Н. Н. ЖУКОВ

Вероятностно-статистические методы анализа геолого-геофизической информации

Н. Н. ЖУКОВ

ВЕРОЯТНОСТНО-СТАТИСТИЧЕСКИЕ МЕТОДЫ АНАЛИЗА ГЕОЛОГО-ГЕОФИЗИЧЕСКОЙ ИНФОРМАЦИИ

Допущено
Министерством высшего и среднего
специального образования УССР
в качестве учебного пособия
для студентов
геологических специальностей вузов

ИЗДАТЕЛЬСКОЕ ОБЪЕДИНЕНИЕ «ВИЩА ШКОЛА»
ГОЛОВНОЕ ИЗДАТЕЛЬСТВО
КИЕВ — 1975

55
Ж86

УДК 55:51

Вероятностно-статистические методы анализа геолого-геофизической информации. Жуков Н. Н. Издательское объединение «Вища школа». 1975, с. 304.

В пособии изложены теоретические основы вероятностно-статистического анализа геолого-геофизической информации, описаны практические приемы и методы решения типовых задач, возникающих при геологических, геофизических и геохимических исследованиях, приводятся примеры их решения. Изложены общепринятые методы применения математической статистики в геологии, а также результаты исследований автора. Пособие рассчитано на студентов геологических специальностей вузов. Им могут воспользоваться специалисты, занимающиеся обработкой геолого-геофизической информации.

Ил. 46. Табл. 21. Библиогр. 28.

Редакция литературы по биологии и географии
Зав. редакцией *Ю. Ф. Кирьяков*

Ж $\frac{20804—194}{M211(04)—75}$ 183—75

В числе первоочередных задач развития научных исследований решениями XXIV съезда КПСС предусматривается дальнейшая разработка проблем теоретической и прикладной математики для более широкого применения математических методов и электронно-вычислительной техники в народном хозяйстве. В геологических исследованиях на современном этапе математические методы занимают все более значительное место.

Геологические объекты, сформировавшиеся в разнообразных физико-химических и термодинамических условиях в течение значительных промежутков времени, характеризуются параметрами, значения которых являются результатом действия сложных, растянутых во времени вероятностных процессов. Такие параметры подвержены влиянию случая, поэтому для их изучения должны применяться вероятностно-статистические методы. В первую очередь это относится к физико-химическим характеристикам пород. Так, концентрации химических элементов в отдельных точках геологических тел — результат совокупного воздействия многих факторов: условий первичной кристаллизации, эндогенных и экзогенных процессов и др., влияние каждого из которых точно учесть невозможно. В лучшем случае, можно выделить основные причины и следствия, определяющие распределение вещества в пределах геологического тела, но и здесь случайным факторам принадлежит немалая роль. Это связано с тем, что основные процессы распределения вещества реализуются не иначе, как вместе с взаимодействием многих второстепенных, так что в конечном счете они проявляют себя лишь в виде тенденции на фоне случайных флуктуаций.

Значительную роль играют и случайные факторы, действующие непосредственно в процессе получения количественной информации: при отборе проб, измерениях, количественном анализе результатов.

Внедрение вероятностно-статистических методов в современные геолого-геофизические исследования обусловливается также и возрастающим увеличением объемов получаемой информации. Полноценное и систематическое применение этих, как и вообще любых математических методов, для ее анализа и обобщения предусматривает прежде всего широкое использование электронно-вычислительной техники, обеспечивающей оперативность и высокую производительность труда при массовой количественной обработке материалов.

В последние годы появилось много работ, посвященных применению вероятностно-статистических методов в геологии [3, 10, 11, 12, 13, 15, 18, 19, 20, 21, 22, 23, 24, 26] и многие другие. В них показана эффективность применения этих методов, разработан ряд вопросов, связанных с их использованием при решении разнообразных геологических задач. Кроме математической статистики все более широко применяются методы разделов математики, развившихся из теории вероятностей: теория игр и статистических решений, теория случайных функций и случайных полей, теория информации, математическое моделирование (на базе стохастических моделей). Насущная задача современности — овладение основным математическим аппаратом вероятностно-статистического анализа всеми специалистами геологической службы.

В первых пяти главах книги даны основные сведения из теории вероятностей и математической статистики, необходимые при изложении последующего материала. В последующих главах содержится описание методов, непосредственно применяемых в процессе количественного анализа геолого-геофизических данных, в объеме, предусмотренном программами. Изложение сопровождается примерами.

Автор благодарит проф. А. Г. Тархова, доц. А. А. Никитина, докт. техн. наук А. Е. Кулинковича за большую помощь в работе над книгой, оказанную в виде ценных указаний и замечаний.

Отзывы, замечания и пожелания будут приняты с благодарностью. Просим направлять их по адресу: г. Киев, ГСП-127, Киевский госуниверситет, геологический факультет, кафедра геофизики.

ОСНОВНЫЕ ПОНЯТИЯ ТЕОРИИ ВЕРОЯТНОСТЕЙ

В этой главе изложены основные понятия теории вероятностей и некоторые дополнительные сведения, необходимые для дальнейшего изложения. Более подробную информацию по этим вопросам можно получить из [5, 7, 9].

§ 1. Вероятность и ее свойства

События. Вероятность. В повседневной практике мы часто сталкиваемся со *случайными событиями*. В отличие от *достоверных* событий, которые являются обязательным следствием определенных условий опыта, случайные события таким детерминированным следствием условий не являются. Иначе говоря, случайные события в заданных условиях могут произойти или не произойти. Кроме достоверных и случайный событий выделяют еще *невозможные* события, которые в данных условиях не могут произойти.

Рассмотрим одну из главных закономерностей, которым подчиняются изучаемые теорией вероятностей случайные события. Если многократно воспроизводить одни и те же условия испытания (опыта) и каждый раз отмечать результат — появление или отсутствие определенного события, то обнаруживается такая тенденция: частота появлений события (отношение количества испытаний, при которых происходило событие, к общему числу испытаний) стремится к вполне определенной величине. Так, при повторении большого числа опытов с подбрасыванием монеты частота выпадения ее «гербом» близка к 0,5.

Предел, к которому стремится частота появлений события при увеличении количества испытаний, называется *вероятностью* события. Вероятность служит естественной мерой возможности появления события: чем больше вероятность, тем более вероятно, что событие произойдет. Таким образом, это свойство событий можно количественно измерить, анализировать и, в конечном счете, вполне строго научно изучать.

Приведенное определение вероятности, носящее название *статистического*, выглядит односторонним, неконструктивным. Следуя ему, нельзя определить вероятность события без проведения большого количества испытаний. Известен иной подход к определению вероятности — так называемое *классическое* определение, вводимое

в том случае, когда опыт имеет конечное число равновозможных элементарных исходов, а интересующее нас событие происходит при появлении каких-либо из них. Вероятность $\mathbf{P}(A)$ события A определяется как отношение числа N_A сопутствующих ему исходов к общему количеству всех возможных исходов N: $\mathbf{P}(A) = \frac{N_A}{N}$. Например, если урна содержит n белых и m черных шаров, то вероятность того, что извлеченный из урны наугад шар будет белым, $\mathbf{P}(A) = \frac{n}{m+n}$ ($N_A = n$, $N = m + n$). Вероятность того, что подброшенная монета выпадет «гербом» — 0,5 ($N_A = 1$, $N = 2$).

Установлено, что для тех событий, для которых допускается такое определение вероятности, частота появлений стремится с увеличением числа испытаний к вероятности в ее классическом определении. Иными словами, вероятности в статистическом и классическом определениях совпадают в тех случаях, когда ввести последнее возможно.

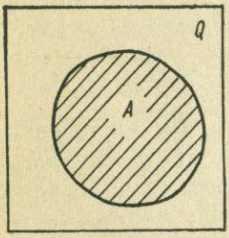


Рис. 1.

Вероятность допускает простую геометрическую интерпретацию — так называемую *геометрическую* вероятность, которой обобщается классическое определение на случай бесконечного множества исходов. Пусть на некоторую ограниченную плоскую область Q площадью S падает брошенная наугад точка (малое материальное тело) так, что все положения ее падения на Q равновозможны. Если определить событие как попадание точки в некоторую часть A области Q, то вероятность его $\mathbf{P}(A)$ пропорциональна площади этой части $S(A)$ и определяется как $\mathbf{P}(A) = \frac{S(A)}{S}$ (рис. 1).

Подобная интерпретация возможна также в одномерном случае — если бросать точку на отрезок прямой, а под событиями понимать попадание ее в различные множества точек этого отрезка, и в трехмерном — если бросать точку в некоторый объем V, где она может оказаться с равными возможностями в произвольном месте. В последнем случае вероятность попадания в некоторую часть A этого объема определяется как $\mathbf{P}(A) = \frac{V(A)}{V}$, где $V(A)$ — объем A.

Свойства вероятности. Событие, состоящее в том, что одновременно происходят события A и B, называют *произведением* (иногда *пересечением*) этих событий и обозначают AB или $A \cap B$ (рис. 2, *а*). Если AB — невозможное событие, т. е. может произойти только одно из них, то A и B называют *несовместными* (рис. 2, *б*).

Событие, состоящее в том, что происходит хотя бы одно из двух событий — A или B, называют *суммой* (иногда *объединением*) этих событий и обозначают $A + B$ или $A \cup B$. *Противоположным* событию A называется событие, заключающееся в том, что не происходит A. Такое событие обозначается $\bar{A}$. *Разностью* двух событий A

и B, обозначаемой $A \setminus B$, является событие, состоящее в том, что одновременно происходят A и $\bar{B}$ (рис. 3). Например, рассмотрим операции над такими событиями: A состоит в том, что образец отобран из гранитов, B — образец отобран из измененных пород. Результатам операций будут соответствовать: AB — «измененный гранит»; $A \cup B$ — «или гранит, или измененная порода, или измененный гранит»; $\bar{A}$ — «не гранит»; $A \setminus B = A\bar{B}$ — «неизмененный гранит».

О событиях $A_1, A_2, \ldots, A_k$ говорят, что они образуют *полную группу событий*, если A_i попарно несовместны и в сумме образуют достоверное событие. Иными словами, какое-либо событие из полной группы обязательно происходит, причем только одно. Обозначив U достоверное событие, V — невозможное, имеем: $A_i A_j = \mathrm{V}$ $(i \neq j)$, $A_1 \cup A_2 \cup \ldots \cup A_k = \mathrm{U}$. Полную группу, например, составляют A и $\bar{A}$.

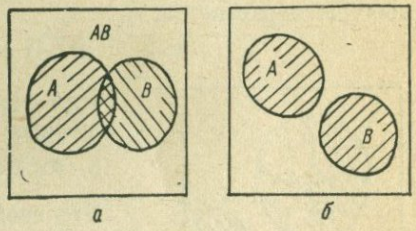


Рис. 2.

Вероятность обладает следующими свойствами, часть из которых очевидна, а в справедливости других можно убедиться, пользуясь геометрической интерпретацией (рис. 3).

1. Вероятность достоверного события равна единице, а невозможного — нулю:

$$\mathbf{P}(\mathrm{U}) = 1, \quad \mathbf{P}(\mathrm{V}) = 0. \tag{1.1}$$

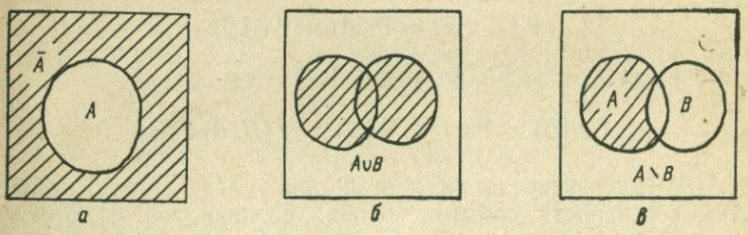


Рис. 3.

2. Вероятность всегда — неотрицательное число, не превышающее единицы: $0 \leqslant \mathbf{P}(A) \leqslant 1$.

3. Если события $A_1, A_2, \ldots, A_k$ несовместны, то вероятность их суммы равна сумме вероятностей каждого:

$$\mathbf{P}(A_1 \cup A_2 \cup \ldots \cup A_k) = \sum_{i=1}^{k} \mathbf{P}(A_i). \tag{1.2}$$

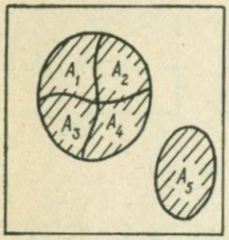
На рис. 4 изображено событие $A = A_1 \cup A_2 \cup \ldots \cup A_k$. По геометрической интерпретации вероятности

$$\mathbf{P}(A_1 \cup A_2 \cup \cdots \cup A_k) = \frac{S(A)}{S} = \frac{1}{S}[S(A_1) + S(A_2) + \cdots + S(A_k)] =$$
$$= \mathbf{P}(A_1) + \mathbf{P}(A_2) + \cdots + \mathbf{P}(A_k).$$

Из (1.2) следует, что, если $A_1, A_2, \ldots, A_k$ — полная группа событий, сумма их вероятностей равна единице: $\sum_{i=1}^{k} \mathbf{P}(A_i) = \mathbf{P}(A_1 \cup A_2 \cup \cdots \cup A_k) = 1$. В частности, $\mathbf{P}(A) + \mathbf{P}(\bar{A}) = 1$.

4. Вероятность противоположного события, как следует из предыдущего равенства,

$$\mathbf{P}(\bar{A}) = 1 - \mathbf{P}(A). \tag{1.3}$$



$\bigcup_{i=1}^{n} A_i; A_i A_j = V$

Рис. 4.

Для любых двух событий A и B справедливы следующие соотношения.

5. $$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(AB). \tag{1.4}$$

Это свойство, именуемое *формулой сложения вероятностей*, следует из представления события $A \cup B$ в виде суммы двух несовместных событий (рис. 3, *б*) — $A \cup B = (A \setminus B) \cup B$, так что по (1.2)

$$\mathbf{P}(A \cup B) = \mathbf{P}(A \setminus B) + \mathbf{P}(B). \tag{1.5}$$

Так как $A = (A \setminus B) \cup AB$ (рис. 2, *а*) и оба слагаемых несовместны,

$$\mathbf{P}(A) = \mathbf{P}(A \setminus B) + \mathbf{P}(AB), \tag{1.6}$$

откуда

$$\mathbf{P}(A \setminus B) = \mathbf{P}(A) - \mathbf{P}(AB) \tag{1.7}$$

и, подставив (1.7) в (1.5), получим (1.4).

6. $$\mathbf{P}(AB) = \mathbf{P}(A/B)\,\mathbf{P}(B) = \mathbf{P}(B/A)\,\mathbf{P}(A), \tag{1.8}$$

где $\mathbf{P}(A/B)$ — вероятность события A при условии, что B произошло; $\mathbf{P}(B/A)$ — вероятность события B при условии, что A произошло. Вероятности $\mathbf{P}(A/B)$, $\mathbf{P}(B/A)$ называются *условными*. Формула (1.8) известна под названием *формулы умножения вероятностей*. Например, если B — попадание пункта опробования в зону рудоносности, A — наличие некоторого признака, то $\mathbf{P}(B/A)$ характеризует его эффективность как признака рудоносности, $\mathbf{P}(A/B)$ — распространенность его в зоне рудоносности, и они вычисляются в виде $\mathbf{P}(A/B) = \frac{\mathbf{P}(AB)}{\mathbf{P}(B)}$, $\mathbf{P}(B/A) = \frac{\mathbf{P}(AB)}{\mathbf{P}(A)}$.

В справедливости формулы (1.8) нетрудно убедиться, пользуясь геометрической интерпретацией вероятности:

$$\mathbf{P}(AB) = \frac{S(AB)}{S} = \frac{S(AB)}{S(A)} \frac{S(A)}{S} = \mathbf{P}(B/A)\,\mathbf{P}(A), \tag{1.9}$$

так как $\mathbf{P}(B/A)$ — отношение площади той части B, которая принадлежит и A, т. е. площади пересечения AB, к площади A: $\mathbf{P}(B/A) = \frac{S(AB)}{S(A)}$. Если

$$\mathbf{P}(AB) = \mathbf{P}(A)\,\mathbf{P}(B), \tag{1.10}$$

(тогда по (1.9) $\mathbf{P}(A/B) = \mathbf{P}(A)$, $\mathbf{P}(B/A) = \mathbf{P}(B)$), события A и B называются *независимыми*.

Если $A_1, A_2, \ldots, A_k$ попарно несовместны, то $BA_1, BA_2, \ldots, BA_k$ тоже несовместны и

$$\mathbf{P}\{(A_1 \cup A_2 \cup \cdots \cup A_k)/B\} = \frac{\mathbf{P}\{(A_1 \cup A_2 \cup \cdots \cup A_k)B\}}{\mathbf{P}(B)} =$$

$$= \frac{\sum_{i=1}^{k} \mathbf{P}(A_i B)}{\mathbf{P}(B)} = \sum_{i=1}^{k} \mathbf{P}(A_i/B). \tag{1.11}$$

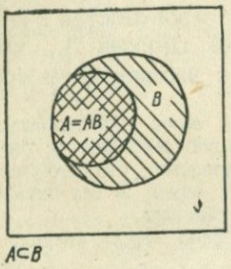


Рис. 5.

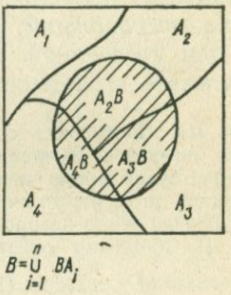


Рис. 6.

7. Если $B \supset A$, т. е. B включает A (когда происходит A, всегда происходит и B), то $AB = A$ и $\mathbf{P}(B) \geqslant \mathbf{P}(A)$. Это видно из того, что при $AB = A$ (рис. 5)

$$\mathbf{P}(A) = \mathbf{P}(AB) = \mathbf{P}(B)\,\mathbf{P}(A/B) \leqslant \mathbf{P}(B). \tag{1.12}$$

8. Если $A_1, A_2, \ldots, A_k$ — полная группа событий, то для любого события B справедливо соотношение

$$\mathbf{P}(B) = \sum_{i=1}^{k} \mathbf{P}(B/A_i)\,\mathbf{P}(A_i), \tag{1.13}$$

которое называется *формулой полной вероятности*. Событие B представляется в виде суммы несовместных событий (рис. 6): $B = BA_1 \cup \cup BA_2 \cup \cdots \cup BA_k$. По свойству (1.2) и (1.8) $\mathbf{P}(B) = \sum_{i=1}^{k} \mathbf{P}(BA_i) = \sum_{i=1}^{k} \mathbf{P}(A_i)\,\mathbf{P}(B/A_i)$.

9. Пусть $A_1, A_2, \ldots, A_k$ — полная группа событий, так что событие B может произойти не иначе, как с одним из них. Если B

произошло, это некоторым образом уточняет информацию о вероятностях событий $A_1, A_2, \ldots, A_k$. Из рис. 6, например, видно, что A_1 не могло произойти, если произошло B. События $A_1, A_2, \ldots, A_k$ естественно назвать *гипотезами* по отношению к B. *Формула Бейеса* и даёт выражения для вероятностей гипотез A_i $(i = 1, 2, \ldots, k)$ при условии, что B произошло:

$$\mathbf{P}(A_i/B) = \frac{\mathbf{P}(A_i)\,\mathbf{P}(B/A_i)}{\sum_{j=1}^{k} \mathbf{P}(A_j)\,\mathbf{P}(B/A_j)}. \qquad (1.14)$$

Эта формула непосредственно следует из формулы полной вероятности (1.13) и формулы умножения вероятностей (1.8): $\mathbf{P}(A_iB) = \mathbf{P}(A_i/B)\,\mathbf{P}(B) = \mathbf{P}(B/A_i)\,\mathbf{P}(A_i)$. Отсюда $\mathbf{P}(A_i/B) = \mathbf{P}(B/A_i)\,\mathbf{P}(A_i) \times (\mathbf{P}(B))^{-1}$, а подставив вместо $\mathbf{P}(B)$ (1.13), получим (1.14).

Формула Бейеса еще именуется *формулой вероятностей гипотез*. Входящие в неё вероятности $\mathbf{P}(A_i)$ называют *априорными*, $\mathbf{P}(A_i/B)$ — *апостериорными*. Переход от вероятностей $\mathbf{P}(A_i)$ к $\mathbf{P}(A_i/B)$ и является тем уточнением вероятностей гипотез A_i, которое получено в результате проведенного опыта с наблюдением события B.

Пример 1.1. Протолочка содержит 65% зерен плагиоклаза, 25% кварца и 10% других минералов. Вычислить: 1) условную вероятность того, что наугад отобранное зерно будет зерном кварца, если известно, что это зерно не плагиоклаза; вероятность того, что из пяти отобранных зерен: 2) все пять будут зернами плагиоклаза; 3) одно зерно кварца, четыре плагиоклаза.

Решение. 1) Обозначим событие A — отобрано зерно кварца, B — отобрано зерно не плагиоклаза. Согласно (1.8), условная вероятность $\mathbf{P}(A/B) = \frac{\mathbf{P}(AB)}{\mathbf{P}(B)} = \frac{\mathbf{P}(A)}{\mathbf{P}(B)} = \frac{0{,}25}{0{,}35} \approx 0{,}714$. $AB = A$ ввиду того, что $A \subset B$.

2) Обозначим событие A_i — i-м отобрано зерно плагиоклаза ($i = \overline{1,5}$ *). Вероятность того, что отобрано пять зерен плагиоклаза, учитывая независимость A_i между собой — $\mathbf{P}(A_1A_2A_3A_4A_5) = \prod_{i=1}^{5} \mathbf{P}(A_i) = 0{,}65^5 \approx 0{,}116$.

3) Событие B_i — i-м отобрано зерно кварца ($i \leqslant 5$), остальные четыре — зёрна плагиоклаза. Событие D — одно зерно кварца, четыре плагиоклаза, состоит из суммы событий B_i. Поскольку B_i несовместны между собой, по формуле (1.2) имеем: $\mathbf{P}(D) = \sum_{i=1}^{5} \mathbf{P}(B_i)$. B_1 состоит из произведения событий C_1, A_2, A_3, A_4, A_5; C_1 — первым отобрано зерно кварца. $\mathbf{P}(C_1) = 0{,}25$ и так как C_1, A_2, A_3, A_4, A_5 независимы, $\mathbf{P}(B_1) = \mathbf{P}(C_1)\,\mathbf{P}(A_2)\,\mathbf{P}(A_3)\,\mathbf{P}(A_4)\,\mathbf{P}(A_5) = 0{,}25 \cdot 0{,}65^4 \approx 0{,}045$. Аналогично, $\mathbf{P}(B_i) = 0{,}25 \cdot 0{,}65^4$. Таким образом, $\mathbf{P}(D) = 5 \cdot 0{,}25 \cdot 0{,}65^4 \approx 0{,}223$.

Пример 1.2. Объект исследований разделен на n одинаковых блоков и содержит m малых рудных тел ($m < n$). Считая, что положение одного тела не влияет на размещение остальных, вычислить вероятность того, что: 1) все m тел размещаются в одном определенном блоке; 2) все m тел размещаются в каком-либо одном блоке; 3) все тела размещаются в разных блоках (каждый блок или не имеет ни одного тела или содержит только одно); 4) все тела размещаются в определённой части объекта, которая состоит из k блоков; 5) хотя бы одно тело находится в данном блоке.

---

* Здесь и дальше $i = \overline{1, n}$ обозначает: $i = 1, 2, \ldots, n$.

Решение. 1) Пронумеруем блоки — $i = 1, 2, \ldots, n$ и тела — $j = 1, 2, \ldots, m$. Событие B_i, которое нас интересует — размещение всех тел в i-м блоке — является произведением событий A_j — «j-е тело находится в i-м блоке». $\mathbf{P}(A_j) = \frac{1}{n}$ и поскольку A_j независимы. $\mathbf{P}(B_i) = \prod_{j=1}^{m} \mathbf{P}(A_j) = \left(\frac{1}{n}\right)^m$.

2) Событие B — «все тела размещаются в одном блоке» представляет собой сумму несовместных событий B_i. По формуле (1.2) имеем: $\mathbf{P}(B) = \sum_{i=1}^{n} \mathbf{P}(B_i) = n \frac{1}{n^m} = \frac{1}{n^{m-1}}$.

3) Подсчитаем количество всех вариантов размещения тел. Первое тело может находиться в одном из n блоков — n вариантов. При каждом положении первого тела второе может оказаться тоже в каждом из n блоков. Всего вариантов размещения двух тел — n^2. Аналогично, число вариантов размещения трех тел — n^3, m тел — n^m. Эти n^m вариантов составляют все возможные элементарные исходы.

Найдем теперь количество вариантов, когда тела размещаются в разных блоках. Для первого тела имеем n возможностей; для второго, при каждом полсжении первого, $n - 1$. Всего возможных размещений двух тел в разных блоках $n(n-1)$. Аналогично, количество размещений m тел по разным блокам $n(n-1) \ldots (n-m+1)$. Искомая вероятность, по классическому определению вероятности, $P = n(n-1) \ldots (n-m+1) \frac{1}{n^m} = \frac{n!}{m!\, n^m}$.

4) Вероятность расположения тела в части из k блоков $\frac{k}{n}$. Искомая вероятность $p = \left(\frac{k}{n}\right)^m$.

5) Для вычисления вероятности $\mathbf{P}(C)$ того, что в данном блоке находится хотя бы одно тело, найдем вероятность противоположного события $\mathbf{P}(\overline{C})$, т. е. вероятность того, что в данном блоке нет ни одного тела. Количество вариантов размещения m тел в остальных $n-1$ блоках, аналогично п. 3, будет $(n-1)^m$. $\mathbf{P}(\overline{C}) = \frac{(n-1)^m}{n^m}$. Искомая вероятность $\mathbf{P}(C) = 1 - \mathbf{P}(\overline{C}) = 1 - \frac{(n-1)^m}{n^m}$.

Пример 1.3. На территории распространены три типа пород: грейзены, грейзенизированные граниты и слабо измененные граниты (без признаков грейзенизации), которые занимают соответственно площади в 5, 20 и 75%. Аномально высокие содержания олова приурочены к этим породам с вероятностями, соответственно: 0,7; 0,3; 0,05. Вычислить вероятность того, что: 1) отобранная проба будет иметь аномально высокое содержание олова; 2) проба будет принадлежать к грейзенизированным гранитам или слабо измененным гранитам, если она не имеет аномально высокого содержания олова.

Решение. 1) Введем полную группу событий: A_1 — отбор пробы из грейзенов; A_2 — из грейзенизированных гранитов; A_3 — из слабо измененных гранитов. Очевидно, $\mathbf{P}(A_1) = 0{,}05$, $\mathbf{P}(A_2) = 0{,}20$, $\mathbf{P}(A_3) = 0{,}75$. Обозначим B — событие, состоящее в получении аномального содержания олова. По условию задачи $\mathbf{P}(B/A_1) = 0{,}7$; $\mathbf{P}(B/A_2) = 0{,}3$; $\mathbf{P}(B/A_3) = 0{,}05$. По формуле полной вероятности $\mathbf{P}(B) = \sum_{i=1}^{3} \mathbf{P}(A_i)\,\mathbf{P}(B/A_i) \approx 0{,}133$. Вероятность $\mathbf{P}(B)$ может служить своеобразной мерой перспективности территории в отношении рудоносности.

2) Вероятность того, что проба с содержанием олова, не являющимся аномальным, принадлежит к грейзенизированным гранитам, по формуле Бейеса будет

$$\mathbf{P}(A_2/\overline{B}) = \frac{\mathbf{P}(A_2)\,\mathbf{P}(\overline{B}/A_2)}{\sum_{i=1}^{3} \mathbf{P}(A_i)\,\mathbf{P}(\overline{B}/A_i)} = \frac{\mathbf{P}(A_2)\,\mathbf{P}(\overline{B}/A_2)}{\mathbf{P}(\overline{B})} = \frac{\mathbf{P}(A_2)\,\mathbf{P}(\overline{B}/A_2)}{1 - \mathbf{P}(B)} \approx \frac{0{,}2 \quad 0{,}7}{0{,}867} \approx 0{,}162.$$

Аналогично, $\mathbf{P}(A_3/\overline{B}) \approx 0{,}822$. Искомая вероятность — $\mathbf{P}\{(A_2 \cup A_3)/\overline{B}\} = \mathbf{P}(A_2/\overline{B}) + \mathbf{P}(A_3/\overline{B}) \approx 0{,}162 + 0{,}822 = 0{,}984$.

Пример 1.4. В районе исследований фундамент могут слагать породы трех типов: гипербазиты, граниты, гнейсы. По данным сейсмозондирования установлен нижний предел a для значений скорости продольных волн V_P с целью последующего выделения по ним гипербазитов. При сопоставлении результатов сейсмозондирования и опробования фундамента значения $V_P \geqslant a$ наблюдались: на гипербазитах — в 90, на гранитах — в 7, на гнейсах — в 5% случаев. На одном из участков проведено измерение V_P и результат оказался выше уровня a. Вычислить вероятность того, что фундамент участка сложен гипербазитами.

Решение. Введем полную группу событий: A_1 — фундамент участка гипербазитовый, A_2 — гранитный, A_3 — гнейсовый; событие B — измеренное значение V_P превышает a. Очевидно, $\mathbf{P}(B/A_1) = 0{,}9$; $\mathbf{P}(B/A_2) = 0{,}07$; $\mathbf{P}(B/A_3) = 0{,}05$. Так как заранее никакому типу пород нельзя отдать предпочтение, полагаем $\mathbf{P}(A_1) = \mathbf{P}(A_2) = \mathbf{P}(A_3) = \frac{1}{3}$.

По формуле Бейеса $\mathbf{P}(A_1/B) = \dfrac{\frac{1}{3} \cdot 0{,}9}{\frac{1}{3}(0{,}9 + 0{,}07 + 0{,}05)} \approx 0{,}88.$

§ 2. Случайная величина. Функция и плотность распределения

Случайные величины. *Случайной величиной* в теории вероятностей называется величина, которая при одних и тех же условиях может принимать различные числовые значения. Значения, которые получает случайная величина, называют ее *наблюдениями*.

С различными случайными величинами обычно имеют дело, когда исследуют количественные показатели объектов геологических исследований. Результаты измерений геофизических параметров, характеристик минерального и химического состава горных пород, их физических свойств в определенных условиях могут рассматриваться как наблюдения соответствующих случайных величин.

Среди случайных величин в математической статистике выделяют *непрерывные* и *дискретные*. Непрерывная случайная величина может принимать любое значение на числовой оси, интервала или группы интервалов на ней, причем вероятность попадания ее значения в произвольный интервал δ из области возможных значений стремится к нулю при $\delta \to 0$. Дискретные получают лишь вполне определенные, дискретные значения $x_1, x_2, \ldots, x_n, \ldots$ с вероятностями $p_1, p_2, \ldots, p_n, \ldots$. Возможные значения дискретной случайной величины могут составлять конечное или бесконечное множество; сумма вероятностей p_i равна единице.

Непрерывной случайной величиной можно считать результат всякого измерения с пренебрежимой, т. е. незначительной по сравнению с вариацией значений измеряемого показателя, дискретностью шкалы измерительного прибора. С дискретными случайными величинами можно связать, например, *качественные признаки*, имеющие два состояния — «есть», «нет», или несколько — «есть», «частично выражен»,

«нет» и т. д. Такое соответствие можно построить, пронумеровав состояния начиная с 0: 0, 1, 2, ... и приписав случайной величине значение i, если признак находится в i-м состоянии. Дискретной случайной величиной является и результат измерения прибором со значительной дискретностью шкалы.

Со случайными величинами можно производить алгебраические операции. Суммой двух случайных величин ξ и ζ является случайная величина, значения которой образуются сложением значений, получаемых при испытаниях величинами ξ и ζ. Например, если ξ — истинное значение показателя, Δ — ошибка его измерения, взятая со знаком (отклонение измеренного значения от истинного), то результат измерения η является случайной величиной вида:

$$\eta = \xi + \Delta. \tag{1.15}$$

Аналогично вводятся и другие операции — произведение, частное, разность и вообще функции от случайных величин. Функция $\eta = f(\xi_1, \xi_2, \ldots, \xi_k)$ случайных величин $\xi_1, \xi_2, \ldots, \xi_k$ определяется как случайная величина, принимающая значение $y = f(x_1, x_2, \ldots, x_k)$, если $\xi_1, \xi_2, \ldots, \xi_k$ получили при испытании значения соответственно $x_1, x_2, \ldots, x_k$.

Функция распределения и плотность распределения. Одним из основных понятий вероятностно-статистического анализа случайных величин является *функция распределения*. Функцией распределения случайной величины ξ называется функция одной переменной x, имеющая смысл:

$$F(x) = \mathbf{P}\{\xi < x\}, \tag{1.16}$$

$\mathbf{P}\{\xi < x\}$ — вероятность того, что ξ получит значение, меньшее x. Переменная x получает все возможные значения на числовой оси.

Функция распределения произвольного количественного показателя ξ пород, слагающих геологический объект объемом V, имеет простое содержание:

$$F(x) = \frac{V\{\xi < x\}}{V}, \tag{1.17}$$

где $V\{\xi < x\}$ — объем части объекта, в точках которой значения показателя ξ меньше x. Если речь идет о распределении показателя на поверхности площадью S, смысл функции распределения аналогичен: $F(x) = \frac{S\{\xi < x\}}{S}$, $S\{\xi < x\}$ — площадь той части объекта исследований, где $\xi < x$.

Функция распределения обладает следующими общими свойствами.

1. $\lim\limits_{x \to \infty} F(x) = 1$, $\lim\limits_{x \to -\infty} F(x) = 0$. Это непосредственно вытекает из определения функции распределения (1.16). Если случайная величина ограничена сверху и снизу, $a \leqslant \xi \leqslant b$ (например, концентрации химических элементов в процентах), то $F(x) = 1$, если $x > b$ и $F(x) = 0$, если $x < a$.

2. Вероятность того, что случайная величина получит значение, заключенное между заданными пределами x_1, x_2,

$$\mathbf{P}\{x_1 \leqslant \xi < x_2\} = F(x_2) - F(x_1). \tag{1.18}$$

Это равенство — следствие того, что событие $\{\xi < x_2\}$ состоит из суммы двух несовместных событий — $B = \{\xi < x_1\}$ и $C = \{x_1 \leqslant \xi < x_2\}$: по формуле (1.2) $F(x_2) = \mathbf{P}(B) + \mathbf{P}(C) = \mathbf{P}\{\xi < x_1\} + \mathbf{P}(C) = F(x_1) + \mathbf{P}\{x_1 \leqslant \xi < x_2\}$, откуда и следует (1.18).

Согласно (1.18), $\mathbf{P}\{\xi \geqslant x\}$ определяется с помощью функции распределения как

$$\mathbf{P}\{\xi \geqslant x\} = \mathbf{P}\{x \leqslant \xi < \infty\} = 1 - F(x). \tag{1.19}$$

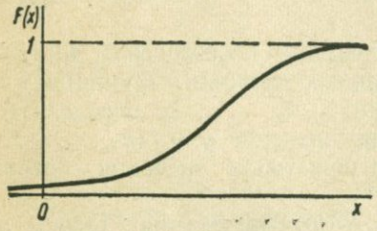


Рис. 7.

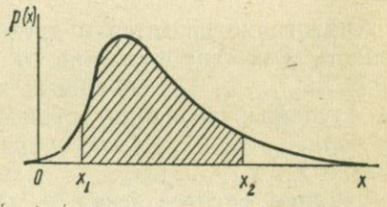


Рис. 8.

Если, например, x_0 — заданный нижний предел промышленной концентрации элемента ξ, то по (1.17) и (1.19) объем пород с концентрацией ξ, большей или равной x_0, составит $V\{\xi \geqslant x_0\} = V[1 - F(x_0)]$, где $F(x)$ — функция распределения ξ в пределах геологического тела объемом V.

3. Функция распределения неубывающая, т. е. $F(x_2) \geqslant F(x_1)$, если $x_2 > x_1$; неотрицательна — $F(x) \geqslant 0$ и ее значения не превышают единицы: $F(x) \leqslant 1$. Эти соотношения вытекают из свойств, приведенных выше. На рис. 7 изображен общий вид функции распределения.

4. Функция распределения непрерывна справа*.

В дальнейшем при изучении непрерывных случайных величин мы будем часто использовать *плотность распределения* случайной величины. Плотность распределения — это производная от функции распределения:

$$p(x) = \frac{dF(x)}{dx} = F'(x). \tag{1.20}$$

Она также обладает рядом общих свойств.

1. Плотность распределения определена при тех же значениях, что и функция распределения, за исключением тех точек, в которых производная (1.20) не существует.

---

* Доказательство этого утверждения приведено в [7].

2. Плотность распределения — всегда неотрицательная функция, причем

$$\lim_{x \to \pm\infty} p(x) = 0; \tag{1.21}$$

$$\int_{-\infty}^{\infty} p(x)\,dx = 1. \tag{1.22}$$

Если случайная величина ξ ограничена сверху и снизу, $a \leqslant \xi \leqslant b$, то ее плотность распределения $p(x) = 0$ при $x > b$ или $x < a$.

3. Вероятность события $\{x_1 \leqslant \xi < x_2\}$ определяется через плотность распределения случайной величины ξ по формуле

$$\mathbf{P}\{x_1 \leqslant \xi < x_2\} = \int_{x_1}^{x_2} p(x)\,dx. \tag{1.23}$$

Если длина интервала $[x_1, x_2]$, $x_2 - x_1 = \delta$ мала, а $p(x)$ непрерывна, то *

$$\mathbf{P}\{x_1 \leqslant \xi < x_2\} = p(x)\,\delta + o(\delta). \tag{1.23'}$$

Величину $p(x)\,\delta$ называют *элементом вероятности*.

Вообще, вероятность попадания значения ξ в заданную область числовых значений L — $\mathbf{P}\{\xi \in L\} = \int_L p(x)\,dx$; ($\{\xi \in L\}$ обозначает: «ξ принадлежит L»).

Приведенные свойства непосредственно следуют из того, что функция распределения является первообразной для плотности

$$F(x) = \int_{-\infty}^{x} p(t)\,dt. \tag{1.24}$$

Поэтому $F(x)$ называют еще *интегральной функцией распределения*.

4. Плотность распределения монотонной дифференцируемой функции $\eta = f(\xi)$ случайной величины ξ

$$p_\eta(x) = p_\xi(g(x))\left|\frac{dg(x)}{dx}\right|, \tag{1.25}$$

где $g(x)$ — функция, обратная $f(x)$, $p_\xi(x)$ — плотность распределения ξ. Действительно, если $f(x)$ — монотонно возрастающая функция, то функция распределения η имеет вид: $F_\eta(x) = \mathbf{P}\{\eta < x\} = \mathbf{P}\{f(\xi) < x\} = \mathbf{P}\{\xi < g(x)\} = F_\xi(g(x))$, где $F_\xi(x)$ — функция распределения ξ. Плотность $p_\eta(x) = F'_\eta(x) = \frac{dF_\xi(g(x))}{dx} = p_\xi(g(x))\,\frac{dg(x)}{dx}$. Если функция $f(x)$ монотонно убывающая, то $F_\eta(x) = \mathbf{P}\{f(\xi) < x\} = \mathbf{P}\{\xi > g(x)\} = 1 - F_\xi(g(x))$ и плотность распределения

$$p_\eta(x) = \frac{d}{dx}[1 - F_\xi(g(x))] = -p_\xi(g(x))\,\frac{dg(x)}{dx} = p_\xi(g(x))\left|\frac{dg(x)}{dx}\right|.$$

* В (1.23') $o(\delta)$ обозначает величину более высокого порядка малости, чем δ.

Это свойство можно сформулировать и так: если ξ распределена с плотностью $p_\xi(x)$ и является монотонной дифференцируемой функцией $\xi = g(\eta)$ случайной величины η, то плотность распределения η имеет вид (1.25).

В дальнейшем понадобится понятие *независимых случайных величин*. Независимыми являются такие случайные величины $\xi_1, \xi_2, \ldots, \xi_n$, для которых при произвольных $x_1, x_2, \ldots, x_n$ вероятность совместного осуществления событий $\xi_1 < x_1, \xi_2 < x_2, \ldots, \xi_n < x_n$ равна произведению значений функций распределения ξ_i в точках x_i:

$$\mathbf{P}\{\xi_1 < x_1, \xi_2 < x_2, \ldots, \xi_n < x_n\} = F_1(x_1) F_2(x_2) \ldots F_n(x_n). \quad (1.26)$$

В (1.26) $F_i(x_i)$ — функции распределения ξ_i $(i = \overline{1, n})$. Из (1.26) вытекает, в частности, взаимная независимость двух произвольных случайных величин ξ_i, ξ_j из $\xi_1, \xi_2, \ldots, \xi_n$: $\mathbf{P}\{\xi_i < x_i, \xi_j < x_j\} = F_i(x_i) F_j(x_j)$. Это равенство получим, положив в (1.26) $x_k = \infty$ при $k \neq i, j$.

Независимость ξ_i и ξ_j в форме (1.26) означает, что функция распределения ξ_i не зависит от того, какое значение принимает ξ_j: условная вероятность $\mathbf{P}\{\xi_i < x_j/(\xi_i < x_i)\} = \mathbf{P}\{\xi_i < x_i, \xi_j < x_j\}/\mathbf{P}\{\xi_i < x_i\} = F_j(x_j)$. Тот факт, что величина ξ_j получила некоторое значение x_j, не несет никакой дополнительной информации о распределении ξ_i. В противном случае ξ_i и ξ_j будут *зависимыми*.

Для независимых величин справедливы соотношения, являющиеся следствием (1.26). Так, если $a_i < b_i$ $(i = \overline{1, n})$, то вероятность того, что события $\{a_i \leqslant \xi_i < b_i\}$ $(i = \overline{1, n})$ произойдут совместно, будет

$$P = \prod_{i=1}^{n} [F_i(b_i) - F_i(a_i)]. \quad (1.27)$$

Действительно, вероятность

$$\mathbf{P}\{a_1 \leqslant \xi_1 < b_1, \xi_2 < b_2, \ldots, \xi_m < b_m\} = \mathbf{P}\{\xi_1 < b_1, \xi_2 < b_2, \ldots, \xi_m < b_m\} - \mathbf{P}\{\xi_1 < a_1, \xi_2 < b_2, \ldots, \xi_m < b_m\} = [F_1(b_1) - F_1(a_1)] F_2(b_2) \ldots F_m(b_m).$$

$$\mathbf{P}\{a_1 \leqslant \xi_1 < b_1, a_2 \leqslant \xi_2 < b_2, \xi_3 < b_3, \ldots, \xi_m < b_m\} = \mathbf{P}\{a_1 \leqslant \xi_1 < b_1, \xi_2 < b_2, \ldots, \xi_m < b_m\} - \mathbf{P}\{a_1 \leqslant \xi_1 < b_1, \xi_2 < a_2, \xi_3 < b_3, \ldots, \xi_m < b_m\} = [F_1(b_1) - F_1(a_1)] F_2(b_2) \ldots F_m(b_m) - [F_1(b_1) - F_1(a_1)] \times F_2(a_2) F_3(b_3) \ldots F_m(b_m) = [F_1(b_1) - F_1(a_1)] [F_2(b_2) - F_2(a_2)] \times F_3(b_3) \ldots F_m(b_m)$$

и вообще $\mathbf{P}\{a_i \leqslant \xi_i < b_i, i = \overline{1, m}\} = \prod_{i=1}^{m} [F_i(b_i) - F_i(a_i)]$.

5. Плотность распределения $p(x)$ суммы двух независимых случайных величин ξ_1 и ξ_2 выражается через их плотности распределения $p_1(x)$ и $p_2(x)$ в виде

$$p(x) = \int_{-\infty}^{\infty} p_1(x - t) p_2(t)\, dt = \int_{-\infty}^{\infty} p_2(x - t) p_1(t)\, dt. \quad (1.28)$$

Это следует из того, что функция распределения $\xi_1 + \xi_2$

$$F(x) = \mathbf{P}\{\xi_1 + \xi_2 < x\} = \mathbf{P}\{\xi_1 < x - \xi_2\} = \lim_{\max \Delta t_i \to 0} \sum_{i=-\infty}^{\infty} \mathbf{P}\{\xi_1 < x - t_i\} \times$$

$$\times \mathbf{P}\{t_i < \xi_2 < t_i + \Delta t_i\} = \lim_{\max \Delta t_i \to 0} \sum_{i=-\infty}^{\infty} F_1(x - t_i)\, p_2(t_i)\, \Delta t_i =$$

$$= \int_{-\infty}^{\infty} F_1(x - t)\, p_2(t)\, dt$$

(t_i — значения, которыми числовая ось разбивается на малые интервалы Δt_i). Продифференцировав $F(x)$ по x, получим (1.28).

Последовательно применяя формулу (1.28) для случая трех, четырех и т. д. случайных величин, получим плотность распределения суммы n независимых величин $\xi_1, \xi_2, \ldots, \xi_n$, имеющих плотности распределения, соответственно, $p_1(x), p_2(x), \ldots, p_n(x)$:

$$p(x) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} p_1(x - t_1 - \cdots - t_{n-1})\, p_2(t_1) \ldots p_n(t_{n-1}) \times$$

$$\times\, dt_1 \ldots dt_{n-1}. \qquad (1.28')$$

Плотность распределения функции независимых случайных величин $\eta = g(\xi_1, \xi_2, \ldots, \xi_n)$ имеет общий вид $p(x) = \frac{dF_g(x)}{dx}$, где $F_g(x)$ — функция распределения η:

$$F_g(x) = \mathbf{P}\{g(\xi_1, \xi_2, \ldots, \xi_n) < x\} =$$

$$= \int \int_{L(x)} \ldots \int p_1(t_1)\, p_2(t_2) \ldots p_n(t_n)\, dt_1 dt_2 \ldots dt_n.$$

$L(x) = \{g(t_1, \ldots, t_n) < x\}$ — область в n-мерном пространстве значений $\{t_1, t_2, \ldots, t_n\}$, для которых $g(t_1, t_2, \ldots, t_n) < x$.

Для независимых величин $\xi_1, \xi_2, \ldots, \xi_n$ функция распределения их максимума $\eta = \max\limits_i \xi_i$ равна произведению функций распределения ξ_i:

$$F_\eta(x) = \prod_{i=1}^{n} F_i(x). \qquad (1.29)$$

Это следует из того, что $F_\eta(x) = \mathbf{P}\{\max\limits_i \xi_i < x\} = \mathbf{P}\{\xi_i < x, i = \overline{1, n}\} =$

$$= \prod_{i=1}^{n} \mathbf{P}\{\xi_i < x\}.$$

Функция распределения минимума независимых величин $\xi_1, \xi_2, \ldots, \xi_n$, $\zeta = \min\limits_i \xi_i$, имеет вид

$$F_\zeta(x) = 1 - \prod_{i=1}^{n} [1 - F_i(x)], \tag{1.30}$$

так как $F_\zeta(x) = 1 - \mathbf{P}\{\zeta \geqslant x\} = 1 - \mathbf{P}\{\xi_i \geqslant x, \quad i = \overline{1, n}\} = 1 - \prod\limits_{i=1}^{n} \mathbf{P}\{\xi_i \geqslant x\} = 1 - \prod\limits_{i=1}^{n} [1 - F_i(x)]$.

Если ξ_i распределены по одному и тому же закону, например, они являются независимыми наблюдениями одной и той же случайной величины с функцией распределения $F_0(x)$, то функция распределения $\eta = \max \xi_i$, согласно (1.29), будет

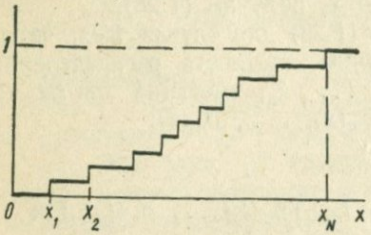


Рис. 9.

$$F_\eta(x) = [F_0(x)]^n, \tag{1.31}$$

а $\zeta = \min\limits_i \xi_i$, по (1.30),

$$F_\zeta(x) = 1 - [1 - F_0(x)]^n. \tag{1.32}$$

Вероятность попадания всех одинаково распределеных ξ_i в одни и те же пределы $[a, b)$

$$p = [F_0(b) - F_0(a)]^n. \tag{1.33}$$

При изучении распределения дискретной случайной величины обычно ограничиваются указанием значений $x_1, x_2, \ldots, x_N$, принимаемых ею, и вероятностей, с которыми она их принимает: $p_1, p_2, \ldots, p_N$. Функция распределения такой величины кусочно-постоянна. Если значения дискретной случайной величины пронумерованы в порядке их возрастания, $x_1 < x_2 < \cdots < x_N$, то ее функция распределения

$$F(x) = \sum_{i=1}^{n(x)} p_i, \tag{1.34}$$

где $n(x)$ — количество значений x_i, меньших x (рис. 9).

§ 3. Числовые характеристики распределений случайных величин

Квантили. Помимо плотности и функции распределения, в статистическом анализе широко используются числовые характеристики распределений случайных величин. К числу таких характеристик принадлежат *квантили*. Квантиль U_q порядка q случайной величины ξ, называемый еще q-квантилем, определяется условием:

$$\mathbf{P}\{\xi < U_q\} = F(U_q) = q. \tag{1.35}$$

Величину $1 - q$ иногда выражают в процентах — $Q = (1 - q)\,100\%$, а квантиль U_q называют Q-процентной точкой распределения.

