. В. Таубвурцель*
Технический редактор *М. Э. Карлайтис*
Корректоры *Г. А. Александрова, С. В. Добрянская* и *Н. П. Яковлева*

Сдано в набор 9/XII 1974 г. Подписано к печати 19/VI 1975 г. Формат бумаги 60×90$^{1}/_{16}$. Бумага № 2. Печ. л. 14$^{3}/_{4}$=14.75 усл. печ. л. Уч.-изд. л. 15.04. Изд. № 5839. Тип. зак. № 1661. М-31671. Тираж 1900. Цена 90 коп.

Ленинградское отделение издательства «Наука»
199164, Ленинград, В-164, Менделеевская линия, д. 1

1-я тип. издательства «Наука».
199034, Ленинград, В-34, 9 линия, д. 12

ИЗДАТЕЛЬСТВО «НАУКА»

В магазинах конторы «Академкнига» имеются в наличии книги:

Борисенко Л. Ф. Скандий. Основные черты геохимии, минералогии и генетические типы месторождений. 1961. 130 стр. 55 к.

Вистелиус А. В. Красноцветные отложения п-ва Челекен. Литология. Опыт стохастического моделирования процессов слоеобразования. 1966. 303 стр. 2 р. 47 к.

Геология, минералогия и геохимия Комсомольского рудного района. 1971. 335 стр. 3 р. 24 к.

Гуляева Л. А. и др. Геохимия доманиковых отложений Волго-Уральской области. 1961. 104 стр. 60 к.

Дорфман М. Д. Минералогия пегматитов и зон выветривания в ийолит-уртитах горы Юкспор Хибинского массива. 1962. 168 стр. 1 р. 20 к.

Зеленова О. И. Литология, фации и геохимические особенности отложений Алтайского яруса Таджикской депрессии. Труды Ин-та геол. рудн. месторожд., петрогр., минерал. и геохимии. Вып. 53. 1961. 128 стр. 1 р. 03 к.

Иванов Н. С. Моделирование тепловых процессов в горных породах. 1972. 138 стр. 85 к.

Кепежинскас К. Б., Хлестов В. В. Статистический анализ минералов группы эпидота и их парагенетические типы. (Труды Ин-та геологии и геофизики. Вып. 103). 1971. 310 стр. 2 р. 10 к.

Левенсон В. Э. Геохимическая битуминология и ее проблемы. Том I. 1960. 192 стр. 1 р. 25 к.

Моисеенко В. Г. Геохимические особенности распределения в породах Тихоокеанского пояса. 1971. 199 стр. 1 р. 26 к.

Применение методов физической химии в петрологии и геохимии. 1972. 216 стр. 1 р. 47 к.

ИЗДАТЕЛЬСТВО «НАУКА»

По плану на 1975 год готовятся к выпуску книги:

Миронов Ю. П. Теоретико-множественные модели гранитоидов. 15 л. 1 р. 50 к.

В работе приводятся доказательства происхождения изверженных пород габбро-гранитного ряда. Каждая минеральная ассоциация исследуется как формальный язык с использованием аппарата теории множеств, в основном аналитической (алгебраической) лингвистики и теории графов. Картирование пород и их математических параметров позволяет выявлять структуры эпохи активизации: сквозные и сводовокупольные структуры, имеющие важное металлогеническое значение.

Бондаренко П. М. Моделирование надвиговых дислокаций в складчатых областях. 12 л. 1 р. 20 к.

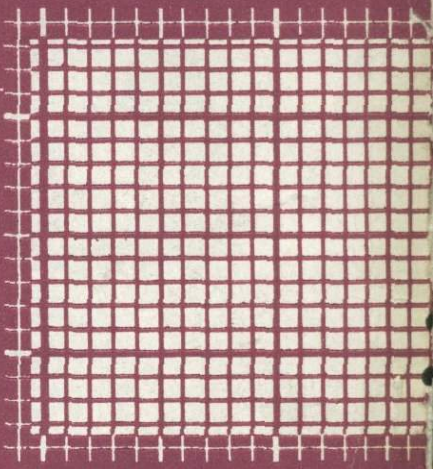
На основе структурных исследований выяснены главные черты морфологии Акташского надвига и разработано представление о механизме его образования вследствие продольного сжатия и изгиба слоистых толщ. Реальность предполагаемого механизма подтверждена данными экспериментов по моделированию деформаций в условиях продольного сжатия и изгиба слоистых толщ с применением поляризационно-оптического исследования напряжений.

ЗАКАЗЫ ПРОСИМ ПРИСЫЛАТЬ ПО АДРЕСУ:

117464, Москва, Мичуринский проспект, дом № 12
Магазин «Книга — почтой»
Центральной конторы «Академкнига»

197110, Ленинград, Петрозаводская улица, дом № 7
Магазин «Книга — почтой»
Северо-Западной конторы «Академкнига»

90 коп.



ИЗДАТЕЛЬСТВО
«НАУКА»
ЛЕНИНГРАДСКОЕ
ОТДЕЛЕНИЕ