При значении аргумента, равном квантилю U_q, функция распределения получает значение, равное его порядку q (рис. 10). Иначе говоря, если $F^{-1}(x)$ — функция, обратная функции распределения $F(x)$, то

$$U_q = F^{-1}(q). \tag{1.36}$$

Вероятность получения случайной величиной значения, большего или равного U_q

$$\mathbf{P}\{\xi \geqslant U_q\} = 1 - F(U_q) = 1 - q, \tag{1.37}$$

и если величина $1 - q$ мала, событие $\{\xi \geqslant U_q\}$ маловероятно. Это дает основание использовать U_q при близких к единице значениях q в качестве нижнего предела для аномальных наблюдений.

Квантиль порядка 0,5 имеет собственное название — *медиана*. Медиана $M_e = U_{0,5}$ нередко используется для оценки уровня «местного фона» показателей. Вероятности получения значений, меньших M_e и превышающих или равных M_e одинаковы и составляют по 0,5:

$$F(M_e) = \mathbf{P}\{\xi < M_e\} = \\ = \mathbf{P}\{\xi \geqslant M_e\} = 0{,}5. \tag{1.38}$$

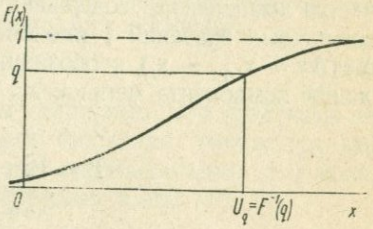


Рис. 10.

Математическое ожидание. Важнейшей числовой характеристикой случайной величины является *математическое ожидание*, или *среднее*. Математическое ожидание дискретной случайной величины определяется формулой

$$\mathbf{M}\xi = \sum_{i=1}^{N} x_i p_i, \tag{1.39}$$

где x_i — значения, получаемые случайной величиной ξ; p_i — вероятности, с которыми она их принимает; N — количество всех возможных ее значений. Если $N = \infty$, математическое ожидание представляется в виде суммы бесконечного ряда: $\mathbf{M}\xi = \sum_{i=1}^{\infty} p_i x_i$.

Математическое ожидание непрерывной случайной величины ξ определяется в виде

$$\mathbf{M}\xi = \int_{-\infty}^{\infty} x p(x)\, dx, \tag{1.40}$$

где $p(x)$ — плотность распределения ξ. Это определение* является естественным обобщением (1.39). Разбив числовую ось на малые

* Математическое ожидание можно определить непосредственно через функцию распределения в виде интеграла Стильтьеса [7] для случаев, когда функция распределения не имеет производной в некоторых точках. Тогда оно будет определено одинаково для непрерывных и дискретных случайных величин. Для наших задач вполне достаточно определение (1.40).

интервалы $[x_i, x_{i+1})$ и опираясь на (1.39), вычислим математическое ожидание:

$$\mathbf{M}\xi = \lim_{\max \Delta x_i \to 0} \sum_{-\infty}^{\infty} x_i \mathbf{P}\{x_i \leqslant \xi < x_{i+1}\} =$$

$$= \lim_{\max \Delta x_i \to 0} \sum_{-\infty}^{\infty} x_i p(x_i)(x_{i+1} - x_i) = \int_{-\infty}^{\infty} xp(x)\,dx, \qquad (1.40')$$

откуда и получим (1.40).

Чтобы уяснить смысл этой характеристики, рассмотрим следующий пример. Пусть ξ — случайная величина, представляющая собой содержание (в долях единицы объема) некоторого компонента в выбираемой наугад точке геологического тела, имеющего объем V; T — общее количество компонента, содержащееся в этом объеме. Разобьем интервал возможных значений ξ точками x_i на малые интервалы равной длины Δx ($\Delta x = x_{i+1} - x_i$) и обозначим V_i объем той части тела, где содержание компонента меньше x_i. Величину T можно представить в виде

$$T = \lim_{\substack{\Delta x \to 0, \\ n \to \infty}} \sum_{i=1}^{n-1} x_i (V_{i+1} - V_i),$$

учитывая, что объем той части тела, где содержание компонента заключено между x_i, x_{i+1}, составит $V_{i+1} - V_i$. В соответствии с (1.17), $V_i = VF(x_i)$, $F(x)$ — функция распределения ξ. Предположив, что $F(x)$ дифференцируема, обозначим плотность распределения $p(x) = F'(x)$. Тогда

$$T = \lim_{\substack{\Delta x \to 0, \\ n \to \infty}} \sum_{i=1}^{n-1} Vx_i [F(x_{i+1}) - F(x_i)] = V \lim_{\substack{\Delta x \to 0, \\ n \to \infty}} \sum_{i=1}^{n-1} x_i p(x_i) \Delta x =$$

$$= V \int_{-\infty}^{\infty} xp(x)\,dx = V\mathbf{M}\xi, \qquad (1.41)$$

откуда видно, что математическое ожидание имеет смысл общего количества компонента в пределах исследуемого объекта, отнесенного к его объему.

Аналогично, если наблюдения случайной величины ξ — значения компонента на некоторой поверхности площадью S, математическим ожиданием будет то значение, которое получилось бы при равномерном распределении компонента, содержащегося в поверхностном слое, на весь этот слой.

Свойства математического ожидания. 1. Математическое ожидание постоянной величины равно этой величине: $\mathbf{M}c = c$. Это следует из формулы (1.39) и того, что постоянную c можно рассматривать как случайную величину, принимающую единственное значение c с вероятностью, равной единице.

2. Математическое ожидание непрерывной случайной величины $\eta = g(\xi)$, являющейся функцией другой непрерывной случайной величины ξ, вычисляется в виде

$$\mathbf{M}g(\xi) = \int_{-\infty}^{\infty} g(x)\, p(x)\, dx, \tag{1.42}$$

где $p(x)$ — плотность распределения ξ.

Аналогичная формула справедлива для дискретных случайных величин:

$$\mathbf{M}g(\xi) = \sum_{i=1}^{N} g(x_i)\, p_i, \tag{1.43}$$

где x_i — как и прежде, возможные значения случайной величины ξ, p_i — вероятности, с которыми она их принимает.

Из этого свойства непосредственно следует, что постоянный множитель можно выносить за знак математического ожидания:

$$\mathbf{M}c\xi = c\mathbf{M}\xi. \tag{1.44}$$

Если все значения случайной величины умножить или разделить на одно и то же число, то и математическое ожидание умножится или разделится на это число. Точно так же математическое ожидание линейной функции $\alpha\xi + \beta$ (α и β — постоянные) имеет вид

$$\mathbf{M}(\alpha\xi + \beta) = \mathbf{M}(\alpha\xi) + \beta = \alpha\mathbf{M}\xi + \beta. \tag{1.45}$$

3. Математическое ожидание суммы конечного числа случайных величин равно сумме их математических ожиданий:

$$\mathbf{M}(\xi_1 + \xi_2 + \cdots + \xi_n) = \mathbf{M}\xi_1 + \mathbf{M}\xi_2 + \cdots + \mathbf{M}\xi_n, \tag{1.46}$$

причем этим важным свойством обладают как зависимые, так и независимые случайные величины. Не приводя общего доказательства (1.46), рассмотрим случай, когда ξ_i независимы. Для двух случайных величин, согласно (1.28), имеем:

$$\mathbf{M}(\xi_1 + \xi_2) = \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} p_1(x-t)\, p_2(t)\, dt\, dx = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (u+t) p_1(u)\, p_2(t)\, dt\, du =$$

$$= \int_{-\infty}^{\infty} u p_1(u)\, du + \int_{-\infty}^{\infty} t p_2(t)\, dt = \mathbf{M}\xi_1 + \mathbf{M}\xi_2.$$

По этой же формуле

$$\mathbf{M}(\xi_1 + \xi_2 + \xi_3) = \mathbf{M}(\xi_1 + \xi_2) + \mathbf{M}\xi_3 = \mathbf{M}\xi_1 + \mathbf{M}\xi_2 + \mathbf{M}\xi_3$$

и так добавляя каждый раз $\xi_4, \xi_5, \ldots, \xi_n$, придем к (1.46).

Как следует из (1.46), если ξ — значение показателя в выбираемой наугад точке геологического объекта, Δ — ошибка измерения (взятое со знаком отклонение результата измерения η от истинного значения), математическое ожидание $\eta = \xi + \Delta$ (1.15) — $\mathbf{M}(\xi + \Delta) = \mathbf{M}\xi + \mathbf{M}\Delta$. Среднее значение показателя по объекту отличается от среднего по

результатам его измерений на величину систематического смещения $\mathbf{M}\Delta$ метода измерений. Оба средних совпадают при отсутствии смещения — $\mathbf{M}\Delta = 0$.

4. Математическое ожидание произведения попарно независимых случайных величин равно произведению их математических ожиданий:

$$\mathbf{M}(\xi_1\xi_2 \ldots \xi_n) = \mathbf{M}\xi_1\mathbf{M}\xi_2 \ldots \mathbf{M}\xi_n. \tag{1.47}$$

Еще одно важное свойство, носящее название закона больших чисел, будет рассмотрено ниже, после введения понятия дисперсии и описания ее свойств.

Дисперсия. *Дисперсия* — числовая характеристика, служащая мерой вариации (рассеяния) значений случайной величины вокруг ее математического ожидания и определяемая формулой

$$\mathbf{D}\xi = \mathbf{M}(\xi - \mathbf{M}\xi)^2. \tag{1.48}$$

Бóльшему значению дисперсии соответствует бóльшая вариация случайной величины. Это и позволяет, сопоставляя дисперсии, сравнивать случайные величины по степени их вариации.

Если случайная величина ξ непрерывна и $p(x)$ — плотность ее распределения, то, согласно (1.42),

$$\mathbf{D}\xi = \int_{-\infty}^{\infty} (x - \mathbf{M}\xi)^2 p(x)\, dx. \tag{1.49}$$

Для дискретной случайной величины

$$\mathbf{D}\xi = \sum_{i=1}^{N} (x_i - \mathbf{M}\xi)^2 p_i, \tag{1.49'}$$

где x_i — значения случайной величины, p_i — вероятности, с которыми она их принимает.

Свойства дисперсии. 1. Дисперсия всегда положительна, за исключением того случая, когда $\xi = c = \text{const}$. Дисперсия постоянной равна нулю:

$$\mathbf{D}c = 0. \tag{1.50}$$

2. Дисперсию можно представить в виде, эквивалентном (1.48):

$$\mathbf{D}\xi = \mathbf{M}\xi^2 - (\mathbf{M}\xi)^2. \tag{1.48'}$$

Это следует из того, что $\mathbf{M}(\xi - \mathbf{M}\xi)^2 = \mathbf{M}[\xi^2 - 2\xi\mathbf{M}\xi + (\mathbf{M}\xi)^2] = \mathbf{M}\xi^2 - 2\mathbf{M}\xi\mathbf{M}\xi + (\mathbf{M}\xi)^2 = \mathbf{M}\xi^2 - (\mathbf{M}\xi)^2$.

3. Если a — произвольное число, то величина $\mathbf{M}(\xi - a)^2$ — математическое ожидание квадрата отклонения ξ от a, получает минимальное значение при $a = \mathbf{M}\xi$, и это значение равно дисперсии $\mathbf{D}\xi$:

$$\mathbf{M}(\xi - a)^2 = \mathbf{M}(\xi - \mathbf{M}\xi + \mathbf{M}\xi - a)^2 = \mathbf{M}[(\xi - \mathbf{M}\xi)^2 + 2(\xi - \mathbf{M}\xi)(\mathbf{M}\xi - a) + (\mathbf{M}\xi - a)^2] = \mathbf{M}(\xi - \mathbf{M}\xi)^2 + (\mathbf{M}\xi - a)^2 = \mathbf{D}\xi + (\mathbf{M}\xi - a)^2 \geqslant \mathbf{D}\xi.$$

Таким образом,

$$\mathbf{D}\xi = \min_{a} \mathbf{M}(\xi - a)^2. \qquad (1.48'')$$

4. Дисперсия суммы попарно независимых случайных величин равна сумме их дисперсий:

$$\mathbf{D}(\xi_1 + \xi_2 + \cdots + \xi_n) = \mathbf{D}\xi_1 + \mathbf{D}\xi_2 + \cdots + \mathbf{D}\xi_n. \qquad (1.51)$$

Действительно, учитывая, что для независимых случайных величин ξ_i и ξ_j

$$\mathbf{M}[(\xi_i - \mathbf{M}\xi_i)(\xi_j - \mathbf{M}\xi_j)] = \mathbf{M}(\xi_i - \mathbf{M}\xi_i)\,\mathbf{M}(\xi_j - \mathbf{M}\xi_j) = 0,$$

имеем:

$$\mathbf{D}\left(\sum_{i=1}^{n}\xi_i\right) = \mathbf{M}\left(\sum_{i=1}^{n}\xi_i - \mathbf{M}\sum_{i=1}^{n}\xi_i\right)^2 = \mathbf{M}\left[\sum_{i=1}^{n}(\xi_i - \mathbf{M}\xi_i)\right]^2 = \mathbf{M}\left[\sum_{i=1}^{n}(\xi_i - \mathbf{M}\xi_i)^2 + \right.$$

$$\left. + \sum_{i\neq j}^{n}(\xi_i - \mathbf{M}\xi_i)(\xi_j - \mathbf{M}\xi_j)\right] = \sum_{i=1}^{n}\mathbf{M}(\xi_i - \mathbf{M}\xi_i)^2 = \sum_{i=1}^{n}\mathbf{D}\xi_i.$$

Возвращаясь к примеру с ошибками измерений, о котором шла речь при описании третьего свойства математического ожидания, имеем: $\mathbf{D}(\xi + \Delta) = \mathbf{D}\xi + \mathbf{D}\Delta$, если Δ не зависит от ξ. Отсюда видно, что для получения дисперсии показателя необходимо из дисперсии результатов измерений вычесть дисперсию ошибок.

5. Если c — постоянная, то

$$\mathbf{D}(c\xi) = \mathbf{M}(c\xi - \mathbf{M}(c\xi))^2 = c^2\mathbf{M}(\xi - \mathbf{M}\xi)^2 = c^2\mathbf{D}\xi \qquad (1.52)$$

(при умножении на c случайной величины, т. е. всех принимаемых ею значений дисперсия возрастает в c^2 раз).

Следствиями перечисленных свойств являются следующие:

$$\mathbf{D}(\xi + c) = \mathbf{D}(\xi - c) = \mathbf{D}\xi \quad (c = \text{const}) — \qquad (1.53)$$

от добавления или вычитания постоянной дисперсия не изменяется;

$$\mathbf{D}(\xi - \eta) = \mathbf{D}\xi + \mathbf{D}\eta, \qquad (1.54)$$

если ξ и η — независимые случайные величины;

$$\mathbf{D}(\alpha\xi + \beta) = \alpha^2\mathbf{D}\xi \quad (\alpha = \text{const},\ \beta = \text{const}). \qquad (1.55)$$

Следствие (1.54) можно использовать, например, для нахождения дисперсии ошибок *воспроизводимого* метода по данным измерений этим методом, не зная истинных значений измеряемого показателя. Под воспроизводимым подразумевается метод, который дает возможность проводить неоднократные измерения в одних и тех же условиях заведомо одного и того же значения показателя, без систематического смещения между показаниями прибора.

Обозначим η_1 — первое, η_2 — второе измерение одного и того же неизвестного значения показателя ξ; Δ_1 — ошибку (отклонение со знаком) при первом, Δ_2 — при втором измерении. Так как $\eta_1 - \eta_2 = (\xi + \Delta_1) - (\xi + \Delta_2) = \Delta_1 - \Delta_2$, дисперсия $\mathbf{D}(\eta_1 - \eta_2) = \mathbf{D}(\Delta_1 - \Delta_2) = \mathbf{D}\Delta_1 + \mathbf{D}\Delta_2 = 2\mathbf{D}\Delta$ ($\mathbf{D}\Delta$ — дисперсия ошибок). Вычислив половину

дисперсии случайной величины, значениями которой являются разности между двумя кратными измерениями, получим дисперсию ошибок.

Закон больших чисел. Одно из основных положений теории вероятностей и математической статистики — *закон больших чисел* состоит в следующем. Если $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ — последовательность попарно независимых случайных величин с дисперсиями, ограниченными одним и тем же числом, то для произвольного малого $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbf{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i - \frac{1}{n}\sum_{i=1}^{n}\mathbf{M}\xi_i\right| < \varepsilon\right\} = 1, \tag{1.56}$$

(знаком $\mathbf{P}\{\ldots\}$ здесь, как и раньше, обозначена вероятность события, заключенного в фигурные скобки)*.

Отсюда следует, что среднее арифметическое, вычисленное по независимым наблюдениям одной и той же случайной величины, с увеличением числа наблюдений становится все более близким к математическому ожиданию. Поэтому за счет увеличения кратности измерений по методу, имеющему высокую чувствительность и не дающему систематического смещения, мы получаем возможность более точно определять истинное значение показателя. Среднее арифметическое из результатов таких кратных измерений будет приближаться к истинному значению с увеличением числа измерений, хотя ошибка отдельного измерения была бы значительной.

Закон больших чисел приведен в *форме Чебышева*. Он является следствием известного *неравенства Чебышева*: для произвольной случайной величины ξ

$$\mathbf{P}\{|\xi - \mathbf{M}\xi| \geqslant \varepsilon\} \leqslant \frac{\mathbf{D}\xi}{\varepsilon^2} \tag{1.57}$$

при любом $\varepsilon > 0$. Пусть $p(x)$ — плотность распределения ξ. Неравенство следует из того, что

$$\mathbf{P}\{|\xi - \mathbf{M}\xi| \geqslant \varepsilon\} = \int\limits_{|x-\mathbf{M}\xi|>\varepsilon} p(x)\,dx \leqslant \int\limits_{|x-\mathbf{M}\xi|>\varepsilon} \frac{(x-\mathbf{M}\xi)^2}{\varepsilon^2} p(x)\,dx$$

(учитывая, что в области интегрирования $(x - \mathbf{M}\xi)^2 \varepsilon^{-2} \geqslant 1$) и того, что

$$\int\limits_{|x-\mathbf{M}\xi|>\varepsilon} \frac{(x-\mathbf{M}\xi)^2}{\varepsilon^2} p(x)\,dx \leqslant \frac{1}{\varepsilon^2}\int\limits_{-\infty}^{\infty} (x-\mathbf{M}\xi)^2 p(x)\,dx = \frac{\mathbf{D}\xi}{\varepsilon^2}.$$

---

* Если для некоторой последовательности случайных величин η_n и любого $\varepsilon > 0$ $\mathbf{P}\{|\eta_n - \eta| < \varepsilon\} \to 1$ при $n \to \infty$, то η_n *сходится по вероятности* к случайной величине η. Таким образом, (1.56) означает сходимость по вероятности $\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i - \frac{1}{n}\sum_{i=1}^{n}\mathbf{M}\xi_i\right)$ к нулю.

Положив в (1.57) $\xi = \frac{1}{n}\sum_{i=1}^{n}\xi_i$ и учитывая, что

$$\mathbf{M}\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{M}\xi_i, \quad \mathbf{D}\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbf{D}\xi_i < \frac{nc}{n^2} = \frac{c}{n},$$

получим (1.56).

Нормированное уклонение. Из свойств (1.45) и (1.55) вытекает следствие: математическое ожидание и дисперсия *нормированного уклонения*

$$\xi_н = \frac{\xi - \mathbf{M}\xi}{\sqrt{\mathbf{D}\xi}} \tag{1.58}$$

случайной величины ξ равны, соответственно, нулю и единице:

$$\mathbf{M}\left(\frac{\xi - \mathbf{M}\xi}{\sqrt{\mathbf{D}\xi}}\right) = \frac{1}{\sqrt{\mathbf{D}\xi}}(\mathbf{M}\xi - \mathbf{M}\xi) = 0, \quad \mathbf{D}\left(\frac{\xi - \mathbf{M}\xi}{\sqrt{\mathbf{D}\xi}}\right) = \frac{1}{\mathbf{D}\xi}\mathbf{D}\xi = 1. \tag{1.59}$$

Преобразование (1.58) приводит к смещению плотности распределения вдоль оси абсцисс до нулевого математического ожидания и растяжению или сжатию ее с сохранением площади до единичной дисперсии.

Среднее квадратическое отклонение. Величина, равная квадратному корню из дисперсии, $\sigma = \sqrt{\mathbf{D}\xi}$ носит наименование *среднего квадратического отклонения*, или *стандарта*. Как и дисперсия, оно характеризует степень вариации случайной величины, однако в отличие от дисперсии имеет ту же размерность, что и случайная величина.

Асимметрия и эксцесс. К часто употребляемым числовым характеристикам случайных величин принадлежат асимметрия и эксцесс. *Асимметрия (коэффициент асимметрии)* случайной величины ξ определяется, как

$$A = \mathbf{M}\left(\frac{\xi - \mathbf{M}\xi}{\sqrt{\mathbf{D}\xi}}\right)^3; \tag{1.60}$$

эксцесс (коэффициент эксцесса) —

$$E = \mathbf{M}\left(\frac{\xi - \mathbf{M}\xi}{\sqrt{\mathbf{D}\xi}}\right)^4 - 3. \tag{1.61}$$

Асимметрия служит характеристикой асимметричности плотности распределения. Если последняя симметрична, асимметрия обращается в нуль. При правой асимметричности распределения (правая ветвь плотности более вытянута по сравнению с левой) асимметрия положительна, при левой — отрицательна (рис. 11). Однако, равенство коэффициента асимметрии нулю ещё не свидетельствует о симметричности плотности распределения. Тем не менее, при изучении распределений геолого-геофизических показателей в относительно однородных объектах, как показывает практика, асимметрия служит вполне достаточным индикатором симметричности. Положительная асимметрия обычно свидетельствует о том, что значительные положительные отклонения случайной

величины от математического ожидания более вероятны, чем такие же по абсолютной величине отрицательные отклонения от него.

Эксцесс используется для характеристики большей или меньшей «островершинности» плотности распределения (рис. 12). Плотностям распределения с вытянутыми концами (одним или обоими) соответствуют бо́льшие значения эксцесса. Для таких распределений обычно характерна возможность получения со сравнительно небольшой вероятностью значительных отклонений случайной величины от своего математического ожидания, при сосредоточении основной массы ее возможных значений в относительно небольшом интервале.

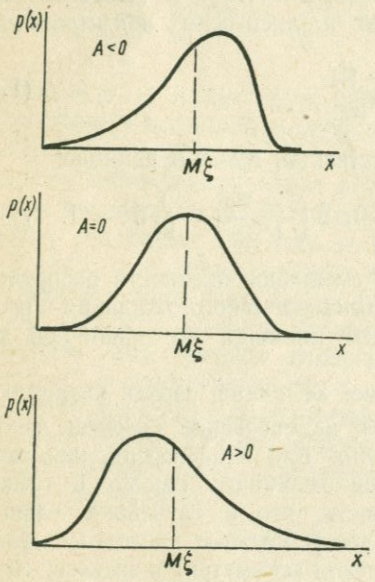


Рис. 11.

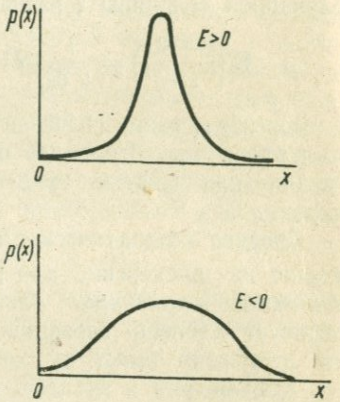


Рис. 12.

Чаще всего асимметрия и эксцесс применяются для проверки гипотез о виде распределения. В дальнейшем мы более подробно познакомимся с этим их применением.

Начальные и центральные моменты. Начальным моментом k-го порядка случайной величины ξ называется математическое ожидание k-й степени ξ:

$$m_k = \mathbf{M}\xi^k; \tag{1.62}$$

центральный момент k-го порядка —

$$\mu_k = \mathbf{M}(\xi - \mathbf{M}\xi)^k. \tag{1.63}$$

В частности, математическое ожидание — начальный момент первого порядка, а дисперсия — центральный момент второго порядка. С использованием (1.63) коэффициенты асимметрии и эксцесса записываются в виде

$$A = \frac{\mu_3}{\sqrt{\mu_2^3}}, \quad E = \frac{\mu_4}{\mu_2^2} - 3.$$

Центральные и начальные моменты связаны соотношением

$$\mu_k = \mathbf{M}\left[\sum_{i=0}^{k}(-1)^i C_k^i \xi^i (\mathbf{M}\xi)^{k-i}\right] = \sum_{i=0}^{k}(-1)^i C_k^i \mathbf{M}\xi^i (\mathbf{M}\xi)^{k-i} =$$

$$= \sum_{i=0}^{k}(-1)^i C_k^i m_i m_1^{k-i}, \qquad (1.64)$$

где C_k^i — число сочетаний из k элементов по i.

Если $\xi_1, \xi_2, \ldots, \xi_n$ независимы, то

$$\mathbf{M}(\xi_1^{k_1}\xi_2^{k_2} \ldots \xi_n^{kn}) = m_{k_1}^{(1)} m_{k_2}^{(2)} \ldots m_{kn}^{(n)}$$

($m_i^{(j)}$ — начальный момент ξ_j i-го порядка);

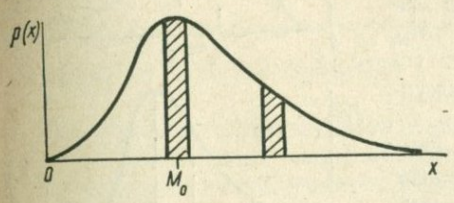


Рис. 13.

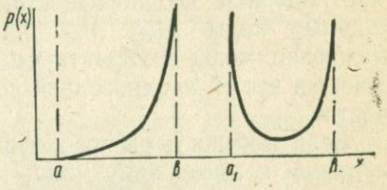


Рис. 14.

$$\mathbf{M}[(\xi_1 - \mathbf{M}\xi_1)^{k_1}(\xi_2 - \mathbf{M}\xi_2)^{k_2} \ldots (\xi_n - \mathbf{M}\xi_n)^{kn}] = \mu_{k_1}^{(1)}\mu_{k_2}^{(2)} \ldots \mu_{kn}^{(n)}$$

($\mu_i^{(j)}$ — центральный момент ξ_j i-го порядка).

Мода. При статистическом анализе распределений геолого-геофизических показателей часто используется *мода*, или *модальное значение*, M_0. Для дискретной случайной величины мода определяется как значение, которое та получает с наибольшей вероятностью. Мода непрерывной случайной величины — значение, при котором плотность распределения обращается в максимум. Так определяемая мода обладает следующим свойством. Если перемещать интервал малой постоянной длины δ по оси абсцисс, то вероятность попадания в него значения непрерывной случайной величины окажется наибольшей, когда средина интервала займет положение M_δ, близкое к моде, причем $M_\delta \to M_0$ при $\delta \to 0$ (рис. 13). В этом смысле мода непрерывной случайной величины есть ее наиболее вероятное значение.

В соответствии с приведенным выше определением, для непрерывных распределений с дифференцируемой плотностью $p(x)$ мода обычно является одним из решений уравнения

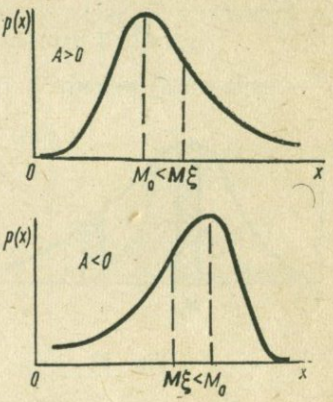
$$\frac{dp(x)}{dx} = 0. \qquad (1.65)$$

Мода распределения с плотностью J- или U-образной формы (рис. 14) равна одному из граничных значений интервала, на котором задана

плотность. Такие распределения могут иметь антимоду A_0 (рис. 14). Они весьма специфичны и в практике геологических исследований встречаются редко.

Как правило, для плотностей распределений показателей в однородных геологических объектах решение уравнения (1.65) бывает одно, причем мода совпадает с этим решением. Иногда плотность распределения имеет два или более локальных максимума (рис. 15). Тогда говорят о *полимодальных распределениях*, обусловленных неоднородностью поведения показателя в пределах изучаемого объекта (встречаются различные типы пород, каждому из которых соответствует свой закон распределения показателя, различные степени измененности и т. п.). Полимодальность часто и указывает на наличие подобной неоднородности, а локальные моды M_{01}, M_{02}, ..., M_{0k} оценивают моды соответствующих составных частей неоднородной совокупности.

Если плотность распределения симметрична и имеет одну моду, последняя совпадает с математическим ожиданием. При правой асимметрии ($A > 0$)



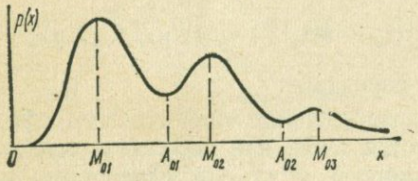


Рис. 15.

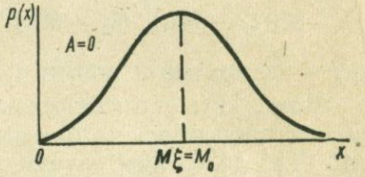


Рис. 16.

мода, как правило, смещается влево от математического ожидания: $M_0 < \mathbf{M}\xi$; при левой ($A < 0$) — вправо: $M_0 > \mathbf{M}\xi$ (рис. 16).

При сопоставлении геологических объектов по распределениям количественных показателей слагающих их пород модальное значение в определенном смысле имеет преимущество перед математическим ожиданием (средним). По своему содержанию мода более объективно характеризует уровень местного фона, так как меньше зависит от факторов, которые могут искажать фоновое распределение показателя в основной массе пород (измененность части пород, наличие инородных включений, аномалий и т. п.). При сравнении математических ожиданий фактически сопоставляются средние значения показателя, приходящиеся на единицу всего объема или площади, как было показано при введении понятия математического ожидания.

Коэффициент вариации. Эта числовая характеристика тоже принадлежит к числу наиболее употребляемых. Коэффициент вариации определяется по формуле

$$V = \frac{\sigma}{M\xi} = \frac{\sqrt{D\xi}}{M\xi}. \qquad (1.66)$$

В отличие от дисперсии и стандартного отклонения он является безразмерной относительной характеристикой вариации случайной величины. Иногда коэффициент вариации выражают в процентах: $V \cdot 100\%$. Коэффициент вариации ошибок измерений характеризует относительную погрешность метода измерений.

Глава 2

ЗАКОНЫ РАСПРЕДЕЛЕНИЯ

Как было показано в первой главе, распределение случайной величины полностью описывается функцией или плотностью распределения. В теории вероятностей известен ряд *законов распределения*, обобщающих различные схемы образования случайных величин. Определенным законам соответствуют параметрические семейства функций распределения. Рассмотрим наиболее распространенные из них.

§ 1. Дискретные распределения

Биномиальное распределение. Этот закон соответствует схеме, известной под названием *схемы Бернулли*.

Производится n независимых испытаний, в каждом из которых некоторое событие A может появиться с вероятностью p. Количество испытаний этой серии, в которых происходит A, является дискретной случайной величиной ξ_n, которая может принимать одно из целочисленных значений: 0, 1, 2, ..., n. Выясним, чему равна вероятность $P_n(m)$ того, что ξ_n получит значение, равное m.

Вероятность того, что A произойдет в испытаниях с номерами i_1, i_2, ..., i_m, а в остальных не произойдет, равно произведению m вероятностей отдельных событий A и $n-m$ вероятностей событий $\overline{A}$: $p^m(1-p)^{n-m}$. Событие $\{\xi_n = m\}$ состоит из суммы всех несовместных вариантов осуществления m раз события A (при разных комбинациях $i_1, i_2, \ldots, i_m$). По формуле (1.2) вероятность $P_n(m)$ равна сумме вероятностей этих вариантов, которых будет столько, сколькими способами можно разместить на n местах m одинаковых элементов. Это количество равно C_n^m — числу сочетаний из n элементов по m. Так как слагаемые вероятности одинаковы и каждая составляет $p^m(1-p)^{n-m}$, имеем

$$P_n(m) = C_n^m p^m (1-p)^{n-m} = \frac{n!}{m!\,(n-m)!} p^m (1-p)^{n-m}. \qquad (2.1)$$

Эти вероятности и определяют биномиальное распределение*.

Рассмотрим пример величины, распределенной по биномиальному закону. Пусть η — результат измерения некоторого показателя и производится n измерений одной и той же пробы. Событие A заключается в том, что результат измерения η превышает заданный уровень z или равен ему. Вероятность события A определяется функцией распределения $F(x)$ ошибок измерений: $p = \mathbf{P}\{\eta \geqslant z\} = 1 - F(z)$. Тогда количество результатов измерений со значениями, большими или равными z, будет следовать биномиальному распределению, определяемому вероятностями (2.1) при $p = 1 - F(z)$.

Функция биномиального распределения, согласно (1.34) и (2.1), имеет вид

$$F_n(x) = \begin{cases} \sum\limits_{0 \leqslant i < x} \frac{n!}{i!\,(n-i)!}\, p^i (1-p)^{n-i} & n \geqslant x > 0, \\ 0 & x \leqslant 0, \\ 1 & x > n. \end{cases} \tag{2.2}$$

Это выражение становится громоздким при больших n, тогда используют приближение, о котором будет идти речь несколько позже.

Математическое ожидание и дисперсия равны соответственно

$$\mathbf{M}\xi_n = np, \quad \mathbf{D}\xi_n = np(1-p). \tag{2.3}$$

Чтобы в этом убедиться, представим ξ_n в виде суммы независимых дискретных случайных величин η_i:

$$\xi_n = \sum_{i=1}^{n} \eta_i; \tag{2.4}$$

$\eta_i = 1$, если в i-м испытании событие A произошло, и $\eta_i = 0$, если не произошло. Математическое ожидание и дисперсия η_i, по (1.39) и (1.49′),

$$\mathbf{M}\eta_i = 1 \cdot p + 0 \cdot (1-p) = p, \quad \mathbf{D}\eta_i = (1-p)^2 p + (0-p)^2 (1-p) = (1-p)\,p,$$

(учитывая, что η_i получает значение 1 с вероятностью p и 0 с вероятностью $1-p$). Математическое ожидание ξ_n, по (2.4) и (1.46), $\mathbf{M}\xi_n = \sum\limits_{i=1}^{n} \mathbf{M}\eta_i = np$; дисперсия, согласно (1.51), $\mathbf{D}\xi_n = \sum\limits_{i=1}^{n} \mathbf{D}\eta_i = np(1-p)$.

Сравнивая отношение вероятностей $\frac{P_n(m+1)}{P_n(m)} = \frac{(n-m)\,p}{(m+1)(1-p)}$ с единицей при последовательно возрастающих m и определив m, при котором оно впервые становится меньшим единицы, найдем, что мода равна $[(n+1)p]$, т. е. целой части $(n+1)p$, если $(n+1)p$ — дробное число.

---

* Отметим, что вероятность появления события A впервые в i-м испытании составит $p_i = p(1-p)^{i-1}$ (т. н. *геометрическое распределение*). Вероятность того, что событие наступит не позже, чем в i-м испытании, равна $\sum\limits_{k=1}^{i} p_k = 1-(1-p)^i$.

Если $(n+1)p$ — целое, то распределение имеет два модальных значения: $(n+1)p-1$ и $(n+1)p$. Отсюда видно, что биномиальное распределение, как правило, имеет одну моду.

Асимметрия биномиального распределения

$$A_n = \frac{1-2p}{\sqrt{np(1-p)}}, \tag{2.5}$$

так что при $p \neq \frac{1}{2}$ оно асимметрично, причем при $p > \frac{1}{2}$ наблюдается левая асимметрия, а при $p < \frac{1}{2}$ — правая.

Эксцесс выражается формулой

$$E_n = \frac{1-6p+6p^2}{np(1-p)}. \tag{2.6}$$

Выражения (2.5), (2.6) показывают, что с увеличением n асимметрия и эксцесс стремятся к нулю. Это вызвано, как мы увидим далее, приближением биномиального распределения к закону, при котором асимметрия и эксцесс равны нулю.

Математическое ожидание частоты $\nu_n = \xi_n/n$ появлений события A в серии из n испытаний, согласно (1.45) и (2.3), $\mathbf{M}\nu_n = p$, дисперсия, по (1.52) и (2.3), $\mathbf{D}\nu_n = \frac{p(1-p)}{n}$.

По неравенству Чебышева (1.57), для любого $\varepsilon > 0$, $\mathbf{P}\left\{\left|\frac{\xi_n}{n} - p\right| \geqslant \varepsilon\right\} \leqslant \frac{\mathbf{D}\xi_n}{\varepsilon^2 n^2} = \frac{p(1-p)}{\varepsilon^2 n} \to 0$ при $n \to \infty$, что полностью соответствует статистическому определению вероятности. Эта форма закона больших чисел дает основание для оценки вероятности события частотой его появления в независимых испытаниях. Дисперсия $\mathbf{D}\nu_n$ характеризует достоверность такой оценки.

Схема Бернулли обобщается на тот случай, когда в каждом испытании может произойти одно из нескольких несовместных событий $A_1, A_2, \ldots, A_k$, образующих полную группу. Вероятность того, что в n испытаниях A_1 произойдет n_1 раз, A_2 — n_2 раз и т. д., A_k — n_k раз, имеет вид

$$P_n(n_1, n_2, \ldots, n_k) = \frac{n!}{n_1!\, n_2! \ldots n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}, \tag{2.1'}$$

где $p_1, p_2, \ldots, p_k$ — вероятности событий $A_1, A_2, \ldots, A_k$; $\sum_{i=1}^{k} p_i = 1$, $\sum_{i=1}^{k} n_i = n$. Эту схему называют *обобщенной схемой Бернулли*, а распределение — *полиномиальным*.

Пример 2.1. На территории бурят шесть скважин. Вероятность встречи рудной залежи одной скважиной 0,3. Считая бурение отдельных скважин независимыми испытаниями, вычислить вероятность того, что: 1) две скважины встретят рудную залежь; 2) ни одна ее не встретит; 3) встретит хотя бы одна; 4) встретят не менее двух скважин; 5) выяснить наиболее вероятное количество скважин, которые встретят рудную залежь.

Решение. 1) По формуле (2.1) вероятность того, что две скважины встретят рудную залежь,

$$P_6(2) = \frac{6!}{2!\,4!}\,0{,}3^2 \cdot 0{,}7^4 = \frac{6 \cdot 5}{2}\,0{,}3^2 \cdot 0{,}7^4 \approx 0{,}324.$$

2) Вероятность того, что ни одна скважина не принесет успеха,

$$P_6(0) = \frac{6!}{0!\,6!}\,0{,}3^0 \cdot 0{,}7^6 \approx 0{,}118.$$

3) Вероятность того, что рудную залежь встретит хотя бы одна скважина, состоит из суммы вероятностей $P_1 = P_6(1) + P_6(2) + \cdots + P_6(6)$. Так как сумма всех вероятностей $P_6(i)$ от $i = 0$ до $i = 6$ равна единице, имеем $P_1 = 1 - P_6(0) \approx \approx 1 - 0{,}118 = 0{,}882$.

4) Аналогично предыдущему

$$P_2 = 1 - P_6(0) - P_6(1) = 1 - 0{,}118 - \frac{6!}{1!\,5!}\,0{,}3 \cdot 0{,}7^5 \approx 0{,}58.$$

5) Наиболее вероятное количество успешно пробуренных скважин — модальное значение $M_0 = [(n+1)\,p] = [7 \cdot 0{,}3] = 2$. Вероятность, соответствующая этому значению $P_6(2) \approx 0{,}324$.

Пример 2.2. При разбуривании структур нефтегазоносного региона вероятность встречи среди них продуктивных (коэффициент удачи) — 0,1. Сколько структур нужно разбурить, чтобы с вероятностью не менее 0,9 встретить хотя бы одну продуктивную?

Решение. Аналогично п. 3 предыдущего примера получим условие $1 - P_n(0) = 1 - 0{,}1^0 \cdot 0{,}9^n \geqslant 0{,}9$, откуда $0{,}9^n \leqslant 0{,}1$, $n \geqslant \lg 0{,}1/\lg 0{,}9 \approx 21{,}74$. Наименьшее число испытаний, удовлетворяющее условию задачи, $n = 22$.

Закон Пуассона. Случайная величина, распределенная по *закону Пуассона*, принимает целочисленные значения с вероятностями:

$$\mathbf{P}\{\xi = m\} = e^{-\lambda}\frac{\lambda^m}{m!}, \quad m = 0,\ 1,\ 2,\ \ldots \tag{2.7}$$

Вероятностная схема, реализующая этот тип распределения, дается следующим известным положением теории вероятностей. Если число n независимых испытаний в схеме Бернулли возрастает, а вероятность p_n события A в отдельном испытании уменьшается так, что $p_n n$ стремится к некоторой постоянной λ, то вероятность того, что A произойдет m раз, стремится к $e^{-\lambda}\frac{\lambda^m}{m!}$ при каждом $m = 0,\ 1,\ 2,\ \ldots$. Так как из условия $p_n n \to \lambda$ следует $p_n \to 0$, то закон Пуассона описывает распределение числа появлений *редкого события*, т. е. такого, которое в отдельном испытании или в условиях, интерпретирующих такое испытание, имеет малую вероятность*.

Функция распределения, в соответствии с (1.34), представляется в виде

$$F(x) = \begin{cases} \displaystyle\sum_{0 \leqslant i < x} e^{-\lambda}\frac{\lambda^i}{i!}, & x > 0, \\ 0, & x \leqslant 0. \end{cases} \tag{2.8}$$

В отличие от биномиального закона, который охватывает двухпараметрическое семейство функций распределения (параметры n и p),

* Аппроксимацию биномиального распределения распределением Пуассона применяют при $p < 0{,}1$, полагая $\lambda = np$.

семейство распределений Пуассона является однопараметрическим (параметр λ). Величина λ имеет смысл математического ожидания: по определению (1.39)

$$\mathbf{M}\xi = \sum_{i=0}^{\infty} i e^{-\lambda} \frac{\lambda^i}{i!} = \sum_{i=1}^{\infty} \lambda e^{-\lambda} \frac{\lambda^{i-1}}{(i-1)!} = e^{-\lambda}\lambda \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = \lambda. \qquad (2.9)$$

Дисперсия также равна λ:

$$\mathbf{D}\xi = \mathbf{M}\xi^2 - (\mathbf{M}\xi)^2 = \sum_{i=0}^{\infty} i^2 e^{-\lambda} \frac{\lambda^i}{i!} - \lambda^2 = e^{-\lambda}\lambda \sum_{i=1}^{\infty} (i-1) \frac{\lambda^{i-1}}{(i-1)!} +$$

$$+ e^{-\lambda}\lambda \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!} - \lambda^2 = \lambda^2 + \lambda - \lambda^2 = \lambda. \qquad (2.10)$$

Если λ — дробное число, распределение Пуассона имеет одно модальное значение, равное $[\lambda]$ — целой части λ; если λ — целое число, то у распределения две моды, λ и $\lambda - 1$. В этом нетрудно убедиться, рассмотрев отношение вероятностей (2.7) для двух последовательных значений m: $\dfrac{\mathbf{P}\{\xi = m+1\}}{\mathbf{P}\{\xi = m\}} = \dfrac{\lambda}{m+1}$.

Асимметрия и эксцесс имеют вид

$$A = \frac{1}{\sqrt{\lambda}}, \quad E = \frac{1}{\lambda}, \qquad (2.11)$$

откуда видно, что распределение Пуассона правоасимметрично и имеет положительный эксцесс.

Можно показать, что распределение суммы двух независимых случайных величин ξ_1 и ξ_2, распределенных по закону Пуассона с параметрами λ_1 и λ_2, также подчиняется закону Пуассона с параметром $\lambda = \lambda_1 + \lambda_2$:

$$\mathbf{P}\{\xi_1 + \xi_2 = m\} = e^{-\lambda_1 - \lambda_2} \frac{(\lambda_1 + \lambda_2)^m}{m!}, \quad m = 0, 1, 2, \ldots \qquad (2.12)$$

В практике лабораторных исследований закон Пуассона используется при анализе результатов измерений, в процессе которых суммируются количества импульсов, фиксирующих радиоактивные распады. Обоснованием этому служит следующее.

Вероятность появления импульса в промежутке времени от t до $t + \Delta t$, очевидно, не зависит от t, а лишь от величины Δt. Кроме того, длительность импульса можно считать малой по сравнению с временем между двумя последовательными импульсами. Вполне естественно полагать, что вероятность появления импульса в малом промежутке времени Δt с точностью до $o(\Delta t)$* пропорциональна величине Δt, т. е. равна $\lambda \Delta t + o(\Delta t)$, а вероятность появления двух или более импульсов — $o(\Delta t)$. Обозначим $P_m(t)$ вероятность того, что

* $o(\Delta t)$ — величина более высокого порядка малости по сравнению с Δt.

с начального момента времени $t_0 = 0$ до момента t появится m импульсов. $P_m(t + \Delta t)$ можно подсчитать как вероятность суммы следующих событий: A_0 — до момента t появится m импульсов, а от t до $t + \Delta t$ ни одного импульса; A_1 — к моменту t появится $m - 1$ импульсов, а от t до $t + \Delta t$ 1 импульс; A_2 — к моменту t $m - 2$ импульсов, а от t до $t + \Delta t$ 2 импульса и т. д. до A_m.

Так как события $A_1, A_2, \ldots, A_m$ несовместны, по формуле (1.2) имеем $P_m(t + \Delta t) = \sum_{i=0}^{m} \mathbf{P}(A_i)$. Используя формулу умножения вероятностей независимых событий (1.10), получим: $\mathbf{P}(A_0) = P_m(t)\,[1 - \lambda\Delta t - o(\Delta t)]$; $\mathbf{P}(A_1) = P_{m-1}(t)\,[\lambda\Delta t + o(\Delta t)]$; $\mathbf{P}(A_2) = P_{m-2}(t) \cdot o(\Delta t), \ldots,$ $\mathbf{P}(A_m) = P_0(t) \cdot o(\Delta t)$, так что

$$P_m(t + \Delta t) = P_m(t)(1 - \lambda\Delta t) + \lambda P_{m-1}(t)\,\Delta t + o(\Delta t) \quad (m \geqslant 0)$$

и

$$\frac{P_m(t + \Delta t) - P_m(t)}{\Delta t} = -\lambda P_m(t) + \lambda P_{m-1}(t) + \frac{o(\Delta t)}{\Delta t}.$$

Перейдя к пределу при $\Delta t \to 0$, получим

$$\frac{dP_m(t)}{dt} = -\lambda P_m(t) + \lambda P_{m-1}(t). \tag{2.13}$$

При $m = 0$ полученное дифференциальное уравнение имеет вид

$$\frac{dP_0(t)}{dt} = -\lambda P_0(t),$$

решение которого с учетом $P_0(0) = 1$ даст вероятность того, что к моменту t не будет ни одного импульса: $P_0(t) = e^{-\lambda t}$. При $m = 1$ решение уравнения (2.13) запишется в виде $P_1(t) = \lambda t e^{-\lambda t}$.

Пользуясь методом математической индукции, легко показать, что решение уравнения (2.13) имеет вид

$$P_m(t) = \frac{(\lambda t)^m}{m!}\, e^{-\lambda t}, \tag{2.14}$$

т. е. распределение количества импульсов за время t подчиняется закону Пуассона с параметром λt. Так как математическое ожидание распределения, определяемого (2.14), равно λt, величина λ имеет смысл среднего количества импульсов в единицу времени.

Пример 2.3. Вероятность встречи аномалии на площади опробования составляет 0,015. Вычислить вероятность того, что из 100 отобранных проб: 1) две будут аномальны; 2) не менее двух будут аномальны.

Решение. Учитывая, что вероятность аномалии мала, используем для описания распределения числа ξ аномальных проб закон Пуассона. В данном случае параметр λ равен математическому ожиданию количества аномальных проб. Согласно (2.3) $\lambda = np = 100 \cdot 0{,}015 = 1{,}5$.

1) По формуле (2.7) вероятность того, что две пробы окажутся аномальными — $\mathbf{P}\{\xi = 2\} = e^{-1,5}\,\frac{1{,}5^2}{2!} \approx 0{,}251$ (см. Приложение, табл. 1).

2) Вероятность того, что не менее двух проб будут аномальными — $\mathbf{P}\{\xi \geqslant 2\} = \sum_{i=2}^{\infty} \mathbf{P}\{\xi = i\}$. Для вычисления $\mathbf{P}\{\xi \geqslant 2\}$ воспользуемся тем, что $\mathbf{P}\{\xi \geqslant 2\} =$

$= 1 - \mathbf{P}\{\xi < 2\}$, где $\mathbf{P}\{\xi < 2\}$ вероятность противоположного события. Очевидно,

$$\mathbf{P}\{\xi < 2\} = \sum_{i=0}^{1} \mathbf{P}\{\xi = i\} = e^{-1,5}\left(\frac{1,5^0}{0!} + \frac{1,5}{1!}\right) \approx 0,558; \quad \mathbf{P}\{\xi \geqslant 2\} \approx 0,442.$$

Пример 2.4. Сколько нужно отобрать независимых проб, чтобы получить хотя бы одну аномальную пробу с вероятностью, не меньшей 0,9, если вероятность встречи аномалии на данном объекте равна 0,05?

Решение. Аналогично предыдущему примеру $\lambda = np = 0,05n$. Вероятность получения хотя бы одной аномальной пробы

$$\mathbf{P}\{\xi \geqslant 1\} = 1 - \mathbf{P}\{\xi = 0\} = 1 - e^{-0,05n}\frac{(0,05n)^0}{0!} = 1 - e^{-0,05n} \geqslant 0,9,$$

откуда $n \geqslant \dfrac{-\lg 0,1}{0,05 \lg e} \approx 46,05$.

Наименьшее количество проб, удовлетворяющее условию задачи, $n = 47$.

Пример 2.5. Какой должна быть минимальная экспозиция, чтобы обеспечить коэффициент вариации ошибок измерений общей радиоактивности не более 5%, если среднее число импульсов в секунду для данных пород не менее десяти?

Решение. При экспозиции t секунд математическое ожидание количества импульсов $\mathbf{M}\xi = \lambda t \geqslant 10t$; дисперсия $\mathbf{D}\xi = \mathbf{M}\xi = \lambda t$. Коэффициент вариации

$$V = \sqrt{\mathbf{D}\xi}/\mathbf{M}\xi = \frac{1}{\sqrt{\lambda t}} \leqslant \frac{1}{\sqrt{10t}}.$$

Условие задачи $V \leqslant 0,05$ будет соблюдено, если $\dfrac{1}{\sqrt{10t}} \leqslant 0,05$. Из этого неравенства следует $t \geqslant \dfrac{1}{10 \cdot 0,05^2} = 40$ *сек*. Минимальная экспозиция — 40 *сек*.

Гипергеометрическое распределение. Этому дискретному распределению соответствует схема, именуемая *выбором без возвращения*. Пусть совокупность N элементов состоит из M элементов, обладающих определенным свойством (элементов первого типа) и $N - M$, не обладающим этим свойством (элементов второго типа). Из совокупности отбирается выборка — группа n элементов, причем каждый выбирается произвольно. При таком выборе все имеющиеся в совокупности элементы равновозможны. Вероятность того, что в выборке окажется m элементов первого типа и $n - m$ второго, составит

$$P_n(m) = \frac{C_M^m C_{N-M}^{n-m}}{C_N^n} = \frac{(N-n)!\, n!\, M!\, (N-M)!}{N!\, (M-m)!\, m!\, (n-m)!\, (N-M-n+m)!}. \quad (2.15)$$

Эти вероятности и определяют гипергеометрическое распределение.

Формула (2.15) доказывается с помощью классического определения вероятности. $P_n(m)$ определяется как отношение количества различных вариантов такого выбора n элементов из N, чтобы среди них было m элементов первого типа и $n - m$ второго, к числу всех возможных вариантов выбора n элементов из N. Всего таких вариантов столько, сколько существует сочетаний из N элементов по n: C_N^n. Число вариантов выбора m элементов из M первого типа C_M^m; при каждом из них C_{N-M}^{n-m} вариантов выбора $n - m$ элементов из

$N - M$ второго типа. Число интересующих нас вариантов, следовательно, $C_M^m C_{N-M}^{n-m}$.

Математическое ожидание и дисперсия случайной величины ξ_n, принимающей значения, равные количествам m элементов первого типа в выборке n элементов:

$$\mathbf{M}\xi_n = \frac{nM}{N}, \quad \mathbf{D}\xi_n = \frac{nM(N-n)(N-M)}{N^2(N-1)}; \tag{2.16}$$

асимметрия

$$A = \frac{(N-2n)(N-2M)}{N(N-2)\sqrt{\mathbf{D}\xi_n}}. \tag{2.16'}$$

Гипергеометрическое распределение находит применение при статистическом обследовании больших партий образцов и для описания некоторых специфических распределений. В промышленности оно широко используется при статистическом контроле качества продукции.

При планировании выборочного контроля используется формула (2.15): определяют набор правил контроля всей партии образцов по результатам анализа выборки и сопоставляют вероятности различных исходов в зависимости от n и количества образцов данного типа (пород, обладающих определенным признаком) в партии. Например, если правило состоит в проверке условия $\xi_n \geqslant k$, т. е. в выборке должно быть не менее k образцов данного типа, то при различных n и k строят зависимость вероятности $\mathbf{P}\{\xi_n \geqslant k\}$ от числа M образцов этого типа в партии. Из условия $\mathbf{P}\{\xi_n \geqslant k\} \leqslant q$ при малом q и фиксированном $M = M_0$ находят оптимальные n и k, обеспечивающие контроль нижнего предела M_0 для $M \geqslant M_0$.

Если число N элементов во всей совокупности намного превышает объем n выборки, удовлетворительную аппроксимацию гипергеометрического распределения дают биномиальный закон или закон Пуассона. Первый рекомендуется использовать при $n < 0{,}1 \cdot N$, полагая параметр $p = \frac{M}{N}$; второй — при $n < 0{,}1N$ и $M < 0{,}1N$, полагая параметр $\lambda = \frac{nM}{N}$. Для приближенных вычислений можно использовать *формулу Стирлинга*: $n! \approx n^n e^{-n}\sqrt{2\pi n}$. При расчетах по (2.15) удобно пользоваться таблицами значений $\lg n!$, приведенными в [2].

Пример 2.6. В партии из 100 образцов гранитов 20 имеют урановую минерализацию. Для анализа шлифов из нее отобрано произвольно 10 образцов. Вычислить вероятность того, что анализ шлифов обнаружит урановую минерализацию: 1) в двух образцах $P_{10}(2)$; 2) хотя бы в одном образце $\mathbf{P}\{\xi_n \geqslant 1\}$; 3) вычислить вероятность $P_{10}(2)$ с помощью аппроксимации биномиальным распределением и сравнить ее с полученной по формуле (2.15).

Решение. 1) По формуле (2.15) имеем

$$P_{10}(2) = \frac{C_{20}^2 C_{80}^8}{C_{100}^{10}} = \frac{90!\,10!\,20!\,80!}{100!\,2!\,18!\,8!\,72!} \approx 0{,}318.$$

2) Вычислим вероятность противоположного события, т. е. того, что ни в одном шлифе урановая минерализация не будет обнаружена:

$$P_{10}(0) = \frac{C_{20}^{0} C_{80}^{10}}{C_{100}^{10}} = \frac{C_{80}^{10}}{C_{100}^{10}} \approx 0{,}095.$$

Искомоя вероятность $\mathbf{P}\{\xi \geqslant 1\} = 1 - P_{10}(0) \approx 0{,}905$.

3) Положив $p = \frac{20}{100} = 0{,}2$ — вероятность того, что выбранный наудачу образец будет иметь урановую минерализацию, по формуле (2.1) получим: $P_{10}(2) = C_{10}^{2} 0{,}2^{2} 0{,}8^{8} \approx 0{,}302$.

Сравнение этой величины с рассчитанной по формуле (2.15) показывает эффективность аппроксимации гипергеометрического распределения биномиальным.

§ 2. Непрерывные распределения

Нормальный закон распределения. *Нормальному закону распределения (Гаусса)* следуют случайные величины, имеющие плотность распределения общего вида

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}, \qquad (2.17)$$

заданную на всей действительной оси (m и σ — параметры). Функция нормального распределения

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-\frac{(t-m)^2}{2\sigma^2}} dt. \qquad (2.17')$$

В математической статистике нормальный закон распределения имеет фундаментальное значение. Ее разветвленный аппарат включает множество статистических правил и методов, имеющих к этому закону непосредственное отношение. Многие из них относятся к задачам, постановка которых предполагает именно нормальное распределение случайных величин, служащих предметом изучения. Его ведущее положение объясняется и широким распространением в технических приложениях, возможностью интерпретации с его помощью распределений ошибок разнообразных методов измерений.

Общие предпосылки, обусловливающие доминирующее положение нормального закона, дает важнейший вывод теории вероятностей, известный под названием *центральной предельной теоремы*. Сущность ее сводится к следующему. Распределение случайной величины, являющейся результатом суммарного воздействия большого числа независимых случайных факторов, интенсивность влияния каждого из которых в отдельности пренебрежимо мала, описывается (по крайней мере, приближенно) нормальным законом. Поскольку эта теорема имеет важное значение, приведем ее точную формулировку.

Если последовательность взаимно независимых случайных величин $\xi_1, \xi_2, \ldots, \xi_n \ldots$ при любом значении $\tau > 0$ удовлетворяет *условию Линдеберга* —

$$\lim_{n\to\infty} B_n^{-2} \sum_{k=1}^{n} \int\limits_{|x-\mathbf{M}\xi_k|>\tau B_n} (x-\mathbf{M}\xi_k)^2 p_k(x)\,dx = 0, \tag{2.18}$$

где $p_k(x)$ — плотность распределения* ξ_k, $k = 1, 2, \ldots$, $B_n^2 = \mathbf{D}\left(\sum\limits_{k=1}^{n} \xi_k\right) = \sum\limits_{k=1}^{n} \mathbf{D}\xi_k$, то при $n \to \infty$

$$\mathbf{P}\left\{B_n^{-1} \sum_{k=1}^{n} (\xi_k - \mathbf{M}\xi_k) < x\right\} \to \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-\frac{z^2}{2}}\,dz \tag{2.19}$$

равномерно относительно x, т. е. функция распределения линейно преобразованной в виде нормированного уклонения суммы $\sum\limits_{k=1}^{n} \xi_k$ стремится с ростом числа слагаемых n к функции нормального распределения (2.17′) с параметрами $m = 0$ и $\sigma = 1$. Важность и содержание теоремы иллюстрируют такие следствия.

1. Если $\eta_1, \eta_2, \ldots, \eta_n, \ldots$ независимы, одинаково распределены и имеют конечную отличную от нуля дисперсию D, а $c_1, c_2, \ldots, c_n, \ldots$ — последовательность чисел, ограниченных снизу и сверху положительными числами, $0 < a \leqslant c_i \leqslant A$, $i = 1, 2, \ldots, n, \ldots$, то для величин $\xi_i = c_i\eta_i$ справедливо (2.19).

Имеем: $B_n^2 = \mathbf{D}\left(\sum\limits_{i=1}^{n} c_i\eta_i\right) = \sum\limits_{i=1}^{n} c_i^2 \mathbf{D}\eta_i \geqslant a^2 nD$. Плотность распределения ξ_k, согласно (1.25), $p_k(x) = p\left(\frac{x}{c_k}\right)\frac{1}{c_k}$, где $p(z)$ — плотность распределения η_k. Проверим условие Линдеберга:

$$B_n^{-2} \sum_{k=1}^{n} \int\limits_{|x-\mathbf{M}\xi_k|>\tau B_n} (x-\mathbf{M}\xi_k)^2 \frac{1}{c_k} p\left(\frac{x}{c_k}\right) dx \leqslant \frac{1}{a^2 nD} \sum_{k=1}^{n} \int\limits_{c_k|y-\mathbf{M}\eta_k|>\tau B_n} c_k^2 \times$$

$$\times (y-\mathbf{M}\eta_k)^2 p(y)\,dy \leqslant A^2 D^{-1} a^{-2} \int\limits_{|y-\mathbf{M}\eta_k|>\tau a\frac{\sqrt{Dn}}{A}} (y-\mathbf{M}\eta_k)^2 p(y)\,dy \to$$

$$\to 0 \text{ при } n \to \infty.$$

Так как дисперсия D конечна, то согласно ее определению (1.49) последний интеграл стремится к нулю.

Из приведенного следствия вытекает, что теорема справедлива для сумм независимых и одинаково распределенных величин, имеющих

---

* Величины ξ_k не обязательно должны иметь плотность распределения. Здесь $p_k(x)$ вводятся лишь для облегчения формулировки теоремы.

отличную от нуля конечную дисперсию, в частности для средних арифметических, составленных из независимых наблюдений одной и той же величины.

2. Если $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ независимы, имеют конечные дисперсии, ограниченные снизу положительным числом — $\mathbf{D}\xi_k \geqslant a > 0$ и четвертые центральные моменты, ограниченные сверху — $\mu_{k4} = \mathbf{M}(\xi_k - \mathbf{M}\xi_k)^4 \leqslant A$, то распределение $\eta_n = \sum_{i=1}^{n} \xi_i$ при $n \to \infty$ приближается к нормальному закону, т. е. справедливо (2.19).

Имеем $B_n^{-2} = \left(\sum_{i=1}^{n} \mathbf{D}\xi_i\right)^{-1} \leqslant (an)^{-1}$ и $B_n^{-2} \sum_{k=1}^{n} \int\limits_{|x-\mathbf{M}\xi_k| > \tau B_n} (x - \mathbf{M}\xi_k)^2 \times$

$$\times p_k(x)\,dx \leqslant B_n^{-2} \sum_{k=1}^{n} \int\limits_{\frac{(x-\mathbf{M}\xi_k)^2}{\tau^2 B_n^2} > 1} (x - \mathbf{M}\xi_k)^2 \frac{(x-\mathbf{M}\xi_k)^2}{\tau^2 B_n^2} p_k(x)\,dx \leqslant$$

$$\leqslant \frac{1}{n^2 a^2 \tau^2} \sum_{k=1}^{n} \int\limits_{-\infty}^{\infty} (x - \mathbf{M}\xi_k)^4 p_k(x)\,dx \leqslant \frac{A}{na^2\tau^2} \to 0$$

(учитывая, что в области интегрирования $|x - \mathbf{M}\xi_k| > \tau B_n$, $1 < \frac{(x-\mathbf{M}\xi_k)^2}{\tau^2 B_n^2}$).

Более общие условия, обеспечивающие сходимость (2.19), дает *теорема Ляпунова*: если $\mathbf{M}\xi_k$, $\mathbf{D}\xi_k$, $T_k = \mathbf{M}|\xi_k - \mathbf{M}\xi_k|^{2+\delta}$ при $\delta > 0$ конечны и $\lim\limits_{n\to\infty} \sum\limits_{k=1}^{n} T_k / B_n^{1+\delta/2} = 0$, то имеет место (2.19).

Распределения ошибок многих методов измерений, согласуясь с центральной предельной теоремой, подчиняются нормальному закону распределения. К таким методам относятся: весовой метод химического анализа проб, пикнометрический метод измерения удельного веса, метод гидростатического взвешивания измерения плотности и др. Локальные распределения содержаний химических элементов и минералов (в пределах малых участков), по А. Б. Вистелиусу, также согласуются с нормальным законом. Центральная предельная теорема обеспечивает возможность аппроксимации распределений многих величин, употребляемых при статистическом анализе геолого-геофизических данных, нормальным законом.

Параметры m и σ, участвующие в формуле плотности нормального распределения (2.17), имеют смысл математического ожидания и среднего квадратического отклонения: $m = \mathbf{M}\xi$, $\sigma = \sqrt{\mathbf{D}\xi}$. В этом можно убедиться, вычислив интегралы

$$\mathbf{M}\xi = \frac{1}{\sqrt{2\pi}\sigma} \int\limits_{-\infty}^{\infty} x e^{-\frac{(x-m)^2}{2\sigma^2}} dx = m, \quad \mathbf{D}\xi = \frac{1}{\sqrt{2\pi}\sigma} \int\limits_{-\infty}^{\infty} (x-m)^2 e^{-\frac{(x-m)^2}{2\sigma^2}} dx = \sigma^2; \tag{2.20}$$

первый — с помощью замены $u = (x - m)/\sigma$, второй — тоже с этой заменой, интегрируя по частям и учитывая, что $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz = 1$ по свойству (1.22).

Поскольку семейство (2.17) двухпараметрическое, функция нормального распределения, а следовательно, и все его числовые характеристики полностью определяются математическим ожиданием и дисперсией. Из выражения плотности (2.17) видно, что она изображается кривой, симметричной относительно m. Поэтому асимметрия нормального распределения равна нулю. Непосредственным вычислением можно убедиться в том, что четвертый центральный момент

$$\mu_4 = 3\sigma^4, \tag{2.21}$$

следовательно, коэффициент эксцесса тоже равен нулю: $E = \frac{\mu_4}{\sigma^4} - 3 = 0$.

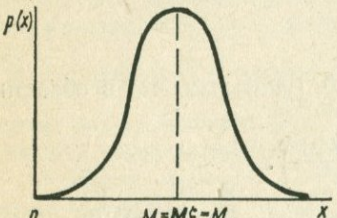


Рис. 17.

Вообще моменты нечетных порядков равны нулю — $\mu_{2k+1} = 0$, а четных — $\mu_{2k} = (2k-1)(2k-3)\ldots 3 \cdot 1 \cdot \sigma^{2k}$ $(k = 1, 2, \ldots)$.

По величине асимметрии и эксцесса любой случайной величины судят о степени отклонения ее распределения от нормального закона.

Так как кривая плотности нормального распределения симметрична, медиана равна математическому ожиданию, т. е. тоже параметру m: $M_e = \mathbf{M}\xi = m$ (рис. 17). Приравняв производную от (2.17) нулю, увидим, что мода совпадает с математическим ожиданием и медианой:

$$M_0 = M_e = \mathbf{M}\xi = m. \tag{2.22}$$

Любая невырожденная линейная комбинация независимых нормально распределенных случайных величин также распределена нормально*. В частности, если ξ нормально распределена, то и распределение $\eta = \alpha\xi + \beta$ ($\alpha \neq 0$, α и β — постоянные) будет гауссовским. Это непосредственно следует из (1.25):

$$p_\eta(x) = \frac{1}{|\alpha|\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{1}{2\sigma^2}\left[\frac{x-\beta}{\alpha} - m\right]^2\right\} = \frac{1}{\sqrt{2\pi}\,|\alpha|\,\sigma} \times$$
$$\times \exp\left\{\frac{-1}{2(\alpha\sigma)^2}(x - \alpha m - \beta)^2\right\}. \tag{2.23}$$

Будучи результатом линейного преобразования, нормированное уклонение $(\xi - \mathbf{M}\xi)/\sqrt{\mathbf{D}\xi}$ нормально распределенной случайной величины ξ

* Это свойство справедливо и для зависимых гауссовских случайных величин — тех, которые в совокупности подчиняются многомерному нормальному закону распределения (гл. 3).

тоже распределяется по нормальному закону, с нулевым математическим ожиданием и дисперсией, равной единице (1.59). Плотность распределения нормированного уклонения, согласно (2.17) и (2.20), с учетом того, что для него $m = 0$, $\sigma = 1$,

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}; \tag{2.24}$$

функция распределения —

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt. \tag{2.25}$$

С помощью функции (2.25) нетрудно построить функцию любого нормального распределения по величинам его математического ожидания $\mathbf{M}\xi$ и дисперсии $\mathbf{D}\xi$:

$$F(x) = \mathbf{P}\{\xi < x\} = \mathbf{P}\left\{\frac{\xi - \mathbf{M}\xi}{\sqrt{\mathbf{D}\xi}} < \frac{x - \mathbf{M}\xi}{\sqrt{\mathbf{D}\xi}}\right\} = \Phi\left(\frac{x - \mathbf{M}\xi}{\sqrt{\mathbf{D}\xi}}\right). \tag{2.26}$$

Квантиль z_q порядка q для любой нормально распределенной случайной величины ξ можно определить с помощью таблицы значений функции $\Phi(x)$ (2.25) по формуле

$$z_q = \mathbf{M}\xi + u_q \sqrt{\mathbf{D}\xi} = \mathbf{M}\xi + u_q \sigma, \tag{2.27}$$

где $u_q = \Phi^{-1}(q)$ — q-квантиль функции распределения $\Phi(x)$ ($\Phi(u_q) = q$), так как

$$F(z_q) = \Phi\left(\frac{z_q - \mathbf{M}\xi}{\sqrt{\mathbf{D}\xi}}\right) = \Phi(u_q) = q$$

($F(z)$ — функция распределения ξ).

Значения $\Phi(z)$ для z с шагом 0,001 сведены в табл. 2 (Приложение). В дальнейшем мы часто будем обращаться к этой таблице и использовать формулу (2.27).

Можно пользоваться также таблицей значений функции

$$\varphi_0(x) = \frac{1}{\sqrt{2\pi}} \int_{0}^{x} e^{-\frac{t^2}{2}} dt, \quad x \geqslant 0, \tag{2.28}$$

учитывая очевидное свойство

$$\Phi(x) = \begin{cases} 0{,}5 - \varphi_0(-x), & x < 0, \\ 0{,}5 + \varphi_0(x), & x \geqslant 0. \end{cases} \tag{2.29}$$

Используя (1.18) и (2.26), легко получить вероятность $\mathbf{P}\{x_0 \leqslant \xi < x_1\}$ с помощью той же таблицы значений $\Phi(z)$:

$$\mathbf{P}\{x_0 \leqslant \xi < x_1\} = F(x_1) - F(x_0) = \Phi\left(\frac{x_1 - \mathbf{M}\xi}{\sqrt{\mathbf{D}\xi}}\right) - \Phi\left(\frac{x_0 - \mathbf{M}\xi}{\sqrt{\mathbf{D}\xi}}\right). \tag{2.30}$$

Часто для случайной величины нужно определить два предела — нижний и верхний, вероятность получения значения между которыми

равна заданной величине q (как правило, близкой к единице). Равенством $\mathbf{P}\{x_0 \leqslant \xi < x_1\} = q$ пределы x_0 и x_1 определяются неоднозначно. В качестве нижнего предела x_0 можно взять любой квантиль порядка $\alpha < 1 - q$, а в качестве верхнего x_1 — квантиль порядка $\beta = q + \alpha$: $\mathbf{P}\{x_0 \leqslant \xi < x_1\} = q + \alpha - \alpha = q$. Поэтому обычно используется такое дополнительное условие:

$$\mathbf{P}\{\xi < x_0\} = \mathbf{P}\{\xi \geqslant x_1\} = \frac{1-q}{2}, \tag{2.31}$$

т. е. вероятности выхода в ту или другую сторону от интервала $[x_0, x_1)$ одинаковы и равны по $\frac{1-q}{2}$. Тогда вероятность того, что случайная величина получит значение из этого интервала, равна $1 - 2\frac{1-q}{2} = q$.

Для нормально распределенной случайной величины, по (2.31)

$$\mathbf{P}\{\xi < x_0\} = \Phi\left(\frac{x_0 - \mathbf{M}\xi}{\sqrt{\mathbf{D}\xi}}\right) = \frac{1-q}{2}, \quad \mathbf{P}\{\xi \geqslant x_1\} = 1 - \Phi\left(\frac{x_1 - \mathbf{M}\xi}{\sqrt{\mathbf{D}\xi}}\right) = \frac{1-q}{2},$$

поэтому пределы x_0 и x_1, для которых $\mathbf{P}\{x_0 \leqslant \xi < x_1\} = q$ при условии (2.31), определяются формулами

$$x_0 = \mathbf{M}\xi + u_{\frac{1-q}{2}}\sqrt{\mathbf{D}\xi}, \quad x_1 = \mathbf{M}\xi + u_{\frac{1+q}{2}}\sqrt{\mathbf{D}\xi}, \tag{2.32}$$

где $u_{\frac{1-q}{2}}$, $u_{\frac{1+q}{2}}$ — квантили (0; 1)-нормального распределения* порядков $\frac{1-q}{2}$, $\frac{1+q}{2}$, определяемые из табл. 2 (Приложение). Так как это распределение симметрично относительно нуля, $u_\alpha = -u_{1-\alpha}$ (рис. 18). Поэтому $u_{\frac{1-q}{2}} = -u_{\frac{1+q}{2}}$, причем $u_{\frac{1+q}{2}} \geqslant 0$ как квантиль порядка $\frac{1+q}{2}$, не меньшего 0,5. Таким образом, пределы (2.32) для нормально распределенной случайной величины симметричны относительно $\mathbf{M}\xi$:

$$x_0 = \mathbf{M}\xi - u_{\frac{1+q}{2}}\sqrt{\mathbf{D}\xi}, \quad x_1 = \mathbf{M}\xi + u_{\frac{1+q}{2}}\sqrt{\mathbf{D}\xi}, \tag{2.32'}$$

где q — вероятность того, что ξ окажется в этих пределах.

Для часто употребляемых значений этой вероятности 0,9, 0,95 и 0,99 они выражаются в виде

$$\begin{aligned} &\mathbf{M}\xi \pm 1{,}645\sqrt{\mathbf{D}\xi}, \quad q = 0{,}9; \\ &\mathbf{M}\xi \pm 1{,}96\sqrt{\mathbf{D}\xi}, \quad q = 0{,}95; \\ &\mathbf{M}\xi \pm 2{,}576\sqrt{\mathbf{D}\xi}, \quad q = 0{,}99, \end{aligned} \tag{2.33}$$

так как по табл. 2 (Приложение) $u_{0,95} = 1{,}645$, $u_{0,975} = 1{,}96$, $u_{0,995} = 2{,}576$.

---

* Здесь и дальше (0; 1)-нормальное распределение обозначает нормальное распределение с нулевым математическим ожиданием и дисперсией, равной единице.

Наоборот, если задан интервал $[x_0, x_1)$, то вероятность q, которая ему соответствует, определяется по значениям функции (0; 1)-нормального распределения $\Phi(x)$ в точках $u_0 = \frac{x_0 - \mathbf{M}\xi}{\sqrt{\mathbf{D}\xi}}$, $u_1 = \frac{x_1 - \mathbf{M}\xi}{\sqrt{\mathbf{D}\xi}}$:

$$q = \Phi\left(\frac{x_1 - \mathbf{M}\xi}{\sqrt{\mathbf{D}\xi}}\right) - \Phi\left(\frac{x_0 - \mathbf{M}\xi}{\sqrt{\mathbf{D}\xi}}\right). \tag{2.34}$$

Например, для пределов $x_0 = \mathbf{M}\xi - \sqrt{\mathbf{D}\xi}$, $x_1 = \mathbf{M}\xi + \sqrt{\mathbf{D}\xi}$ по табл. 2 (Приложение) $q = \Phi(1) - \Phi(-1) = 0{,}841 - 0{,}159 = 0{,}682$. Для характеристики ошибок измерений в технических приложениях нередко используют пределы $\mathbf{M}\xi \pm 3\sigma$, полагая $\sigma = \sqrt{\mathbf{D}\xi}$ («правило 3σ»). По (2.34) и табл. 2 (Приложение) с учетом интерполяции четвертого знака, им соответствует вероятность $q = 0{,}997$.

Пределы (2.33) употребляют в качестве критических границ для отбраковки наблюдений, аномальных по отношению к некоторой нормально распределенной величине ξ. Если наблюдение превышает $\mathbf{M}\xi + u_{\frac{1+q}{2}}\sqrt{\mathbf{D}\xi}$ или оказывается меньшим $\mathbf{M}\xi - u_{\frac{1+q}{2}}\sqrt{\mathbf{D}\xi}$, то его считают аномальным, так как вероятность такого события для ξ мала и равна $1 - q$.

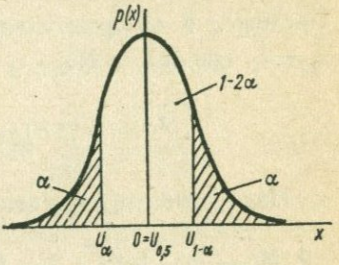


Рис. 18.

Это можно использовать, например, при контроле метода измерений на отсутствие систематического смещения. Обозначим ξ_1 и ξ_2 независимые результаты измерения одной и той же величины с интервалом времени между ними, на протяжении которого возможно появление смещения, и предположим, что ошибки распределены нормально с дисперсией D_0. Как было показано ранее, при отсутствии систематического смещения разность $\xi_1 - \xi_2$ имеет нулевое математическое ожидание и дисперсию $\mathbf{D}(\xi_1 - \xi_2) = \mathbf{D}\xi_1 + \mathbf{D}\xi_2 = 2D_0$. Величина $\eta = \xi_1 - \xi_2$ как линейная комбинация независимых гауссовских случайных величин, распределена нормально. Пределы, в которых она должна находиться с вероятностью q при $\mathbf{M}\xi_1 = \mathbf{M}\xi_2$, можно определить по (2.32′):

$$x_{\min} = -u_{\frac{1+q}{2}}\sqrt{2D_0},\ x_{\max} = u_{\frac{1+q}{2}}\sqrt{2D_0}. \tag{2.35}$$

Если абсолютная величина разности $|\xi_1 - \xi_2|$ результатов измерений превышает $u_{\frac{1+q}{2}}\sqrt{2D_0}$, следует определить наличие систематического смещения или грубую ошибку хотя бы одного из них. Вероятность того, что такой вывод в этой ситуации ошибочен, равна $1 - q$.

Для повышения достоверности делают две серии по n независимых измерений, отделенные интервалом времени, и вычисляют средние

арифметические по каждой серии $\bar{\xi}_1 = \frac{1}{n}\sum_{i=1}^{n} \xi_i^{(1)}$, $\bar{\xi}_2 = \frac{1}{n}\sum_{i=1}^{n} \xi_i^{(2)}$. Величины $\bar{\xi}_1$, $\bar{\xi}_2$, $\bar{\xi}_1 - \bar{\xi}_2$ нормально распределены, причем, если между $\bar{\xi}_1$ и $\bar{\xi}_2$ нет систематического смещения, то

$$\mathbf{M}(\bar{\xi}_1 - \bar{\xi}_2) = \mathbf{M}\bar{\xi}_1 - \mathbf{M}\bar{\xi}_2 = \frac{1}{n}\mathbf{M}\sum_{i=1}^{n} \xi_i^{(1)} - \frac{1}{n}\mathbf{M}\sum_{i=1}^{n} \xi_i^{(2)} = 0.$$

$$\mathbf{D}(\bar{\xi}_1 - \bar{\xi}_2) = \mathbf{D}\bar{\xi}_1 + \mathbf{D}\bar{\xi}_2 = \frac{1}{n^2}\mathbf{D}\left(\sum_{i=1}^{n} \xi_i^{(1)}\right) + \frac{1}{n^2}\mathbf{D}\left(\sum_{i=1}^{n} \xi_i^{(2)}\right) =$$

$$= nD_0 n^{-2} + nD_0 n^{-2} = \frac{2D_0}{n}.$$

Пределы, в которых с вероятностью q должна находиться разность $\bar{\xi}_1 - \bar{\xi}_2$ при $\mathbf{M}\bar{\xi}_1 = \mathbf{M}\bar{\xi}_2$:

$$\bar{x}_{\min} = -u_{\frac{1+q}{2}}\sqrt{\frac{2D_0}{n}}, \quad \bar{x}_{\max} = u_{\frac{1+q}{2}}\sqrt{\frac{2D_0}{n}}.$$

Надо отметить еще такое обстоятельство. Так как

$$\mathbf{P}\{\mathbf{M}\xi - u_{\frac{1+q}{2}}\sqrt{\mathbf{D}\xi} \leqslant \xi < \mathbf{M}\xi + u_{\frac{1+q}{2}}\sqrt{\mathbf{D}\xi}\} = q = \mathbf{P}\{\xi + u_{\frac{1+q}{2}}\sqrt{\mathbf{D}\xi} \geqslant$$

$$\geqslant \mathbf{M}\xi > \xi - u_{\frac{1+q}{2}}\sqrt{\mathbf{D}\xi}\},$$

в пределах

$$\xi - u_{\frac{1+q}{2}}\sqrt{\mathbf{D}\xi}, \quad \xi + u_{\frac{1+q}{2}}\sqrt{\mathbf{D}\xi} \tag{2.32″}$$

находится математическое ожидание с вероятностью q. Это дает возможность по наблюдению ξ указать пределы, которые ограничивают с заданной вероятностью q неизвестное математическое ожидание.

Нормальный закон распределения часто используется для аппроксимации распределений с целью упрощения производимых при обработке данных вычислений. При изучении биномиального распределения упоминалось о возможности приближенного его описания для вычисления вероятностей (2.1) и функции распределения (2.2). Это приближение как раз и осуществляется с помощью нормального распределения.

Пусть ξ_n — число появлений события A в схеме Бернулли из n испытаний, p — вероятность события A. Представим ξ_n, как и раньше, в виде $\xi_n = \sum_{i=1}^{n} \eta_i$; η_i — случайная величина, принимающая значение, равное единице, если в i-м испытании произошло событие A, и нулю, если не произошло. Величины η_i одинаково распределены и имеют дисперсии $\mathbf{D}\eta_i = p(1-p)$. По первому следствию центральной предель-

ной теоремы функция распределения величины $\frac{\xi_n - \mathbf{M}\xi_n}{\sqrt{\mathbf{D}\xi_n}}$ с увеличением n становится все более близкой к функции нормального распределения. Так как по (2.3) математическое ожидание $\mathbf{M}\xi_n = np$ и дисперсия $\mathbf{D}\xi_n = np(1-p)$, согласно (2.19)

$$\lim_{n\to\infty} \mathbf{P}\left\{\frac{\xi_n - np}{\sqrt{np(1-p)}} < z\right\} = \Phi(z). \tag{2.36}$$

Вероятность того, что число ξ_n появлений события A в схеме Бернулли из n испытаний будет заключено между заданными пределами x_0, x_1, можно приближенно вычислить по *формуле Муавра — Лапласа*

$$\mathbf{P}\{x_0 \leqslant \xi_n < x_1\} \approx \Phi\left(\frac{x_1 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{x_0 - np}{\sqrt{np(1-p)}}\right). \tag{2.36'}$$

Хорошее приближение в этой формуле обеспечивается при $n > \frac{9}{p(1-p)}$. Используется также несколько измененная формула, дающая уточнение по сравнению с (2.36')

$$\mathbf{P}\{x_0 \leqslant \xi_n \leqslant x_1\} \approx \Phi\left(\frac{x_1 - np + 0{,}5}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{x_0 - np - 0{,}5}{\sqrt{np(1-p)}}\right). \tag{2.36''}$$

Описанная аппроксимация позволяет вычислить и вероятность $\mathbf{P}\{\xi_n = m\}$ того, что ξ_n примет заданное значение m. Используя формулу конечных приращений*,

$$\mathbf{P}\{\xi_n = m\} = \mathbf{P}\{m \leqslant \xi_n < m+1\} \approx \Phi\left(\frac{m+1-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{m-np}{\sqrt{np(1-p)}}\right) =$$

$$= p\left(\frac{m-np+\theta}{\sqrt{np(1-p)}}\right)\frac{1}{\sqrt{np(1-p)}} \approx \frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{m-np}{\sqrt{np(1-p)}}\right)^2\right]\frac{1}{\sqrt{np(1-p)}}, \tag{2.37}$$

где $p(x)$ — плотность $(0; 1)$-нормального распределения (2.24), $0 \leqslant \theta \leqslant 1$.

При больших n формулы (2.36'), (2.37) значительно упрощают расчеты по сравнению с (2.2), (2.1).

В практике геологических исследований иногда используется *круговое нормальное распределение* — для таких показателей, как ориентировка частиц в породах, направление трещиноватости, простирание тектонических нарушений и т. п. Плотность этого распределения имеет вид

$$p(x) = K(\theta)\, e^{\theta\cos(x-\gamma)},\ \gamma - \pi \leqslant x \leqslant \gamma + \pi,\ K(\theta) = 1\Big/\int_{-\pi}^{\pi} e^{\theta\cos x}\, dx.$$

Как указано в [15], различия между этой моделью и обычным нормальным распределением несущественны, если среднее квадратическое отклонение $\sigma \leqslant \frac{\pi}{6}$. Более подробное описание этого распределения можно найти в [15].

---

* Формула конечных приращений для функции $F(x)$ имеет вид: $F(x+\Delta x) - F(x) = F'(x+\theta\Delta x)\,\Delta x$, $0 \leqslant \theta \leqslant 1$.

Усеченное нормальное распределение. Пусть ξ — нормально распределенная случайная величина с математическим ожиданием m и дисперсией σ^2. Введем случайную величину η, равную ξ, если ξ принимает значение, большее или равное a и будем считать ее ненаблюдаемой, если ξ получает значение, меньшее a. Такая величина имеет *односторонне усеченное* нормальное распределение с *точкой усечения* a. Аналогично вводится распределение, усеченное справа (наблюдаемое при $\xi \leqslant a$). При усечении слева *степень усечения* составит $\alpha_1 = \Phi\left(\frac{a-m}{\sigma}\right)$, функция и плотность распределения имеют вид

$$F(x) = \frac{1}{1-\alpha_1}\left[\Phi\left(\frac{x-m}{\sigma}\right) - \alpha_1\right], \quad f(x) = F'(x) =$$

$$= [\sqrt{2\pi}\sigma(1-\alpha_1)]^{-1} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right), \quad x \geqslant a;$$

$$F(x) = f(x) = 0, \quad x < a.$$

При усечении справа степень усечения $\alpha_2 = 1 - \Phi\left(\frac{a-m}{\sigma}\right)$; функция и плотность распределения —

$$F(x) = \frac{1}{1-\alpha_2}\Phi\left(\frac{x-m}{\sigma}\right), \quad f(x) = [\sqrt{2\pi}\sigma(1-\alpha_2)]^{-1} \exp\left(-\frac{1}{2}\frac{(x-m)^2}{\sigma^2}\right)$$

$$x \leqslant a;$$

$$F(x) = 1, \quad f(x) = 0, \quad x > a.$$

Оба распределения в практике статистического анализа применяются в тех случаях, когда значения, меньшие (или большие) определенного уровня не используются или не могут быть использованы (если метод измерений обладает порогом чувствительности, сопоставимым с измеряемыми значениями; если точкой усечения отделяется область значений, заведомо не содержащая аномальных, в которой исследуется фоновое распределение).

Пример 2.7. $\xi_1, \xi_2, \ldots, \xi_n$ — независимые наблюдения случайной величины ξ, имеющей конечную дисперсию D и математическое ожидание $\mathbf{M}\xi = M$. Используя центральную предельную теорему, выяснить приближенный вид: 1) функции распределения среднего арифметического, вычисленного по этим наблюдениям; 2) квантиля порядка q распределения среднего арифметического; 3) симметричных относительно M пределов, заключающих среднее арифметическое с вероятностью q.

Решение. Обозначим $\bar{\xi}_n$ среднее арифметическое, составленное из ξ_i: $\bar{\xi}_n = \frac{1}{n}\sum_{i=1}^{n}\xi_i$. Его математическое ожидание $\mathbf{M}\bar{\xi}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{M}\xi_i = \frac{nM}{n} = M$, так как математическое ожидание каждого наблюдения совпадает с математическим ожиданием M наблюдаемой величины. Дисперсия $\mathbf{D}\xi_i = \mathbf{D}\xi = D$ и $\mathbf{D}\bar{\xi}_n = \frac{1}{n^2}\sum_{i=1}^{n}\mathbf{D}\xi_i = \frac{D}{n}$.

1) По первому следствию центральной предельной теоремы функция распределения $\bar{\xi}_n$ приближенно представляется в виде

$$F_n(x) = \Phi\left(\frac{x - \mathbf{M}\bar{\xi}_n}{\sqrt{\mathbf{D}\bar{\xi}_n}}\right) = \Phi\left(\frac{x - M}{\sqrt{Dn^{-1}}}\right),$$

где $\Phi(z)$ — функция (0; 1)-нормального распределения (2.25).

2) Квантиль порядка q величины $\bar{\xi}_n$, по (2.27),

$$z_q = \mathbf{M}\bar{\xi}_n + u_q\sqrt{\mathbf{D}\bar{\xi}_n} = M + u_q\sqrt{\frac{D}{n}},$$

где u_q — q-квантиль (0, 1)-нормального распределения, определяемый по табл. 2 (Приложение).

3) Симметричные пределы для $\bar{\xi}_n$, согласно (2.32′),

$$x_0 = \mathbf{M}\bar{\xi}_n - u_{\frac{1+q}{2}}\sqrt{\mathbf{D}\bar{\xi}_n} = M - u_{\frac{1+q}{2}}\sqrt{\frac{D}{n}},\quad x_1 = M + u_{\frac{1+q}{2}}\sqrt{\frac{D}{n}}.$$

Пример 2.8. Распределение концентраций SiO_2 на некоторой территории описывается нормальным законом с математическим ожиданием $M = 65\%$ и дисперсией $D = 3{,}24\ (\%)^2$. 1) Построить такие симметричные доверительные пределы для концентраций SiO_2, чтобы часть территории, на которой содержание SiO_2 находится в этих пределах, составила 95% от общей. 2) На какой части территории содержание SiO_2 превышает 70%?

Решение. 1) Очевидно, для построения нужных пределов следует задаться вероятностью $q = 0{,}95$. Согласно (2.33) имеем:

$$x_0 = M - 1{,}96\sqrt{D} \approx 61{,}47\%,\quad x_1 = M + 1{,}96\sqrt{D} \approx 68{,}53\%.$$

2) По формуле (2.26) $\mathbf{P}\{SiO_2 < 70\%\} = \Phi\left(\frac{70\% - M}{\sqrt{D}}\right) \approx \Phi(2{,}78)$. По табл. 2 (Приложение) $\Phi(2{,}78) \approx 0{,}997$. Искомая часть составляет $\mathbf{P}\{SiO_2 \geqslant 70\%\} = 1 - \mathbf{P}\{SiO_2 < 70\%\} \approx 0{,}003 = 0{,}3\%$ всей территории.

Пример 2.9. Среднее квадратическое отклонение нормально распределенной ошибки измерения плотности горных пород составляет $\sigma = 0{,}01$ *г/см*^{3}. 1) Построить симметричные пределы вокруг результата измерения ξ, которые с вероятностью 0,95 содержат истинное значение m плотности. 2) С какой вероятностью m заключено в интервале $[(\xi - 0{,}015),\ (\xi + 0{,}015)]$ *г/см*^{3}? 3) Чему равно математическое ожидание n_a количества измерений, отклоняющихся от истинного значения не менее, чем на 0,02 *г/см*^{3}, если планируется сделать 60 измерений?

Решение. 1) Согласно (1.15) результат измерения представляется в виде $\xi = m + \Delta$ (m — истинное значение, Δ — ошибка измерения, имеющая нулевое математическое ожидание и среднее квадратическое отклонение $\sigma = 0{,}01$ *г/см*^{3}). Для одного образца $\mathbf{M}\xi = \mathbf{M}(m + \Delta) = m$, $\mathbf{D}\xi = \mathbf{D}m + \mathbf{D}\Delta = \mathbf{D}\Delta = \sigma^2$. По формуле (2.32″) пределы, заключающие $m = \mathbf{M}\xi$ с вероятностью $q = 0{,}95$, составляют $\xi \mp 1{,}96 \times 0{,}01\ \frac{г}{см^3}\Big) \approx (\xi \mp 0{,}02)$ *г/см*^{3}.

2) Пределам для m $[(\xi - 0{,}015),\ (\xi + 0{,}015)]$ *г/см*^{3} соответствует та же вероятность q, что и пределам для ξ вида $[(\mathbf{M}\xi - 0{,}015),\ (\mathbf{M}\xi + 0{,}015)]$ *г/см*^{3}. Согласно (2.34),

$$q = \Phi\left(\frac{\mathbf{M}\xi + 0{,}015 - \mathbf{M}\xi}{\sigma}\right) - \Phi\left(\frac{\mathbf{M}\xi - 0{,}015 - \mathbf{M}\xi}{\sigma}\right) = \Phi(1{,}5) - \Phi(-1{,}5) \approx$$
$$\approx 0{,}93 - 0{,}07 = 0{,}86.$$

3) Вероятность того, что истинное значение m окажется в пределах $(\xi \pm 0{,}02)$ *г/см*^{3}, составляет 0,95 (по п. 1 этого примера). Вероятность того, что истинное значение окажется вне этих пределов, $1 - 0{,}95 = 0{,}05$. Для описания

распределения количества наблюдений, содержащих такие отклонения, используем схему Бернулли с $p = 0{,}05$ и $n = 60$. Из (2.3) искомое математическое ожидание $n_a = np = 3$.

Пример 2.10. Вероятность того, что отобранная на месторождении проба будет иметь кондиционное содержание олова, составляет $p = 0{,}4$. 1) Вычислить симметричные пределы, в которых с вероятностью $q = 0{,}9$ будет находиться количество ξ_n проб с кондиционным содержанием олова, если всего отобрано $n = 250$ независимых проб. 2) Каким должно быть общее количество проб, чтобы обеспечить отклонение $\frac{\xi_n}{n}$ от p не большее, чем на 10% с вероятностью 0,95? 3) Считая p неизвестным, оценить пределы, заключающие p с вероятностью 0,95, если из 250 независимых проб 125 имеют кондиционное содержание.

Решение. 1) Обозначим искомые границы x_0 и x_1. Ввиду того, что распределение ξ_n приближенно нормально с математическим ожиданием $\mathbf{M}\xi_n = np$ и дисперсией $\mathbf{D}\xi_n = np(1-p)$, согласно (2.32′), имеем: $x_0 = np - u_{0{,}95}\sqrt{np(1-p)} = 250 \times \times 0{,}4 - 1{,}645\sqrt{250 \cdot 0{,}4 \cdot 0{,}6} \approx 82{,}3$; $x_1 = np + u_{0{,}95}\sqrt{np(1-p)} \approx 112{,}7$.

2) В соответствии с поставленным условием $\mathbf{P}\left\{-0{,}1p \leqslant \frac{\xi_n}{n} - p < 0{,}1p\right\} = = q = 0{,}95$. Так как с вероятностью $q = 0{,}95$ значение ξ_n попадает в пределы $x_0 = = np - u_{0{,}975}\sqrt{np(1-p)}$, $x_1 = np + u_{0{,}975}\sqrt{np(1-p)}$, с этой же вероятностью $\frac{\xi_n}{n}$ находится в пределах $\frac{x_0}{n}$, $\frac{x_1}{n}$, а $\left(\frac{\xi_n}{n} - p\right)$ — в пределах $\left(\frac{x_0}{n} - p\right)$, $\left(\frac{x_1}{n} - p\right)$. Приравнивая $\frac{x_0}{n} - p = -0{,}1p$, $\frac{x_1}{n} - p = 0{,}1p$, получим $u_{0{,}975}\sqrt{\frac{p(1-p)}{n}} = = 0{,}1p$, откуда $n = \frac{u^2_{0{,}975}\, p(1-p)}{0{,}1^2 p^2} \approx 576$.

3) По (2.32″) с учетом того, что $\mathbf{M}\frac{\xi_n}{n} = p$ и $\frac{\xi_n}{n} \to p$, искомые пределы имеют приближенный вид: $\frac{\xi_n}{n} \pm u_{0{,}975}\sqrt{\frac{p(1-p)}{n}} \approx \frac{\xi_n}{n} \pm u_{0{,}975}\sqrt{\frac{\xi_n(1-\xi_n/n)}{n^2}} = = \frac{125}{250} \pm 1{,}96\sqrt{\frac{0{,}5 \cdot 0{,}5}{250}} \approx 0{,}5 \pm 0{,}062$.

Логарифмически нормальный и обобщенно-логнормальный законы распределения. *Логарифмически нормальный* или, как его еще называют, *логнормальный закон распределения* является одной из распространенных вероятностных схем, используемых при статистическом анализе геолого-геофизических показателей горных пород. Его часто используют для аппроксимации распределений концентраций химических элементов и минералов в породах, а также характеристик физических свойств горных пород. Логнормальный закон часто оказывается более приемлемым для описания распределений содержаний элементов-примесей в изверженных горных породах по сравнению с нормальным законом. Это в свое время дало повод некоторым исследователям назвать его основным законом геохимии. Однако в настоящее время такая точка зрения совершенно обоснованно отвергнута.

Логнормальному закону следуют случайные величины, логарифмы которых распределены нормально. Условия, при которых он действует, можно определить теоремой, подобной центральной предельной теореме. Если $\xi_1, \xi_2, \ldots, \xi_n, \ldots$ — последовательность независимых случайных величин, логарифмы которых удовлетворяют условиям

центральной предельной теоремы (Линдеберга), то функция распределения $\left(\prod_{i=1}^{\xi_i}\frac{\xi_i}{c_i}\right)^{\alpha_n}\left(\theta = \exp \mathbf{M} \ln \xi_i,\ \alpha_n = \left(\sum_{i=1}^{n} \mathbf{D} \ln \xi_i\right)^{-\frac{1}{2}}\right)$ при $n \to \infty$ стремится к логарифмически нормальной. Иными словами, случайная величина, являющаяся результатом действия большого числа факторов малой интенсивности с пропорциональным эффектом влияния распределена по закону, близкому к логнормальному.

Логарифмически нормальный закон можно рассматривать как частный случай *обобщенно-логнормального* или, как его иногда называют, *трехпараметрического распределения*. Случайная вгличина ξ имеет обобщенно-логнормальное распределение, если $\ln(a + \lambda\xi)$ распределяется нормально (a и λ параметры, причем λ может равняться одному из двух значений: 1 или -1).

Плотность обобщенно-логнормального распределения, в соответствии с формулой (1.25), имеет вид

$$p(x) = \frac{1}{\sqrt{2\pi}\delta(a + \lambda x)} \exp\left\{-\frac{1}{2\delta^2}[\ln(a + \lambda x) - \mu]^2\right\}, \text{ если } a + \lambda x > 0; \tag{2.38}$$

$$p(x) = 0, \text{ если } a + \lambda x \leqslant 0.$$

Она отлична от нуля при $x > -a$, если $\lambda = 1$ и при $x < a$, если $\lambda = -1$. Это значит, что величина $-a$ ограничивает снизу все возможные значения случайной величины ξ при $\lambda = 1$, а величина a — сверху при $\lambda = -1$. Положив в (2.38) $a = 0$, $\lambda = 1$, получим плотность логнормального распределения. Логнормально распределенные случайные величины всегда положительны.

Так как логнормальный закон входит в семейство обобщенно-логнормальных распределений с параметрами $a = 0$ и $\lambda = 1$, приведем вид основных числовых характеристик лишь для последнего. Числовые характеристики логарифмически нормального распределения вычисляются точно так же, только в приводимых ниже формулах надо принять $a = 0$ и $\lambda = 1$.

Плотность распределения (2.38) определяется параметрами: a, λ, $\mu = \mathbf{M} \ln(a + \lambda\xi)$ —математическим ожиданием $\ln(a + \lambda\xi)$ и $\delta^2 = \mathbf{D}\ln(a + \lambda\xi)$ — дисперсией $\ln(a + \lambda\xi)$.

Математическое ожидание ξ имеет вид

$$\mathbf{M}\xi = \left[\exp\left(\mu + \frac{1}{2}\delta^2\right) - a\right]\lambda; \tag{2.39}$$

дисперсия —

$$\mathbf{D}\xi = \exp(2\mu + 2\delta^2) - \exp(2\mu + \delta^2); \tag{2.40}$$

мода —

$$M_0 = [\exp(\mu - \delta^2) - a]\lambda; \tag{2.41}$$

квантиль порядка q —

$$z_q = [\exp(\mu + \lambda u_q \delta) - a]\lambda, \tag{2.42}$$

где u_q — квантиль (0; 1)-нормального распределения;

медиана —

$$M_e = (\exp\mu - a)\,\lambda. \tag{2.43}$$

Формулы (2.39), (2.40) доказываются непосредственным вычислением по общей формуле (1.42). Например, математическое ожидание

$$\mathbf{M}\xi = \int\limits_{a+\lambda x>0} \frac{x}{\sqrt{2\pi}\,\delta\,(a+\lambda x)} \exp\left\{-\frac{1}{2\delta^2}[\ln(a+\lambda x)-\mu]^2\right\} dx =$$

$$= \lambda \int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} (e^{\mu+u\delta} - a)\, e^{-\frac{u^2}{2}}\, du = -\int\limits_{-\infty}^{\infty} \frac{a\lambda}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}\, du +$$

$$+ \lambda e^{\mu} \int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}+u\delta}\, du = -\lambda a + \lambda e^{\mu} \int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} e^{\frac{\delta^2}{2}}\, dz =$$

$$= -a\lambda + \lambda e^{\mu+\frac{\delta^2}{2}}$$

(сделана замена $\frac{1}{\delta}[\ln(a+\lambda x)-\mu] = u$; $\lambda^2 = 1$; $\int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du = 1$ по свойству (1.22) для плотности (0; 1)-нормального распределения).

Выражение (2.41) для моды получим, приравняв нулю производную от плотности (2.38). Квантиль z_q порядка q определяется из условия

$$\mathbf{P}\{\xi < z_q\} = \begin{cases} \mathbf{P}\{\ln(a+\lambda\xi) < \ln(a+\lambda x)\} = \Phi\left(\dfrac{\ln(a+\lambda z_q)-\mu}{\delta}\right) = q \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{при } \lambda = 1, \\ \mathbf{P}\{\ln(a+\lambda\xi) > \ln(a+\lambda x)\} = 1 - \Phi\left(\dfrac{\ln(a+\lambda z_q)-\mu}{\delta}\right) = q \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{при } \lambda = -1. \end{cases}$$

Приравнивая $\frac{1}{\delta}[\ln(a+\lambda z_q)-\mu] = u_q$, если $\lambda = 1$ и $\frac{1}{\delta}[\ln[a+\lambda z_q)-\mu] = u_{1-q} = -u_q$, если $\lambda = -1$, получим: $z_q = \exp(\mu + u_q\delta) - a$, если $\lambda = 1$ и $z_q = -\exp(\mu - u_q\delta) + a$, если $\lambda = -1$. Это и записано формулой (2.42).

Медиана, как квантиль порядка 0,5, определяется по формуле (2.42) при $u_q = u_{0,5} = 0$, что и дает (2.43).

Формулы (2.39) — (2.43) будут использоваться в дальнейшем при изучении методов оценки числовых характеристик распределений.

Можно показать, что асимметрия и эксцесс имеют вид:

$$A = \lambda(\gamma^3 + 3\gamma),\ E = \gamma^8 + 6\gamma^6 + 15\gamma^4 + 16\gamma^2, \tag{2.44}$$

где $\gamma = \sqrt{\exp\delta^2 - 1}$. Формулы (2.44) показывают, что независимо от λ эксцесс E положителен. При $\lambda = -1$ асимметрия A отрицательна, т. е. плотность распределения левоасимметрична (рис. 19). При этом, согласно (2.39), (2.41) и (2.43), $M_0 > M_e > \mathbf{M}\xi$. Если $\lambda = 1$, то $A > 0$— плотность правоасимметрична, причем $M_0 < M_e < \mathbf{M}\xi$.

Пределы, в которые с вероятностью q попадает значение случайной величины, можно определить по условию (2.31) с использованием выражения для квантиля (2.42):

$$x_0 = \exp(\mu - u_{\frac{1+q}{2}}\delta) - a,\ x_1 = \exp(\mu + u_{\frac{1+q}{2}}\delta) - a \text{ при } \lambda = 1; \quad (2.45)$$

$$x_0 = -\exp(\mu + u_{\frac{1+q}{2}}\delta) + a,\ x_1 = -\exp(\mu - u_{\frac{1+q}{2}}\delta) + a \text{ при } \lambda = -1.$$

Эти пределы удовлетворяют условию $\mathbf{P}\{x_0 \leqslant \xi < x_1\} = q$, $\mathbf{P}\{\xi < x_0\} = \mathbf{P}\{\xi \geqslant x_1\} = \frac{1-q}{2}$, но в отличие от пределов (2.32′) для нормального распределения они не будут располагаться симметрично относительно математического ожидания.

Как упоминалось выше, случайные величины, распределенные по обобщенно-логнормальному закону, ограничены с одной стороны — сверху или снизу — в зависимости от λ. Существуют модели распределений, ограниченных одновременно сверху и снизу. Класс подобных распределений образуют, например, случайные величины ξ, для которых $\ln\left(\frac{a-\xi}{b-\xi}\right)$ распределен нормально $(a < \xi < b)$. Такая схема, впрочем, почти не применяется, поэтому мы не будем на ней останавливаться.

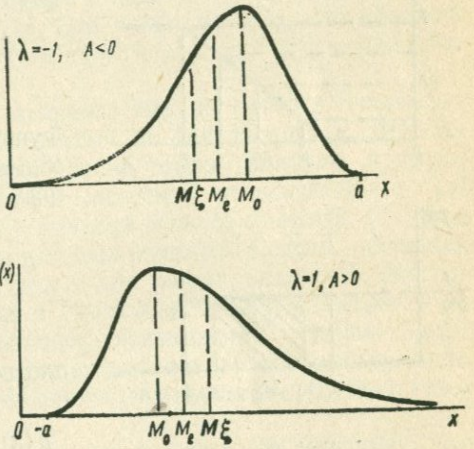


Рис. 19.

Пример 2.11. По данным спектрального анализа проб, отобранных на массиве гранитов, установлено, что концентрация титана распределяется по логнормальному закону с математическим ожиданием логарифма концентрации $\mu = \mathbf{M}\ln\xi = -1{,}944$ и его средним квадратическим отклонением $\delta = 0{,}60$. Вычислить: 1) наиболее вероятную концентрацию титана (моду); 2) среднее значение содержания титана (математическое ожидание); 3) пределы, в которых находится концентрация Ti на 90% территории массива.

Решение. 1) При логнормальном распределении ξ в формуле (2.41) следует положить $a = 0$, $\lambda = 1$. По (2.41)

$$M_0 = \exp(-1{,}944 - 0{,}60^2) = e^{-2{,}304} = 10^{-0{,}434 \cdot 2{,}304} \approx 0{,}100\%.$$

2) Математическое ожидание, по (2.39) при $a = 0$, $\lambda = 1$

$$\mathbf{M}\xi = \exp\left(-1{,}944 + \frac{1}{2} \cdot 0{,}60^2\right) = 10^{-0{,}434 \cdot 1{,}764} \approx 0{,}171\%.$$

3) Для вычисления искомых пределов воспользуемся их выражением (2.45), приняв $q = 0{,}9$, $a = 0$ $\lambda = 1$. С учетом того, что $u_{\frac{1+q}{2}} = u_{0,95} = 1{,}645$.

$$x_0 = \exp(-1{,}944 - 1{,}645 \cdot 0{,}60) \approx 0{,}053\%, \quad x_1 = \exp(-1{,}944 + 1{,}645 \cdot 0{,}60) \approx \approx 0{,}382\%.$$

Равномерное распределение. Простейшим распределением случайной величины, ограниченной сверху и снизу, является *равномерное*. Равномерно распределенная случайная величина принимает значения на отрезке $[a, b]$ числовой оси, причем вероятность попадания наблюдения случайной величины в некоторый интервал $[x_0, x_1]$, находящийся внутри $[a, b]$, пропорциональна длине этого интервала:

$$\mathbf{P}\{x_0 \leqslant \xi \leqslant x_1\} = \frac{x_1 - x_0}{b - a}, \quad a \leqslant x_0 < x_1 \leqslant b. \tag{2.46}$$

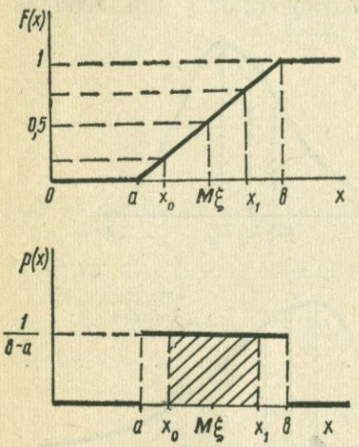


Рис. 20.

Функция распределения, согласно ее общему определению (1.16), имеет вид (рис. 20):

$$F(x) = \begin{cases} \dfrac{x-a}{b-a}, & \text{если } a \leqslant x \leqslant b; \\ 0, & \text{если } x < a; \\ 1, & \text{если } x > b; \end{cases} \tag{2.47}$$

плотность (рис. 20) —

$$p(x) = \begin{cases} \dfrac{1}{b-a}, & \text{если } a \leqslant x \leqslant b; \\ 0, & \text{если } x < a \text{ или } x > b. \end{cases} \tag{2.48}$$

С равномерным распределением мы уже встречались ранее, когда знакомились с геометрической интерпретацией вероятности. В одномерном случае геометрическая вероятность полностью соответствует формуле (2.46).

Квантиль z_q порядка q, согласно (2.47), определяется по формуле

$$z_q = a + q(b - a); \tag{2.49}$$

пределы x_0, x_1, в которые с вероятностью q попадает значение ξ, с соблюдением условия $\mathbf{P}\{\xi < x_0\} = \mathbf{P}\{\xi \geqslant x_1\}$,

$$x_0 = a + \frac{1-q}{2}(b - a), \quad x_1 = a + \frac{1+q}{2}(b - a) = b - \frac{1-q}{2}(b - a). \tag{2.49'}$$

Математическое ожидание —

$$\mathbf{M}\xi = \int_a^b x \frac{1}{b-a}\, dx = \frac{a+b}{2}. \tag{2.50}$$

Оно является серединой отрезка $[a, b]$ и совпадает с медианой. Дисперсия —

$$\mathbf{D}\xi = \mathbf{M}\xi^2 - (\mathbf{M}\xi)^2 = \int_a^b \frac{x^2}{b-a}\,dx - \left(\frac{b+a}{2}\right)^2 = \frac{(b-a)^2}{12}. \qquad (2.51)$$

Моды у этого распределения нет — все значения из $[a, b]$ равновероятны. Асимметрия, ввиду симметричности плотности (2.48), равна нулю, а эксцесс, как в этом нетрудно убедиться, равен $-1{,}2$.

Преобразование произвольного распределения в равномерное и нормальное. Если ξ — случайная величина с функцией распределения $F(x)$, то результат преобразования ξ вида $\eta = F(\xi)$ будет распределен равномерно на отрезке $[0, 1]$. Это видно из того, что положив $x = F(z)$, получим: $x = F(z) = \mathbf{P}\{\xi < z\} = \mathbf{P}\{F(\xi) < F(z)\} = \mathbf{P}\{F(\xi) < x\}$ для $0 \leqslant x \leqslant 1$.

Построив график функции распределения в системе координат с линейным относительно $F(x)$ масштабом по оси абсцисс и обычным по оси ординат, получим прямую. Точно также получим прямую, если по оси ординат выбран масштаб, линейный относительно $F^{-1}(y)$ ($F^{-1}(y)$ — функция, обратная $F(x)$), а по оси абсцисс обычный. На этом свойстве основано использование *вероятностной бумаги*, имеющей соответствующую координатную сетку для нормального или логнормального распределения (рисунки в Приложении). Если построить на такой бумаге функцию распределения, искривление графика будет свидетельствовать об отклонении распределения от соответствующего закона — нормального или логнормального, в зависимости от типа вероятностной бумаги.

Нормализующее преобразование, т. е. функцию $f(x)$ такую, что $f(\xi)$ распределена нормально, очевидно, можно представить в виде $f(x) = \Phi^{-1}(F(x))$ ($\Phi^{-1}(u)$ — функция, обратная (2.25)). Так как линейное преобразование сохраняет нормальность распределения, нормализующим преобразованием будет и $\alpha f(x) + \beta$ ($\alpha \neq 0$). Для обобщенно-логнормального закона нормализующее преобразование $\ln(a + \lambda\xi)$, для логнормального $\ln x$.

Эффективность использования произвольной функции $\varphi(x)$ в качестве нормализующего преобразования можно определить, например, построив на вероятностной бумаге функцию распределения величины $f(\xi)$. Часто для оценки его эффективности используются коэффициенты асимметрии и эксцесса преобразованной величины.

Распределения χ^2, Стьюдента и Фишера. Эти распределения не используются для аппроксимации распределений геолого-геофизических показателей в геологических объектах, однако играют важную роль в математической статистике, так как на них основан ряд ее широко употребляемых методов.

Распределению χ^2, или *распределению Пирсона*, подчиняются случайные величины, которые можно представлять в виде суммы некоторого числа k квадратов независимых нормально распределенных случайных величин, каждая из которых имеет нулевое математическое

ожидание и единичную дисперсию. Такие суммы обычно обозначают χ^2_k:

$$\chi^2_k = \sum_{i=1}^{k} \xi_i^2, \quad \mathbf{M}\xi_i = 0, \quad \mathbf{D}\xi_i = 1, \tag{2.52}$$

причем ξ_i нормально распределены и независимы. Количество k их квадратов в сумме (2.52) называют *числом степеней свободы*.

Как следует из приведенного определения, распределению χ^2 подчиняются случайные величины вида

$$\chi^2_k = \sum_{i=1}^{k} \left(\frac{\eta_i - \mathbf{M}\eta_i}{\sqrt{\mathbf{D}\eta_i}} \right)^2, \tag{2.52'}$$

если η_i произвольные независимые нормально распределенные случайные величины.

Математическое ожидание и дисперсия χ^2_k равны, соответственно,

$$\mathbf{M}\chi^2_k = \mathbf{M}\sum_{i=1}^{k} \xi_i^2 = \sum_{i=1}^{k} \mathbf{D}\xi_i = k, \quad \mathbf{D}\chi^2_k = \mathbf{M}\chi^4_k - (\mathbf{M}\chi^2_k)^2 = \mathbf{M}\left(\sum_{i=1}^{k} \xi_i^2\right)^2 -$$

$$- k^2 = \mathbf{M}\left(\sum_{i=1}^{k} \xi_i^4 + \sum_{i \neq j} \xi_i^2 \xi_j^2\right) - k^2 = 3k + k(k-1) - k^2 = 2k \tag{2.53}$$

($\mathbf{M}\xi_i^4 = 3(\mathbf{D}\xi_i)^2 = 3$ по (2.21)).

Мода χ^2_k

$$M_0 = k - 2 \quad (k > 2). \tag{2.54}$$

Коэффициенты асимметрии и эксцесса

$$A = 2\sqrt{\frac{2}{k}}, \quad E = \frac{12}{k}, \tag{2.55}$$

откуда видно, что кривая плотности одномодальна и правоасимметрична. При $k \to \infty$, $A \to 0$ и $E \to 0$, что объясняется сходимостью распределения χ^2_k к нормальному. В этом можно убедиться, применив центральную предельную теорему.

Распределению Стьюдента, или *t-распределению*, с k степенями свободы следуют случайные величины, представляемые в виде

$$t = \xi \Big/ \sqrt{\sum_{i=1}^{k} \frac{\xi_i^2}{k}} = \xi \Big/ \sqrt{\frac{\chi^2_k}{k}}, \tag{2.56}$$

где ξ, ξ_1, ξ_2, ..., ξ_k — независимые случайные величины, распределенные каждая по нормальному закону с нулевым математическим ожиданием и единичной дисперсией. Математическое ожидание и мода случайной величины, имеющей распределение Стьюдента, совпадают и равны нулю (при $k > 1$); дисперсия —

$$\mathbf{D}t = \frac{k}{k-2} \quad (k > 2); \tag{2.57}$$

асимметрия и эксцесс —

$$A = 0,\ E = \frac{6}{k-4}\ (k > 4). \tag{2.58}$$

Если ξ, ξ_1, ξ_2, ..., ξ_k — произвольные независимые случайные величины, распределенные по нормальному закону, то

$$t_1 = \frac{\xi - \mathbf{M}\xi}{\sqrt{\mathbf{D}\xi}} \Bigg/ \sqrt{\frac{1}{k}\sum_{i=1}^{k}\frac{(\xi_i - \mathbf{M}\xi_i)^2}{\mathbf{D}\xi_i}} \tag{2.56'}$$

также подчиняется распределению Стьюдента.

С увеличением количества степеней свободы распределение Стьюдента приближается к нормальному, причем намного быстрее, чем распределение χ^2. Это видно и по коэффициентам асимметрии и эксцесса (2.58) — первый равен нулю, а второй стремится к нулю, будучи к нему близким уже при небольших значениях k. Это свойство часто используется для аппроксимации распределения Стьюдента нормальным — уже при $k > 30$ приближение удовлетворительно (см. табл. 2, 9 Приложения).

Распределению Фишера, или *F-распределению*, подчиняются случайные величины, которые могут быть представлены в виде

$$F_{m,n} = \frac{1}{m}\sum_{i=1}^{m}\xi_i^2 \Bigg/ \left(\frac{1}{n}\sum_{i=1}^{n}\eta_i^2\right), \tag{2.59}$$

где ξ_i, η_i — независимые одинаково нормально распределенные случайные величины с нулевым математическим ожиданием. Величины m и n называются числами степеней свободы F-распределения.

Если ξ_i и η_j имеют математические ожидания, не равные нулю, но одинаковые дисперсии, то

$$F_{m,n} = \frac{1}{m}\sum_{i=1}^{m}(\xi_i - \mathbf{M}\xi_i)^2 \Bigg/ \left[\frac{1}{n}\sum_{j=1}^{n}(\eta_j - \mathbf{M}\eta_j)^2\right] \tag{2.60}$$

также подчиняется распределению Фишера с m и n степенями свободы.

Математическое ожидание

$$\mathbf{M}F_{m,n} = \frac{n}{n-2}\ (n > 2); \tag{2.61}$$

дисперсия

$$\mathbf{D}F_{m,n} = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}\ (n > 4); \tag{2.62}$$

мода

$$M_0 = \frac{n(m-2)}{m(n+2)}. \tag{2.63}$$

Использование трех упомянутых выше распределений осуществляется с помощью таблиц их квантилей. Квантили $\chi_q^2(k)$ распределения

χ^2 приведены в табл. 3; квантили $t_q(k)$ распределения Стьюдента — в табл. 9; квантили $F_q(m, n)$ распределения Фишера — в табл. 10, 11 (Приложение).

Системы кривых Пирсона и Грама-Шарлье. Семейство *кривых Пирсона* определяется плотностями распределений, которые являются решениями дифференциального уравнения:

$$\frac{1}{p(x)}\frac{dp(x)}{dx} = \frac{x - M}{B_0 + B_1 x + B_2 x^2}, \tag{2.64}$$

где началом отсчета для x служит математическое ожидание. В этой системе семь типов кривых, которые охватывают большое разнообразие плотностей распределений — асимметричные и симметричные распределения, ограниченные справа и слева, кривые плотностей J- и U-образной формы. Типы кривых определяются по специальному критерию, после чего рассчитываются числовые характеристики и плотности распределения. В литературе известны применения кривых Пирсона при описании распределений концентраций химических элементов в гранитоидах [24].

Система кривых Грама-Шарлье основывается на своеобразном согласовании нормального закона с центральными моментами порядков выше второго. Обычно плотность согласовывается по асимметрии и эксцессу:

$$p(z) = f(z) + \frac{A}{6} f'''(z) + \frac{E}{24} f^{IV}(z), \tag{2.65}$$

где $z = \frac{x - M\xi}{\sqrt{D\xi}}$, $f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z^2\right)$, $f^{(j)}(z)$ — производная j-го порядка; A и E — асимметрия и эксцесс.

Кривые Грама-Шарлье просты в использовании, однако имеют недостатки: они не всегда достаточно гладкие, могут пересекать ось абсцисс и иметь неоправданные провалы [14].

По ряду причин (трудность эффективной оценки параметров и интерпретации, возможность аппроксимации распределений более простыми схемами) обе системы не нашли широкого применения в практике анализа геолого-геофизических данных. Поэтому мы не будем их подробно рассматривать, рекомендуя заинтересованному читателю специальную литературу [2, 14, 24].

Глава 3

МНОГОМЕРНЫЕ СЛУЧАЙНЫЕ ВЕЛИЧИНЫ

До сих пор мы рассматривали свойства и характеристики одномерных случайных величин, т. е. величин, принимающих в результате одного испытания одно числовое значение. Однако на практике часто приходится изучать особенности совместного распределения

нескольких величин. Например, если содержание химического элемента A в горной породе оказывает влияние на содержание элемента B, то информацию о таком влиянии можно получить лишь по данным совместных измерений A и B. Результатами наблюдений в этом случае будут пары числовых значений — концентрации A и B. Вообще, породу при количественном анализе можно рассматривать как совокупность всех ее минералогических, химических, физических и других показателей, которые определенным образом связаны между собой и совокупность значений которых образует ее качественную определенность. Этот подход предполагает изучение как распределений отдельных показателей, так и совместного распределения их, взятых в совокупности. Такое исследование и проводится методами многомерного статистического анализа.

§ 1. Понятие многомерной случайной величины

Векторы и матрицы. Напомним сведения из теории матриц и векторного анализа, которые понадобятся в дальнейшем.

Вектор m-мерного векторного пространства, представляющий собой m упорядоченных чисел, обычно записывают в одной из двух форм: в виде строки его компонент

$$\mathbf{a} = \{a_1, a_2, \ldots, a_m\}, \tag{3.1}$$

или в виде столбца

$$\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}. \tag{3.1'}$$

Число m компонент вектора называют размерностью векторного пространства.

Операция *транспонирования* преобразует вектор-строку в вектор-столбец или наоборот и обозначается штрихом. Если $\mathbf{b}$ — вектор-столбец (3.1′), то

$$\mathbf{b}' = \{b_1, b_2, \ldots, b_m\}. \tag{3.2}$$

В дальнейшем будем под вектором подразумевать вектор-столбец, а для удобства его записи пользоваться формой (3.2), либо

$$\mathbf{b} = \{b_1, b_2, \ldots, b_m\}'. \tag{3.2'}$$

В геометрической интерпретации компоненты вектора определяют точку m-мерного эвклидового пространства, координаты которой равны этим компонентам.

Сумма двух m-мерных векторов $\mathbf{a}$ и $\mathbf{b}$ есть вектор, компоненты которого равны суммам соответствующих компонент a_i и b_i:

$$\mathbf{a} + \mathbf{b} = \{a_1 + b_1, a_2 + b_2, \ldots, a_m + b_m\}'.$$

Скалярное произведение двух векторов $\mathbf{a} = \{a_1, a_2, \ldots, a_m\}'$ и $\mathbf{b} = \{b_1, b_2, \ldots, b_m\}'$, обозначаемое $\mathbf{a}'\mathbf{b}$, представляет собой сумму попарных произведений компонент:

$$\mathbf{a}'\mathbf{b} = \sum_{i=1}^{m} a_i b_i = \mathbf{b}'\mathbf{a}. \tag{3.3}$$

Таблицу чисел, состоящую из k строк и m столбцов, называют *матрицей* размера $k \times m$ или $(k \times m)$-матрицей:

$$\mathbf{B} = \begin{pmatrix} b_{11} & b_{12} & \ldots & b_{1m} \\ b_{21} & b_{22} & \ldots & b_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ b_{k1} & b_{k2} & \ldots & b_{km} \end{pmatrix} = \{b_{ij}\}, \quad i = \overline{1, k}, \quad j = \overline{1, m}. \tag{3.4}$$

Произведение прямоугольной $(k \times m)$-матрицы $\mathbf{B}$ на $(m \times l)$-матрицу $\mathbf{C}$ есть $(k \times l)$-матрица вида

$$\mathbf{BC} = \left\{\sum_{s=1}^{m} b_{is} c_{sj}\right\}, \quad i = \overline{1, k}, \quad j = \overline{1, l}. \tag{3.5}$$

Поскольку m-мерные векторы-столбцы можно рассматривать как матрицы $m \times 1$, произведением двух векторов $\mathbf{ab}'$ является $(m \times m)$-матрица вида

$$\mathbf{ab}' = \{a_i b_j\}, \quad i = \overline{1, m}, \quad j = \overline{1, m}. \tag{3.5'}$$

Матрица, *обратная* данной квадратной матрице $\mathbf{A}$, обозначаемая $\mathbf{A}^{-1}$, определяется соотношением

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}, \tag{3.6}$$

$\mathbf{I}$ — *единичная матрица* (по ее диагонали единицы, остальные нули).

Транспонированная матрица по отношению к данной $\mathbf{B}$ имеет вид

$$\mathbf{B}' = \{b'_{ij}\}, \quad i = \overline{1, m}, \quad j = \overline{1, k}, \quad b'_{ij} = b_{ji}, \tag{3.7}$$

если $\mathbf{B}$ записана в форме (3.4), т. е $\mathbf{B}'$ имеет размер $m \times k$ и строки $\mathbf{B}'$ являются столбцами $\mathbf{B}$. Если $\mathbf{C}$ — матрица $m \times l$, $(\mathbf{BC})' = \mathbf{C}'\mathbf{B}'$.

Квадратная матрица $\mathbf{A}$ *симметрическая*, если $\mathbf{A}' = \mathbf{A}$.

Произведение $(k \times m)$-матрицы на m-мерный вектор — k-мерный вектор вида

$$\mathbf{Ba} = \left\{\sum_{j=1}^{m} b_{1j} a_j, \sum_{j=1}^{m} b_{2j} a_j, \ldots, \sum_{j=1}^{m} b_{kj} a_j\right\}'. \tag{3.8}$$

Система линейных уравнений с m неизвестными $\{x_1, x_2, \ldots x_m\} = \mathbf{x}'$ записывается в векторно-матричной форме:

$$\mathbf{Ax} = \mathbf{b}, \tag{3.9}$$

где $\mathbf{A}$ — матрица коэффициентов системы, $\mathbf{b}$ — вектор свободных членов. Домножив слева на $\mathbf{A}^{-1}$, получим решение системы в таком виде

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}. \tag{3.10}$$

Отсюда, в частности, следует, что $\mathbf{b}'\mathbf{A}^{-1}\mathbf{b} = \mathbf{b}'\mathbf{x} = \sum_{i=1}^{m} b_i x_i$, где $\mathbf{x}$ — решение системы (3.9).

Многомерная случайная величина как вектор. m-Мерной случайной величиной $\boldsymbol{\xi}$ называют m-мерный вектор, компонентами которого являются совместно наблюдаемые одномерные случайные величины ξ_1, ξ_2, ..., ξ_m:

$$\boldsymbol{\xi} = \{\xi_1, \xi_2, \ldots, \xi_m\}'. \tag{3.11}$$

Иными словами, m-мерная случайная величина — это m упорядоченных обычных одномерных случайных величин. Наблюдениями $\mathbf{x}$ m-мерной случайной величины являются m-мерные векторы, составленные из наблюдений ее компонент, полученных при одном и том же испытании:

$$\mathbf{x} = \{x_1, x_2, \ldots, x_m\}'. \tag{3.12}$$

В геометрической интерпретации наблюдение $\mathbf{x}$ образует точку m-мерного эвклидового пространства.

Область значений, принимаемых одномерной случайной величиной, располагается на действительной числовой оси. Все возможные наблюдения m-мерной случайной величины образуют область в m-мерном пространстве. В частном случае, ею может быть и все пространство.

Примеры многомерных случайных величин: упорядоченные результаты измерений содержаний химических элементов и минералов, физических свойств пород, а также те и другие вместе; направления действия физико-химических процессов; упорядоченные геологические и геоморфологические признаки в их количественном выражении; наблюденные значения и результаты трансформаций физических полей и т. п.

Над многомерными случайными величинами можно выполнять математические операции. Если $\mathbf{f}(\mathbf{x})$ — k-мерная вектор-значная функция, заданная на m-мерном пространстве —

$$\mathbf{f}(\mathbf{x}) = \{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_k(\mathbf{x})\}', \tag{3.13}$$

то $\mathbf{f}(\boldsymbol{\xi})$ будет m-мерной случайной величиной вида

$$\mathbf{f}(\boldsymbol{\xi}) = \{f_1(\boldsymbol{\xi}), f_2(\boldsymbol{\xi}), \ldots, f_k(\boldsymbol{\xi})\}', \tag{3.14}$$

где $f_i(\boldsymbol{\xi})$ — обычные (одномерные) случайные величины, получающие значения $f_i(\mathbf{x})$, если $\boldsymbol{\xi}$ получает значение $\mathbf{x}$.

Суммой двух m-мерных случайных величин $\boldsymbol{\xi}$ и $\boldsymbol{\eta}$ является m-мерная случайная величина, составленная из сумм соответствующих компонент:

$$\boldsymbol{\xi} + \boldsymbol{\eta} = \{\xi_1 + \eta_1, \xi_2 + \eta_2, \ldots, \xi_m + \eta_m\}'. \tag{3.15}$$

Скалярное произведение m-мерных случайных величин $\boldsymbol{\xi}$ и $\boldsymbol{\eta}$ — одномерная случайная величина вида

$$\boldsymbol{\xi}'\boldsymbol{\eta} = \sum_{i=1}^{m} \xi_i \eta_i, \tag{3.16}$$

произведение

$$\xi\eta' = \{\xi_i\eta_j\},\ i = \overline{1,\ m},\quad j = \overline{1,\ m} - \tag{3.17}$$

случайная матрица размером $m \times m$.

Если $\mathbf{B}$ — $(m \times m)$-матрица с неслучайными элементами, $\mathbf{B} = \{b_{ij}\}$, $i = \overline{1,\ m}$, $j = \overline{1,\ m}$, то $\mathbf{B}\xi$ — m-мерная случайная величина, представляющая собой вектор-столбец вида

$$\mathbf{B}\xi = \{\sum_{j=1}^{m} \xi_j b_{1j},\ \sum_{j=1}^{m} \xi_j b_{2j},\ \ldots,\ \sum_{j=1}^{m} \xi_j b_{mj}\}', \tag{3.18}$$

а $(\mathbf{B}\xi)' = \xi'\mathbf{B}'$ — вектор-строка, составленный из тех же компонент.

§ 2. Функция и плотность многомерного распределения

Функция распределения многомерной случайной величины. Функция m переменных, определяемая в виде

$$F(\mathbf{x}) = F(x_1, x_2, \ldots, x_m) = \mathbf{P}\{\xi_1 < x_1,\ \xi_2 < x_2,\ \ldots,\ \xi_m < x_m\}, \tag{3.19}$$

называется *функцией распределения m-мерной случайной величины* $\xi = \{\xi_1, \xi_2, \ldots, \xi_m\}'$.

Перечислим ее наиболее важные свойства.

1. $0 \leqslant F(x_1, x_2, \ldots, x_m) \leqslant 1$, $F(-\infty, -\infty, \ldots, -\infty) = 0$,

$$F(\infty, \infty, \ldots, \infty) = 1. \tag{3.20}$$

2. $F(x_1, x_2, \ldots, x_{j-1}, \infty, x_{j+1}, \ldots, x_m) = F_{m-1}(x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_m)$, (3.21)

где $F_{m-1}(x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_m)$ — функция распределения $(m-1)$-мерной случайной величины $\xi_{m-1} = \{\xi_1, \xi_2, \ldots, \xi_{j-1}, \xi_{j+1}, \ldots, \xi_m\}'$. Отсюда следует также свойство:

$$F(x_1, x_2, \ldots, x_k, \infty, \ldots, \infty) = F_k(x_1, x_2, \ldots, x_k), \tag{3.21'}$$

где $F_k(x_1, x_2, \ldots, x_k)$ — функция распределения k-мерной случайной величины $\xi_k = \{\xi_1, \xi_2, \ldots, \xi_k\}'$. В частности,

$$F(\infty, \infty, \ldots, \infty, x_k, \infty, \ldots, \infty) = F_k(x_k), \tag{3.21''}$$

где $F_k(x_k)$ — функция распределения ξ_k.

3. Вероятность события $\{a_1 \leqslant \xi_1 < b_1,\ a_2 \leqslant \xi_2 < b_2,\ \ldots,\ a_m \leqslant \xi_m < b_m\}$ вычисляется с помощью функции распределения по приведенной ниже формуле (3.23).

Пусть

$$\Delta_1(x_2, x_3, \ldots, x_m) = F(b_1, x_2, x_3, \ldots, x_m) - F(a_1, x_2, x_3, \ldots, x_m);$$
$$\Delta_2(x_3, x_4, \ldots, x_m) = \Delta_1(b_2, x_3, x_4, \ldots, x_m) - \Delta_1(a_2, x_3, x_4, \ldots, x_m);$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$\Delta_{m-1}(x_m) = \Delta_{m-2}(b_{m-1}, x_m) - \Delta_{m-2}(a_{m-1}, x_m),$$
$$\Delta_m = \Delta_{m-1}(b_m) - \Delta_{m-1}(a_m). \tag{3.22}$$

Тогда

$$\mathbf{P}\{a_1 \leqslant \xi_1 < b_1,\ \xi_2 < b_2,\ \ldots,\ \xi_m < b_m\} = \mathbf{P}(\xi_1 < b_1,\ \xi_2 < b_2,\ \ldots,\ \xi_m < b_m) - \mathbf{P}\{\xi_1 < a_1,\ \xi_2 < b_2,\ \ldots,\ \xi_m < b_m\} = \Delta_1(b_2, b_3,\ \ldots,\ b_m); \quad (3.22')$$

$$\mathbf{P}\{a_1 \leqslant \xi_1 < b_1,\ a_2 \leqslant \xi_2 < b_2,\ \xi_3 < b_3,\ \ldots,\ \xi_m < b_m\} = \mathbf{P}(a_1 \leqslant \xi_1 < b_1,\ \xi_2 < b_2,\ \ldots,\ \xi_m < b_m\} - \mathbf{P}\{a_1 \leqslant \xi_1 < b_1,\ \xi_2 < a_2,\ \xi_3 < b_3,\ \ldots,\ \xi_m < b_m\} = \Delta_1(b_2, b_3,\ \ldots,\ b_m) - \Delta_1(a_2, b_3,\ \ldots,\ b_m) = \Delta_2(b_3, b_4,\ \ldots,\ b_m).$$

Продолжая этот процесс, получим окончательно:

$$\mathbf{P}\{a_i \leqslant \xi_i < b_i,\ i = \overline{1, m}\} = F(b_1, b_2,\ \ldots,\ b_m) - \sum_{i=1}^{m} p_i + \sum_{i<j}^{m} p_{ij} - \ldots + (-1)^m F(a_1, a_2,\ \ldots,\ a_m), \quad (3.23)$$

где $p_{ij\ldots k} = F(c_1, c_2, \ldots c_k)$ при $c_i = a_i$, $c_j = a_j$, $\ldots$, $c_k = a_k$ и остальных c_s, равных b_s.

4. Если компоненты m-мерной случайной величины $\xi = \{\xi_1, \xi_2, \ldots, \xi_m\}'$ взаимно независимы, то функция ее распределения, согласно (1.26), имеет вид

$$F_m(x_1, x_2,\ \ldots,\ x_m) = \prod_{i=1}^{m} F_i(x_i), \quad (3.24)$$

где $F_i(x_i)$ — функция распределения ξ_i.

Плотность многомерного распределения. Аналогом плотности распределения для многомерных непрерывных* случайных величин служит функция

$$p(x_1, x_2,\ \ldots,\ x_m) = \frac{\partial F(x_1, x_2,\ \ldots,\ x_m)}{\partial x_1 \partial x_2 \ldots \partial x_m}, \quad (3.25)$$

которую называют *плотностью многомерного распределения*. Из (3.25) —

$$F(x_1, x_2,\ \ldots,\ x_m) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \ldots \int_{-\infty}^{x_m} p(t_1, t_2,\ \ldots,\ t_m)\, dt_1 dt_2 \ldots dt_m. \quad (3.25')$$

Из определения плотности (3.25) следуют такие ее свойства.

1. Как и в одномерном случае, плотность многомерного распределения неотрицательна и на бесконечности обращается в нуль.

2. Если Q — некоторая область в m-мерном пространстве, то вероятность попадания наблюдения m-мерной случайной величины ξ в эту область представляется в виде

$$\mathbf{P}\{\xi \in Q\} = \iint\limits_{Q} \ldots \int p(x_1, x_2,\ \ldots,\ x_m)\, dx_1 dx_2 \ldots dx_m = \int\limits_{Q} p(\mathbf{x})\, d\mathbf{x}. \quad (3.26)$$

---

* *Непрерывными* называются многомерные случайные величины, функции распределения которых непрерывны.

Чтобы убедиться в этом, разобьем Q на N малых m-мерных параллелепипедов $v_j = \{\delta_1^{(j)}, \delta_2^{(j)}, \ldots, \delta_m^{(j)}\}$ ($\delta_i^{(j)}$ — проекция v_j на ось $0x_i$, $x_i^{(j)}$ — середина $\delta_i^{(j)}$). Такое расчленение Q возможно с точностью до слоя, прилегающего к границе Q, которым при переходе к пределу можно пренебречь. Очевидно,

$$\mathbf{P}\{\xi \in Q\} = \sum_{j=1}^{N} \mathbf{P}\{\xi \in v_j\}. \tag{3.26'}$$

Обозначим $\Delta_k^{(j)}(x_{k+1}, x_{k+2}, \ldots, x_m)$ ($k = \overline{1, m-1}$) разности (3.22) для параллелепипеда v_j. По (3.22) и (3.22')

$$\mathbf{P}\{\xi_1 \in \delta_1^{(j)}, \xi_2 < x_2, \ldots, \xi_m < x_m\} = \Delta_1^{(j)}(x_2, x_3, \ldots, x_m) = \frac{\partial F(x_1^{(j)}, x_2, \ldots, x_m)}{\partial x_1}\delta_1^{(j)} + o(\delta_1^{(j)}),$$

$$\mathbf{P}\{\xi_1 \in \delta_1^{(j)}, \xi_2 \in \delta_2^{(j)}, \xi_3 < x_3, \ldots, \xi_m < x_m\} = \Delta_2^{(j)}(x_3, x_4, \ldots, x_m) = \frac{\partial F(x_1^{(j)}, x_2^{(j)}, x_3, \ldots, x_m)}{\partial x_1 \partial x_2}\delta_1^{(j)}\delta_2^{(j)} + o(\delta_1^{(j)}\delta_2^{(j)})$$

и вообще

$$\mathbf{P}\{\xi \in v_j\} = \frac{\partial F(x_1^{(j)}, x_2^{(j)}, \ldots, x_m^{(j)})}{\partial x_1 \partial x_2 \ldots \partial x_m}\delta_1^{(j)}\delta_2^{(j)} \ldots \delta_m^{(j)} + o(\delta_1^{(j)}\delta_2^{(j)} \ldots \delta_m^{(j)}). \tag{3.26''}$$

Перейдя к пределу в (3.26') при $N \to \infty$, $\max\limits_{i,\,j} \delta_i^{(j)} \to 0$, получим (3.26).

Если $[a_i, b_i]$ — интервалы для компонент ξ_i, $i = \overline{1, m}$, то по (3.26)

$$\mathbf{P}\{a_i \leqslant \xi_i < b_i,\ i = \overline{1, m}\} = \int_{a_1}^{b_1}\int_{a_2}^{b_2} \ldots \int_{a_m}^{b_m} p(x_1, x_2, \ldots, x_m)\, dx_1 dx_2 \ldots dx_m. \tag{3.26'''}$$

3. Плотность распределения k-мерной случайной величины $\xi_k = \{\xi_1, \xi_2, \ldots, \xi_k\}'$, составленной из k компонент $\xi_1, \xi_2, \ldots, \xi_k$ m-мерной величины ξ ($k < m$), имеет вид

$$p(x_1, x_2, \ldots, x_k) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} p(x_1, x_2, \ldots, x_m)\, dx_{k+1}\, dx_{k+2} \ldots dx_m. \tag{3.21'''}$$

В частности, по этой формуле можно определить плотность распределения отдельной компоненты ξ_k, интегрируя в (3.21''') по всем переменным, кроме x_k.

4. Плотность распределения многомерной случайной величины ξ с независимыми компонентами ξ_i, согласно (3.24) и (3.25), имеет вид

$$p(\mathbf{x}) = p(x_1, x_2, \ldots, x_m) = \prod_{i=1}^{m} p_i(x_i), \tag{3.27}$$

где $p_i(x_i)$ — плотность распределения ξ_i. Если q_i — вероятности попа-

дания независимых ξ_i в интервалы $[a_i, b_i)$, $q_i = \mathbf{P}\{a_i \leqslant \xi_i < b_i\}$, то из (3.26^m) и (3.27) вероятность одновременного попадания всех ξ_i в свои пределы a_i, b_i равна произведению вероятностей q_i:

$$\mathbf{P}\{a_i \leqslant \xi_i < b_i, \ i = \overline{1, m}\} = \prod_{i=1}^{m} \mathbf{P}\{a_i \leqslant \xi_i < b_i\} = \prod_{i=1}^{m} q_i. \quad (3.28)$$

Эту формулу мы получили раньше другим способом (1.27).

Если нужно обеспечить вероятность q одновременного попадания величин ξ_i в свои интервалы $[a_i, b_i)$, последние следует строить с таким расчетом, чтобы для соответствующих им вероятностей q_i выполнялось условие

$$\prod_{i=1}^{m} q_i = q. \quad (3.29)$$

Если q_i считать одинаковыми, $q_i = q_0$, то

$$q_0 = \sqrt[m]{q}. \quad (3.29')$$

В частности, если ξ_i нормально распределены, то пределы для ξ_i, удовлетворяющие условию $\mathbf{P}\{a_i \leqslant \xi_i < b_i, \ i = \overline{1, m}\} = q$, имеют вид:

$$a_i = \mathbf{M}\xi_i - u_{\frac{1+q_0}{2}} \sqrt{\mathbf{D}\xi_i}, \ b_i = \mathbf{M}\xi_i + u_{\frac{1+q_0}{2}} \sqrt{\mathbf{D}\xi_i},$$

где q_0 вычисляются по (3.29′).

5. С помощью формулы (3.26) доказывается такое свойство: функция распределения суммы случайных величин $\eta = \xi_1 + \xi_2 + \cdots + \xi_m$, являющихся компонентами m-мерной величины $\boldsymbol{\xi}$ с плотностью распределения $p(\mathbf{x})$, представляется в виде

$$F_\eta(z) = \int\int\limits_{\sum_1^m x_j < z} \cdots \int p(\mathbf{x})\, d\mathbf{x} = \int\limits_{-\infty}^{z - \sum_{i=2}^{m} x_i} \int\limits_{-\infty}^{\infty} \cdots \int\limits_{-\infty}^{\infty} p(\mathbf{x})\, d\mathbf{x}; \quad (3.30)$$

плотность распределения суммы $\eta = \sum_{i=1}^{m} \xi_i$

$$p_\eta(z) = \frac{dF_\eta(z)}{dz} = \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} \cdots \int\limits_{-\infty}^{\infty} p\left(z - \sum_{i=2}^{m} x_i, \ x_2, \ \ldots, x_m\right) dx_2 dx_3 \ldots dx_m. \quad (3.31)$$

Из формулы (3.31), в частности, следует (1.28).

6. Пусть m-мерная случайная величина $\boldsymbol{\eta} = \mathbf{f}(\boldsymbol{\xi})$ является результатом преобразования данной m-мерной случайной величины $\boldsymbol{\xi}$. Обозначим $\left|\frac{D\mathbf{f}(\mathbf{x})}{D\mathbf{x}}\right|$ якобиан преобразования (определитель матрицы

$\left\{\frac{\partial f_i}{\partial x_j}\right\}$, $i = \overline{1, m}$, $j = \overline{1, m}$). Если преобразование устанавливает взаимно однозначное соответствие* между совокупностями значений ξ и η, то плотности распределения ξ и η связаны соотношением

$$p_\xi(\mathbf{x}) = \left|\frac{D\mathbf{f}(\mathbf{x})}{D\mathbf{x}}\right| p_\eta(\mathbf{f}(\mathbf{x})). \tag{3.32}$$

Если $\eta = \mathbf{A}\xi$, $\mathbf{A}$ — невырожденная матрица (т. е. её определитель $|\mathbf{A}|$ не равен нулю), то

$$p_\xi(\mathbf{x}) = |\mathbf{A}|\, p_\eta(\mathbf{Ax}), \quad p_\eta(\mathbf{x}) = |\mathbf{A}|^{-1} p_\xi(\mathbf{A}^{-1}\mathbf{x}). \tag{3.32'}$$

Понятие условного распределения. Пусть $\Delta_1, \Delta_2, \ldots, \Delta_m$ — интервалы вокруг $x_1, x_2, \ldots, x_m$. Условная вероятность события $\{\xi_1 \in \Delta_1, \xi_2 \in \Delta_2, \ldots, \xi_k \in \Delta_k\}$ при условии, что $\{\xi_{k+1} \in \Delta_{k+1}, \xi_{k+2} \in \Delta_{k+2}, \ldots, \xi_m \in \Delta_m\}$, согласно (1.8) составит

$$\mathbf{P}\{\xi_i \in \Delta_i,\ i = \overline{1, k} \,/\, \xi_j \in \Delta_j,\ j = \overline{k+1, m}\} = \frac{\mathbf{P}\{\xi_i \in \Delta_i,\ i = \overline{1, m}\}}{\mathbf{P}\{\xi_j \in \Delta_j,\ j = \overline{k+1, m}\}}.$$

Естественно определить условную плотность распределения $\{\xi_1, \xi_2, \ldots, \xi_k\}' = \xi_k$ при фиксированных $\xi_{k+1} = x_{k+1}, \xi_{k+2} = x_{k+2}, \ldots, \xi_m = x_m$ в виде

$$p(x_1, x_2, \ldots, x_k \,/\, x_{k+1}, x_{k+2}, \ldots, x_m) =$$

$$= \lim_{\max\Delta_i \to 0} \frac{\mathbf{P}\{\xi_i \in \Delta_i,\ i = \overline{1, k} \,/\, \xi_j \in \Delta_j, j = \overline{k+1, m}\}}{\Delta_1\Delta_2 \ldots \Delta_k} =$$

$$= \lim_{\max\Delta_i \to 0} \frac{p(x_1, x_2, \ldots, x_m)\,\Delta_1\Delta_2 \ldots \Delta_m}{p_{m-k}(x_{k+1}, x_{k+2}, \ldots, x_m)\,\Delta_1\Delta_2 \ldots \Delta_m} =$$

$$= \frac{p(x_1, x_2, \ldots, x_m)}{p_{m-k}(x_{k+1}, x_{k+2}, \ldots, x_m)}, \tag{3.33}$$

где $p_{m-k}(x_{k+1}, \ldots, x_m)$ — плотность распределения $\{\xi_{k+1}, \xi_{k+2}, \ldots, \xi_m\} = \xi'_{m-k}$. Использование в приведенных выше формулах этого выражения плотности дает возможность вычислять условные вероятности событий $\{\xi_k \in Q\}$, где Q — некоторая область значений ξ_k.

Аналогично, если условием является $\xi_{m-k} \in Q_0$, Q_0 — некоторая область значений ξ_{m-k}, то условная плотность распределения k-мерной величины ξ_k —

$$p(x_1, x_2, \ldots, x_k \,/\, \xi_{m-k} \in Q_0) = \iint\limits_{Q_0} \cdots \int p(x_1, x_2, \ldots, x_k, t_{k+1}, t_{k+2},$$

$$\ldots, t_m)\, dt_{k+1} dt_{k+2} \ldots dt_m \,/ \iint\limits_{Q_0} \cdots \int p_{m-k}(t_{k+1}, t_{k+2}, \ldots, t_m) \times$$

$$\times\, dt_{k+1}\, dt_{k+2} \ldots dt_m. \tag{3.33'}$$

---

* Таким будет преобразование, якобиан которого не обращается в нуль.

§ 3. Числовые характеристики многомерных распределений

Основными числовыми характеристиками многомерных случайных величин являются: математическое ожидание, представляющее собой вектор математических ожиданий отдельных компонент

$$\mathbf{M}\boldsymbol{\xi} = \{\mathbf{M}\xi_1, \mathbf{M}\xi_2, \ldots, \mathbf{M}\xi_m\}' \tag{3.34}$$

и *ковариационная матрица*

$$\mathbf{B}(\boldsymbol{\xi}) = \mathbf{B} = \{b_{ij}\}, \quad i = \overline{1, m}, \quad j = \overline{1, m}; \quad b_{ij} = \mathbf{M}[(\xi_i - \mathbf{M}\xi_i)(\xi_j - \mathbf{M}\xi_j)] = \mathbf{M}\xi_i\xi_j - \mathbf{M}\xi_i\mathbf{M}\xi_j. \tag{3.35}$$

Величина b_{ij} называется *ковариацией* ξ_i и ξ_j. По диагонали ковариационной матрицы, как это следует из (3.35), расположены дисперсии компонент ξ_i: $b_{ii} = \mathbf{M}(\xi_i - \mathbf{M}\xi_i)^2 = \mathbf{D}\xi_i$.

С использованием векторно-матричной записи ковариационная матрица представляется в виде

$$\mathbf{B} = \mathbf{M}[(\boldsymbol{\xi} - \mathbf{M}\boldsymbol{\xi})(\boldsymbol{\xi} - \mathbf{M}\boldsymbol{\xi})']. \tag{3.35'}$$

При этом по определению полагают, что математическое ожидание матрицы со случайными элементами представляет собой матрицу из математических ожиданий соответствующих элементов.

Свойства математического ожидания и ковариационной матрицы многомерной случайной величины. 1. Математическое ожидание функции $f(\boldsymbol{\xi}) = f(\xi_1, \xi_2, \ldots, \xi_m)$ m-мерной непрерывной случайной величины $\boldsymbol{\xi}$ вычисляется по формуле*:

$$\mathbf{M}f(\xi_1, \xi_2, \ldots, \xi_m) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_m)\, p(x_1, x_2, \ldots, x_m)\, dx_1 dx_2 \ldots dx_m. \tag{3.36}$$

2. Математическое ожидание суммы m-мерных случайных величин $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_k$ равно сумме их математических ожиданий:

$$\mathbf{M}(\boldsymbol{\xi}_1 + \boldsymbol{\xi}_2 + \ldots + \boldsymbol{\xi}_k) = \mathbf{M}\boldsymbol{\xi}_1 + \mathbf{M}\boldsymbol{\xi}_2 + \ldots + \mathbf{M}\boldsymbol{\xi}_k. \tag{3.37}$$

3. Если $\mathbf{A}$ — некоторая $(m \times m)$-матрица с неслучайными элементами, $\boldsymbol{\xi}$ — m-мерная случайная величина, $\mathbf{b}$ — неслучайный m-мерный вектор, то математическое ожидание величины $\mathbf{A}\boldsymbol{\xi} + \mathbf{b}$, $\mathbf{M}(\mathbf{A}\boldsymbol{\xi} + \mathbf{b}) = \mathbf{A}\mathbf{M}\boldsymbol{\xi} + \mathbf{b}$, а ее ковариационная матрица

$$\mathbf{B}(\mathbf{A}\boldsymbol{\xi} + \mathbf{b}) = \mathbf{M}[(\mathbf{A}\boldsymbol{\xi} + \mathbf{b} - \mathbf{A}\mathbf{M}\boldsymbol{\xi} - \mathbf{b})(\mathbf{A}\boldsymbol{\xi} + \mathbf{b} - \mathbf{A}\mathbf{M}\boldsymbol{\xi} - \mathbf{b})'] = \mathbf{A}\mathbf{M}[(\boldsymbol{\xi} - \mathbf{M}\boldsymbol{\xi})(\boldsymbol{\xi} - \mathbf{M}\boldsymbol{\xi})'\mathbf{A}'] = \mathbf{A}\mathbf{B}\mathbf{A}', \tag{3.38}$$

где $\mathbf{B}$ — ковариационная матрица $\boldsymbol{\xi}$.

---

* Функция $f(x_1, x_2, \ldots, x_m)$ предполагается непрерывной.

4. Математическое ожидание произведений $\xi'\eta$ и $\xi\eta'$ двух независимых* m-мерных случайных величин ξ и η равно произведениям их математических ожиданий:

$$\mathbf{M}\xi'\eta = (\mathbf{M}\xi)'\,\mathbf{M}\eta, \quad \mathbf{M}\xi\eta' = \mathbf{M}\xi\,(\mathbf{M}\eta)'. \tag{3.39}$$

5. Ковариационная матрица суммы или разности двух независимых m-мерных случайных величин ξ и η равна сумме их ковариационных матриц:

$$\mathbf{B}(\xi \pm \eta) = \mathbf{M}\{[\xi \pm \eta - (\mathbf{M}\xi \pm \mathbf{M}\eta)]\,[\xi \pm \eta - (\mathbf{M}\xi \pm \mathbf{M}\eta)]'\} =$$
$$= \mathbf{M}[(\xi - \mathbf{M}\xi)(\xi - \mathbf{M}\xi)'] + \mathbf{M}[(\eta - \mathbf{M}\eta)(\eta - \mathbf{M}\eta)'] \pm \mathbf{M}[(\xi - \mathbf{M}\xi)(\eta - \mathbf{M}\eta)'] \mp \mathbf{M}[(\eta - \mathbf{M}\eta)(\xi - \mathbf{M}\xi)'] = \mathbf{B}(\xi) + \mathbf{B}(\eta). \tag{3.40}$$

6. Ковариационная матрица симметрическая и *неотрицательно определена*, т. е. обладает свойствами:

$$b_{ij} = \mathbf{M}[(\xi_i - \mathbf{M}\xi_i)(\xi_j - \mathbf{M}\xi_j)] = b_{ji}; \tag{3.41}$$

$\mathbf{x}'\mathbf{B}\mathbf{x} \geqslant 0$ для произвольного вектора $\mathbf{x}$.

Коэффициент корреляции. При анализе многомерных распределений часто используется *корреляционная матрица*

$$\mathbf{R} = \{r_{ij}\}, \quad i = \overline{1, m}, \quad j = \overline{1, m}, \tag{3.42}$$

элементы которой вычисляются по элементам ковариационной матрицы многомерной случайной величины ξ в виде

$$r_{ij} = b_{ij}/\sqrt{\mathbf{D}\xi_i \mathbf{D}\xi_j} = b_{ij}/\sqrt{b_{ii}b_{jj}}. \tag{3.43}$$

Величина r_{ij} безразмерная и носит название *коэффициента корреляции* случайных величин ξ_i и ξ_j. По диагонали $\mathbf{R}$, как следует из (3.43), расположены единицы: $r_{ii} = 1$. Коэффициент корреляции используется для оценки силы и характера линейной *статистической связи* между двумя случайными величинами. Прежде, чем перейти к характеристике этого содержания коэффициента корреляции, рассмотрим само понятие статистической связи.

Компоненты ξ_i многомерной случайной величины — одномерные случайные величины — могут быть независимыми или статистически связанными. Под статистической связью мы будем понимать связь двух зависимых случайных величин. При функциональной зависимости значение одной величины полностью определяется значением другой. В отличие от нее при статистической связи, в зависимости от того, какое значение принимает одна случайная величина, изменяется функция распределения другой. Согласно (3.33), плотность условного распределения ξ_i при условии $\xi_j = x_j$ будет $p_1(x_i/x_j) = p(x_i, x_j)/p_2(x_j)$, где $p(x_i, x_j)$ — плотность двумерного распределения $\xi = \{\xi_i, \xi_j\}'$, $p_2(x_j)$ — плотность распределения ξ_j. Лишь когда $p(x_i, x_j) = p_1(x_i) \times$

* Точно так же, как и одномерные случайные величины, m-мерные величины ξ и η независимы, если функция совместного распределения ξ и η равна произведению функций распределения ξ и η.

$\times p_2(x_j)$, т. е. ξ_i и ξ_j независимы, условное распределение ξ_i совпадает с обычным — $p_1(x_i/x_j) = p_1(x_i)$ и не зависит от того, какое значение получает ξ_j.

Если попытаться изобразить статистическую связь графически, откладывая на координатных осях наблюдаемые при каждом испытании значения ξ_i и ξ_j, мы не получим четко обозначенной кривой, на которую ложились бы нанесенные точки (рис. 21). Зависимость будет сказываться на расположении точек лишь в виде тенденции к группировке их определенным образом, которая проявляется тем четче, чем сильнее связь, чем ближе она к функциональной. При этом точки будут рассеиваться вокруг функции *условного математического ожидания*

$$f(x) = \mathbf{M}[\xi_i/(\xi_j = x)] = \int_{-\infty}^{\infty} t p_1(t/x)\, dt = \int_{-\infty}^{\infty} t \frac{p(t, x)}{p_2(x)}\, dt. \qquad (3.44)$$

Эта функция описывает функциональную составляющую статистической связи и носит название *регрессии* или *функции регрессии* ξ_i на ξ_j. Самое же статистическую связь между двумя случайными величинами ξ_i и ξ_j, подобную изображенной на рис. 21, можно выразить уравнением *

$$\xi_i = f(\xi_j) + \Delta_{ij}, \qquad (3.45)$$

где $f(\xi_j)$ — функция регрессии; Δ_{ij} — случайная величина с нулевым математическим ожиданием.

$$\mathbf{M}[\xi_i/(\xi_j = x)] = \mathbf{M}(f(x) + \Delta_{ij}) = f(x) + \mathbf{M}\,\Delta_{ij} = f(x).$$

Величина Δ_{ij} имеет смысл взятого со знаком отклонения ξ_i от значения $f(\xi_j)$, предписываемого регрессией. Наличие случайной добавки Δ_{ij} как раз и приводит к тому, что связь проявляется лишь в виде тенденции на фоне случайных отклонений. На рис. 21 точкам, лежащим ниже кривой регрессии $f(x)$ соответствуют наблюдения, для которых значения Δ_{ij} оказались отрицательными. Точкам, расположенным выше $f(x)$, соответствуют положительные значения Δ_{ij}.

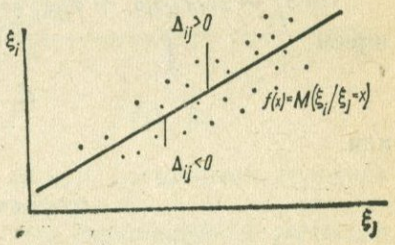


Рис. 21.

Линейная статистическая зависимость выражается уравнением

$$\xi_i = \alpha_{ij}\xi_j + \beta_{ij} + \Delta_{ij}, \qquad (3.46)$$

где α_{ij}, β_{ij} — коэффициенты связи. Будем считать, что Δ_{ij} не зависит от ξ_j. Силу связи между ξ_i и ξ_j можно охарактеризовать дисперсией $\mathbf{D}\,\Delta_{ij}$: чем меньше $\mathbf{D}\,\Delta_{ij}$ по сравнению с дисперсией $\mathbf{D}\,\xi_i$, тем более выражена

* Такое представление не является самым общим; полное описание статистической связи дает лишь функция двумерного распределения величин ξ_i и ξ_j. Однако, для задач анализа геолого-геофизической информации это удобное представление оказывается, как правило, вполне достаточным.

связь между ξ_i и ξ_j, тем ближе она к функциональной. Как показывают формулы (3.45), (3.46), $\mathbf{D}\,\Delta_{ij}$ — дисперсия ошибки прогноза значения ξ_i по известному значению ξ_j. Если она равна нулю, $\mathbf{D}\,\Delta_{ij}=0$, связь функциональна: $\xi_i=\alpha_{ij}\xi_j+\beta_{ij}$.

Коэффициенты α_{ij}, β_{ij} связаны с коэффициентом корреляции r_{ij} величин ξ_i, ξ_j, их средними квадратическими отклонениями σ_i, σ_j и математическими ожиданиями M_i, M_j следующими соотношениями

$$\alpha_{ij}=r_{ij}\frac{\sigma_i}{\sigma_j},\quad \beta_{ij}=M_i-\alpha_{ij}M_j. \tag{3.47}$$

В этом нетрудно убедиться: так как $\mathbf{M}\,\Delta_{ij}=0$, $\mathbf{M}\,(\xi_i-\alpha_{ij}\xi_j-\beta_{ij})=\mathbf{M}\,\Delta_{ij}=M_i-\alpha_{ij}M_j-\beta_{ij}=0$, откуда $\beta_{ij}=M_i-\alpha_{ij}M_j$. Учитывая, что Δ_{ij} не зависит от ξ_j, по свойству математического ожидания произведения независимых случайных величин $\mathbf{M}\,[(\xi_j-M_j)\,\Delta_{ij}]=[M_j-M_j)\,\mathbf{M}\,\Delta_{ij}=0$. Так как $\mathbf{M}\,[(\xi_j-M_j)\,\Delta_{ij}]=\mathbf{M}\,[(\xi_j-M_j)\,(\xi_i-\alpha_{ij}\xi_j-\beta_{ij})]=\mathbf{M}\,[(\xi_j-M_j)\,(\xi_i-\alpha_{ij}\xi_j-M_i+\alpha_{ij}M_j)]=r_{ij}\sigma_i\sigma_j-\alpha_{ij}\sigma_j^2=0$ (учитывая (3.43), (3.35)), получим $\alpha_{ij}=r_{ij}\sigma_i/\sigma_j$.

Отметим, что регрессии ξ_i на ξ_j и ξ_j на ξ_i, построенные в одной и той же системе координат, в общем случае не совпадают:

$$\alpha_{ji}=\frac{\alpha_{ij}\sigma_j^2}{\sigma_i^2}\neq\frac{1}{\alpha_{ij}};\quad \beta_{ji}=M_j-\alpha_{ji}M_i\neq-\frac{\beta_{ij}}{\alpha_{ij}}.$$

Смысл коэффициента корреляции как характеристики силы линейной связи становится ясным из следующего. Так как дисперсия Δ_{ij},

$$\mathbf{D}\,\Delta_{ij}=\mathbf{M}\,(\xi_i-\alpha_{ij}\xi_j-\beta_{ij})^2=\mathbf{M}\,[\xi_i-M_i-\alpha_{ij}\,(\xi_j-M_j)]^2=$$
$$=\sigma_i^2-2\alpha_{ij}r_{ij}\sigma_i\sigma_j+\alpha_{ij}^2\sigma_j^2=\sigma_i^2-2r_{ij}^2\sigma_i^2+r_{ij}^2\sigma_i^2=\sigma_i^2\,(1-r_{ij}^2), \tag{3.48}$$

имеем

$$r_{ij}^2=1-\frac{\mathbf{D}\,\Delta_{ij}}{\sigma_i^2} \tag{3.48'}$$

или

$$|r_{ij}|=\sqrt{1-\frac{\mathbf{D}\,\Delta_{ij}}{\sigma_i^2}}. \tag{3.48''}$$

Так как $1-r_{ij}^2=\mathbf{D}\,\Delta_{ij}/\sigma_i^{-2}\geqslant 0$, $r_{ij}^2\leqslant 1$ и $|r_{ij}|\leqslant 1$ — коэффициент корреляции всегда находится в пределах от -1 до 1. $\mathbf{D}\,\Delta_{ij}\leqslant\mathbf{D}\xi_j$ — рассеивание значений ξ_j вокруг функции регрессии не превосходит рассеивания их вокруг математического ожидания $\mathbf{M}\xi_j$. При $|r_{ij}|=1$ дисперсия $\mathbf{D}\,\Delta_{ij}=0$ и связь между ξ_i и ξ_j будет детерминированной и линейной.

Если коэффициент корреляции положителен, связь между ξ_i и ξ_j называют прямой, так как тогда, согласно (3.47), $\alpha_{ij}>0$ и бо́льшим значениям ξ_j соответствуют бо́льшие в среднем значения ξ_i. Если $r_{ij}<0$, то $\alpha_{ij}<0$ и связь между ξ_i и ξ_j обратная — бо́льшим значениям ξ_j соответствуют в среднем меньшие значения ξ_i.

Коэффициент корреляции не изменится, если к величинам ξ_i и ξ_j применить невырожденные линейные преобразования с коэффициентами одного знака при ξ_i и ξ_j. Пусть $\eta_i = a_i\xi_i + b_i$, $\eta_j = a_j\xi_j + b_j$ ($a_i \neq 0$, $a_j \neq 0$). Коэффициент корреляции η_i и η_j

$$r^0_{ij} = \frac{\mathbf{M}\left[(a_i\xi_i + b_i - a_iM_i - b_i)\,(a_j\xi_j + b_j - a_jM_j - b_j)\right]}{|a_i|\,\sigma_i \cdot |a_j|\cdot\sigma_j} = \\ = \frac{a_ia_j\mathbf{M}\left[(\xi_i - \mathbf{M}\xi_i)\,(\xi_j - \mathbf{M}\xi_j)\right]}{|a_i||a_j|\,\sigma_i\sigma_j} = \frac{a_ia_jr_{ij}}{|a_ia_j|}. \qquad (3.49)$$

В частности, нормированные величины $(\xi_i - M_i)/\sigma_i$ и $(\xi_j - M_j)/\sigma_j$, как следует из (3.47), а также из того, что они имеют нулевые математические ожидания и единичные дисперсии, связаны соотношением

$$\frac{\xi_i - M_i}{\sigma_i} = r_{ij}\frac{\xi_j - M_j}{\sigma_j} + \delta_{ij}, \qquad (3.50)$$

где δ_{ij} — случайная величина, не зависящая от ξ_j, с нулевым математическим ожиданием и дисперсией, по (3.48),

$$\mathbf{D}\delta_{ij} = 1 - r^2_{ij}. \qquad (3.51)$$

Если коэффициент корреляции равен нулю, связь между величинами ξ_i и ξ_j в линейной форме (3.46) отсутствует. Это следует из того, что тогда $\alpha_{ij} = 0$, $\xi_i = \Delta_{ij} + \beta_{ij}$ и ξ_i не зависит от ξ_j, так как Δ_{ij} не зависит от ξ_j. Следует подчеркнуть, что такой вывод справедлив лишь по отношению к линейной связи. Можно привести примеры криволинейных зависимостей между случайными величинами, для которых коэффициент корреляции близок к нулю или даже равен нулю. Это происходит от того, что прямолинейная регрессия, описывающая зависимость при корреляционном анализе, оказывается слишком далеким приближением для реально существующей связи (рис. 22). При таком приближении отклонения этой зависимости от прямолинейной формально относятся к Δ_{ij}, увеличивая $\mathbf{D}\,\Delta_{ij}$, что влечет уменьшение абсолютной величины коэффициента корреляции. Тем не менее обратный вывод справедлив всегда: если две случайные величины ξ_i и ξ_j независимы, то их ковариация и коэффициент корреляции равны нулю. Это следует из того, что

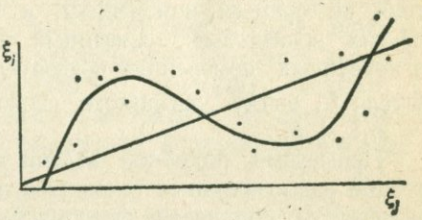


Рис. 22.

$$\mathbf{M}\left[(\xi_i - M_i)\,(\xi_j - M_j)\right] = (M_i - M_i)\,(M_j - M_j) = 0,$$

если ξ_i и ξ_j независимы, в соответствии с (1.47).

Коэффициенты (3.47) обладают следующим важным свойством: среди всех линейных функций $\eta = a\xi_j + b$ случайной величины ξ_j, ближайшие (в среднем) значения к наблюдениям ξ_i дает регрессия $\xi_i = \alpha_{ij}\xi_j + \beta_{ij}$, где α_{ij}, β_{ij} определяются по (3.47). Мерой близости ξ_i и η является средний квадрат $\mathbf{M}(\xi_i - \eta)^2$ разности $\xi_i - \eta$. Эта величина имеет вид

$$\mathbf{M}(\xi_i - \eta)^2 = \mathbf{M}(\xi_i - a\xi_j - b)^2 = \mathbf{M}[\xi_i - M_i - a(\xi_j - M_j) - b + M_i - aM_j]^2 = \sigma_i^2 + a^2\sigma_j^2 - 2ab_{ij} + (b - M_i + aM_j)^2,$$

где $b_{ij} = r_{ij}\sigma_i\sigma_j$ — ковариация ξ_i и ξ_j. Приравнивая нулю производные по a и b, получим

$$\begin{cases} a\sigma_j^2 + (b - M_i + aM_j)M_j - r_{ij}\sigma_i\sigma_j = 0, \\ b - M_i + aM_j = 0, \end{cases}$$

откуда, решая систему уравнений относительно a и b,

$$a = r_{ij}\sigma_i / \sigma_j = \alpha_{ij}, \quad b = M_i - \alpha_{ij}M_j = \beta_{ij}.$$

Таким образом, средний квадрат отклонения значения ξ_i, прогнозируемого в виде $\eta = a\xi_j + b$, от действительного минимален при $a = \alpha_{ij}$, $b = \beta_{ij}$. Это свойство используется для решения задачи *косвенных измерений*. Сущность ее состоит в определении значений показателя не прямым измерением, а по величинам другого или группы других показателей, связанных с ним. В данном случае речь идет о косвенных измерениях показателя ξ_i по значениям ξ_j. Значения показателя ξ_i, ввиду указанного свойства, оцениваются по формуле $\tilde{\xi}_i = \alpha_{ij}\xi_j + \beta_{ij}$.

Примерами подобной задачи могут служить: определение концентраций радиоактивных элементов по полевым измерениям общей радиоактивности; количества ферромагнитных минералов по намагниченности насыщения; плотности горных пород по скорости упругих продольных волн; содержания кремнекислоты по удельному весу эффузивных пород; глинистости осадочных пород по кажущемуся сопротивлению и т. д.

Как показывает формула (3.46), дисперсия $\mathbf{D}\Delta_{ij}$ величины Δ_{ij} будет иметь смысл дисперсии ошибки такого прогноза значения ξ_i по известному наблюдению ξ_j. В соответствии с (3.48″), чем больше по абсолютной величине коэффициент корреляции, тем меньше дисперсия $\mathbf{D}\Delta_{ij}$ ошибки косвенных измерений величины ξ_i по сравнению с ее собственной дисперсией σ_i^2, тем, следовательно, точнее косвенные измерения. В предельном случае, при $|r_{ij}| = 1$, косвенные измерения безошибочны.

Коэффициент корреляции можно использовать для вычисления дисперсий сумм или разностей зависимых величин. В отличие от (1.51) и (1.54), дисперсия суммы или разности двух таких величин имеет вид

$$\mathbf{D}(\xi_i \pm \xi_j) = \mathbf{M}[\xi_i - M_i \pm (\xi_j - M_j)]^2 = \mathbf{M}(\xi_i - M_i)^2 \pm 2\mathbf{M}[(\xi_i - M_i) \times (\xi_j - M_j)] + \mathbf{M}(\xi_j - \mathbf{M}_j)^2 = \sigma_i^2 \pm 2r_{ij}\sigma_i\sigma_j + \sigma_j^2 = \mathbf{D}\xi_i \pm 2r_{ij}\sqrt{\mathbf{D}\xi_i\mathbf{D}\xi_j} + \mathbf{D}\xi_j. \quad (3.52)$$

Вообще, дисперсия линейной комбинации n величин,

$$\mathbf{D}\left(\sum_{j=1}^{n}\alpha_j\xi_j\right)=\sum_{j=1}^{n}\alpha_j^2\mathbf{D}\xi_j+2\sum_{i<j}^{n}\alpha_i\alpha_j r_{ij}\sqrt{\mathbf{D}\xi_i\,\mathbf{D}\xi_j}. \qquad (3.52')$$

§ 4. Многомерное нормальное распределение

Основным объектом изучения математической статистики многомерных случайных величин является многомерное нормальное распределение. Для него разработан разветвленный математический аппарат.

Плотность многомерного нормального распределения. Нормальному m-мерному распределению подчиняются m-мерные случайные величины, плотность многомерного распределения которых имеет вид:

$$p(\mathbf{x})=p(x_1,\ x_2,\ \ldots,\ x_n)=$$
$$=(2\pi)^{-\frac{m}{2}}|\mathbf{B}|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{m})'\mathbf{B}^{-1}(\mathbf{x}-\mathbf{m})\right], \qquad (3.53)$$

где $|\mathbf{B}|$ — определитель ковариационной матрицы $\mathbf{B}$; $\mathbf{m}=\{M_1,\ M_2,\ \ldots,\ M_m\}'$ — вектор математических ожиданий; $\mathbf{x}$ — вектор-столбец независимых переменных $\{x_1,\ x_2,\ \ldots,\ x_m\}$; $\mathbf{B}^{-1}$ — матрица, обратная $\mathbf{B}$.

Как показывает формула (3.53), плотность многомерного нормального распределения полностью определяется вектором средних $\mathbf{m}$ и ковариационной матрицей $\mathbf{B}$. Каждая компонента ξ_i многомерной нормально распределенной случайной величины $\boldsymbol{\xi}$ распределена нормально*.

В двумерном случае $\boldsymbol{\xi}=\{\xi_1,\ \xi_2\}'$ и

$$|\mathbf{B}|=\begin{vmatrix}b_{11} & b_{12}\\ b_{21} & b_{22}\end{vmatrix}=b_{11}b_{22}-b_{12}^2=\sigma_1^2\sigma_2^2(1-r^2),$$

где σ_1, σ_2 — средние квадратические отклонения компонент ξ_1, ξ_2; r — коэффициент корреляции ξ_1 и ξ_2. Квадратичную форму $T(\mathbf{x},\ \mathbf{x})=(\mathbf{x}-\mathbf{m})'\,\mathbf{B}^{-1}(\mathbf{x}-\mathbf{m})$ можно представить в виде $T(\mathbf{x},\ \mathbf{x})=(\mathbf{x}-\mathbf{m})'\,\mathbf{y}$, где $\mathbf{y}=\{y,\ y_2\}'=\mathbf{B}^{-1}(\mathbf{x}-\mathbf{m})$ — решение системы уравнений $\mathbf{B}\mathbf{y}=\mathbf{x}-\mathbf{m}$, т. е.

$$\begin{cases}b_{11}y_1+b_{12}y_2=x_1-M_1,\\ b_{12}y_1+b_{22}y_2=x_2-M_2,\end{cases}$$

---

* Обратное утверждение в общем случае неверно, хотя для независимых компонент оно справедливо. Можно привести примеры, когда многомерное распределение не является нормальным, хотя каждая компонента в отдельности распределена нормально.

(по (3.41) $b_{12} = b_{21}$). Решив систему, получим:

$$y_1 = \frac{b_{22}(x_1 - M_1) - b_{12}(x_2 - M_2)}{b_{11}b_{22} - b_{12}^2}, \quad y_2 = \frac{b_{11}(x_2 - M_2) - b_{12}(x_1 - M_1)}{b_{11}b_{22} - b_{12}^2};$$

$$T(\mathbf{x}, \mathbf{x}) = (x_1 - M_1)y_1 + (x_2 - M_2)y_2 =$$

$$= \frac{1}{b_{11}b_{22} - b_{12}^2}\{b_{22}(x_1 - M_1)^2 - 2b_{12}(x_1 - M_1)(x_2 - M_2) + b_{11}(x_2 - M_2)^2\} =$$

$$= \frac{1}{1 - r^2}\left\{\frac{(x_1 - M_1)^2}{\sigma_1^2} - 2r\frac{(x_1 - M_1)(x_2 - M_2)}{\sigma_1\sigma_2} + \frac{(x_2 - M_2)^2}{\sigma_2^2}\right\}.$$

Таким образом, плотность двухкомпонентного (двумерного) нормального распределения представляется в виде

$$p(x, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1 - r^2}} \exp\left\{-\frac{1}{2(1 - r^2)}\left[\frac{(x_1 - M_1)^2}{\sigma_1^2} - \right.\right.$$

$$\left.\left. - 2r\frac{(x_1 - M_1)(x_2 - M_2)}{\sigma_1\sigma_2} + \frac{(x_2 - M_2)^2}{\sigma_2^2}\right]\right\}. \qquad (3.54)$$

Форма поверхности $z = p(x_1, x_2)$, уравнение которой имеет вид (3.54), показана в изолиниях на рис. 23. Ее горизонтальные сечения, как видно из уравнения, представляют собой эллипсы, а максимум находится в точке с координатами (M_1, M_2).

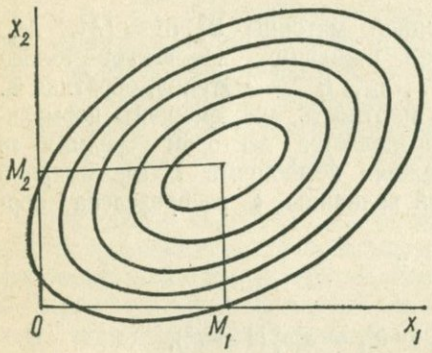


Рис. 23.

Можно доказать такое свойство: если ξ подчиняется многомерному нормальному закону распределения, то линейное преобразование $\eta = \mathbf{A}\xi$ ($\mathbf{A}$ — матрица преобразования) также следует нормальному закону распределения с математическим ожиданием $\mathbf{M}\eta = \mathbf{A}\mathbf{M}\xi$ и ковариационной матрицей $\mathbf{M}[(\eta - \mathbf{M}\eta)(\eta - \mathbf{M}\eta)'] = \mathbf{A}\mathbf{B}\mathbf{A}'$ ($\mathbf{B}$ — ковариационная матрица ξ). Свойство справедливо как для вырожденных, так и для невырожденных преобразований; для последних оно следует из (3.32′). Из него следует также, что линейная комбинация одномерных нормально распределенных случайных величин распределена нормально, если они зависимы и являются компонентами многомерной нормально распределенной случайной величины.

Сумма двух нормально распределенных многомерных случайных величин также распределена нормально.

Равенство нулю коэффициента корреляции двух нормально распределенных случайных величин — компонент многомерной нормально распределенной случайной величины — свидетельствует о их независимости. Это следует из вида плотности многомерного нормального распределе-

ния (3.53): если все компоненты нормально распределенного вектора ξ некоррелированы, т. е. $r_{ij} = b_{ij} = 0$ при $i \neq j$, то его ковариационная матрица $\mathbf{B}$ становится диагональной. Тогда плотность запишется в виде произведения плотностей распределения отдельных компонент (3.27):

$$p(\mathbf{x}) = (2\pi)^{-\frac{m}{2}} \prod_{l=1}^{m} \frac{1}{\sigma_l} \exp\left\{-\frac{1}{2}\frac{(x_l - M_l)^2}{\sigma_l^2}\right\} =$$

$$= \prod_{l=1}^{m} \frac{1}{\sqrt{2\pi}\,\sigma_l} \exp\left\{-\frac{1}{2}\frac{(x_l - M_l)^2}{\sigma_l^2}\right\}. \qquad (3.55)$$

Связь двух зависимых случайных величин, двумерное распределение которых нормально, выражается линейной регрессией. Это следует из того, что условную плотность $p_1(x_1/x_2)$ распределения ξ_1 при $\xi_2 = x_2$ можно представить, используя (3.54), в виде

$$p_1(x_1/x_2) = \frac{p(x_1, x_2)}{p(x_2)} = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \exp\left\{-\frac{1}{2(1-r^2)}\left[\frac{(x_1 - M_1)^2}{\sigma_1^2} - \right.\right.$$

$$\left.\left. - 2r\frac{(x_1 - M_1)(x_2 - M_2)}{\sigma_1\sigma_2} + \frac{(x_2 - M_2)^2}{\sigma_2^2}\right]\right\} \Big/ \left(\frac{1}{\sqrt{2\pi}\,\sigma_2} \exp\left\{-\frac{1}{2}\frac{(x_2 - M_2)^2}{\sigma_2^2}\right\}\right) =$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma_1\sqrt{1-r^2}} \exp\left\{\frac{\left[x_1 - r\frac{\sigma_1}{\sigma_2}(x_2 - M_2) - M_1\right]^2}{-2\sigma_1^2(1-r^2)}\right\},$$

так что при условии $\xi_2 = x_2$ величина ξ_1 распределена нормально с математическим ожиданием $\mathbf{M}(\xi_1/\xi_2 = x_2) = r\frac{\sigma_1}{\sigma_2}(x_2 - M_2) + M_1$ и дисперсией $\mathbf{D}(\xi_1/\xi_2 = x_2) = \sigma_1^2(1 - r^2)$. Иными словами, связь представляется в виде (3.46) с коэффициентами (3.47): $\xi_1 = \alpha_{12}\xi_2 + \beta_{12} + \Delta_{12}$, причем $\mathbf{M}\,\Delta_{12} = 0$. Независимость Δ_{12} и ξ_2 следует из их некоррелированности — $\mathbf{M}[\Delta_{12}(\xi_2 - M_2)] = \mathbf{M}[(\xi_1 - \alpha_{12}\xi_2 - \beta_{12})(\xi_2 - M_2)] = r\sigma_1\sigma_2 - \alpha_{12}\sigma_2^2 = 0$ и их совместного нормального распределения.

Множественная связь. *Множественной* называется связь случайной величины с группой нескольких случайных величин. Если эта случайная величина ξ_{k+1} и группа связываемых с нею $\xi_1, \xi_2, \ldots, \xi_k$ нормальны, образуя $(k+1)$-мерное нормальное распределение, множественная связь линейна и представляется в форме, аналогичной (3.46):

$$\xi_{k+1} = \sum_{j=1}^{k} \alpha_j \xi_j + \alpha_{k+1} + \Delta_{k+1}, \qquad (3.56)$$

где α_j — коэффициенты, Δ_{k+1} — нормально распределенная случайная величина с нулевым математическим ожиданием, не зависящая от $\xi_1, \xi_2, \ldots, \xi_k$. Как и в рассмотренной выше связи двух случайных

величин, Δ_{k+1} — взятое со знаком отклонение ξ_{k+1} от соответствующего значения функции

$$\tilde{\xi}_{k+1} = \sum_{j=1}^{k} \alpha_j \xi_j + \alpha_{k+1}, \tag{3.56'}$$

называемой функцией регрессии ξ_{k+1} на $\xi_1, \xi_2, \ldots, \xi_k$. Дисперсия отклонения $\Delta_{k+1} = \xi_{k+1} - \tilde{\xi}_{k+1}$ определяет точность косвенных измерений величины ξ_{k+1} с помощью функции (3.56').

Выясним вид коэффициентов в уравнении (3.56). Взяв математическое ожидание от обеих частей (3.56), получим: $M_{k+1} = \sum_{j=1}^{k} \alpha_j M_j + \alpha_{k+1}$ (обозначив $M_j = \mathbf{M}\xi_j$, $j = \overline{1, k+1}$), откуда

$$\alpha_{k+1} = M_{k+1} - \sum_{j=1}^{k} \alpha_j M_j. \tag{3.57}$$

Используя это соотношение, зависимость (3.56) можно записать в виде

$$\xi_{k+1} - M_{k+1} = \sum_{j=1}^{k} \alpha_j (\xi_j - M_j) + \Delta_{k+1}. \tag{3.56''}$$

Введем коэффициенты β_j вида

$$\beta_j = \frac{\alpha_j \sigma_j}{\sigma_{k+1}} \quad (j = \overline{1, k}). \tag{3.58}$$

Тогда (3.56'') примет форму

$$\frac{\xi_{k+1} - M_{k+1}}{\sigma_{k+1}} = \sum_{j=1}^{k} \beta_j \frac{\xi_j - M_j}{\sigma_j} + \delta_{k+1}, \quad \delta_{k+1} = \frac{\Delta_{k+1}}{\sigma_{k+1}}. \tag{3.56'''}$$

Умножив это равенство последовательно на $\frac{\xi_1 - M_1}{\sigma_1}$, $\frac{\xi_2 - M_2}{\sigma_2}$, $\ldots$, $\frac{\xi_k - M_k}{\sigma_k}$ и взяв математическое ожидание от обеих частей каждого полученного равенства, получим систему уравнений для вычисления β_j

$$\begin{cases} r_{1k+1} = \sum_{j=1}^{k} \beta_j r_{1j}, \\ r_{2k+1} = \sum_{j=1}^{k} \beta_j r_{2j}, \\ \ldots\ldots\ldots \\ r_{kk+1} = \sum_{j=1}^{k} \beta_j r_{kj}, \end{cases} \tag{3.59}$$

или в векторно-матричной форме

$$\mathbf{R}\beta = \mathbf{r}_{k+1}, \tag{3.59'}$$

где $\beta = \{\beta_1, \beta_2, \ldots, \beta_k\}'$; $\mathbf{r}_{k+1} = \{r_{1k+1}, r_{2k+1}, \ldots, r_{kk+1}\}'$ — вектор коэффициентов корреляции ξ_{k+1} и $\xi_1, \xi_2, \ldots, \xi_k$; $\mathbf{R} = \{r_{ij}\}_{i,\,j=1}^{k}$ — корреляционная матрица $\xi_1, \xi_2, \ldots, \xi_k$; r_{ij} — коэффициенты корреляции ξ_i и ξ_j $(i, j = \overline{1, k})$. Решение этой системы уравнений

$$\beta = \mathbf{R}^{-1}\mathbf{r}_{k+1} \tag{3.60}$$

дает коэффициенты $\beta_1, \beta_2, \ldots, \beta_k$, а с использованием (3.58) и коэффициенты α_j:

$$\alpha_j = \frac{\beta_j \sigma_{k+1}}{\sigma_j} \quad (j = \overline{1, k}). \tag{3.61}$$

Найденные коэффициенты α_j обладают свойством: среди всех линейных комбинаций случайных величин $\xi_1, \xi_2, \ldots, \xi_k$ вида $\eta_{k+1} = \sum_{j=1}^{k} a_j\xi_j + a_{k+1}$ наилучший прогноз значений ξ_{k+1} по известным величинам $\xi_1, \xi_2, \ldots, \xi_k$ достигается при $a_j = \alpha_j$ $(j = \overline{1, k+1})$. Иными словами, математическое ожидание квадрата отклонения ξ_{k+1} от η_{k+1} обращается в минимум при $a_j = \alpha_j$ $(j = \overline{1, k+1})$.

Это вытекает из следующего. В случае минимума формы $S = \mathbf{M}\left(\xi_{k+1} - \sum_{j=1}^{k} a_j\xi_j - a_{k+1}\right)^2$ должно выполняться условие

$$\frac{\partial S}{\partial a_{k+1}} = \frac{\partial}{\partial a_{k+1}}\left[\mathbf{M}\left(\xi_{k+1} - \sum_{j=1}^{k} a_j\xi_j\right)^2 - 2\mathbf{M}\left(\xi_{k+1} - \sum_{j=1}^{k} a_j\xi_j\right)a_{k+1} + a_{k+1}^2\right] =$$

$$= \frac{\partial}{\partial a_{k+1}}\left[-2a_{k+1}\left(M_{k+1} - \sum_{j=1}^{k} M_j a_j\right) + a_{k+1}^2\right] =$$

$$= -2M_{k+1} + 2\sum_{j=1}^{k} M_j a_j + 2a_{k+1} = 0,$$

откуда $a_{k+1} = M_{k+1} - \sum_{j=1}^{k} M_j a_j$. Вводя теперь, по аналогии с (3.58), коэффициенты $b_j = a_j\sigma_j/\sigma_{k+1}$ $(j = \overline{1, k})$ и используя полученное равенство, преобразуем форму S к виду

$$S = \sigma_{k+1}^2\,\mathbf{M}\left(\frac{\xi_{k+1} - M_{k+1}}{\sigma_{k+1}} - \sum_{j=1}^{k} b_j\frac{\xi_j - M_j}{\sigma_j}\right)^2 =$$

$$= \sigma_{k+1}^2\left(1 + \sum_{j=1}^{k} b_j^2 - 2\sum_{j=1}^{k} b_j r_{jk+1} + \sum_{j,\,l;\,j\neq l} b_j b_l r_{jl}\right).$$

Взяв частные производные по $b_1, b_2, \ldots, b_k$ и приравняв их нулю (необходимое условие минимума S), получим для b_j ту же систему уравнений (3.59), что и для β_j. Это и доказывает указанное свойство, которое, как видно из приведенных рассуждений, справедливо и для

случайных величин, отклоняющихся от нормального закона распределения. Оно служит естественным обоснованием решения задачи косвенных измерений с помощью уравнения множественной регрессии (3.56′) при статистическом анализе геолого-геофизических данных.

Множественный коэффициент корреляции. Определение этой характеристики силы множественной связи основывается на рассмотренном экстремальном свойстве коэффициентов α_j. *Множественным коэффициентом корреляции* случайной величины ξ_{k+1} и группы других $\xi_1, \xi_2, \ldots, \xi_k$ называют коэффициент R_{k+1} парной корреляции ξ_{k+1} и такой линейной комбинации $\tilde{\eta}_{k+1}$ величин $\xi_1, \xi_2, \ldots, \xi_k$, при которой дисперсия отклонения ξ_{k+1} от $\tilde{\eta}_{k+1}$ минимальна*. Так как эта линейная комбинация имеет вид (3.56′), где α_j определяются по (3.57), (3.61), нетрудно получить выражение R_{k+1} в явном виде.

По свойству (3.49) коэффициент корреляции не изменится, если вместо ξ_{k+1} и $\tilde{\xi}_{k+1} = \sum_{j=1}^{k} \alpha_j \xi_j + \alpha_{k+1}$ взять, соответственно, $(\xi_{k+1} - M_{k+1})/\sigma_{k+1}$ и $\tilde{\xi}_{k+1}/\sigma_{k+1} - M_{k+1}/\sigma_{k+1}$. Ковариация этих величин

$$B_{k+1} = \mathbf{M}\left(\frac{\xi_{k+1} - M_{k+1}}{\sigma_{k+1}} \sum_{j=1}^{k} \beta_j \frac{\xi_j - M_j}{\sigma_j}\right) = \sum_{j=1}^{k} \beta_j r_{jk+1} = \boldsymbol{\beta}' \mathbf{r}_{k+1};$$

дисперсия первой равна единице, а второй — $\mathbf{M}\left[\sum_{j=1}^{k} \beta_j \sigma_j^{-1} (\xi_j - M_j)\right]^2 =$

$= \sum_{i,j=1}^{k} \beta_i r_{ij} \beta_j = \boldsymbol{\beta}' \mathbf{R} \boldsymbol{\beta} = \boldsymbol{\beta}' \mathbf{r}_{k+1}$, учитывая (3.59′). Итак,

$$R_{k+1} = \frac{B_{k+1}}{\sqrt{\boldsymbol{\beta}' \mathbf{r}_{k+1}}} = \sqrt{\boldsymbol{\beta}' \mathbf{r}_{k+1}} = \sqrt{\sum_{j=1}^{k} \beta_j r_{jk+1}}. \qquad (3.62)$$

Дисперсия случайной добавки Δ_{k+1}, определяющей ошибку косвенного измерения, учитывая (3.56‴),

$$\mathbf{D}\,\Delta_{k+1} = \sigma_{k+1}^2 \mathbf{M}\left(\frac{\xi_{k+1} - M_{k+1}}{\sigma_{k+1}} - \sum_{j=1}^{k} \beta_j \frac{\xi_j - M_j}{\sigma_j}\right)^2 =$$

$$= \sigma_{k+1}^2 \left[1 - 2\sum_{j=1}^{k} r_{jk+1}\beta_j + \mathbf{M}\left(\sum_{j=1}^{k} \beta_j \frac{\xi_j - M_j}{\sigma_j}\right)^2\right] =$$

$$= \sigma_{k+1}^2 (1 - 2\boldsymbol{\beta}' \mathbf{r}_{k+1} + \boldsymbol{\beta}' \mathbf{r}_{k+1}) = \sigma_{k+1}^2 (1 - R_{k+1}^2).$$

Таким образом, между коэффициентом множественной корреляции и средним квадратическим отклонением ошибки Δ_{k+1} косвенного изме-

---

* С учетом (3.48′) множественный коэффициент корреляции можно определить как максимальный среди коэффициентов парной корреляции величины ξ_{k+1} и линейных комбинаций величин $\xi_1, \xi_2, \ldots, \xi_k$. Ясно, что R_{k+1} будет не меньше, чем модуль любого из коэффициентов парной корреляции r_{ik+1} $(i = \overline{1, k})$, а множественный коэффициент корреляции величины ξ_{k+1} и любой группы из величин $\xi_1, \xi_2, \ldots, \xi_k$ не будет превышать R_{k+1}.

рения устанавливается соотношение, аналогичное (3.48″) для случая парной корреляции:

$$R_{k+1} = \sqrt{1 - \frac{\mathbf{D}\,\Delta_{k+1}}{\sigma^2_{k+1}}}. \qquad (3.63)$$

Эта формула наглядно иллюстрирует смысл коэффициента множественной корреляции как количественной характеристики силы линейной связи между случайной величиной и группой других случайных величин. Следует подчеркнуть, что приведенные формулы справедливы и для случайных величин, отклоняющихся от нормального закона, но связанных линейной зависимостью вида (3.56). Равенство нулю множественного коэффициента корреляции R_{k+1} величин, распределение которых в совокупности нормально, свидетельствует об отсутствии связи между ξ_{k+1} и любой величиной из группы $\xi_1, \xi_2, \ldots, \xi_k$. Обратное справедливо для любых распределений: если ξ_{k+1} не зависит от $\xi_1, \xi_2, \ldots, \xi_k$, то, как следует из (3.62), множественный коэффициент корреляции равен нулю.

Частный коэффициент корреляции. При статистическом анализе геолого-геофизических данных иногда возникает необходимость оценки собственной связи двух показателей, при исключенном влиянии группы других. Коэффициент корреляции двух случайных величин ξ_i и ξ_j характеризует силу линейной связи, существующей не только за счет их собственной зависимости, но и за счет других компонент, с которыми может быть связана каждая из этих величин. В этой ситуации, очевидно, следует анализировать условное распределение ξ_i и ξ_j при фиксированных значениях тех компонент $\xi_1, \xi_2, \ldots, \xi_k$, влияние которых необходимо исключить, т. е. положить $\xi_s = x_s$, $s = \overline{1, k}$. Сила этой связи характеризуется коэффициентом корреляции ξ_i и ξ_j при фиксированных значениях $\xi_s = x_s$ ($s = \overline{1, k}$), который и называется *частным коэффициентом корреляции*. Обозначим его $r^{(k)}_{ij}$, полагая i и j номерами коррелируемых величин ξ_i, ξ_j; $\xi_1, \xi_2, \ldots, \xi_k$ — величины, влияние которых необходимо исключить. Для простоты мы считаем этими компонентами именно $\xi_1, \xi_2, \ldots, \xi_k$, так как такое допущение не ограничивает общности дальнейших рассуждений.

Запишем уравнения связей ξ_i и ξ_j с $\xi_1, \xi_2, \ldots, \xi_k$ в форме (3.56‴).

$$\frac{\xi_i - M_i}{\sigma_i} = \sum_{s=1}^{k} \beta'_s \frac{\xi_s - M_s}{\sigma_s} + \delta', \quad \frac{\xi_j - M_j}{\sigma_j} = \sum_{s=1}^{k} \beta''_s \frac{\xi_s - M_s}{\sigma_s} + \delta'', \qquad (3.64)$$

где δ', δ'' — случайные величины с нулевыми средними, не зависящие от $\xi_1, \xi_2, \ldots, \xi_k$; $\{\beta'_1, \beta'_2, \ldots, \beta'_k\}' = \beta_i$, $\{\beta''_1, \beta''_2, \ldots, \beta''_k\}' = \beta_j$ — решения систем уравнений $\mathbf{R}\beta_i = \mathbf{r}_i$, $\mathbf{R}\beta_j = \mathbf{r}_j$, т. е. $\beta_i = \mathbf{R}^{-1}\mathbf{r}_i$, $\beta_j = \mathbf{R}^{-1}\mathbf{r}_j$; $\mathbf{R}$ — корреляционная матрица величин $\xi_1, \xi_2, \ldots, \xi_k$; $\mathbf{r}_i = \{r_{i1}, r_{i2}, \ldots, r_{ik}\}'$, $\mathbf{r}_j = \{r_{j1}, r_{j2}, \ldots, r_{jk}\}'$; r_{st} — коэффициенты кор-

реляции ξ_s и ξ_t. Положив в (3.64) ξ фиксированными, $\xi_s = x_s$, $s = \overline{1, k}$, получим при этом условии:

$$\left.\frac{\xi_i - M_i}{\sigma_i}\right|_{\substack{\xi_s = x_s, \\ s=\overline{1,k}}} = \sum_{s=1}^{k} \beta_s' \frac{x_s - M_s}{\sigma_s} + \delta', \quad \left.\frac{\xi_j - M_j}{\sigma_j}\right|_{\substack{\xi_s = x_s, \\ s=\overline{1,k}}} = \sum_{s=1}^{k} \beta_s'' \frac{x_s - M_s}{\sigma_s} + \delta''. \tag{3.65}$$

Условные математические ожидания величин ξ_i и ξ_j составят

$$M_i^{(k)} = \sum_{s=1}^{k} \beta_s' \frac{\sigma_i}{\sigma_s}(x_s - M_s) + M_i, \quad M_j^{(k)} = \sum_{s=1}^{k} \beta_s'' \frac{\sigma_j}{\sigma_s}(x_s - M_s) + M_j. \tag{3.65'}$$

Используя свойство (3.49), будем вычислять частный коэффициент корреляции в виде $r_{ij}^{(k)} = \frac{b_{ij}^{(k)}}{\sigma_i^{(k)}\sigma_j^{(k)}}$, где $b_{ij}^{(k)}$ — условная ковариация величин (3.64), $\sigma_i^{(k)}$, $\sigma_j^{(k)}$ — их условные средние квадратические отклонения. Используя (3.65), имеем

$$b_{ij}^{(k)} = \mathbf{M}\left\{\left[\frac{\xi_i - M_i}{\sigma_i} - \sum_{s=1}^{k} \beta_s' \frac{x_s - M_s}{\sigma_s}\right]\left[\frac{\xi_j - M_j}{\sigma_j} - \sum_{s=1}^{k} \beta_s'' \frac{x_s - M_s}{\sigma_s}\right]\right\} = \mathbf{M}(\delta'\delta''),$$

$$\sigma_i^{(k)} = \sqrt{\mathbf{M}(\delta')^2}, \quad \sigma_j^{(k)} = \sqrt{\mathbf{M}(\delta'')^2}.$$

Из (3.64)

$$\delta' = \frac{\xi_i - M_i}{\sigma_i} - \sum_{s=1}^{k} \beta_s' \frac{\xi_s - M_s}{\sigma_s}, \quad \delta'' = \frac{\xi_j - M_j}{\sigma_j} - \sum_{s=1}^{k} \beta_s'' \frac{\xi_s - M_s}{\sigma_s},$$

следовательно, учитывая, что $\mathrm{R}\beta_j = \mathrm{r}_j$, $\mathrm{R}\beta_i = \mathrm{r}_i$,

$$(\sigma_i^{(k)})^2 = \mathbf{M}(\delta')^2 = 1 - \sum_{t,s=1}^{k} \beta_s' r_{st} \beta_t' = 1 - \beta_i' \mathrm{R}\beta_i = 1 - \beta_i' \mathrm{r}_i =$$

$$= 1 - \sum_{s=1}^{k} \beta_s' r_{si},$$

$$(\sigma_j^{(k)})^2 = \mathbf{M}(\delta'')^2 = 1 - \sum_{t,s=1}^{k} \beta_s'' r_{st} \beta_t'' = 1 - \beta_j' \mathrm{R}\beta_j = 1 - \beta_j' \mathrm{r}_j =$$

$$= 1 - \sum_{s=1}^{k} \beta_s'' r_{sj},$$

$$b_{ij}^{(k)} = r_{ij} - \sum_{s=1}^{k} \beta_s'' r_{si} - \sum_{s=1}^{k} \beta_s' r_{sj} + \sum_{s,t=1}^{k} \beta_s' r_{st} \beta_t'' = r_{ij} - \beta_j' \mathrm{r}_i - \beta_i' \mathrm{r}_j +$$

$$+ \beta_i' \mathrm{R}\beta_j = r_{ij} - \sum_{s=1}^{k} \beta_s' r_{js} = r_{ij} - \sum_{s=1}^{k} \beta_s'' r_{is}.$$

Итак, частный коэффициент корреляции ξ_i и ξ_j при фиксированных $\xi_1, \xi_2, \ldots, \xi_k$ вычисляется в виде

$$r_{ij}^{(k)} = \frac{r_{ij} - \sum_{s=1}^{k} \beta_s' r_{js}}{\sqrt{\left(1 - \sum_{s=1}^{k} \beta_s' r_{si}\right)\left(1 - \sum_{s=1}^{k} \beta_s'' r_{sj}\right)}} = \frac{r_{ij} - \sum_{s=1}^{k} \beta_s'' r_{is}}{\sqrt{\left(1 - \sum_{s=1}^{k} \beta_s' r_{si}\right)\left(1 - \sum_{s=1}^{k} \beta_s'' r_{sj}\right)}}, \tag{3.66}$$

где r_{ij} — коэффициент корреляции ξ_i и ξ_j; r_{is}, r_{js} $(s = \overline{1, k})$ — коэффициенты корреляции соответственно ξ_i и $\xi_1, \xi_2, \ldots, \xi_k$, ξ_j и $\xi_1, \xi_2, \ldots, \xi_k$; β_s', β_s'' — коэффициенты множественных регрессий нормированных уклонений $\frac{\xi_i - M_i}{\sigma_i}$ и $\frac{\xi_j - M_j}{\sigma_j}$ соответственно на нормированные же уклонения $\frac{\xi_s - M_s}{\sigma_s}$ $(s = \overline{1, k})$.

Нетрудно выразить частный коэффициент корреляции через коэффициенты α_s', α_s'' множественной регрессии непосредственно величин ξ_i, ξ_j на $\xi_1, \xi_2, \ldots, \xi_k$. Для этого, согласно (3.58), в формулу (3.66) следует подставить

$$\beta_s' = \frac{\alpha_s' \sigma_s}{\sigma_i}, \quad \beta_s'' = \frac{\alpha_s'' \sigma_s}{\sigma_j}.$$

Глава 4

ОЦЕНКА ПАРАМЕТРОВ И ЧИСЛОВЫХ ХАРАКТЕРИСТИК РАСПРЕДЕЛЕНИЙ

В предыдущих главах изложены основные понятия и другие необходимые сведения из теории вероятностей. Теперь познакомимся с теми вопросами математической статистики, в форме которых будут формулироваться и решаться задачи количественного анализа геолого-геофизических данных. Как и в предыдущих главах, ограничимся кратким изложением необходимых сведений. Читатель, желающий получить более подробную информацию, может найти ее в литературе по методам математической статистики [2, 4, 9, 14, 22].

§ 1. Понятие статистической оценки. Свойства оценок

При количественном анализе данных измерений любых показателей, как правило, имеют дело с *выборками наблюдений* тех или иных случайных величин. Множество всех возможных значений случайной величины, в котором распределение признака совпадает с ее распределением, рассматривают как *генеральную совокупность*, из которой извлекаются

выборки. Генеральная совокупность служит моделью соответствующего действительного множества. В выборки группируются наблюдения по координатам пунктов отбора проб, типам пород, методам измерений, показателям и т. д. При этом обычно нет точной информации о распределениях изучаемых случайных величин. Её дают функции и плотности распределения, а также различные числовые характеристики распределений. Поэтому одна из задач статистического анализа состоит в оценке функции или плотности распределения по выборке наблюдений. Методы решения этой задачи основываются на том, что наблюдения случайной величины определенным образом отражают закон распределения. Другая необходимая задача — оценка числовых характеристик распределений.

Оценки. Иногда имеются основания для того, чтобы заранее указать закон распределения, которому подчиняется случайная величина. Иными словами, известно параметрическое семейство $F(x, \alpha_1, \alpha_2, \ldots, \alpha_k)$ функций, к которому принадлежит функция распределения данной случайной величины ξ. В этом случае задача оценки функции распределения сводится к отысканию *оценок параметров* $\alpha_1, \alpha_2, \ldots, \alpha_k$ по выборке наблюдений $x_1, x_2, \ldots, x_n$ величины ξ, по возможности наиболее точным способом. Под оценкой параметра или числовой характеристики α понимают величину, являющуюся функцией наблюдений:

$$\check{\alpha} = f(x_1, x_2, \ldots, x_n), \tag{4.1}$$

для которой гарантируется близость (в определенном смысле) к оцениваемому параметру. Подставив в $F(x, \alpha_1, \alpha_2, \ldots, \alpha_k)$ оценки параметров $\check{\alpha}_1, \check{\alpha}_2, \ldots, \check{\alpha}_k$, получим оценку функции распределения. Задача оценки параметров обычно имеет и самостоятельное значение при описании распределений с помощью числовых характеристик.

Статистикой называется любая функция наблюдений; таким образом, оценка — это построенная по определенному правилу статистика. Будучи функцией наблюдений случайной величины, она является также случайной величиной, для которой можно определить функцию распределения, математическое ожидание, дисперсию и вообще употребить весь математический аппарат изучения случайных величин.

Попутно с оценкой параметров и числовых характеристик распределений возникает задача определения точности их оценки. Она решается в математической статистике построением *доверительных интервалов* для оцениваемых параметров, которое сводится к вычислению *доверительных пределов (границ)*, ограничивающих эти интервалы. Доверительный интервал $[\alpha_q^-, \alpha_q^+]$ для неизвестного параметра α обладает свойством: вероятность того, что α окажется в доверительных пределах α_q^-, α_q^+, не меньше заданной близкой к единице величины q, называемой *доверительной вероятностью* (коэффициентом доверия, или уровнем) доверительного интервала. Обычно ее полагают равной 0,9, 0,95, иногда 0,99. Таким образом,

$$\mathbf{P}\{\alpha_q^- \leqslant \alpha \leqslant \alpha_q^+\} \geqslant q, \quad 0{,}9 \leqslant q < 1. \tag{4.2}$$

Состоятельность и несмещенность оценок. Для оценок определен ряд характеризующих их свойств. Оценка $\check{a}_n$ некоторого параметра a, вычисляемая по независимым наблюдениям $x_1, x_2, \ldots, x_n$ случайной величины ξ, *состоятельна*, если

$$\mathbf{P}\{|\check{a}_n - a| > \varepsilon\} \to 0 \text{ при } n \to \infty \tag{4.3}$$

для произвольного малого $\varepsilon > 0$. Состоятельность оценки свидетельствует о том, что с увеличением числа независимых наблюдений она приближается (сходится по вероятности) к оцениваемой величине.

Другое важное желательное свойство оценок — *несмещенность*. Оценка $\check{a}_n$ является *несмещенной*, если ее математическое ожидание равно оцениваемой ею величине:

$$\mathbf{M}\check{a}_n = a \tag{4.4}$$

и *асимптотически несмещенной*, если $\lim\limits_{n\to\infty} \mathbf{M}\check{a}_n = a$.

Эффективность оценки. Для одного и того же параметра можно предложить много несмещенных и состоятельных оценок. Вопрос о том, каким из них отдать предпочтение, решают, сравнивая их дисперсии. Наиболее точными надо считать оценки, обладающие минимальными дисперсиями. В частности, несмещенность оценки обеспечивает ей определенное преимущество в этом смысле по сравнению со смещенными. Если $\check{a}_n$ и $\check{a}_n$ — две оценки параметра a с одинаковыми дисперсиями, причем первая несмещенная $\mathbf{M}\check{a}_n = a$, а вторая смещенная $\mathbf{M}\check{a}_n = a + b \neq a$, то математическое ожидание квадрата отклонения $\check{a}_n$ от a будет превышать эту же величину для $\check{a}_n$ — дисперсию $\mathbf{D}\check{a}_n$:

$$\mathbf{M}(\check{a}_n - a)^2 = \mathbf{M}(\check{a}_n - a + a - a)^2 = \mathbf{M}(\check{a}_n - a)^2 + (a - a)^2 = \\ = \mathbf{D}\check{a}_n + (a - a)^2 = \mathbf{D}\check{a}_n + (a - a)^2 > \mathbf{D}\check{a}_n.$$

В математической статистике доказано важнейшее положение, известное под названием *неравенства Крамера — Рао:* если плотность распределения некоторой случайной величины ξ имеет вид $p(x, a)$, то дисперсия произвольной несмещенной оценки $\check{a}_n$, полученной по n независимым наблюдениям ξ, не может быть меньше, чем* $\varphi_n = \left[n\,\mathbf{M}\left(\frac{\partial \ln p(\xi, a)}{\partial a}\right)^2\right]^{-1}$:

$$\mathbf{D}\check{a}_n \geqslant \varphi_n. \tag{4.5}$$

---

* Неравенство Крамера — Рао обобщается на тот случай, когда плотность распределения зависит от нескольких параметров: $p(x) = p(x, a_1, a_2, \ldots, a_k)$. Если $\check{a}_{1n}, \check{a}_{2n}, \ldots, \check{a}_{kn}$ — несмещенные оценки параметров $a_1, a_2, \ldots, a_k$, вычисленные по выборке n независимых наблюдений, $\mathbf{B}_n$ — ковариационная матрица оценок — $\mathbf{B}_n = \{\mathbf{M}[(\check{a}_{in} - a_i)(\check{a}_{jn} - a_j)]\}$, $i = \overline{1, k}$, $j = \overline{1, k}$, Φ_n — *матрица Фишера* —

$$\Phi_n = n\left\{\mathbf{M}\left(\frac{\partial \ln p(\xi, a_1, a_2, \ldots, a_k)}{\partial a_i}\,\frac{\partial \ln p(\xi, a_1, a_2, \ldots, a_k)}{\partial a_j}\right)\right\},\ i = \overline{1, k},\ j = \overline{1, k},$$

то матрица $\mathbf{B}_n - \Phi_n^{-1}$ неотрицательно определена [16].

Если несмещенная оценка $\check{\alpha}_n^0$ обладает дисперсией, равной φ_n,

$$\mathbf{M}(\check{\alpha}_n^0 - \alpha)^2 = \mathbf{D}\check{\alpha}_n^0 = \varphi_n, \tag{4.6}$$

то она будет наилучшей среди всех возможных при данном числе наблюдений, так как $\mathbf{D}\check{\alpha}_n^0 \leqslant \mathbf{D}\check{\alpha}_n$ для любой оценки $\check{\alpha}_n$, вычисленной по не более чем n наблюдениям. Оценки, обладающие свойством (4.6), называются *эффективными*. *Асимптотически эффективны* оценки, для которых

$$\lim_{n\to\infty} \frac{\mathbf{M}(\check{\alpha}_n - \alpha)^2}{\varphi_n} = 1. \tag{4.7}$$

Величина $e = \frac{\varphi_n}{\mathbf{M}(\check{\alpha}_n - \alpha)^2}$ называется *эффективностью* оценки $\check{\alpha}_n$. Эффективность любой оценки не превышает единицы. Эффективные оценки имеют эффективность, равную единице, а у асимптотически эффективных оценок она стремится к единице с увеличением количества используемых при их построении наблюдений.

Асимптотическая нормальность оценок. Оценка $\check{\alpha}_n$ распределена *асимптотически нормально*, если с увеличением числа независимых наблюдений в выборке ее распределение приближается к нормальному: $\mathbf{P}\left\{\frac{\check{\alpha}_n - \mathbf{M}\check{\alpha}_n}{\sqrt{\mathbf{D}\check{\alpha}_n}} < x\right\} \to \Phi(x)$ при $n \to \infty$ ($\Phi(x)$ — функция (0; 1)-нормального распределения). Будучи установленным для несмещенной оценки $\check{\alpha}_n$, это свойство позволяет по величине ее дисперсии приближенно оценивать доверительные пределы (4.2) для оцениваемой величины. Если $\mathbf{D}\check{\alpha}_n$ — дисперсия $\check{\alpha}_n$, то при количестве наблюдений, обеспечивающем эффективную аппроксимацию распределения $\check{\alpha}_n$ нормальным законом, доверительные пределы для оцениваемого параметра α, в которых он будет находиться с заданной вероятностью q, можно оценивать, согласно (2.32″), в виде

$$\alpha_q^- = \check{\alpha}_n - u_{\frac{1+q}{2}}\sqrt{\mathbf{D}\check{\alpha}_n}, \quad \alpha_q^+ = \check{\alpha}_n + u_{\frac{1+q}{2}}\sqrt{\mathbf{D}\check{\alpha}_n}, \tag{4.8}$$

где $u_{\frac{1+q}{2}}$ — как и раньше, квантиль (0; 1)-нормального распределения порядка $\frac{1+q}{2}$, определяемый по табл. 2 (Приложение).

Свойство асимптотической нормальности оценок обычно доказывается с помощью центральной предельной теоремы и ее следствий (см. гл. 2; пример 2.7).

§ 2. Методы нахождения оценок

В математической статистике выработано несколько общих методов построения оценок.

Метод усреднения. Независимые наблюдения $x_1, x_2, \ldots, x_n$ одной и той же случайной величины ξ распределены по тому же закону,

что и ξ. Поэтому, если можно указать такую функцию $f(x)$, что математическое ожидание $f(\xi)$ равно оцениваемому параметру α, $\mathbf{M}f(\xi)=\alpha$, а дисперсия конечна, $\mathbf{D}f(\xi)=D$, то $\mathbf{M}f(x_i)=\alpha$ и $\mathbf{D}f(x_i)=D$. Состоятельной и несмещенной оценкой α будет

$$\check{\alpha}_n=\frac{1}{n}\sum_{i=1}^{n}f(x_i). \tag{4.9}$$

Несмещенность следует из того, что $\mathbf{M}\check{\alpha}_n=\frac{1}{n}\mathbf{M}\left[\sum_{i=1}^{n}f(x_i)\right]=\frac{1}{n}n\alpha=\alpha$.

Так как дисперсия оценки $\check{\alpha}_n$

$$\mathbf{D}\check{\alpha}_n=\mathbf{M}\left[\frac{1}{n}\sum_{i=1}^{n}f(x_i)-\alpha\right]^2=\frac{1}{n^2}\mathbf{M}\left\{\sum_{i,\,j=1}^{n}[f(x_i)-\alpha][f(x_j)-\alpha]\right\}=$$

$$=\frac{1}{n^2}\sum_{i=1}^{n}\mathbf{M}[f(x_i)-\alpha]^2=\frac{\mathbf{D}f(\xi)}{n}, \tag{4.10}$$

по неравенству Чебышева получим: для любого $\varepsilon>0$

$$\mathbf{P}\{|\check{\alpha}_n-\alpha|\geqslant\varepsilon\}\leqslant\frac{\mathbf{D}\check{\alpha}_n}{\varepsilon^2}=\frac{\mathbf{D}f(\xi)}{\varepsilon^2 n}\to 0 \text{ при } n\to\infty, \tag{4.11}$$

что и доказывает состоятельность оценки $\check{\alpha}_n$.

Как показывает формула (4.11), для состоятельности произвольной несмещенной оценки $\check{\alpha}_n$ достаточно, чтобы ее дисперсия с увеличением количества наблюдений n стремилась к нулю:

$$\mathbf{D}\check{\alpha}_n=\mathbf{M}(\check{\alpha}_n-\alpha)^2\to 0 \text{ при } n\to\infty. \tag{4.12}$$

Оценки по методу усреднения вида (4.9) обладают и важным свойством асимптотической нормальности своего распределения. Оно вытекает из первого следствия центральной предельной теоремы (гл. 2). Это свойство дает возможность вычислять приближенные доверительные пределы для α по формуле (4.8).

Оценка по методу усреднения (4.9) имеет наименьшую дисперсию среди всех несмещенных оценок вида $\tilde{\alpha}=\sum_{i=1}^{n}a_i f(x_i)$ с произвольными коэффициентами a_i, удовлетворяющими условию несмещенности. В этом нетрудно убедиться. Из условия несмещенности $\mathbf{M}\tilde{\alpha}=\mathbf{M}\left[\sum_{i=1}^{n}a_i f(x_i)\right]=$

$=\alpha\sum_{i=1}^{n}a_i=\alpha$, откуда $\sum_{i=1}^{n}a_i=1$. Дисперсия оценки $\tilde{\alpha}$ —

$$\mathbf{D}\tilde{\alpha}=\sum_{i=1}^{n}a_i^2\,\mathbf{D}f(x_i)=\sum_{i=1}^{n-1}a_i^2\,\mathbf{D}f(x_i)+\left(1-\sum_{i=1}^{n-1}a_i\right)^2\mathbf{D}f(x_n).$$

Дифференцируя по a_i, получим

$$a_i \mathbf{D}f(x_i) = \left(1 - \sum_{i=1}^{n-1} a_i\right) \mathbf{D}f(x_n) = a_n \mathbf{D}f(x_n)$$

и

$$a_i = \frac{1}{\mathbf{D}f(x_i)} \Big/ \sum_{i=1}^{n} \frac{1}{\mathbf{D}f(x_i)}. \tag{4.13}$$

Так как $\mathbf{D}f(x_i) = D$, $a_i = \frac{1}{n}$.

Подобным образом можно действовать и при вычислении оценки по n *неравноточным наблюдениям*, представляющим собой n независимых, но не одинаково распределенных случайных величин $\xi_1, \xi_2, \ldots, \xi_n$. Если для каждой из них можно указать функцию $f_i(\xi_i)$ такую, что $\mathbf{M}f_i(\xi_i) = \alpha$ при возможно наименьшей дисперсии $\mathbf{D}f_i(\xi_i) = D_i$, то наилучшей оценкой α среди всех оценок вида $\tilde{\alpha} = \sum_{i=1}^{n} a_i f_i(\xi_i)$ при условии $\sum_{i=1}^{n} a_i = 1$, по аналогии с (4.13), будет

$$\check{\alpha} = \sum_{i=1}^{n} a_i^0 f_i(\xi_i), \quad \text{где} \quad a_i^0 = \frac{\frac{1}{D_i}}{\sum_{i=1}^{n} \frac{1}{D_i}}, \quad i = \overline{1, n}. \tag{4.13'}$$

По формуле (4.9), оценка математического ожидания произвольной случайной величины ξ с конечной дисперсией по ее независимым наблюдениям x_i, с учетом того, что $\mathbf{M}x_i = \mathbf{M}\xi$, представляет собой среднее арифметическое из наблюдений *(выборочное среднее)*

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{4.14}$$

причем она будет состоятельной и несмещенной. Дисперсия этой оценки, как следует из (4.10),

$$\mathbf{D}\bar{x} = \frac{\mathbf{D}\xi}{n}. \tag{4.15}$$

Состоятельной и несмещенной оценкой дисперсии $D = \mathbf{D}\xi$ при известном математическом ожидании будет

$$\hat{D} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mathbf{M}\xi)^2. \tag{4.16}$$

Дисперсия этой оценки

$$\mathbf{D}\check{D} = \mathbf{M}\check{D}^2 - D^2 = \frac{1}{n^2}\mathbf{M}\left[\sum_{i=1}^{n}(x_i - \mathbf{M}\xi)^2\right]^2 - D^2 = \frac{1}{n^2}\mathbf{M}\left[\sum_{i=1}^{n}(x_i - \mathbf{M}\xi)^4 + \right.$$

$$\left. + \sum_{i \neq j}^{n}(x_i - \mathbf{M}\xi)^2 (x_j - \mathbf{M}\xi)^2\right] - D^2 = \frac{1}{n^2}[n\mu_4 + n(n-1)D^2 - n^2 D] = \frac{\mu_4 - D^2}{n}, \tag{4.17}$$

где μ_4 — четвёртый центральный момент ξ. Если математическое ожидание неизвестно, вместо $\mathbf{M}\xi$ в (4.16) используется его оценка (4.14) и оценка дисперсии приобретает вид

$$\check{D} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\left(\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right). \tag{4.16'}$$

Оценку (4.16′) называют *выборочной дисперсией*. Чаще используют несколько измененную оценку

$$\bar{D} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right). \tag{4.16''}$$

В отличие от оценки (4.16′), которая лишь асимптотически несмещенна —

$$\mathbf{M}\check{D} = \mathbf{M}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - M + M - \bar{x})^2\right] = \frac{1}{n}\mathbf{M}\left[\sum_{i=1}^{n}(x_i - M)^2 - \right.$$

$$\left. - 2\sum_{i=1}^{n}(x_i - M)(\bar{x} - M) + n(\bar{x} - M)^2\right] =$$

$$= D - \frac{2}{n}\mathbf{M}\left[\sum_{i=1}^{n}(x_i - M)\sum_{i=1}^{n}\frac{x_i - M}{n}\right] + \frac{D}{n} = D - \frac{2}{n^2}\sum_{i=1}^{n}\mathbf{M}(x_i - M)^2 +$$

$$+ \frac{D}{n} = \frac{n-1}{n}D \to D \text{ при } n \to \infty,$$

оценка (4.16″) несмещенная: $\mathbf{M}\bar{D} = \frac{n}{n-1}\mathbf{M}\check{D} = D$.

Можно показать, что дисперсия оценки $\bar{D}$

$$\mathbf{D}\bar{D} = \frac{\mu_4 - D^2}{n-1} - \frac{\mu_4 - 3D^2}{(n-1)^2} + \frac{\mu_4 - 3D^2}{n(n-1)^2} = \frac{\mu_4}{n} - \frac{n-3}{n(n-1)}D^2; \tag{4.17'}$$

при нормальном распределении ξ, учетом (2.21),

$$\mathbf{D}\bar{D} = \frac{2D^2}{n-1}. \tag{4.17''}$$

Начальные моменты оцениваются *выборочными начальными моментами*

$$\overline{m}_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k, \tag{4.18}$$

а центральные — оценками

$$\overline{\mu}_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mathbf{M}\xi)^k, \quad \check{\mu}_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k. \tag{4.19}$$

Эти оценки состоятельны; оценки $\overline{m}_k$, $\overline{\mu}_k$ несмещенны, а $\check{\mu}_k$ — асимптотически несмещенны. Дисперсии оценок $\overline{m}_k$ и $\overline{\mu}_k$, по (4.10) —

$$\mathbf{D}\overline{m}_k = \frac{1}{n}\mathbf{M}(\xi^k - m_k)^2 = \frac{m_{2k} - m_k^2}{n}, \quad \mathbf{D}\overline{\mu}_k = \frac{1}{n}\mathbf{M}[(\xi - \mathbf{M}\xi)^k - \mu_k]^2 = \frac{\mu_{2k} - \mu_k^2}{n}; \tag{4.19'}$$

дисперсия оценки* $\check{\mu}_k$ (называемой *выборочным центральным моментом*) [16]

$$\mathbf{D}\check{\mu}_k = \frac{1}{n}(\mu_{2k} - 2k\mu_{k-1}\mu_{k+1} - \mu_k^2 + k^2\mu_2\mu_{k-1}^2) + O\left(\frac{1}{n^2}\right). \tag{4.19''}$$

Метод моментов. Сущность метода сводится к следующему. Если параметр α можно представить в виде функции центральных моментов, то в качестве его оценки используют значение этой функции при значениях аргументов, равных выборочным центральным моментам тех же порядков. При не очень обременительных условиях эта оценка оказывается состоятельной и асимптотически несмещенной. Аналогично можно использовать и начальные моменты.

Сформулируем теорему, относящуюся к обоснованию метода моментов. Пусть $\alpha = f(\mu_{k_1}, \mu_{k_2}, \ldots, \mu_{k_p})$ — функция p центральных моментов случайной величины ξ, непрерывная вместе с частными производными до второго порядка включительно и обладающая свойством $|f(\check{\mu}_{k_1}, \check{\mu}_{k_2}, \ldots, \check{\mu}_{k_p})| \leqslant cn^t$ $(c \geqslant 0,\ t \geqslant 0)$ для всех возможных значений $\check{\mu}_{k_i}$ $(i = \overline{1, p})$. Тогда оценка

$$\check{\alpha} = f(\check{\mu}_{k_1}, \check{\mu}_{k_2}, \ldots, \check{\mu}_{k_p}) \tag{4.20}$$

имеет математическое ожидание

$$\mathbf{M}\check{\alpha} = f(\mu_{k_1}, \mu_{k_2}, \ldots, \mu_{k_p}) + O\left(\frac{1}{n}\right) = \alpha + O\left(\frac{1}{n}\right) \tag{4.21}$$

и дисперсию

$$\mathbf{D}\check{\alpha} = \sum_{i=1}^{p}\left(\frac{\partial f}{\partial i}\right)^2 \mathbf{M}(\check{\mu}_{k_i} - \mu_{k_i})^2 + 2\sum_{i<j}^{p}\frac{\partial f}{\partial i}\frac{\partial f}{\partial j}\mathbf{M}(\check{\mu}_{k_i} - \mu_{k_i})(\check{\mu}_{k_j} - \mu_{k_j}) + o\left(\frac{1}{n}\right), \tag{4.22}$$

* В (4.19″) и дальше $O(a_n)$ обозначает величину такого же порядка малости, как и a_n; $o(a_n)$ — величина более высокого порядка малости по сравнению с a_n.

где $\frac{\partial f}{\partial i} = \frac{\partial f(\mu_{k_1}, \mu_{k_2}, \ldots, \mu_{k_p})}{\partial \mu_{k_i}}$; $\mathbf{M}(\check{\mu}_k - \mu_k)^2 \approx \mathbf{D}\check{\mu}_k$ определяется по (4.19″);

$$\mathbf{M}[(\check{\mu}_k - \mu_k)(\check{\mu}_t - \mu_t)] = \frac{1}{n}(\mu_{k+t} - k\mu_{k-1}\mu_{t+1} - t\mu_{k+1}\mu_{t-1} -$$

$$- \mu_k\mu_t + kt\mu_2\mu_{k-1}\mu_{t-1}) + O\left(\frac{1}{n^2}\right).$$

Аналогичную теорему можно сформулировать и для начальных моментов.

Следующее соображение, хотя и не является строгим доказательством, показывает естественность такого вывода. Выборочные центральные или начальные моменты порядка k_i имеют дисперсию порядка $\frac{1}{n}$ (4.19″), если моменты порядка $2k_i$ конечны. Поэтому отклонения оценок $\check{\mu}_{k_i}$ от μ_{k_i} соответственно малы и, согласно известной формуле дифференциального исчисления,

$$\check{\alpha} - \alpha = f(\check{\mu}_{k_1}, \check{\mu}_{k_2}, \ldots, \check{\mu}_{k_p}) - f(\mu_{k_1}, \mu_{k_2}, \ldots, \mu_{k_p}) =$$

$$= \sum_{i=1}^{p} \frac{\partial f}{\partial i}(\check{\mu}_{k_i} - \mu_{k_i}) + o(\rho),$$

$\rho = \sqrt{\sum_{i=1}^{p}(\check{\mu}_{k_i} - \mu_{k_i})^2}$. Учитывая, что $\mathbf{M}(\check{\mu}_{k_i} - \mu_{k_i})^2 = O\left(\frac{1}{n}\right)$, ρ в среднем — величина $O\left(\frac{1}{\sqrt{n}}\right)$. Дисперсия $\check{\alpha}$ составит

$$\mathbf{D}\check{\alpha} = \mathbf{M}\left[\sum_{i=1}^{n} \frac{\partial f}{\partial i}(\check{\mu}_{k_i} - \mu_{k_i})\right]^2 + o\left(\frac{1}{n}\right),$$

откуда и следует (4.22).

Формула (4.22) используется для определения точности оценок в виде доверительных пределов, обычно на основании свойства асимптотической нормальности. Для асимптотически нормальных оценок доверительные пределы вычисляются по формулам (4.8). Разумеется, нужно иметь в виду, что они дают лишь приближённые величины, уточняющиеся с увеличением количества наблюдений.

Примером применения метода моментов может служить оценка среднего квадратического отклонения* —

$$\check{\sigma} = \sqrt{\check{D}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (4.23)$$

* В приводимых примерах предполагается, что распределение ξ удовлетворяет условиям теоремы либо его можно заменить удовлетворяющей им аппроксимацией.

(математическое ожидание неизвестно). Эта оценка асимптотически несмещенная, а ее дисперсия с точностью до $o\left(\frac{1}{n}\right)$ имеет вид

$$\mathbf{D}\check{\sigma} = \mathbf{D}\left(\sqrt{\check{\mu}_2}\right) \approx \left(\frac{1}{2}\frac{1}{\sqrt{\mu_2}}\right)^2 \mathbf{D}\check{\mu}_2 \approx \frac{\mu_4 - D^2}{4nD} = \frac{D}{4n}(E+2), \quad (4.24)$$

где E — эксцесс распределения ξ.

Другой пример — оценка коэффициента вариации V —

$$\check{V} = \frac{1}{\bar{x}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sqrt{\check{D}}}{\bar{x}}. \quad (4.25)$$

Дисперсия этой оценки

$$\mathbf{D}\check{V} \approx \frac{1}{4n}[4V^4 + V^2(E+2) - 4AV^3] \quad (4.26)$$

с точностью до $o\left(\frac{1}{n}\right)$. В (4.26) A — коэффициент асимметрии распределения, E — коэффициент эксцесса.

Оценка $\check{B}_q = \bar{x} + u_q\check{\sigma}$ величины $B_q = M + u_q\sigma$ (M — математическое ожидание, σ^2 — дисперсия, u_q — q-квантиль (0; 1)-нормального распределения) также состоятельна и асимптотически несмещенная; ее дисперсия

$$\mathbf{D}\check{B}_q \approx \frac{\sigma^2}{n}\left[1 + u_q A + \frac{u_q^2(E+2)}{4}\right]. \quad (4.27)$$

Для нормального распределения величина B_q служит квантилем порядка q. Дисперсия $\check{B}_q$ в этом случае

$$\mathbf{D}\check{B}_q = \frac{\sigma^2}{n}\left(1 + \frac{u_q^2}{2}\right). \quad (4.27')$$

Оценки коэффициентов асимметрии и эксцесса

$$\check{A} = \frac{\check{\mu}_3}{\check{\sigma}^3}, \quad \check{E} = \frac{\check{\mu}_4}{\check{\sigma}^4} - 3. \quad (4.28)$$

Можно показать [16], что дисперсии оценок $\check{A}$ и $\check{E}$ представляются в виде

$$\mathbf{D}(\check{A}) = \frac{1}{n}\left(\frac{\mu_6}{\mu_2^3} - 3\frac{\mu_3\mu_5}{\mu_2^4} - 6\frac{\mu_4^2}{\mu_2^5} + \frac{9}{4}\frac{\mu_3^2\mu_4}{\mu_2^5} + \frac{35}{4}\frac{\mu_3^4}{\mu_2^6} + 9\right) + o\left(\frac{1}{n}\right), \quad (4.29)$$

$$\mathbf{D}(\check{E}) = \frac{1}{n}\left(\frac{\mu_8}{\mu_2^4} - 4\frac{\mu_4\mu_6}{\mu_2^5} - 8\frac{\mu_3\mu_5}{\mu_2^4} + 4\frac{\mu_4^3}{\mu_2^6} - \frac{\mu_4^2}{\mu_2^4} + 16\frac{\mu_3^2\mu_4}{\mu_2^5} + 16\frac{\mu_3^4}{\mu_2^6}\right) + o\left(\frac{1}{n}\right).$$

В случае нормального распределения случайной величины ξ

$$\mathbf{D}\check{A} = \frac{6}{n} + o\left(\frac{1}{n}\right), \quad \mathbf{D}\check{E} = \frac{24}{n} + o\left(\frac{1}{n}\right). \quad (4.30)$$

Более точная оценка коэффициента эксцесса основана на учете смещения оценки $\check{E}$ [2]:

$$\check{E} = \frac{1}{n\check{\sigma}^4}\sum_{i=1}^{n}(x_i - \bar{x})^4 - 3 + \frac{6}{n+1}. \qquad (4.29')$$

Уточненные выражения дисперсий оценок асимметрии и эксцесса (при нормальном распределении случайной величины):

$$\mathbf{D}\check{A} \approx \frac{6}{n}\left(1 - \frac{12}{2n-7}\right), \quad \mathbf{D}\check{E} \approx \frac{24}{n}\left(1 - \frac{225}{15n+124}\right). \qquad (4.30')$$

По аналогии с (4.9), в качестве оценки коэффициента корреляции r случайных величин ξ и η можно взять

$$\check{r} = \frac{1}{n\sigma_\xi\sigma_\eta}\sum_{i=1}^{n}(x_i - \mathbf{M}\xi)(y_i - \mathbf{M}\eta), \qquad (4.31)$$

где x_i и y_i — наблюдения ξ и η; $\mathbf{M}\xi$ и $\mathbf{M}\eta$ — математические ожидания ξ и η; σ_ξ, σ_η — средние квадратические отклонения. Поскольку

$$\mathbf{M}\frac{(\xi - \mathbf{M}\xi)(\eta - \mathbf{M}\eta)}{\sigma_\xi\sigma_\eta} = r = \mathbf{M}\frac{(x_i - \mathbf{M}\xi)(y_i - \mathbf{M}\eta)}{\sigma_\xi\sigma_\eta},$$

математическое ожидание $\check{r}$

$$\mathbf{M}\check{r} = \frac{1}{n\sigma_\xi\sigma_\eta}\sum_{i=1}^{n}\mathbf{M}[(x_i - \mathbf{M}\xi)(y_i - \mathbf{M}\eta)] = \frac{nr\sigma_\xi\sigma_\eta}{n\sigma_\xi\sigma_\eta} = r. \qquad (4.32)$$

Так как математические ожидания $\mathbf{M}\xi$, $\mathbf{M}\eta$ и средние квадратические отклонения σ_ξ, σ_η обычно неизвестны, используют оценку коэффициента корреляции, называемую *выборочным коэффициентом корреляции*

$$\check{r} = \frac{1}{n\check{\sigma}_x\check{\sigma}_y}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n\check{\sigma}_x\check{\sigma}_y}\left(\sum_{i=1}^{n}x_iy_i - n\bar{x}\bar{y}\right). \qquad (4.33)$$

Здесь

$$\check{\sigma}_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \check{\sigma}_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n}y_i.$$

Дисперсия этой оценки, при нормальном распределении $\{\xi, \eta\}$,

$$\mathbf{D}\check{r} = \frac{(1-r^2)^2}{n} + O\left(\frac{1}{n^{3/2}}\right); \tag{4.34}$$

математическое ожидание $\mathbf{M}\check{r} = r + O\left(\frac{1}{n}\right)$.

Метод максимума правдоподобия. Сущность метода максимума правдоподобия — наиболее эффективного, хотя и нередко трудоемкого способа нахождения оценок, состоит в следующем. Пусть $x_1, x_2, \ldots, x_n$ — независимые* наблюдения случайной величины ξ, плотность распределения которой принадлежит к некоторому параметрическому семейству $p(x) = p(x, \alpha_1, \alpha_2, \ldots, \alpha_k)$; $\alpha_1, \alpha_2, \ldots, \alpha_k$ — параметры, подлежащие оценке. *Оценками наибольшего правдоподобия*, или *наиболее правдоподобными оценками*, параметров $\alpha_1, \alpha_2, \ldots, \alpha_k$ будут такие их значения, при которых обращается в максимум функция

$$L(x_1, x_2, \ldots, x_n; \alpha_1, \alpha_2, \ldots, \alpha_k) = \sum_{i=1}^{n} \ln p(x_i, \alpha_1, \alpha_2, \ldots, \alpha_k). \tag{4.35}$$

Функция $f(x_1, x_2, \ldots, x_n; \alpha_1, \alpha_2, \ldots, \alpha_k) = \prod_{i=1}^{n} p(x_i, \alpha_1, \alpha_2, \ldots, \alpha_k)$ называется *функцией правдоподобия*.

Такой принцип вполне естественен. Действительно, пусть $\Delta_1, \Delta_2, \ldots, \Delta_n$ малые интервалы длиной δ вокруг значений случайной величины, равных $x_1, x_2, \ldots, x_n$. Вероятность того, что n наблюдений случайной величины попадут: первое — в Δ_1, второе — в Δ_2 и т. д., n-е — в Δ_n (как это и имело место) —

$$\mathbf{P}\{x_1 \in \Delta_1, x_2 \in \Delta_2, \ldots, x_n \in \Delta_n\} = \prod_{i=1}^{n} \mathbf{P}\{x_i \in \Delta_i\} =$$

$$= \delta^n \prod_{i=1}^{n} p(x_i, \alpha_1, \alpha_2, \ldots, \alpha_k) + o(\delta^n), \tag{4.36}$$

ввиду того, что $\mathbf{P}\{x_i \in \Delta_i\} = \int_{\Delta_i} p(t, \alpha_1, \alpha_2, \ldots, \alpha_k)\, dt = p(x_i, \alpha_1, \alpha_2, \ldots, \alpha_k)\,\delta + o(\delta)$.

Целесообразно определить оценки параметров α_j так, чтобы они наилучшим образом согласовывались с наблюдениями x_i. Такое согласие и обеспечивается требованием, чтобы вероятность (4.36) была максимальной, т. е. оценки $\check{\alpha}_1, \check{\alpha}_2, \ldots, \check{\alpha}_k$ должны обращать в максимум $\prod_{i=1}^{n} p(x_i, \alpha_1, \alpha_2, \ldots, \alpha_k)$. Для упрощения используется логарифмичес-

---

* Метод максимума правдоподобия можно распространить и на случай зависимых наблюдений. Однако, такая его реализация на практике, как правило, оказывается сложной, поэтому при построении оценок используется предположение о независимости наблюдений.

кое преобразование функции правдоподобия — оно монотонно и не изменяет точек максимума.

По условию максимума функции правдоподобия, оценки наибольшего правдоподобия находят из системы *уравнений правдоподобия*

$$\frac{\partial L}{\partial \alpha_j} = 0, \quad j = \overline{1, k}. \tag{4.37}$$

При необременительных условиях*, налагаемых на плотность распределения $p(x, \alpha_1, \alpha_2, \ldots, \alpha_k)$, которые, как правило, выполняются для встречающихся на практике величин, оценки наибольшего правдоподобия обладают таким свойством. Решение системы уравнений правдоподобия (4.37) существует и дает состоятельные, асимптотически несмещенные, асимптотически эффективные и асимптотически нормальные оценки. Если эффективная оценка (или совокупность совместно эффективных оценок) существует, то она будет единственным решением уравнения (или системы уравнений) (4.37). Асимптотическая эффективность гарантирует, по крайней мере при больших n, оптимальность наиболее правдоподобных оценок. Асимптотическая нормальность используется для определения точности оценивания — в виде доверительных интервалов (4.8) для оцениваемых параметров.

В качестве примера рассмотрим нахождение оценок наибольшего правдоподобия параметров нормального распределения по n наблюдениям $x_1, x_2, \ldots, x_n$. Логарифмическое преобразование функции правдоподобия —

$$L(x_1, x_2, \ldots, x_n, m, D) = \sum_{i=1}^{n} \ln\left[\frac{1}{\sqrt{2\pi D}} e^{-\frac{(x_i - m)^2}{2D}}\right] =$$

$$= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln D - \frac{1}{2}\sum_{i=1}^{n}\frac{(x_i - m)^2}{D}, \tag{4.38}$$

где m и D подлежат оценке и имеют смысл соответственно матема-

---

* В простейшем случае $p(x) = p(x, \alpha)$ (однопараметрическое семейство) эти условия сводятся к следующему: 1) при всяком α существуют $\frac{\partial^j \ln p}{\partial \alpha^j}$ $(j = \overline{1,3})$, причем $\left|\frac{\partial^j p}{\partial \alpha^j}\right| \leqslant F(x)$, $i = 1, 2$; $F(x)$ — ограниченная функция, не зависящая от α и такая, что $\int_{-\infty}^{\infty} F(x)\,dx < \infty$; $\left|\frac{\partial^3 p(x, \alpha)}{\partial x^3}\right| \leqslant F_1(x)$, причем $\mathbf{M}F_1(\xi) = \int_{-\infty}^{\infty} F_1(x) \times$ $\times p(x, \alpha)\,dx < t < \infty$; 2) $\mathbf{M}\left(\frac{\partial \ln p(\xi, \alpha)}{\partial \alpha}\right)^2$ положительно и конечно; 3) при $\alpha_1 \neq \alpha_2$ плотности $p(x, \alpha_1)$ и $p(x, \alpha_2)$ отличаются на интервале ненулевой длины [16].

тического ожидания и дисперсии. Система уравнений (4.37) для вычисления оценок m и D имеет вид

$$\begin{cases} \dfrac{\partial L}{\partial m} = \sum\limits_{i=1}^{n} \dfrac{x_i - m}{D} = 0, \\ \dfrac{\partial L}{\partial D} = -\dfrac{n}{2D} + \dfrac{1}{2} \sum\limits_{i=1}^{n} \dfrac{(x_i - m)}{D^2} = 0, \end{cases} \tag{4.39}$$

откуда $\check{m} = \dfrac{1}{n} \sum\limits_{i=1}^{n} x_i = \bar{x}$, $\check{D} = \dfrac{1}{n} \sum\limits_{i=1}^{n} (x_i - \check{m})^2$. Среднее арифметическое $\bar{x}$ и выборочная дисперсия, вычисленные по независимым наблюдениям нормально распределенной случайной величины, как оценки наибольшего правдоподобия будут асимптотически эффективными оценками математического ожидания и дисперсии. Более того, в этом случае эффективность оценки $\bar{x}$

$$e_m = [\mathbf{M}(\bar{x} - m)^2]^{-1} \varphi_n(m) = \frac{n}{D} \frac{1}{n} \left[\mathbf{M} \left\{ \frac{\partial}{\partial m} \left[-\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln D - \right. \right. \right.$$

$$\left. \left. \left. - \frac{1}{2} \frac{(\xi - m)^2}{D} \right] \right\}^2 \right]^{-1} = \frac{n}{D} \cdot \frac{D}{n} = 1,$$

т. е. она эффективна. Аналогично оценка $\check{D}$ (4.16) дисперсии при известном математическом ожидании также эффективна: с учетом (4.17)

$$e_D = \left(\frac{\mu_4 - D^2}{n} \right)^{-1} \varphi_n(D) = \frac{n}{2D^2} \cdot \frac{1}{n} \left\{ \mathbf{M} \left[-\frac{1}{2D} + \frac{(\xi - M)^2}{2D^2} \right]^2 \right\}^{-1} = \frac{n}{2D^2} \frac{2D^2}{n} = 1.$$

Для дискретных случайных величин метод максимума правдоподобия применяется в такой форме. Пусть определяющие дискретное распределение наборы вероятностей $\mathbf{P}\{\xi = y_j\} = P(y_j)$ принадлежат некоторому параметрическому семейству:

$$P(y_j) = P(y_j, \alpha_1, \alpha_2, \ldots, \alpha_k), \quad j = 1, 2, \ldots$$

Оценки параметров $\alpha_1, \alpha_2, \ldots, \alpha_k$ определяются условием максимума функции

$$L(x_1, x_2, \ldots, x_n; \alpha_1, \alpha_2, \ldots, \alpha_k) = \sum_{i=1}^{n} \ln P(x_i, \alpha_1, \alpha_2, \ldots, \alpha_k) \tag{4.35'}$$

(x_i — наблюдения ξ).

Применение метода максимума правдоподобия для дискретных случайных величин демонстрирует такой пример. Пусть необходимо по результатам n испытаний оценить вероятность p события, которое может происходить в каждом испытании с этой вероятностью. Введем дискретную случайную величину ξ, которая получает значение, равное единице, если в результате испытания событие произошло, и нулю, если не произошло. Функция, (4.35′) получит вид

$$L(x_1, x_2, \ldots, x_n, p) = \ln [p^{\nu}(1-p)^{n-\nu}] = \nu \ln p + (n - \nu) \ln (1 - p),$$

где $p = \mathbf{P}\{\xi = 1\}$, $1 - p = \mathbf{P}\{\xi = 0\}$, ν — число испытаний, в которых событие осуществилось. В данном случае функция правдоподобия зависит от одного параметра p и его оценка $\check{p}$ определяется из уравнения

$$\frac{\partial L}{\partial p} = \frac{\nu}{p} - \frac{n - \nu}{1 - p} = 0,$$

откуда $\check{p} = \frac{\nu}{n}$ *.

Следующее свойство оценок максимального правдоподобия может оказаться полезным при их вычислении. Пусть плотность распределения $p(x, \alpha_1, \alpha_2, \ldots, \alpha_k)$ случайной величины ξ зависит от k параметров $\alpha_1, \alpha_2, \ldots, \alpha_k$, для которых известны оценки наибольшего правдоподобия $\check{\alpha}_1, \check{\alpha}_2, \ldots, \check{\alpha}_k$, являющиеся решением системы уравнений (4.37); $\theta_1, \theta_2, \ldots, \theta_k$ — параметры или числовые характеристики, являющиеся дифференцируемыми функциями $\alpha_1, \alpha_2, \ldots, \alpha_k$,

$$\theta_i = \theta_i(\alpha_1, \alpha_2, \ldots, \alpha_k), \quad i = \overline{1, k}, \tag{4.40}$$

устанавливающими взаимно однозначное соответствие между совокупностями значений $\{\alpha_1, \alpha_2, \ldots, \alpha_k\}$ и $\{\theta_1, \theta_2, \ldots, \theta_k\}$. Логарифм функции правдоподобия, определяемой относительно параметров $\theta_1, \theta_2, \ldots, \theta_k$

$$L(x_1, x_2, \ldots, x_n, \theta_1, \theta_2, \ldots, \theta_k) = \sum_{i=1}^{n} \ln p(x_i, \alpha_1(\theta_1, \theta_2, \ldots, \theta_k),$$

$$\ldots, \alpha_k(\theta_1, \theta_2, \ldots, \theta_k)) = L_0(x_1, x_2, \ldots, x_n, \alpha_1(\theta_1, \theta_2, \ldots, \theta_k),$$

$$\ldots, \alpha_k(\theta_1, \theta_2, \ldots, \theta_k)),$$

где L_0 — логарифм функции правдоподобия относительно параметров $\alpha_1, \alpha_2, \ldots, \alpha_k$; $\alpha_i(\theta_1, \theta_2, \ldots, \theta_k)$ $(i = \overline{1, k})$ — функции, обратные (4.40).

Очевидно, $\frac{\partial L}{\partial \theta_j} = \sum_{i=1}^{k} \frac{\partial L_0}{\partial \alpha_i} \frac{\partial \alpha_i}{\partial \theta_j}$. Определим $\theta_j = \check{\theta}_j$ так, чтобы $\alpha_i(\check{\theta}_1, \check{\theta}_2, \ldots, \check{\theta}_k) = \check{\alpha}_i$ при $i = \overline{1, k}$. Тогда $\left.\frac{\partial L}{\partial \theta_j}\right|_{\check{\theta}_1, \check{\theta}_2, \ldots, \check{\theta}_k} = 0$, так как $\left.\frac{\partial L_0}{\partial \alpha_i}\right|_{\check{\alpha}_1, \check{\alpha}_2, \ldots, \check{\alpha}_k} = 0$. Ввиду взаимной однозначности отображения, задаваемого функциями (4.40), значениями $\check{\theta}_i$ являются

$$\check{\theta}_i = \theta_i(\check{\alpha}_1, \check{\alpha}_2, \ldots, \check{\alpha}_k) \tag{4.41}$$

и они, следовательно, будут оценками максимального правдоподобия параметров θ_i.

---

* Ранее мы получили эту оценку, опираясь на закон больших чисел. Теперь, однако, можно утверждать, что она обладает определенными оптимальными свойствами.

Пусть, например, необходимо оценить параметр $\theta = \theta(\alpha_1, \alpha_2, \ldots, \alpha_k)$, являющийся дифференцируемой функцией $\alpha_1, \alpha_2, \ldots, \alpha_k$, причем эта функция имеет отличную от нуля и не изменяющую знака частную производную по какой-либо переменной — скажем, по α_1. Введем новые параметры $\theta_1 = \theta$, $\theta_2 = \alpha_2$, $\theta_3 = \alpha_3$, $\ldots$, $\theta_k = \alpha_k$. По известной теореме дифференциального исчисления параметр α_1 определится из условия $\theta = \theta(\alpha_1, \alpha_2, \ldots, \alpha_k)$ как однозначная функция параметров $\theta = \theta_1, \alpha_2, \alpha_3, \ldots, \alpha_k$: $\alpha_1 = g(\theta_1, \alpha_2, \alpha_3, \ldots, \alpha_k)$, которая вместе с равенствами $\alpha_i = \theta_i$, $i = \overline{2, k}$ определит взаимно-однозначное соответствие (4.40). Поэтому оценкой наибольшего правдоподобия параметра θ будет функция оценок наибольшего правдоподобия $\check{\alpha}_i$ вида

$$\check{\theta} = \theta(\check{\alpha}_1, \check{\alpha}_2, \ldots, \check{\alpha}_k). \tag{4.41'}$$

Примером может служить оценка среднего квадратического отклонения. Если $x_1, x_2, \ldots, x_n$ — независимые наблюдения нормально распределенной случайной величины, то наиболее правдоподобной оценкой среднего квадратического отклонения будет

$$\sigma = \sqrt{D} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{4.42}$$

Это следует из того, что между значениями $\{\mathbf{M}\xi, \mathbf{D}\xi\}$ и $\{\mathbf{M}\xi, \sigma\}$ существует взаимно-однозначное соответствие: производная $\frac{\partial\sigma}{\partial D} = \frac{1}{2\sqrt{D}} > 0$ $(D = \mathbf{D}\xi, \sigma = \sqrt{D})$.

Точно также максимально правдоподобными оценками верхних и нижних пределов $A^{\pm} = \mathbf{M}\xi \pm u\sqrt{\mathbf{D}\xi}$ будут

$$A^{\pm} = \bar{x} \pm u\sqrt{\check{D}} = \bar{x} \pm u\check{\sigma}, \tag{4.43}$$

где $\check{\sigma}$ вычисляется по (4.42), $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Рассмотрим еще один пример — построение оценок математического ожидания, моды и медианы обобщенно-логнормального распределения. Максимально правдоподобные оценки получим, подставив в формулы (2.39), (2.41), (2.43) вместо $\mu = \mathbf{M}\ln(a + \lambda\xi)$, $\delta^2 = \mathbf{D}\ln(a + \lambda\xi)$ оценки максимального правдоподобия. Последние, ввиду нормального распределения $\ln(a + \lambda\xi)$, вычисляются в виде

$$\check{\mu} = \frac{1}{n}\sum_{i=1}^{n}\ln(a + \lambda x_i), \tag{4.44}$$

$$\check{\delta}^2 = \frac{1}{n}\sum_{i=1}^{n}[\ln(a + \lambda x_i) - \check{\mu}]^2. \tag{4.45}$$

Оценка математического ожидания —

$$\check{M} = \left[\exp\left(\check{\mu} + \frac{1}{2}\check{\delta}^2\right) - a\right]\lambda; \tag{4.46}$$

моды —

$$\check{M}_0 = [\exp(\check{\mu} - \check{\delta}^2) - a]\,\lambda; \tag{4.47}$$

медианы —

$$\check{M}_e = (\exp\check{\mu} - a)\,\lambda. \tag{4.48}$$

§ 3. Метод наименьших квадратов

Постановка задачи. Рассмотрим применение метода максимума правдоподобия при анализе множественной связи показателя с группой неслучайных компонент. Важность этой задачи обусловливается, как будет показано ниже, широким использованием метода ее решения для анализа самых разнообразных связей и зависимостей, в том числе и криволинейных, при обработке геолого-геофизической информации.

Формулировка задачи сводится к следующему. Имеем некоторую группу k неслучайных величин $X_1, X_2, \ldots, X_k$, т. е. таких, значения которых получают с пренебрежимой погрешностью, и величину η, которую, напротив, нельзя измерять точно, так что результаты ее измерений являются случайными величинами. Между η и $X_1, X_2, \ldots, X_k$ существует зависимость вида

$$\eta = \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_k X_k, \tag{4.49}$$

называемая, подобно (3.56′), функцией регрессии. Значения $X_1, X_2, \ldots, X_k$ фиксируются, после чего измеряется величина η, затем снова фиксируются некоторые значения $X_1, X_2, \ldots, X_k$, измеряется η и так эта процедура повторяется n раз. В каждом опыте взятое со знаком отклонение Δ_i результата измерения величины η от действительного значения представляет собой нормально распределенную случайную величину с нулевым математическим ожиданием. Дисперсия $\mathbf{D}\Delta_i$ может быть неизвестной, но известно, как она изменяется от опыта к опыту — во сколько раз уменьшается или увеличивается. Таким образом, наблюдения величины η могут быть равноточными и неравноточными, причем можно считать, что при i-м испытании дисперсия

$$D_i = \mathbf{D}\Delta_i = \frac{D}{p_i}, \quad i = \overline{1, n}, \tag{4.50}$$

где p_i заданы из каких-либо априорных предпосылок. В случае равноточных наблюдений дисперсии D_i постоянны: $p_i = 1$, $i = \overline{1, n}$.

Задача состоит в получении оценки функции регрессии (4.49), связывающей истинные значения η с $X_1, X_2, \ldots, X_k$. Фактически она сводится к отысканию оценок коэффициентов $\alpha_1, \alpha_2, \ldots, \alpha_k$ по наблюдениям y_i величины η ($i = \overline{1, n}$) и соответствующих им группам

фиксированных значений $\{x_{i1}, x_{i2}, \ldots, x_{ik}\}$ $(i = \overline{1, n})$ величин $X_1, X_2, \ldots, X_k$, причем

$$y_i = \sum_{j=1}^{k} \alpha_j x_{ij} + \Delta_i, \quad \mathbf{M}\Delta_i = 0, \quad \mathbf{D}\Delta_i = \frac{D}{p_i}. \tag{4.51}$$

В такой форме, в частности, можно ставить задачу оценки криволинейной зависимости между случайной и неслучайной величинами. Например, если связь между некоторым показателем η и точно измеряемым параметром X существенно криволинейна, можно использовать полиномиальную аппроксимацию: $\eta = \alpha_0 + \alpha_1 X + \alpha_2 X^2 + \ldots + \alpha_m X^m$. Тогда приняв $X_1 = 1$, $X_2 = X$, $\ldots$, $X_k = X^m$ $(k = m + 1)$, получим задачу оценки параметров α_j в сформулированной выше форме.

В другом случае, когда речь идет об оценке коэффициентов линейной связи некоторого показателя ξ с параметрами $Z_1, Z_2, \ldots, Z_m$,

$$\xi = \alpha_0 + \alpha_1 Z_1 + \alpha_2 Z_2 + \ldots + \alpha_m Z_m + \Delta,$$

(Δ — нормально распределенная случайная величина с нулевым средним), приняв $X_1 = 1$, $X_2 = Z_1$, $X_3 = Z_2$, $\ldots$, $X_k = Z_m$ $(k = m + 1)$, снова получим задачу оценки коэффициентов регрессии (4.49). Если дисперсия отклонения Δ мала, этими оценками можно воспользоваться в дальнейшем для прогноза (т. е. косвенных измерений) величины ξ по значениям $Z_1, Z_2, \ldots, Z_m$ или, в случае (4.51), η — по X с помощью функции регрессии η на X, подобно тому, как это делалось при использовании парных или множественных связей между случайными величинами. Прогнозные значения η и ξ вычисляются по формулам: $\tilde{\eta} = \sum_{j=0}^{m} \check{\alpha}_j X^j$, $\tilde{\xi} = \check{\alpha}_0 + \sum_{j=1}^{m} \check{\alpha}_j Z_j$.

Оценки коэффициентов регрессии по методу наименьших квадратов. Согласно (4.51) $\Delta_i = y_i - \sum_{j=1}^{k} \alpha_j x_{ij}$, так что величины

$$\sqrt{p_i}\,(y_i - \sum_{j=1}^{k} \alpha_j x_{ij}) = \Delta_i \sqrt{p_i} \tag{4.52}$$

распределены нормально с нулевым математическим ожиданием и одной и той же дисперсией: $\mathbf{M}(\Delta_i \sqrt{p_i}) = 0$; $\mathbf{D}(\Delta_i \sqrt{p_i}) = D$. Это значит, что нормированные по (4.52) отклонения можно рассматривать как независимые наблюдения нормально распределенной случайной величины. Логарифмическое преобразование функции правдоподобия запишется по (4.38) с учетом (4.52) в виде

$$L(\alpha_1, \alpha_2, \ldots, \alpha_k) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln D - \frac{1}{2D}\sum_{i=1}^{n} p_i (y_i - \sum_{j=1}^{k} \alpha_j x_{ij})^2. \tag{4.53}$$

Уравнения для вычисления максимально правдоподобных оценок коэффициентов α_j определяются условием максимума функции (4.53), т. е. фактически минимума формы

$$L_0 = \sum_{i=1}^{n} p_i \left(y_i - \sum_{j=1}^{k} \alpha_j x_{ij}\right)^2. \tag{4.54}$$

Иными словами, наиболее правдоподобными оценками коэффициентов α_j будут служить такие их значения, при которых сумма нормированных умножением на p_i квадратов отклонений наблюденных величин y_i от значений, предписываемых функцией (4.49), окажется наименьшей. Поэтому описываемый метод носит название *метода наименьших квадратов**.

Система уравнений для определения оценок α_j получит вид:

$$\frac{\partial L_0}{\partial \alpha_l} = 0, \quad l = \overline{1, k}.$$

Подставив вместо L_0 выражение (4.54), получим

$$\sum_{i=1}^{n} p_i \left(y_i - \sum_{j=1}^{k} \alpha_j x_{ij}\right) x_i = 0, \quad l = \overline{1, k}, \tag{4.55}$$

или

$$\begin{cases} \sum\limits_{j=1}^{k} \alpha_j m_{j1} = s_1 \\ \sum\limits_{j=1}^{k} \alpha_j m_{j2} = s_2 \\ \quad \cdots \cdots \\ \sum\limits_{j=1}^{k} \alpha_j m_{jk} = s_k, \end{cases} \tag{4.55'}$$

где $m_{jl} = \sum\limits_{i=1}^{n} x_{ij} p_i x_{il}$, $s_l = \sum\limits_{i=1}^{n} y_i p_i x_{il}$ $(j = \overline{1, k}, l = \overline{1, k})$. Полученная система уравнений линейна относительно $\alpha_1, \alpha_2, \ldots, \alpha_k$. В матричной форме решение, представляющее собой вектор искомых оценок $\check{\boldsymbol{\alpha}} = \{\check{\alpha}_1, \check{\alpha}_2, \ldots, \check{\alpha}_k\}'$, имеет вид

$$\boldsymbol{\alpha} = (\mathbf{X}'\mathbf{P}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}\mathbf{y}, \tag{4.56}$$

где $\mathbf{X}$ — $(n \times k)$-матрица вида $\mathbf{X} = \{x_{ij}\}$, $i = \overline{1, n}$, $j = \overline{1, k}$; $\mathbf{X}'$ —

* Подробное изложение метода с разнообразными приложениями и примерами можно найти в [17].

транспонированная матрица $\mathbf{X}$; $\mathbf{P}$ — диагональная матрица, по диагонали которой расположены p_i:

$$\mathbf{P} = \begin{pmatrix} p_1 & 0 & 0 & \dots & 0 \\ 0 & p_2 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & p_n \end{pmatrix};$$

$\mathbf{y}$, $\boldsymbol{\alpha}$ — векторы вида: $\mathbf{y} = \{y_1, y_2, \dots, y_n\}'$, $\boldsymbol{\alpha} = \{\check{\alpha}_1, \check{\alpha}_2, \dots, \check{\alpha}_k\}'$.

Оценки $\check{\alpha}_1, \check{\alpha}_2, \dots, \check{\alpha}_k$ несмещенны и эффективны. Несмещенность следует из того, что математическое ожидание

$$\mathbf{M}\check{\boldsymbol{\alpha}} = \mathbf{M}[(\mathbf{X}'\mathbf{P}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}\mathbf{y}] = (\mathbf{X}'\mathbf{P}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}\mathbf{X}\boldsymbol{\alpha} = \boldsymbol{\alpha}, \tag{4.57}$$

учитывая, что согласно (4.51), $\mathbf{M}\mathbf{y} = \mathbf{X}\boldsymbol{\alpha}$.

Ковариационная матрица вектора оценок $\check{\boldsymbol{\alpha}}$

$$\begin{aligned}\mathbf{M}[(\check{\boldsymbol{\alpha}} - \boldsymbol{\alpha})(\check{\boldsymbol{\alpha}} - \boldsymbol{\alpha})'] &= \mathbf{M}\{[(\mathbf{X}'\mathbf{P}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha})][\mathbf{X}'\mathbf{P}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}(\mathbf{y} - \\ &- \mathbf{X}\boldsymbol{\alpha})]'\} = \mathbf{M}\{(\mathbf{X}'\mathbf{P}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha})(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha})'\mathbf{P}\mathbf{X}(\mathbf{X}'\mathbf{P}\mathbf{X})^{-1}\} = \\ &= D(\mathbf{X}'\mathbf{P}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}\mathbf{P}^{-1}\mathbf{P}\mathbf{X}(\mathbf{X}'\mathbf{P}\mathbf{X})^{-1} = D(\mathbf{X}'\mathbf{P}\mathbf{X})^{-1}.\end{aligned} \tag{4.58}$$

Дисперсия $\check{\alpha}_j$ — оценки коэффициента при X_j регрессии (4.49) равна j-му диагональному элементу матрицы $\mathbf{C} = (\mathbf{X}'\mathbf{P}\mathbf{X})^{-1}$, умноженному на D: $\mathbf{D}\check{\alpha}_j = Dc_{jj}$. Доверительные пределы для α_j при доверительной вероятности q

$$\check{\alpha}_j - u_{\frac{1+q}{2}}\sqrt{D c_{jj}}, \quad \check{\alpha}_j + u_{\frac{1+q}{2}}\sqrt{Dc_{jj}}, \tag{4.59}$$

$u_{\frac{1+q}{2}}$ — квантиль (0; 1)-нормального распределения порядка $\frac{1+q}{2}$. Задача оценки точности косвенных измерений величины η по значениям $X_1, X_2, \dots, X_k$ решается построением доверительных пределов для значений η:

$$y_q^- = \sum_{j=1}^{k}\alpha_j X_j - u_{\frac{1+q}{2}}\sqrt{\frac{D}{p}}, \quad y_q^+ = \sum_{j=1}^{k}\alpha_j X_j + u_{\frac{1+q}{2}}\sqrt{\frac{D}{p}}, \tag{4.60}$$

где q — доверительная вероятность.

Если величина D неизвестна, вычисляется ее оценка

$$\check{D} = \frac{1}{n-k}\sum_{i=1}^{n} p_i\left(y_i - \sum_{j=1}^{k}\check{\alpha}_j x_{ij}\right)^2. \tag{4.61}$$

Как показано в [17], она несмещенная, причем $(n-k)\check{D}$ подчиняется распределению χ^2 с $n-k$ степенями свободы. Доверительные

пределы для значений прогнозируемого показателя в этом случае —

$$y_q^- = \sum_{j=1}^{k} \check{\alpha}_j X_j - t_{\frac{1+q}{2}} \sqrt{\frac{\check{D}}{p}}, \quad y_q^+ = \sum_{j=1}^{k} \check{\alpha}_j X_j + t_{\frac{1+q}{2}} \sqrt{\frac{\check{D}}{p}}, \qquad (4.61')$$

где $t_{\frac{1+q}{2}}$ — квантиль порядка $\frac{1+q}{2}$ распределения Стьюдента с $n-k$ степенями свободы (табл. 9, Приложение). При $(n-k) \geqslant 30$ не будет большой ошибки, если вместо $t_{\frac{1+q}{2}}$ использовать квантиль того же порядка (0; 1)-нормального распределения.

Подобно этому, доверительные пределы для коэффициента α с использованием того же квантиля:

$$\check{\alpha}_i - t_{\frac{1+q}{2}} \sqrt{\check{D}c_{ii}}, \ \check{\alpha}_i + t_{\frac{1+q}{2}} \sqrt{\check{D}c_{ii}}. \qquad (4.59')$$

§ 4. Расчет количества независимых наблюдений для оценки параметра с заданной точностью

Приведенные в предыдущем разделе формулы для дисперсий оценок и доверительных пределов оценок позволяют решать и обратную задачу — рассчитывать количество n независимых наблюдений для оценки параметра либо числовой характеристики с заданной точностью.

Пусть задано A_q — допустимое с вероятностью q отклонение оценки $\check{\alpha}$ от оцениваемой величины α или a_q — относительное отклонение, $a_q = \frac{A_q}{\alpha}$:

$$\mathbf{P}\{|\check{\alpha} - \alpha| < A_q\} = q, \ \mathbf{P}\{|\check{\alpha} - \alpha| < a_q \alpha\} = q. \qquad (4.62)$$

По (4.8) имеем для несмещенной асимптотически нормальной оценки $\check{\alpha}$

$$u_\beta^2 \mathbf{D}\check{\alpha} \approx A_q^2, \ \frac{u_\beta^2 \mathbf{D}\check{\alpha}}{\alpha^2} \approx a_q^2 \ \left(\beta = \frac{1+q}{2}\right). \qquad (4.63)$$

Если дисперсия оценки $\check{\alpha}$ имеет вид

$$\mathbf{D}\check{\alpha} = \frac{c_\alpha}{n}, \qquad (4.64)$$

где c_α не зависит от n, то минимальное количество наблюдений, обеспечивающее условия (4.62), определится из (4.63) приближенными равенствами

$$n \approx \frac{u_\beta^2 c_\alpha}{A_q^2}, \ n \approx \frac{u_\beta^2 c_\alpha}{\alpha^2 a_q^2}. \qquad (4.65)$$

Приближенные вычислительные формулы для отдельных параметров распределений приведены в табл. 4.1. Обозначения табл. 4.1 : A — асимметрия, E — эксцесс, u_β — β-квантиль (0,1)-нормального распределения ($\beta = (1+q)/2$).

Таблица 4.1. **Формулы расчета количества наблюдений, обеспечивающего оценку параметра с заданной точностью**

Оцениваемый параметр	Необходимое количество наблюдений	
	для максимально допустимого абсолютного отклонения A_q	для максимально допустимого относительного отклонения a_q
Среднее M	$Du_\beta^2 A_q^{-2}$	$V^2 u_\beta^2 a_q^{-2}$
Дисперсия D	$\frac{D^2 (E+2) u_\beta^2}{A_q^2}$	$\frac{(E+2) u_\beta^2}{a_q^2}$
Среднее квадратическое отклонение $\sigma = \sqrt{D}$	$\frac{D (E+2) u_\beta^2}{4A_q^2}$	$\frac{(E+2) u_\beta^2}{4a_q^2}$
Коэффициент вариации V	$\frac{[4V^4 + V^2 (E+2) - 4AV^3] u_\beta^2}{4A_q^2}$	$\frac{(4V^2 + E + 2 - 4AV) u_\beta^2}{4a_q^2}$
Критическая граница $B_\alpha = M + u_\alpha \sigma$	$\frac{D\left[1 + u_\alpha A + \frac{1}{4} u_\alpha^2 (E+2)\right] u_\beta^2}{A_q^2}$	$\frac{\left[1 + u_\alpha A + \frac{1}{4} u_\alpha^2 (E+2)\right] u_\beta^2}{(V^{-1} + u_\alpha)^2 a_q^2}$
Вероятность p события, оцениваемая частотой $\check{p} = \frac{\nu}{n}$	$\frac{u_\beta^2 p (1-p)}{A_q^2}$	$\frac{u_\beta^2 (p^{-1} - 1)}{a_q^2}$
Коэффициент корреляции r двух величин*	$\frac{u_\beta^2 (1-r^2)^2}{A_q^2}$	$\frac{u_\beta^2 (r^{-1} - r)^2}{a_q^2}$

Примечание. * В предположении нормального распределения коррелируемых величин.

Для использования приведенных в табл. 4.1 формул необходимы предварительные оценки входящих в них величин. Для получения этих оценок обычно привлекают литературные или ориентировочные данные.

В табл. 4.2 приведены количества наблюдений, рассчитанные по формуле табл. 4.1, для оценки среднего с заданными относительными отклонениями a_q при доверительной вероятности этих отклонений $q = 0{,}9$ и различных коэффициентах вариации V наблюдаемой величины.

Таблица 4.2. **Количества наблюдений, обеспечивающие оценку среднего с заданной относительной точностью (при $q = 0{,}9$).**

$a_{0,9}$%	V%				
	20	30	50	70	100
50	1	1	3	6	11
30	2	3	8	15	30
20	3	7	17	34	68
10	11	25	68	133	272
5	44	98	272	530	1080

Таблица 4.3. **Количества наблюдений, обеспечивающие оценку дисперсии с заданной относительной точностью (при $q = 0{,}9$).**

$a_{0,9}$%	E				
	−1	0	1	2	3
100	3	6	9	12	14
70	6	12	17	23	28
50	11	22	33	44	55
30	30	60	90	121	151
10	271	541	812	1080	1360

В табл. 4.3 приведены аналогичные данные для оценки дисперсии. Значения коэффициента эксцесса соответствуют определенным приближениям к наблюдаемому распределению: $E = 0$ — нормальному; $E = -1$ — уплощенному; $E = 1, 2, 3$ — островершинному.

Глава 5

ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Методы проверки статистических гипотез широко применяются при количественном анализе геолого-геофизических данных. Это обусловливается тем, что его важнейшими компонентами являются задачи сопоставления распределений, параметров и числовых характеристик последних, а также широкий круг разнообразных задач классификации, которые можно сформулировать как проверку статистических гипотез.

В этой главе мы ознакомимся с некоторыми общими вопросами теории проверки гипотез. Методы решения задач геолого-геофизических исследований, относящихся к этому разделу, будут изложены в гл. 7.

§ 1. Постановка задачи проверки гипотез

Понятие статистической гипотезы. Пусть $x_1, x_2, \ldots, x_n$ — независимые наблюдения одномерной или многомерной случайной величины ξ. Каждое предположение или совокупность непротиворечивых предположений относительно распределения ξ называют *статистической гипотезой*. Наблюдения несут информацию, позволяющую с большей или меньшей уверенностью судить о ее справедливости. Правило, по которому на основании имеющихся наблюдений принимают или отвергают гипотезу H_0, называют *критерием (тестом)* для проверки гипотезы, или просто критерием проверки гипотезы H_0. Каждая задача проверки гипотезы в конечном счете сводится к построению и применению такого критерия. *Параметрические статистические гипотезы* содержат суждения о параметрах распределений: о равенстве параметров заданным значениям либо о попадании их в определенные области значений. Проверяемое утверждение часто называют *нулевой гипотезой*. В форме статистических гипотез формулируются утверждения о равенстве между собой значений параметров распределений нескольких случайных величин, об отличии параметров на величины, не превышающие заданный уровень, и т. п.

Все гипотезы, которые могут быть выдвинуты, но отличаются от проверяемой, называют *конкурирующими*, или *альтернативными (альтернативами)*. Результат проверки гипотезы обычно не является полностью достоверным, а справедлив лишь с определенной вероятностью. Критерий фактически выделяет так называемую *критическую область*

для точек nm-мерного пространства (m — размерность ξ), координаты которых определяются всеми компонентами $\{x_1', x_2', \ldots, x_n'\}$. Если точка с координатами $\{x_1', x_2', \ldots, x_n'\}$ попадает в эту область, то гипотеза отвергается; в противном случае считают, что гипотеза справедлива. Критическую область обычно определяют с помощью специально подбираемой функции $f(x_1, x_2, \ldots, x_n)$ наблюдений неравенством $f(x_1, x_2, \ldots, x_n) \geqslant c$. Те наблюдения, для которых неравенство не выполняется, составляют область принятия гипотезы.

Ошибки I и II рода. Мощность критерия. В процедуре проверки гипотезы бывают ошибки двух родов, возможность появления которых обусловливается характером самой выдвинутой гипотезы, структурой критерия, недостаточным количеством наблюдений.

Ошибка I рода — отклонение гипотезы по результату ее проверки критерием в том случае, когда она верна. *Ошибка II рода* — ошибочное принятие гипотезы, когда в действительности она неверна. Критерий характеризуют вероятности ошибок обоих родов, причем в зависимости от решаемой задачи подбирается оптимальное соотношение этих вероятностей. Пусть, например, гипотеза, состоящая в том, что исследуемый район рудоносен, проверяется по его геолого-геофизическим данным некоторым критерием. Ошибка I рода — вывод о том, что район нерудоносен, когда он в действительности рудоносен, а ошибка II рода — вывод о рудоносности, когда в действительности это не так. В этой задаче критерий следует строить так, чтобы обеспечить малую вероятность α_{01} ошибки I рода, хотя бы и за счет некоторого увеличения вероятности α_{10} ошибки II рода.

Гипотеза называется *простой*, если она состоит в том, что случайная величина ξ распределена с заданной плотностью распределения $p_0(x)$, т. е. простая гипотеза однозначно определяет плотность распределения; в противном случае гипотеза называется *сложной*. Если $p(x, \theta) = p(x, \theta_1, \theta_2, \ldots, \theta_k)$ — параметрическое семейство плотностей распределения, простой будет гипотеза $\theta_i = \theta_i^0$, $i = \overline{1, k}$ (θ_i^0 — заданные значения параметров). В случае проверки простой гипотезы вероятность α_{01} ошибки I рода единственна — она равна вероятности отклонения нулевой гипотезы, когда та в действительности верна. Эту вероятность часто называют *уровнем значимости* критической области критерия. Если альтернатива — простая гипотеза, вероятность α_{10} ошибки II рода также определяется единственным образом. В случае сложной альтернативы вероятность ошибки II рода будет функцией простых гипотез, которые в совокупности составляют альтернативу.

Часто для характеристики критерия используют его *мощность* $q_{10} = 1 - \alpha_{10}$, имеющую смысл вероятности отклонения гипотезы, когда та и в действительности неверна. Мощность характеризует чувствительность критерия, его способность отличать альтернативу от нулевой гипотезы. Как и вероятность ошибки II рода, в случае сложной альтернативы мощность является функцией простых гипотез, составляющих в совокупности альтернативу. Для «близких» к нулевой гипотезе альтернатив, т. е. таких, при которых распределение наблюдений мало отличается от гипотетического, мощность характеризуется небольшими,

близкими к нулю величинами. Случается, однако, что она оказывается малой и для некоторых «далеких» альтернатив, т. е. критерий обеспечивает лишь уверенное отклонение гипотезы.

Пусть, например, необходимо по данным каротажа скважины определить характер насыщенности пласта осадочных пород — проверить гипотезу о том, что пласт нефтенасыщенный. Если рассматриваются глинистые пласты, то для альтернативы «пласт водонасыщенный» характерны распределения геофизических параметров, относительно близкие к их распределениям при нулевой гипотезе. Поэтому мощность критерия при такой альтернативе будет сравнительно невелика. Если же рассматриваются не глинистые пласты, то при отделении среди них водонасыщенных от нефтенасыщенных мощность критерия будет выше за счет более выраженных различий распределений параметров.

Оптимальный критерий для простой гипотезы относительно простой альтернативы называется *наиболее мощным;* он обеспечивает минимальную вероятность ошибки II рода. *Равномерно наиболее мощным* называется наиболее мощный критерий относительно всех допустимых альтернатив. Надо отметить, что такие критерии не всегда существуют.

С расширением критической области критерия вероятность ошибки I рода не может уменьшиться, а вероятность ошибки II рода не увеличивается. Более того, расширение критической области обычно влечет за собой увеличение вероятности ошибки I рода и уменьшение вероятности ошибки II рода. Если критическая область критерия определяется условием общего вида $f(x_1, x_2, \ldots, x_n) \geqslant c$ и $c_1 \geqslant c_2$, то вероятность α'_{01} ошибки I рода критерия $\{f(x_1, x_2, \ldots, x_n) \geqslant c_1\}$ не превысит вероятности α''_{01} ошибки I рода критерия $\{f(x_1, x_2, \ldots, x_n) \geqslant c_2$, а вероятность ошибки II рода $\alpha'_{10} \geqslant \alpha''_{10}$. Это следует из того, что событие $\{f \geqslant c_2\}$ включает в себя событие $\{f \geqslant c_1\}$ и по (1.12) при нулевой гипотезе $\alpha'_{01} = \mathbf{P}_0\{f \geqslant c_1\} \leqslant \mathbf{P}_0\{f \geqslant c_2\} = \alpha''_{01}$; аналогично $\alpha'_{10} = \mathbf{P}_1\{f < c_1\} \geqslant \mathbf{P}_1\{f < c_2\} = \alpha''_{10}$ при любой альтернативе H_1. Поэтому обычно при каждой фиксированной альтернативе вероятности ошибок I и II рода нельзя сделать одновременно меньшими произвольного малого положительного числа за счет выбора критической области. Если при ее расширении α_{01} и α_{10} изменяются непрерывно в интервале (0, 1), то величина $\max(\alpha_{01}, \alpha_{10})$ достигает минимума, когда $\alpha_{01} = \alpha_{10} = \alpha$, и не может быть меньшей α. Безошибочное принятие или отклонение гипотезы возможно лишь в том тривиальном случае, когда для каждой критической области $\alpha_{01} = 0$ либо $\alpha_{10} = 0$, причем можно указать такую область, где $\alpha_{10} = \alpha_{01} = 0$. Критическую область критерия в каждой конкретной задаче определяют с учетом соотношения между вероятностями ошибок I и II рода так, чтобы по возможности обеспечить минимальный риск (минимальные потери) при принятии решения.

Рассмотрим следующий пример. Пусть $x_1, x_2, \ldots, x_n$ — результаты измерений показателя в n пробах; $y_1, y_2, \ldots, y_n$ — результаты таких же измерений этих проб, сделанных на другом приборе. Необходимо проверить гипотезу об отсутствии систематического смещения между показаниями приборов. Подобная задача возникает и тогда, когда необходимо проверить стабильность показаний одного прибора в больших промежутках времени, отсутствие «сползания» системы отсчета его показаний

при продолжительной работе. В этом случае в качестве наблюдений $y_1, y_2, \ldots, y_n$ следует взять результаты измерений тех же проб на том же приборе через длительный промежуток времени.

Несмещенной оценкой смещения τ между первой и второй сериями служит величина

$$\check{\tau} = \frac{1}{n} \sum_{i=1}^{n} (x_i - y_i), \tag{5.1}$$

так как ее математическое ожидание

$$\mathbf{M}\check{\tau} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{M}x_i - \mathbf{M}y_i) = \frac{n\tau}{n} = \tau. \tag{5.2}$$

Оценка $\check{\tau}$ состоятельна — это следует из того, что дисперсия

$$\mathbf{D}\check{\tau} = \mathbf{M}(\check{\tau} - \tau)^2 = \frac{1}{n^2} \sum_{i=1}^{n} \mathbf{M}(x_i - y_i - \tau)^2 = \frac{1}{n^2} \sum_{i=1}^{n} \mathbf{D}(x_i - y_i) = \frac{2D}{n} \tag{5.3}$$

стремится к нулю при $n \to \infty$, и неравенства Чебышева. В формуле (5.3) D — дисперсия ошибок измерений в одной серии *(дисперсия воспроизводимости)*, которая предполагается одинаковой для обеих серий и не зависящей от измеряемой величины. Последнее означает, что средняя абсолютная ошибка не зависит от измеряемых значений в пределах, охватывающих обе серии.

Пользуясь центральной предельной теоремой, можно доказать, что оценка $\check{\tau}$ асимптотически нормальна. Если нулевая гипотеза об отсутствии смещения ($\tau = 0$) справедлива, $\check{\tau}$ распределяется асимптотически нормально с нулевым математическим ожиданием и дисперсией $\frac{2D}{n}$. Пределы, в которых с вероятностью q должна находиться величина $\check{\tau}$ при этой гипотезе, приближенно (см. стр. 44)

$$-u_{\frac{1+q}{2}} \sqrt{\frac{2D}{n}}, \; u_{\frac{1+q}{2}} \sqrt{\frac{2D}{n}}, \tag{5.4}$$

где $u_{\frac{1+q}{2}} - \frac{1+q}{2}$-квантиль (0; 1)-нормального распределения. Если справедлива нулевая гипотеза, вероятность события

$$|\check{\tau}| \geqslant u_{\frac{1+q}{2}} \sqrt{\frac{2D}{n}}. \tag{5.5}$$

составит $\sim (1 - q)$, так что при значениях q, близких к единице, оно будет маловероятным.

Описанные свойства оценки $\check{\tau}$ определяют критерий для проверки гипотезы $\tau = 0$. Критическую область уровня значимости $\alpha_{01} = 1 - q$

образуют все возможные серии $\{x_i\}$, $\{y_i\}$ $(i=\overline{1, n})$, для которых выполняется (5.5). Если $|\overset{*}{\tau}| \geqslant u_{\frac{1+q}{2}} \sqrt{\frac{2D}{n}}$, гипотеза об отсутствии смещения отвергается с вероятностью ошибки $\alpha_{01} \approx 1-q$.

Альтернативами по отношению к проверяемой гипотезе $\tau = 0$ могут быть различные смещения $\tau \neq 0$. Вероятность ошибки II рода при уровне значимости $\alpha_{01} = 1-q$ в данном случае будет функцией τ : $\alpha_{10} = \alpha_{10}(\tau)$, если, конечно, дисперсия воспроизводимости D не изменяется. В соответствии с приведенным выше определением вероятность ошибки II рода $\alpha_{10}(\tau)$ при альтернативном смещении τ имеет смысл

$$\alpha_{10}(\tau) = \mathbf{P}\left\{-u_{\frac{1+q}{2}} \sqrt{\frac{2D}{n}} < \overset{*}{\tau}_\tau < u_{\frac{1+q}{2}} \sqrt{\frac{2D}{n}}\right\}, \tag{5.6}$$

где $\overset{*}{\tau}_\tau$ — случайная величина, распределенная как $\overset{*}{\tau}$ (5.1) при наличии систематического смещения $\tau \neq 0$ между сериями $\{x_i\}$ и $\{y_i\}$. С учетом (5.2) и (5.3) имеем, по (2.30),

$$\alpha_{10}(\tau) = \Phi\left(\frac{u_{\frac{1+q}{2}} \sqrt{\frac{2D}{n}} - \tau}{\sqrt{\frac{2D}{n}}}\right) - \Phi\left(\frac{-u_{\frac{1+q}{2}} \sqrt{\frac{2D}{n}} - \tau}{\sqrt{\frac{2D}{n}}}\right) =$$

$$= \Phi\left(u_{\frac{1+q}{2}} - \frac{\tau}{\sqrt{\frac{2D}{n}}}\right) - \Phi\left(-u_{\frac{1+q}{2}} - \frac{\tau}{\sqrt{\frac{2D}{n}}}\right) \tag{5.7}$$

($\Phi(z)$ — функция (0; 1)-нормального распределения). Формула (5.7) показывает, что мощность критерия (5.5) зависит от величины $\tau \Big/ \sqrt{\frac{2D}{n}}$, причем, как легко убедиться, при фиксированных вероятности ошибки I рода, дисперсии воспроизводимости и количестве n наблюдений в каждой серии вероятность ошибки II рода уменьшается с ростом τ. Эту зависимость иллюстрирует табл. 5.1, которая содержит вычисленные по (5.7) с помощью табл. 2 (Приложение) вероятности ошибок II рода при $q = 0{,}95$ и $q = 0{,}9$. Этим значениям q соответствуют вероятности ошибок I рода 0,05 и 0,1.

Таблица 5.1. **Вероятности ошибок II рода в зависимости от величины** $\tau \Big/ \sqrt{\frac{2D}{n}}$

$\tau \Big/ \sqrt{\frac{2D}{n}}$	0	0,1	0,3	0,5	1,0	1,5	2	2,5	3	4	10
$\alpha_{10}(\tau)$ $(\alpha_{01}=0{,}05)$	0,95	0,949	0,94	0,921	0,830	0,677	0,484	0,295	0,149	0,021	0,000
$\alpha_{10}(\tau)$ $(\alpha_{01}=0{,}1)$	0,9	0,898	0,885	0,858	0,735	0,557	0,361	0,197	0,088	0,009	0,000

Данные табл. 5.1 показывают, что при альтернативах, близких к нулевой гипотезе, т. е. при $\tau \approx 0$, вероятность ошибки II рода становится значительной. С приближением τ к нулю она стремится к $1 - \alpha_{01} = q$. Это значит, что употребление критерия имеет смысл лишь для проверки гипотез при относительно далеких альтернативах. В данном случае мерой различия нулевой и альтернативной гипотез служит отношение $\tau \Big/ \sqrt{\frac{2D}{n}}$. Чем меньше возможное смещение τ, тем большее количество независимых наблюдений необходимо иметь для успешного использования критерия. Увеличение количества наблюдений — естественный, а часто и единственный способ повышения эффективности вообще подавляющего большинства статистических критериев.

В рассматриваемом примере вероятность ошибки II рода зависит и от дисперсии воспроизводимости D: чем меньше последняя, тем меньшее систематическое смещение удается обнаружить при прочих равных условиях.

С увеличением критической области для τ, при фиксированных D и n, увеличивается уровень ее значимости — вероятность ошибки I рода и уменьшается вероятность ошибки II рода. Это наглядно видно из сравнения данных табл. 5.1 при $\alpha_{01} = 0{,}05$ и $\alpha_{01} = 0{,}1$. Те же данные показывают, что вероятность ошибки II рода достигает уровня значимости при довольно большой по сравнению с $\sqrt{\frac{D}{n}}$ величиной τ. При уровне значимости 0,05 это происходит, когда $\tau \Big/ \sqrt{\frac{2D}{n}} \approx 3{,}6$.

§ 2. Критерий Неймана — Пирсона

Для проверки простой гипотезы с простой альтернативой существует наиболее мощный *критерий Неймана — Пирсона*, или *критерий отношения правдоподобия*, который при каждой фиксированной вероятности ошибки I рода и одном и том же объеме выборки обеспечивает минимальную вероятность ошибки II рода. Пусть нулевая гипотеза состоит в том, что случайная величина ξ, независимые наблюдения которой — $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, подчиняется распределению с плотностью $p_0(\mathbf{x})$, а альтернатива — в том, что плотностью распределения ξ является $p_1(\mathbf{x})$. По этому критерию нулевая гипотеза принимается, если

$$\prod_{i=1}^{n} p_0(\mathbf{x}_i) \Big/ \prod_{i=1}^{n} p_1(\mathbf{x}_i) > c(\alpha_{01}), \tag{5.8}$$

где $c(\alpha_{01})$ — предел, выбираемый в зависимости от уровня значимости α_{01}. При выполнении противоположного неравенства принимается гипотеза о распределении ξ с плотностью $p_1(\mathbf{x})$. Такой критерий вполне естественен. При нулевой гипотезе плотность совместного распределения $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ будет $\prod_{i=1}^{n} p_0(\mathbf{x}_i^0)$, где $\mathbf{x}_i^0$ — аргументы плотности. То, что

величина $\prod_{i=1}^{n} p_0(\mathbf{x}_i)$ больше, чем $\prod_{i=1}^{n} p_1(\mathbf{x}_i)$, свидетельствует в пользу нулевой гипотезы.

Обычно используют эквивалентную модификацию критерия (5.8):

$$\sum_{i=1}^{n} \ln p_0(\mathbf{x}_i) - \sum_{i=1}^{n} \ln p_1(\mathbf{x}_i) > \ln c(\alpha_{01}). \tag{5.8'}$$

Предел $c(\alpha_{01})$ можно выбирать также и с учетом *цен ошибок* обоих родов, т. е. потерь, связанных с этими ошибками — так, чтобы обеспечить их наименьшую среднюю величину.

Подобно (5.8) можно построить критерий для проверки простых гипотез при сложных альтернативах. Пусть плотность распределения случайной величины ξ зависит от группы параметров $\beta = \{\beta_1, \beta_2, \ldots, \beta_k\}'$. Необходимо по наблюдениям $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ проверить гипотезу $\beta = \beta_0$ о том, что значения параметров равны заданным величинам $\{\beta_{10}, \beta_{20}, \ldots, \beta_{k0}\} = \beta_0$. При каждой альтернативной гипотезе $\beta_i = \beta_{i1}\,(i = \overline{1, k})$, т. е. $\beta = \beta_1$, критерий будет иметь вид (5.8′):

$$\sum_{i=1}^{n} \ln p(\mathbf{x}_i, \beta_0) > c_1(\alpha_{01}) + \sum_{i=1}^{n} \ln p(\mathbf{x}_i, \beta_1).$$

Взяв постоянную c такую, чтобы при нулевой гипотезе

$$\mathbf{P}\left\{\sum_{i=1}^{n} \ln p(\mathbf{x}_i, \beta_0) \leqslant c\right\} = \alpha_{01}, \tag{5.9}$$

получим искомый критерий: если

$$\sum_{i=1}^{n} \ln p(\mathbf{x}_i, \beta_0) \leqslant c, \tag{5.10}$$

нулевая гипотеза отвергается с вероятностью ошибки α_{01}. Следует учесть, что уверенное принятие нулевой гипотезы возможно лишь при таких альтернативах, для которых вероятность выполнения условия (5.10) близка к единице. Это нужно обязательно принимать во внимание в тех случаях, когда по критерию принимается нулевая гипотеза.

Рассмотрим пример построения критерия вида (5.10). Пусть ξ — нормально распределенная случайная величина с дисперсией D; $x_1, x_2, \ldots, x_n$ — ее независимые наблюдения; необходимо проверить гипотезу о том, что математическое ожидание ξ равно M. Имеем

$$\sum_{i=1}^{n} \ln p(x_i, M) = -\frac{n}{2} \ln(2\pi D) - \frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - M)^2}{D}. \tag{5.11}$$

Отбросив в (5.11) часть, не зависящую от x_i, получим критерий (5.10): $-\frac{1}{2}\sum_{i=1}^{n} \frac{(x_i - M)^2}{D} \leqslant c$ или $\sum_{i=1}^{n} \frac{(x_i - M)^2}{D} \geqslant c_1$. Величину c_1 определяем по

условию (5.9) с учетом того, что при нулевой гипотезе $\sum_{i=1}^{n} \frac{(x_i - M)^2}{D}$ распределяется по закону χ^2 (2.52′) с n степенями свободы: $c_1 = \chi^2_q(n)$ — квантиль распределения χ^2 с n степенями свободы порядка $q = 1 - \alpha_{01}$; α_{01} — вероятность ошибки I рода. Нулевая гипотеза отвергается с вероятностью ошибки α_{01}, если

$$\sum_{i=1}^{n} \frac{(x_i - M)^2}{D} \geqslant \chi^2_q(n).$$

§ 3. Задачи проверки гипотез в геолого-геофизических исследованиях

Методы проверки гипотез используются в геолого-геофизических исследованиях для решения задач следующих основных типов.

1. Сопоставление параметров и числовых характеристик распределений геолого-геофизических показателей, функций и плотностей распределения.

2. Классификация объектов наблюдений (по комплексу показателей и по отдельным показателям).

3. Проверка гипотез о законах распределения.

4. Проверка гипотез о статистических связях показателей.

5. Выделение из выборок аномальных наблюдений.

Надо сказать, что выделение названных типов задач до некоторой степени условно, так как некоторые вопросы могут рассматриваться как задача одного типа, так и другого, а иные могут вообще не принадлежать ни к одному из них. В большей мере выделенные типы отражают специфику задач со стороны метода решения.

Перейдем к краткой характеристике перечисленных типов задач.

Сопоставление распределений по параметрам и числовым характеристикам. Задача сравнения параметров обычно формулируется в виде проверки гипотез об их равенстве по оценкам, построенным по независимым наблюдениям сравниваемых случайных величин. Один из простейших способов состоит в следующем. Пусть $\check{a}_1$, $\check{a}_2$ — несмещенные оценки параметра a, вычисленные по выборкам независимых наблюдений сравниваемых величин ξ_1, ξ_2 соответственно; n_1 и n_2 — количества наблюдений в выборках. Нулевая гипотеза состоит в том, что значение параметра для обеих величин одно и то же: $\mathbf{M}\check{a}_1 = \mathbf{M}\check{a}_2$. Если она справедлива, разность $\check{a}_1 - \check{a}_2$ будет иметь своим математическим ожиданием нуль: $\mathbf{M}(\check{a}_1 - \check{a}_2) = 0$. Обозначим $\mathbf{D}\check{a}_1$, $\mathbf{D}\check{a}_2$ — дисперсии оценок $\check{a}_1$, $\check{a}_2$ при нулевой гипотезе и будем считать, что $\mathbf{D}\check{a}_1 \to 0$ при $n_1 \to \infty$, $\mathbf{D}\check{a}_2 \to 0$ при $n_2 \to \infty$. Дисперсия разности $\check{a}_1 - \check{a}_2$ с учетом независимости $\check{a}_1$ и $\check{a}_2$ составит

$$\mathbf{D}(\check{a}_1 - \check{a}_2) = \mathbf{D}\check{a}_1 + \mathbf{D}\check{a}_2 \to 0 \text{ при } n_1 \to \infty,\ n_2 \to \infty. \qquad (5.12)$$

Пусть распределения $\check{a}_1$ и $\check{a}_2$, по крайней мере, асимптотически нормальны — такой вывод обычно делают на основании центральной

предельной теоремы или теоремы об оценках наибольшего правдоподобия. Критерий выглядит так: нулевая гипотеза должна быть отвергнута, если

$$|\check{a}_1 - \check{a}_2| \geqslant u_{\frac{1+q}{2}} \sqrt{\mathbf{D}\check{a}_1 + \mathbf{D}\check{a}_2} \tag{5.13}$$

$\left(u_{\frac{1+q}{2}} - \frac{1+q}{2}\right.$ - квантиль (0; 1)-нормального распределения$\left.\right)$. Вероятность того, что этот вывод окажется ошибочным (вероятность ошибки I рода), $\alpha_{01} \approx 1 - q$, так как событие (5.13) при нулевой гипотезе может произойти с вероятностью $\sim (1 - q)$.

Чем меньше сумма дисперсий $\mathbf{D}\check{a}_1 + \mathbf{D}\check{a}_2$, тем мощнее критерий — тем более близкие к нулевой гипотезе альтернативы он будет способен отличать. Мощность критерия можно повысить за счет увеличения n_1 и n_2, что вытекает из (5.12).

Иногда возникает задача проверки гипотезы о равенстве параметра a распределения заданному значению a_0. Обычно для ее решения используют состоятельную несмещенную оценку $\check{a}$. Гипотеза $\mathbf{M}\check{a} = a_0$ отвергается с вероятностью ошибки $\alpha_{01} = 1 - q$, если оценка выходит из пределов, в которых она должна находиться с вероятностью q при справедливости нулевой гипотезы.

В случае асимптотической нормальности $\check{a}$ критерий будет выглядеть так: если

$$|\check{a} - a_0| \geqslant u_{\frac{1+q}{2}} \sqrt{\mathbf{D}\check{a}}, \tag{5.14}$$

гипотеза отвергается с вероятностью ошибки* $\alpha_{01} \approx 1 - q$. По такой схеме можно построить, например, критерий для проверки гипотезы о нормальном распределении случайной величины по ее выборочным коэффициентам асимметрии и эксцесса, используя для вычисления дисперсий формулы (4.30), (4.30′) и принимая во внимание, что нулевой гипотезе соответствуют нулевые значения коэффициентов асимметрии и эксцесса. Подобно этому можно построить пределы для выборочного коэффициента корреляции двух нормально распределенных случайных величин, предполагая, что последние независимы. Для этого используют выражение (4.34) дисперсии выборочного коэффициента корреляции, приняв $r = 0$. Сравнение выборочного коэффициента корреляции с этими пределами, т. е. проверка условия

$$|\check{r}| < u_{\frac{1+q}{2}} \sqrt{\frac{1}{n}} \tag{5.15}$$

даст приближенный критерий для проверки гипотезы о равенстве коэффициента корреляции нулю.

---

* Количество наблюдений, конечно, предполагается здесь, как и в (5.13), достаточно большим, чтобы обеспечить возможность аппроксимации распределения $\check{a}$ нормальным законом.

В математической статистике разработан ряд специальных критериев для сопоставления отдельных числовых характеристик распределений. Кроме того, существует группа методов сравнения не параметров, а самих функций распределения — критерии для проверки гипотез о тождественности функций распределения. С ними мы познакомимся в следующих главах.

Задача классификации. Общая постановка задачи такова. Имеем t классов Q_j, $j = \overline{1, t}$, каждому из которых соответствует определенное распределение некоторой, в общем случае, m-мерной случайной величины $\boldsymbol{\xi}$. Компонентами ее обычно бывают результаты измерений геофизических параметров, содержаний химических элементов или минералов в образцах горных пород, их физических свойств и т. п. Каждому классу Q_j соответствует своя плотность распределения $\boldsymbol{\xi}$

$$p_j(\mathbf{x}) = p_j(x_1, x_2, \ldots, x_m), \quad j = \overline{1, t}. \tag{5.16}$$

Дано также наблюдение $\mathbf{y} = \{y_1, y_2, \ldots, y_m\}'$ величины $\boldsymbol{\xi}$, принадлежащее к одному из классов Q_j, причем неизвестно, какому именно. Задача состоит в определении класса, к которому принадлежит вектор $\mathbf{y}$, используя плотности (5.16) или их оценки $\check{p}_j(\mathbf{x})$. Фактически она сводится к выяснению того, какому распределению наиболее соответствует $\mathbf{y}$ и последующему определению вероятного класса.

Если речь идет о классификации лишь на два класса ($t = 2$), причем обе плотности $p_1(\mathbf{x})$ и $p_2(\mathbf{x})$ известны, задача решается по критерию Неймана — Пирсона (5.8), причем этот критерий будет оптимальным, так как имеет наибольшую мощность. Поскольку число испытываемых наблюдений $n = 1$, критерий приобретает следующий вид. Гипотеза о принадлежности $\mathbf{y}$ к классу Q_1, т. е. о распределении $\mathbf{y}$ с плотностью $p_1(\mathbf{x})$, принимается, если

$$\frac{p_1(\mathbf{y})}{p_2(\mathbf{y})} > c(\alpha_{01}). \tag{5.17}$$

В случае выполнения противоположного неравенства принимается гипотеза о принадлежности $\mathbf{y}$ к классу Q_2.

Такой критерий можно получить и из формулы Бейеса (1.14). Пусть A_1 — гипотеза о принадлежности $\mathbf{y}$ к классу Q_1, A_2 — к классу Q_2. Так как $\mathbf{y}$ к одному из двух классов обязательно принадлежит, A_1 и A_2 составляют полную группу событий. Построим малый m-мерный параллелепипед с центром в точке $\mathbf{y}$:

$$\Omega(\mathbf{y}) = \{\Delta_1(y_1), \Delta_2(y_2), \ldots, \Delta_m(y_m)\}, \tag{5.18}$$

где $\Delta_i(y_i)$ — его проекции на соответствующие координатные оси, представляющие собой отрезки малых длин δ_i со срединами y_i. Пусть событие B — попадание $\boldsymbol{\xi}$ в $\Omega(\mathbf{y})$; $\mathbf{P}(A_1) = q_1$, $\mathbf{P}(A_2) = q_2$ — априорные вероятности гипотез A_1, A_2, выражающие исходную информацию о принадлежности к классам Q_1, Q_2 без учета той ее доли, которую содержит само наблюдение $\mathbf{y}$. Если такой информации нет (о принадлежности объекта наблюдения заранее ничего сказать нельзя), априорные вероят-

ности считают равными. В этом случае $q_1 = q_2 = \frac{1}{2}$, поскольку $q_1 + q_2 = 1$.

По формуле Бейеса и используя (3.26″), вероятность того, что у принадлежит Q_j

$$\mathbf{P}\{\mathbf{y} \in Q_j\} = \mathbf{P}(A_j/B) = \frac{\mathbf{P}(A_j)\,\mathbf{P}(B/A_j)}{\sum_{j=1}^{2} \mathbf{P}(A_j)\,\mathbf{P}(B/A_j)} = \frac{q_j \mathbf{P}\{\xi \in \Omega(\mathbf{y})/\xi \in Q_j\}}{\sum_{j=1}^{2} q_j \mathbf{P}\{\xi \in \Omega(\mathbf{y})/\xi \in Q_j\}} \approx$$

$$\approx \frac{q_j p_j(\mathbf{y})\,\delta_1\delta_2 \ldots \delta_m}{\sum_{j=1}^{2} q_j p_j(\mathbf{y})\,\delta_1\delta_2 \ldots \delta_m} = \frac{q_j p_j(\mathbf{y})}{q_1 p_1(\mathbf{y}) + q_2 p_2(\mathbf{y})} \quad (j = 1, 2). \tag{5.19}$$

Наиболее вероятный класс определяется максимальной из вероятностей $\mathbf{P}\{\mathbf{y} \in Q_j\}$, сумма которых равна единице. Другими словами, если

$$\frac{q_1 p_1(\mathbf{y})}{q_1 p_1(\mathbf{y}) + q_2 p_2(\mathbf{y})} > \frac{1}{2}, \tag{5.20}$$

принимается гипотеза A_1 (класс Q_1); в противном случае — гипотеза A_2 (класс Q_2). Условие (5.20) эквивалентно неравенству

$$\frac{p_1(\mathbf{y})}{p_2(\mathbf{y})} > \frac{q_2}{q_1} \tag{5.21}$$

и, таким образом, приходим к решающему правилу вида (5.17).

Аналогично случаю двух классов с помощью формулы Бейеса строится критерий классификации при числе классов более двух. Принадлежность у определяется максимальной вероятностью вида

$$\mathbf{P}\{\mathbf{y} \in Q_j\} = \frac{q_j p_j(\mathbf{y})}{\sum_{j=1}^{t} q_j p_j(\mathbf{y})} \quad (j = \overline{1, t}), \tag{5.22}$$

где $p_j(\mathbf{x})$ — плотность m-мерного распределения ξ в классе Q_j; q_j — априорная вероятность того, что у принадлежит Q_j.

В случае, когда априорной информации нет, т. е. ни одному из классов заранее нельзя отдать предпочтения, критерий приобретает простейший вид. Наиболее вероятный класс определяется максимальной апостериорной вероятностью (5.22) при $q_j = \frac{1}{t}$:

$$\mathbf{P}\{\mathbf{y} \in Q_j\} = \frac{p_j(\mathbf{y})}{\sum_{j=1}^{t} p_j(\mathbf{y})} \tag{5.23}$$

т. е. просто максимальной из величин $p_j(\mathbf{y})$ $(j = \overline{1, t})$. Доказано следующее важное свойство оптимальности бейесовских критериев классификации: средняя величина вероятностей ошибок при использовании правила Бейеса минимальна.

Приведенные критерии не учитывают *цен ошибок* — потерь, связанных с ошибочным принятием или отклонением гипотез. Метод Бейеса обобщается на тот случай, когда цены таких ошибок заданы. Пусть при классификации на два класса, Q_1 и Q_2, c_{21} — величина потерь при ошибочном отнесении $\mathbf{y}$ в класс Q_1 (в то время, когда в действительности $\mathbf{y}$ из Q_2); c_{12} — потери при ошибочном принятии гипотезы $\{\mathbf{y} \in Q_2\}$. Критерий, при котором математическое ожидание потерь минимально, имеет вид [1]: если

$$\frac{p_1(\mathbf{y})}{p_2(\mathbf{y})} > \frac{c_{21}q_2}{c_{12}q_1}, \tag{5.24}$$

то $\mathbf{y}$ относят в класс Q_1; в случае противоположного неравенства — в Q_2.

Если, например, класс Q_1 — «рудоносные участки», Q_2 — «безрудные участки», $\mathbf{y}$ — вектор геолого-геофизических показателей, характеризующих отдельный участок, то критический уровень для $\frac{p_1(\mathbf{y})}{p_2(\mathbf{y})}$ должен быть меньше, чем предел $\frac{q_2}{q_1}$ при правиле (5.21), так как потери c_{12} при ошибочном признании рудоносного участка безрудным превышают потери c_{21} при противоположной ошибке. Конечно, используя правило (5.24), мы будем чаще ошибаться, признавая безрудные участки рудоносными — за счет уменьшения количества противоположных ошибок.

Критерий классификации на несколько классов ($t > 2$), при котором математическое ожидание потерь минимально, имеет вид: если

$$\sum_{i=1,\ i\neq k}^{t} q_i c_{ik} p_i(\mathbf{y}) < \sum_{i=1,\ i\neq j}^{t} q_i c_{ij} p_i(\mathbf{y}) \quad (j = \overline{1,t},\ j \neq k), \tag{5.24'}$$

принимается гипотеза $\{\mathbf{y} \in Q_k\}$. В (5.24′) c_{ij} — цена ошибочного отнесения $\mathbf{y}$ в Q_j, тогда как в действительности $\mathbf{y}$ принадлежит к Q_i.

Учет цен ошибок при решении задач классификации необходим тогда, когда потери, связанные с ошибками отнесения к различным классам, резко неравнозначны — как в рассмотренном выше примере. В других случаях можно ограничиться правилами (5.22) или (5.23). Первое совпадает с критерием (5.24′) при одинаковых c_{ij}, второе — еще и при одинаковых q_j.

Такой подход приводит к эффективному решению задачи классификации сразу группы независимых наблюдений $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\} = \mathbf{Y}$. Если $q_1, q_2, \ldots, q_t$ — априорные вероятности того, что эта группа принадлежит к $Q_1, Q_2, \ldots, Q_t$ соответственно, то

$$\mathbf{P}\{\mathbf{Y} \in Q_j\} = \frac{q_j \prod_{i=1}^{n} p_j(\mathbf{y}_i)}{\sum_{l=1}^{t} q_l \prod_{i=1}^{n} p_l(\mathbf{y}_i)}, \tag{5.25}$$

так как плотность $p_j(\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_n')$ совместного распределения $\{\mathbf{y}_1', \mathbf{y}_2', \ldots, \mathbf{y}_n'\}'$ равна произведению $\prod_{i=1}^{n} p_j(\mathbf{x}_i)$.

Применение описанных критериев часто затрудняется тем, что нельзя дать точные выражения плотностей $p_j(\mathbf{x})$. Обычно вместо $p_j(\mathbf{x})$ используют их оценки, построенные по выборкам наблюдений, заведомо принадлежащих к классам Q_j.

Правила классификации по дискретным распределениям строятся аналогично, с той разницей, что вместо плотностей используются вероятности, соответствующие полученному наблюдению $\mathbf{y}$ при каждой гипотезе.

Задача проверки гипотез о законах распределения. Употребляемые для решения этой задачи *критерии согласия* обычно применяют к одномерным случайным величинам. Исходные данные представляют собой выборку n независимых наблюдений $x_1, x_2, \ldots, x_n$ случайной величины ξ. Нулевая гипотеза состоит в том, что распределение ξ подчиняется определенному закону, т. е. его плотность принадлежит заданному параметрическому семейству $p(x, \alpha_1, \alpha_2, \ldots, \alpha_k)$. Альтернативами служат какие-либо другие параметрические семейства плотностей распределения. Впрочем, часто задача ставится без указания альтернатив. Тогда речь идет фактически об оценке степени соответствия имеющихся наблюдений гипотетическому закону распределения. Оценка плотности имеет вид $p(x, \check{\alpha}_1, \check{\alpha}_2, \ldots, \check{\alpha}_k)$, где $\check{\alpha}_1, \check{\alpha}_2, \ldots, \check{\alpha}_k$ — оценки параметров. Положительный результат проверки гипотезы служит обоснованием для использования такой оценки в дальнейшем анализе.

Проверка гипотез о статистических связях. Один из основных вопросов, решаемых при анализе связей количественных показателей — оценка степени связанности показателей. Часто возникает задача, формулируемая следующим образом: по наблюдениям показателей $\xi, \xi_1, \xi_2, \ldots, \xi_k$ проверить гипотезу о наличии связи между ξ и группой $\xi_1, \xi_2, \ldots, \xi_k$. Обычно выделяют задачу о связи двух показателей ($k = 1$); для ее решения существует ряд различных статистических критериев, на которых мы остановимся в дальнейшем. Другая задача, с которой иногда имеют дело, — сопоставление форм связей (проверка гипотез об идентичности функций регрессии).

Задача выделения аномальных наблюдений. По своему содержанию задача близка к проблеме классификации, однако отличается от нее специфическими особенностями. Она состоит в разделении смешанной совокупности, в общем случае, m-мерных наблюдений $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$, каждое из которых может быть аномальным или фоновым. Плотность распределения случайной величины $\boldsymbol{\xi}$, наблюдениями которой являются $\mathbf{y}_i$, можно представить в виде

$$p(\mathbf{x}) = \alpha u(\mathbf{x}) + (1 - \alpha) t(\mathbf{x}), \tag{5.26}$$

где α — доля фоновой совокупности; $u(\mathbf{x})$, $t(\mathbf{x})$ — плотности распределения фоновой и аномальной совокупностей. Формула (5.26) следует

из того, что для малого m-мерного параллелепипеда $\Omega(\mathbf{x})$ с центром $\mathbf{x}$ и проекциями на координатные оси $\Delta_1, \Delta_2, \ldots, \Delta_m$, имеющими длины, соответственно $\delta_1, \delta_2, \ldots, \delta_m$, в предположении непрерывности $p(x)$, $u(x)$ и $t(x)$:

$$\mathbf{P}\{\xi \in \Omega(\mathbf{x})\} = (p(\mathbf{x}) + O(\rho))\delta_1\delta_2 \ldots \delta_m = \alpha \mathbf{P}_u\{\xi \in \Omega(\mathbf{x})\} + (1-\alpha)\mathbf{P}_t\{\xi \in \Omega(\mathbf{x})\} = (\alpha u(\mathbf{x}) + (1-\alpha)t(\mathbf{x}) + O(\rho))\delta_1\delta_2 \ldots \delta_m,$$

где $\rho = \sqrt{\sum_{i=1}^{m} \delta_i^2}$; $\mathbf{P}_u\{\xi \in \Omega(\mathbf{x})\}$, $\mathbf{P}_t\{\xi \in \Omega(\mathbf{x})\}$ — вероятности попадания ξ в $\Omega(\mathbf{x})$ при фоновом и аномальном распределениях. Разделив на $\prod_{i=1}^{m} \delta_i$ и перейдя к пределу при $\rho \to 0$, получим (5.26).

Если известны плотности $u(\mathbf{x})$ и $t(\mathbf{x})$, задача сводится к проверке простой гипотезы, которую надо делать для каждого y_i. Однако на практике, как правило, невозможно указать точный вид функций $u(\mathbf{x})$ и $t(\mathbf{x})$, а для $t(\mathbf{x})$ часто трудно указать хотя бы удовлетворительную оценку. Кроме того, обычно неизвестна доля α фоновых наблюдений в выборке.

Наиболее полное решение задачи состоит в вычислении для каждого y_i вероятностей принадлежности к каждой совокупности. Для одномерных наблюдений y_i обычно достаточно определить предел A такой, что при $y_i > A$ (или, наоборот, $y_i < A$) наблюдение классифицируется как аномальное с плотностью распределения $t(x)$, а в случае противоположного неравенства — как фоновое, с плотностью $u(x)$. Решающее правило строят так, чтобы обеспечить наилучшее соотношение вероятностей ошибок I и II рода.

Глава 6

КОЛИЧЕСТВЕННАЯ ХАРАКТЕРИСТИКА РАСПРЕДЕЛЕНИЙ ГЕОЛОГО-ГЕОФИЗИЧЕСКИХ ПОКАЗАТЕЛЕЙ

На основании общности методов решения среди задач статистического анализа геолого-геофизической информации можно выделить следующие типы.

1. Количественная характеристика одномерных распределений показателей, включая проверку гипотез о законе распределения, оценку плотностей распределения, числовых характеристик и их доверительных интервалов.

2. Анализ связей показателей и косвенные измерения.

3. Сопоставление геологических объектов по распределениям геолого-геофизических показателей и числовым характеристикам их распределений.

4. Задачи классификации по комплексу показателей.

Попутно с решением этих типовых задач нередко возникает необходимость в решении ряда других, также относящихся к сфере применения статистических методов. Таковы, например: выделение аномальных наблюдений выборки; анализ распределений ошибок измерений и учет их при статистической обработке данных; оценка влияния различных факторов на распределения показателей; построение рациональных способов опробования; оценка *трендов* — функциональных составляющих геохимических и физических полей и др. Перечисленные вопросы будут рассматриваться в этой и последующих главах.

§ 1. Статистические критерии однородности упорядоченных наблюдений

Методы математической статистики как одномерных, так и многомерных случайных величин, как правило, предусматривают статистическую независимость наблюдений, по которым решается та или иная задача. При оценке параметров представляемый для статистического анализа материал предполагается заранее разделенным по выборкам независимых наблюдений. Предполагается также, что наблюдения из одной и той же выборки одинаково распределены, т. е. являются *статистически однородными*. Статистическую однородность необходимо обеспечивать однородностью геологических (структурных, морфологических) признаков, однотипностью минерального состава пород, а независимость — достаточной удаленностью пунктов наблюдений друг от друга в пределах части объекта, относящейся к одной и той же выборке, независимостью ошибок измерений.

Конечно, такое разделение исходного материала — в виде выборок независимых наблюдений — создает упрощенное представление о строении поля значений показателей. Фактически предполагается, что исследуемый объект можно разделить на участки, в которых поле значений показателей в указанном выше смысле однородно. Это предположение представляется не слишком обременительным — распределения показателей в пределах обоснованно выбранных областей можно считать однородными хотя бы приближенно. С другой стороны, одним из основных принципов геологического исследования является классификационный — разделение материала по типам пород, фациальным разновидностям, по степени и характеру измененности, глубине залегания и т. д.

В математической статистике разработаны критерии для проверки гипотезы об однородности результатов измерений, объединенных в выборку. Ориентировочное представление можно получить уже по виду плотности распределения. Если она имеет два и более максимума, то распределение, очевидно, неоднородно и исследуемая совокупность представляет собой смесь нескольких однородных совокупностей. Плотность распределения такой совокупности имеет вид $p(x) = \sum_{j=1}^{t} \alpha_j p_j(x)$, где $p_j(x)$ — плотности отдельных однородных распределений, α_j — доли этих распределений в общей совокупности $\left(\alpha_j \geqslant 0,\ \sum_{j=1}^{t} \alpha_j = 1\right)$. Если

каждое из них предполагается нормальным, однородность совокупности можно проверить одним из критериев согласия с этим законом.

В математической статистике известны критерии для проверки гипотезы о наличии закономерной тенденции в ряду наблюдений, построенном в соответствии с некоторым правилом упорядочения. Мы рассмотрим здесь наиболее употребительные из них.

Критерий Аббе. В соответствии с этим критерием гипотеза об отсутствии закономерной тенденции в ряду наблюдений $\xi_1, \xi_2, \ldots, \xi_n$ отвергается с вероятностью ошибки, близкой к $1-q$, если не выполняется условие

$$1+u_{\frac{1+q}{2}}\sqrt{\frac{n-2}{(n-1)(n+1)}}>\frac{g^2}{\check{D}}>1-u_{\frac{1+q}{2}}\sqrt{\frac{n-2}{(n-1)(n+1)}}, \quad (6.1)$$

где n — количество наблюдений; $g^2=\frac{1}{2(n-1)}\sum_{i=1}^{n-1}(\xi_{i+1}-\xi_i)^2$; $\check{D}=$

$=\frac{1}{n-1}\sum_{i=1}^{n}\left(\xi_i-\frac{1}{n}\sum_{i=1}^{n}\xi_i\right)^2$ — оценка дисперсии; $u_{\frac{1+q}{2}}$ — квантиль порядка $\frac{1+q}{2}$ (0; 1)-нормального распределения*. Распределение ξ_i при нулевой гипотезе (об однородности) предполагается нормальным.

Если тенденция высокочастотного изменения не может быть альтернативой, используется односторонний критерий, т. е. проверяется условие

$$\frac{g^2}{\check{D}}\leqslant 1-u_q\sqrt{\frac{n-2}{(n-1)(n+1)}}. \quad (6.1')$$

Нулевая гипотеза (об однородности) отвергается с вероятностью ошибки $\sim(1-q)$, если это условие выполняется. Аналогично, если альтернативой является тенденция высокочастотного изменения (например, при чередовании пластов), следует проверять условие

$$\frac{g^2}{\check{D}}\geqslant 1+u_q\sqrt{\frac{n-2}{(n-1)(n+1)}}. \quad (6.1'')$$

Упорядочение наблюдений, т. е. выбор номера для каждого измерения, обычно производится сообразно с расположением пунктов наблюдений по профилю или скважине.

Использование статистики g^2 в этом критерии основывается на следующем. При постоянном математическом ожидании, $\mathbf{M}\xi_i=m$,

---

* Ввиду использования статистики g^2 этот метод иногда называют *методом последовательных квадратов.*

$$\mathbf{M}g^2 = \frac{1}{2(n-1)}\mathbf{M}\sum_{i=1}^{n-1}(\xi_{i+1}-\xi_i)^2 = \frac{1}{2(n-1)}\mathbf{M}\sum_{i=1}^{n-1}(\xi_{i+1}-m-\xi_i+m)^2 =$$

$$= \frac{1}{2(n-1)}\mathbf{M}\sum_{i=1}^{n-1}[(\xi_{i+1}-m)^2 - 2(\xi_{i+1}-m)(\xi_i-m)+(\xi_i-m)^2] =$$

$$= \frac{n-1}{2(n-1)}(2D-2r_{ii+1}D) = D - r_{ii+1}D, \qquad (6.2)$$

где r_{ii+1} — коэффициент корреляции между ξ_{i+1} и ξ_i. Таким образом, в случае взаимной независимости ξ_i математическое ожидание отноше-

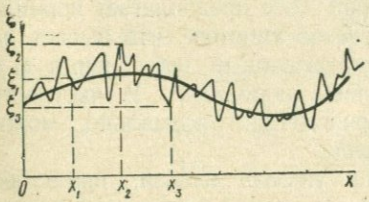


Рис. 24.

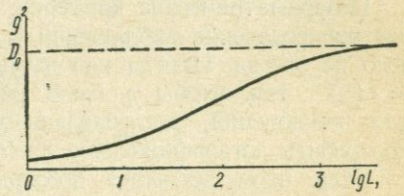


Рис. 25.

ния $\frac{g^2}{D}$ близко к единице. Если же ξ_i зависимы, что влечет за собой, как правило, коррелированность наблюдений, т. е. наблюдается так называемая *автокорреляция*, причем $r_{ii+1} > 0$, то среднее отношения $\frac{g^2}{D}$ будет меньше единицы, причем тем меньше, чем сильнее выражена автокорреляция наблюдений. Точно так же математическое ожидание $\frac{g^2}{D}$ будет меньше единицы при неоднородности, выраженной в виде плавной тенденции (рис. 24) либо редких скачкообразных изменений математического ожидания — причем тем меньше, чем более проявлена эта тенденция на фоне случайных колебаний.

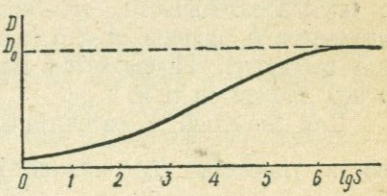


Рис. 26.

Вид полученной экспериментально усредненной зависимости g^2 от расстояния L в *м* между пунктами опробования массива (по содержаниям химических элементов) изображен на рис. 25. Величине D_0, отмеченной на оси ординат, равна дисперсия показателя по всему массиву. Зависимости g^2 от L соответствует и зависимость дисперсии от площади опробования S в $м^2$, изображенная на рис. 26.

Аналогично (6.1) можно построить критерий для проверки однородности по комплексу показателей $\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(m)}$. Если при гипотезе однородности последние предполагаются нормально распределенными

и независимыми, то приближенный односторонний критерий состоит в проверке условия

$$\frac{1}{m}\sum_{j=1}^{m}\frac{g_j^2}{\check{D}_j} \leqslant 1 - u_q\sqrt{\frac{n-2}{m(n-1)(n+1)}},$$

а двухсторонний $-\sum_{j=1}^{m}\left(\frac{g_j^2}{\check{D}_j}-1\right)^2 \geqslant \frac{n-2}{n^2-1}\chi_q^2(m)$, где $\chi_q^2(m)$ — q-квантиль распределения χ^2 с m степенями свободы; g_j^2 и $\check{D}_j$ вычисляются по упорядоченным наблюдениям показателей $\xi^{(j)}$ $(j=\overline{1, m})$.

Непараметрические критерии. Критерий Аббе предполагает нормальное распределение наблюдений ξ_i при нулевой гипотезе, что бывает далеко не всегда. Иногда используют преобразованные наблюдения $\eta_i = f(\xi_i)$ с тем, чтобы η_i были распределены нормально. В частности, для наблюдений, распределенных логарифмически нормально, можно употребить логарифмическое преобразование.

Если нормализующее преобразование указать нельзя, применяют непараметрические критерии, в которых не используется информация о законе распределения. К ним относится *серийный* критерий [2, 9]. Общая схема его применения такова. Предполагается, что результатами наблюдений могут быть элементы только двух типов: a и b. Применительно к рассматриваемой задаче эти элементы могут быть определены так: a — наблюдения, меньшие некоторого уровня A (например, близкого к медиане); b — наблюдения, большие этого уровня или равные ему. После упорядочивания наблюдений по определенному признаку последовательность элементов образует серии элементов одинакового типа: *aabaaabb...abb*. Пусть γ — общее количество серий в последовательности, m — количество элементов a, n — количество элементов b, причем $m \leqslant n$ (в противном случае элементы обозначаются наоборот). Проверяется гипотеза H_0 о случайном расположении в ней элементов a и b.

Для значений γ рассчитаны пределы в предположении справедливости гипотезы H_0: нижние $g\left(\frac{\alpha}{2}, m, n\right)$ и верхние $G\left(\frac{\alpha}{2}, m, n\right)$, соответствующие малым уровням значимости α. Гипотеза H_0 принимается, если справедливо неравенство

$$g\left(\frac{\alpha}{2}, m, n\right) < \gamma < G\left(\frac{\alpha}{2}, m, n\right). \qquad (6.3)$$

В противном случае она отвергается с вероятностью ошибки $\alpha_{01} \leqslant \alpha$.

Если альтернативой может быть только плавная тенденция в изменении значений показателя с резкими скачками в небольшом числе границ, то достаточно проверить лишь условие

$$\gamma > g(\alpha, m, n), \qquad (6.4)$$

так как альтернативе будут соответствовать «слишком длинные» серии. Если это условие не выполняется, гипотеза об однородности отвер-

гается с вероятностью ошибки, не превышающей α. Если наоборот, альтернативой является только тенденция высокочастотного изменения значений показателя, то следует проверять условие

$$\gamma < G(\alpha, m, n) \tag{6.5}$$

(этой альтернативе соответствуют «слишком короткие» серии). Если неравенство (6.5) не выполняется, то гипотеза об однородности отвергается с вероятностью ошибки, не превышающей α.

В табл. 4 (Приложение) даны значения $g\left(\frac{\alpha}{2}, m, n\right)$ и $G\left(\frac{\alpha}{2}, m, n\right)$ для $\alpha = 0{,}10,\ 0{,}05,\ 0{,}02,\ 0{,}01$; $m = 2, 3, 4, \ldots, 20$ и $n = m, m+1, \ldots, 20$, а также для $n = m = 20, 21, 22, \ldots, 100$.

При $n > 20$ и отношении $\frac{m}{n}$, близком к единице, используется нормальная аппроксимация распределения γ:

$$g(\alpha, m, n) \approx [\mathbf{M}\gamma - u_{1-\alpha}\sqrt{\mathbf{D}\gamma} - 0{,}5],$$
$$G(\alpha, m, n) \approx [\mathbf{M}\gamma + u_{1-\alpha}\sqrt{\mathbf{D}\gamma} + 0{,}5], \tag{6.6}$$

где $[y]$ — целая часть y, $u_{1-\alpha}$ — $(1-\alpha)$-квантиль $(0; 1)$-нормального распределения;

$$\mathbf{M}\gamma = 1 + \frac{2mn}{m+n}, \quad \mathbf{D}\gamma = \frac{2mn\,(2mn - m - n)}{(m+n)^2\,(m+n-1)} = \frac{(\mathbf{M}\gamma - 1)\,(2mn - m - n)}{(m+n)\,(m+n-1)}.$$

Существует и другой тип серийного критерия, основанный на сравнении длины максимальных серий с критическими границами, рассчитанными при гипотезе однородности. Этот тип используется в тех случаях, когда предполагаемая неоднородность может выражаться в виде протяженных областей повышенных или пониженных значений показателя.

Обозначим R_{1k} — число серий элементов a длины, большей или равной k, R_{2k} — число серий элементов b длины, большей или равной k, m и n — общие количества элементов a и b. При больших k и $m = n$ для описания распределения R_{1k} и R_{2k} можно с успехом использовать закон Пуассона [9], с параметром $\lambda = \frac{n+m}{2^{k+1}}$.

Если $m \neq n$, то в качестве параметра λ можно использовать: для серий a

$$\lambda_1 = \frac{m\,(m-1)\ldots(m-k+1)\,(n+1)}{N\,(N-1)\ldots(N-k+1)} \quad (N = m+n);$$

для серий b

$$\lambda_2 = \frac{n\,(n-1)\ldots(n-k+1)\,(m+1)}{N\,(N-1)\ldots(N-k+1)}.$$

В соответствии с распределением Пуассона, при $m \neq n$

$$\mathbf{P}\{R_{1k} = \nu\} = \frac{\lambda_1^{\nu}}{\nu!}\,e^{-\lambda_1}, \quad \mathbf{P}\{R_{2k} = \nu\} = \frac{\lambda_2^{\nu}}{\nu!}\,e^{-\lambda_2};$$

при $m = n$

$$\mathbf{P}\{R_{1\kappa} = \nu\} = \mathbf{P}\{R_{2k} = \nu\} = \frac{\lambda^\nu}{\nu!} e^{-\lambda}.$$

Критическую границу уровня значимости α для длины наибольших серий элементов a и b устанавливают из неравенств

$$\mathbf{P}\{R_{ik} \geqslant 1\} \leqslant \alpha \ (i = 1, 2), \tag{6.7}$$

учитывая, что

$$\mathbf{P}\{R_{ik} \geqslant 1\} = 1 - \mathbf{P}\{R_{ik} = 0\}. \tag{6.7'}$$

В частности, при $m = n$ имеем:

$$1 - \exp\left(-\frac{m+n}{2^{k+1}}\right) \leqslant \alpha, \tag{6.8}$$

откуда

$$k \geqslant k_\alpha = \frac{1}{\lg 2} \lg\left[-\frac{m+n}{\ln(1-\alpha)}\right] - 1 \approx \{\lg(n+m) - \\ - \lg[-\ln(1-\alpha)]\}\, 3{,}32 - 1. \tag{6.9}$$

Значение $[k_\alpha]$ можно взять в качестве верхней критической границы уровня значимости α для длины наибольшей серии элементов b. Гипотеза об однородности (отсутствии положительных сигналов) отвергается с вероятностью ошибки, не превышающей α, если длина этой серии R_2 больше $[k_\alpha]$:

$$R_2 > [k_\alpha]. \tag{6.10}$$

Гипотеза об отсутствии отрицательных сигналов отвергается с вероятностью ошибки, не большей α, если

$$R_1 > [k_\alpha] \tag{6.10'}$$

(R_1 — длина наибольшей серии элементов a). Здесь элементы a — наблюдения, меньшие медианы, а элементы b — наблюдения, большие медианы или равные ей.

Для $\alpha = 0{,}05$ величина $k_\alpha = k_{0,05}$ составит $k_{0,05} \approx 3{,}32 \lg(n+m) + 3{,}28$, а для $\alpha = 0{,}10$, $k_{0,10} \approx 3{,}32 \lg(n+m) + 2{,}24 = k_{0,05} - 1{,}04$. Значения $k_{0,05}$ и $k_{0,10}$ даны в табл. 5 (Приложение).

В случае $n \neq m$ критические границы $k_1(\alpha)$ и $k_2(\alpha)$ для R_{1k} и R_{2k} вычисляются из неравенств

$$1 - e^{-\lambda_1} \leqslant \alpha, \quad 1 - e^{-\lambda_2} \leqslant \alpha, \tag{6.11}$$

которые можно решить табулированием функций $1 - e^{-\lambda_i}$ по k.

Описанные методы применяются для анализа наблюдений, упорядоченных в ряд. Приведенный ниже критерий однородности χ^2 позволяет обойтись без такого упорядочения. Сущность критерия состоит в следующем. Объект, на котором получено n наблюдений, разделяют на k компактных областей R_j. Интервал значений показателя ξ, в свою очередь, разбивают на r интервалов Δ_i (k и r могут быть и одинаковыми). Далее, вычисляют: n_{ij} — количества наблюдений, попавших

одновременно в Δ_i и R_j; n_i — количества наблюдений в Δ_i и m_j — количества наблюдений в R_j $(i=\overline{1,r},\ j=\overline{1,k})$.

При статистической однородности наблюдений математическое ожидание n_{ij}, $\mathbf{M}n_{ij}=m_j\mathbf{P}\{\xi\in\Delta_i\}\approx\frac{n_im_j}{n}$. Статистика

$$\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{k}\left(n_{ij}-\frac{n_im_j}{n}\right)^2\frac{n}{n_im_j}=\sum_{i=1}^{r}\sum_{j=1}^{k}\frac{n_{ij}^2n}{n_im_j}-n \qquad (6.12)$$

подчиняется приближенно распределению χ^2 с $(k-1)(r-1)$ степенями свободы. Гипотеза об однородности отвергается с вероятностью ошибки $\alpha\approx 1-q$, если

$$\chi^2\geqslant\chi_q^2(N), \qquad (6.13)$$

где $\chi_q^2(N)$ — q-квантиль распределения χ^2 с $N=(k-1)(r-1)$ степенями свободы, определяемый по табл. 3 (Приложение).

Чтобы обеспечить возможность аппроксимации распределения (6.12) распределением χ^2, для каждой пары R_j, Δ_i нужно выполнить требование минимальной представительности: $\frac{n_im_j}{n}>5$. Отсюда видно, что этот критерий требует значительного увеличения количества наблюдений при увеличении k и r. Следует также иметь в виду, что критерий слабо учитывает неоднородность, проявляющуюся в областях, меньших R_j.

Отметим, что если области R_j выделяются на основе каких-либо геологических предпосылок, для решения задачи можно применить различные методы сопоставления распределений и, в частности, дисперсионный анализ. С ними мы познакомимся в гл. 7.

Описанный критерий применяется и для проверки гипотезы о независимости двух показателей. Области значений каждого показателя разбивают на интервалы R_j и Δ_i соответственно; расчет производят по (6.12) с использованием *таблицы сопряженности признаков*, которую образуют значения n_{ij}.

Пример 6.1. По данным измерений содержаний SiO_2 в 40 пробах гранитов, отобранных по профилю, вычислены следующие величины: $g^2=\frac{1}{2\cdot 39}\sum_{i=1}^{39}(\xi_{i+1}-\xi_i)^2=6{,}5\,(\%)^2$, оценка среднего $\bar{\xi}=\frac{1}{40}\sum_{i=1}^{40}\xi_i=68\%$, оценка дисперсии $\check{D}=\frac{1}{39}\sum_{i=1}^{40}(\xi_i-68)^2=10{,}5\,(\%)^2$. Проверить гипотезу об однородности критерием Аббе, полагая альтернативой наличие резкостных границ либо тенденцию плавного изменения средних содержаний SiO_2.

Решение. Взяв в качестве уровня значимости $\alpha=0{,}05$ $(q=0{,}95)$, имеем

$$1-u_q\sqrt{\frac{n-2}{(n-1)(n+1)}}\approx 1-1{,}645\sqrt{\frac{38}{39\cdot 41}}\approx 0{,}75$$

и так как

$$\frac{g^2}{D} \approx 0{,}62 < 0{,}75.$$

гипотеза об однородности отвергается с вероятностью ошибки 0,05.

Пример 6.2. После измерения методом полуколичественного спектрального анализа содержаний титана в тридцати образцах, упорядоченных по близости друг к другу пунктов опробования, получены следующие результаты.

№ п/п	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Ti, 10^{-1}%	25	25	32	40	32	40	63	40	32	25	12	20	12	25	32
№ п/п	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Ti, 10^{-1}%	32	40	12	12	20	40	25	25	15	25	32	32	25	25	15

Используя серийный критерий, проверить гипотезу об однородности этого ряда при уровне значимости $\alpha = 0{,}05$, предполагая в качестве альтернативы плавную тенденцию изменения содержания Ti.

Решение. Используем в качестве уровня A для определения элементов a и b значение, разделяющее ряд упорядоченных по возрастанию наблюдений на две части, количества наблюдений в которых наименее различны. Таким будет любое значение, находящееся между $25 \cdot 10^{-1}$% и $32 \cdot 10^{-1}$%, например $A = 28 \cdot 10^{-1}$%. Количество наблюдений, меньших $28 \cdot 10^{-1}$%, равно 17, а больших $28 \cdot 10^{-1}$% — 13 ($m = 13$, $n = 17$). Элементы a определим как наблюдения, большие 28×10^{-1}%, а элементы b — как наблюдения, меньшие $28 \cdot 10^{-1}$%. Исходному ряду наблюдений будет соответствовать последовательность

$$\underbrace{bb}\underbrace{aaaaaaa}\underbrace{bbbbb}\underbrace{aaa}\underbrace{bbb}\ \underbrace{a}\ \underbrace{bbb}\underbrace{aa}\underbrace{bbb}.$$

Количество серий $\gamma = 9$. По табл. 4 (Приложение) для $m = 13$, $n = 17$, $2\alpha = 0{,}10$ находим $g(\alpha, m, n) = 10$. Так как не выполняется условие (6.4), гипотеза об однородности отвергается с вероятностью ошибки $\alpha = 0{,}05$.

Пример 6.3. По данным измерения общей радиоактивности горных пород по профилю в 40 точках введены элементы a как значения, меньшие медианы, и b как значения, большие или равные медиане. Количество элементов a $m = 20$, элементов b $n = 20$. Длина наибольшей серии элементов b равна шести. Проверить гипотезу об отсутствии положительного сигнала (т. е. протяженной области повышенных значений общей радиоактивности) при уровне значимости $\alpha = 0{,}05$.

Решение. По табл. 5 (Приложение) находим $k_{0,05} = 8{,}60$. Проверка условия (6.12) показывает, что данные не противоречат гипотезе об однородности (при уровне значимости 0,05).

§ 2. Оценка функций и плотностей распределения

Для проведения статистического анализа составляются выборки независимых наблюдений. Эта подготовка материалов проводится на основании геологического описания с тем, чтобы обеспечить однородность каждой генеральной совокупности, представленной соответствующей выборкой. При этом те или иные статистические критерии однородности должны служить лишь инструментом для проверки правильности

имеющихся представлений о поведении показателей в выделенных группах наблюдений. Впрочем, в дальнейшем мы познакомимся со статистическими методами классификации по комплексу показателей.

Необходимо также иметь в виду, что статистические выводы, получаемые по выборкам, правомерно распространять лишь на представленные ими совокупности. Например, если метод измерений обладает значительной дисперсией ошибок D_0, то оценка дисперсии D, полученная по результатам измерений показателя пород некоторого района, оценивает сумму дисперсии истинных значений показателя и дисперсии ошибок измерений. Поэтому D нельзя интерпретировать как дисперсию самого показателя, а в качестве ее оценки следует взять $D - D_0$, причем ею уже нельзя характеризовать, например, породы региона.

Исходные данные для статистического анализа (выборка) обычно представляются в виде таблицы. Координаты пунктов наблюдений приводятся, если они известны для каждого пункта и их планируется в дальнейшем использовать. Результаты измерений x_{ij} показателей $\xi_1, \xi_2, \ldots, \xi_m$ заносят в соответствующие столбцы таблицы. При статистическом анализе распределений информация о точном положении пунктов наблюдений часто не используется, в этом случае в таблицу заносят только результаты измерений.

Форма представления данных для статистического анализа

№ п/п	Координаты	ξ_1	ξ_2	...	ξ_m
1	x_1, y_1	x_{11}	x_{12}	...	x_{1m}
2	x_2, y_2	x_{21}	x_{22}	...	x_{2m}
⋮	⋮	⋮	⋮	⋮	⋮
n	x_n, y_n	x_{n1}	x_{n2}	...	x_{2m}

Статистический анализ количественных показателей горных пород включает:

1) проверку гипотез о законе распределения каждого показателя, аппроксимацию распределений теми или иными законами, построение оценок функций и плотностей распределения;

2) вычисление оценок числовых характеристик и параметров распределений;

3) определение точности оценок (доверительных пределов);

4) сопоставление распределений с помощью статистических критериев;

5) оценку формы и силы парных связей показателей, сопоставление характеристик связей;

6) анализ множественных связей.

В соответствии с характером проводимого исследования могут быть поставлены и специальные задачи: классификация объектов по комплексу показателей; расчет количества наблюдений для оценки того или иного параметра с заданной точностью; построение тренда — функции, описывающей зависимость математического ожидания от координат; анализ распределений ошибок измерений и их учет при интерпретации результатов статистической обработки; выбор информативных признаков для тех или иных совокупностей; корреляция данных по профилям и др. На этих вопросах мы остановимся ниже.

Выборочные функция и плотность распределения. Пусть x_1, x_2, ..., x_n — независимые наблюдения показателя ξ, расположенные в порядке возрастания, $F(x)$ — функция распределения ξ. Очевидно, вероятности $\mathbf{P}\{\xi < x\}$ при фиксированных x можно оценивать в виде

$$\check{\mathbf{P}}\{\xi < x\} = \frac{n(x)}{n} = F_n(x), \tag{6.14}$$

где $n(x)$ — количество наблюдений x_i, меньших x. Эта возможность определяется законом больших чисел: частота появления события $\{\xi < x\}$ в n независимых испытаниях стремится (сходится по вероятности) к вероятности этого события $\mathbf{P}\{\xi < x\} = F(x)$: $\lim\limits_{n\to\infty} \mathbf{P}\{|F_n(x) - F(x)| > \varepsilon\} \to 0$ для произвольного $\varepsilon > 0$. Более того, как утверж-

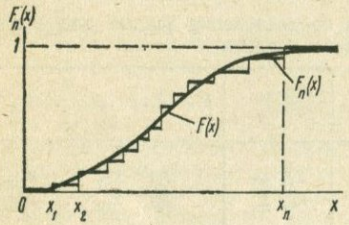


Рис. 27.

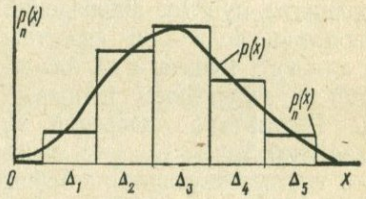


Рис. 28.

дает *теорема Гливенко*, $\lim\limits_{n\to\infty} \mathbf{P}\{\sup\limits_x |F_n(x) - F(x)| > \varepsilon\} = 0$. Таким образом, имея достаточно большое количество независимых наблюдений, можно добиться того, что $F_n(x)$ будет как угодно близка к функции распределения $F(x)$. Это служит обоснованием для использования $F_n(x)$ в качестве оценки функции распределения (рис. 27). Функцию $F_n(x)$ называют *выборочной функцией распределения*, иногда *функцией накопленных частот*.

Для построения *выборочной плотности распределения* отрезок числовой оси, содержащий значения x_1, x_2, ..., x_n, делят на несколько (k) интервалов * Δ_i. Далее, подсчитывают количества n_i наблюдений, попавших в интервалы Δ_i, вычисляют для каждого $\check{p}_i = \frac{n_i}{n\delta_i}$ (δ_i — длина Δ_i) и строят ступенчатую функцию

$$p_n(x) = \frac{n_i}{n\delta_i}, \quad (x \in \Delta_i, \quad i = \overline{1, k}). \tag{6.15}$$

Отношение $\frac{n_i}{n}$ служит оценкой вероятности p_i попадания ξ в i-й

* Разбиение на интервалы может быть неравномерным, однако для удобства построения обычно используют одинаковые интервалы.

интервал. Так как значение плотности $p(x)$ в некоторой точке x можно приближенно представить в виде

$$p(x) = \frac{dF(x)}{dx} \approx \frac{F\left(x + \frac{\delta_i}{2}\right) - F\left(x - \frac{\delta_i}{2}\right)}{\delta_i} \approx \frac{\check{p}_i}{\delta_i}, \qquad (6.16)$$

отношение $\frac{\check{p}_i}{\delta_i}$ может служить оценкой значения плотности распределения в точке x i-го интервала. Полученная описанным способом кусочно-постоянная функция $p_n(x)$ (рис. 28) называется *гистограммой*, или выборочной плотностью распределения. Соединив точки гистограммы, соответствующие серединам интервалов, получим *полигон частот*, иногда употребляемый также в качестве оценки плотности распределения.

Гистограмма обладает тем свойством, что для произвольного $\varepsilon > 0$

$$\lim_{\max \delta_i \to 0} \lim_{n \to \infty} \mathbf{P}\{|p_n(x) - p(x)| > \varepsilon\} = 0. \qquad (6.17)$$

Надо отметить, что даже при одних и тех же наблюдениях гистограмма определяется не единственным образом: ее вид изменяется в зависимости от выбора интервалов Δ_i. Нельзя указать и теоретически строгих обоснований по наилучшему выбору величин и расположению интервалов Δ_i. Однако существуют практические рекомендации, которыми обычно руководствуются при построении гистограмм. При равной длине интервалов обычно определяют их так, чтобы в каждый, за исключением, возможно, крайних попадало не менее пяти наблюдений, причем общее количество интервалов должно быть не менее шести-семи. Иногда число интервалов определяют по формуле Стерджесса: $k_0 = [k]$, где $k = 1 + 3{,}32 \lg n$, $[k]$ — округленное до целого числа значение k.

Параметрическая оценка плотности распределения. Если принята гипотеза о виде (законе) распределения показателя, т. е. известно параметрическое семейство функций, к которому принадлежит плотность распределения, то задача ее оценки сводится к вычислению, по возможности, наилучших оценок параметров. Найденные оценки определяют оценку плотности распределения. Например, если распределение показателя логарифмически нормально, то, в соответствии с (2.38), оценкой плотности распределения будет

$$\check{p}(x) = \frac{1}{\sqrt{2\pi}\,\check{\delta}x} \exp\left[-\frac{1}{2}\frac{(\ln x - \check{\mu})^2}{\check{\delta}^2}\right] (x > 0),$$

где $\check{\mu}$ и $\check{\delta}^2$ — оценки математического ожидания и дисперсии $\ln \xi$.

Для применения подобного метода необходима проверка гипотезы о виде распределения. Эта задача решается с помощью соответствующих статистических правил — критериев согласия.

§ 3. Проверка гипотез о виде распределения

Критерий Пирсона. Наиболее употребительным критерием согласия является *критерий Пирсона*, или, как его еще называют, *критерий* χ^2 *(«хи-квадрат»)*. Сущность его состоит в следующем. Пусть $p(x, \theta_1, \theta_2, \ldots, \theta_k)$ — семейство плотностей распределения; $\theta_1, \theta_2, \ldots, \theta_k$ — неизвестные независимые параметры; $x_1, x_2, \ldots, x_n$ — независимые наблюдения показателя ξ. Необходимо проверить гипотезу о соответствии распределения ξ семейству $p(x, \theta_1, \theta_2, \ldots, \theta_k)$. Для параметров $\theta_1, \theta_2, \ldots, \theta_k$ предварительно находят оценки $\check{\theta}_1, \check{\theta}_2, \ldots, \check{\theta}_k$, которыми определяются оценки плотности $p(x, \check{\theta}_1, \check{\theta}_2, \ldots, \check{\theta}_k)$ и функции распределения $F(x, \check{\theta}_1, \check{\theta}_2, \ldots, \check{\theta}_k) = \int\limits_{-\infty}^{x} p(t, \check{\theta}_1, \check{\theta}_2, \ldots, \check{\theta}_k)dt =$ $= \check{F}(x)$.

Далее, область значений, принимаемых показателем, разбивается на интервалы $\Delta_1, \Delta_2, \ldots, \Delta_m$ (как при построении гистограммы) границами $x_0^0, x_1^0, \ldots, x_m^0 (m > k+1)$, полагая $x_0^0 = -\infty$, $x_m^0 = \infty$. С помощью функции $\check{F}(x)$ определяются оценки теоретических вероятностей попадания ξ в интервалы Δ_i

$$\check{p}_i = F(x_i^0, \check{\theta}_1, \check{\theta}_2, \ldots, \check{\theta}_k) - F(x_{i-1}^0, \check{\theta}_1, \check{\theta}_2, \ldots, \check{\theta}_k) \qquad (6.18)$$

и вычисляется

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - n\check{p}_i)^2}{n\check{p}_i} = \sum_{i=1}^{m} \frac{n_i^2}{n\check{p}_i} - n, \qquad (6.19)$$

где n_i — число наблюдений из интервала Δ_i $(i = \overline{1, m})$.

При справедливости нулевой гипотезы (соответствия наблюдений $x_1, x_2, \ldots, x_n$ плотности распределения вида $p(x, \theta_1, \theta_2, \ldots, \theta_k)$) эта величина подчиняется приближенно распределению χ^2 с $m-k-1$ степенями свободы. Приближение удовлетворительно при $n\check{p}_i > 5$ $(i = 1, m)$. Нулевая гипотеза отвергается с вероятностью ошибки $\alpha \approx 1 - q$, если вычисленная величина χ^2 «слишком велика», т. е. выполняется условие

$$\chi^2 \geqslant \chi_q^2(m-k-1), \qquad (6.20)$$

где $\chi_q^2(m-k-1)$ — q-квантиль распределения χ^2 с $m-k-1$ степенями свободы.

Табл. 3 значений $\chi_q^2(m)$ дана в Приложении.

Если нулевой гипотезой является нормальный закон распределения, то оценки $\check{p}_i$ вычисляются с помощью табл. 2 (Приложение) значений функции (0; 1)-нормального распределения. Для этого вычисляют

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \check{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad t_i = \frac{x_i^0 - \bar{x}}{\check{\sigma}} \quad (i = \overline{0, m}) \qquad (6.21)$$

и находят $\Phi(t_0)$, $\Phi(t_1)$, ..., $\Phi(t_m)$ — значения функции (0; 1)-нормального распределения, соответствующие t_0, t_1, ..., t_m. Полагают t_0 равным $-\infty$, а $t_m = \infty$, так что $\Phi(t_0) = 0$, $\Phi(t_m) = 1$. Оценки $\check{p}_i$ в соответствии с (6.18) определяют в виде

$$\check{p}_i = \Phi(t_i) - \Phi(t_{i-1}) \quad (i = \overline{1, m}). \qquad (6.18')$$

По (6.20) гипотеза о нормальном распределении отвергается с вероятностью ошибки $\sim (1 - q)$, если

$$\chi^2 \geqslant \chi_q^2(m - 3). \qquad (6.20')$$

Аналогично проверяется гипотеза о том, что данное преобразование является нормализующим. Например, при проверке гипотезы о логнормальном законе распределения наблюдения логарифмируются — $y_i = \ln x_i$, а затем проверяется соответствие y_i нормальному закону распределения.

Если проверяется гипотеза о дискретном распределении, при вычислении статистики χ^2 по (6.19) в качестве оценок вероятностей $\check{p}_i$ используются величины вида $\check{p}_i = \sum_{j=k_i+1}^{k_{i+1}} P_j$; для значений $j = k_i + 1$, $k_i + 2$, ..., k_{i+1}, P_j — теоретические вероятности получения случайной величиной значений x_j, попавших в интервал Δ_i.

Пример 6.4. По измерениям магнитной восприимчивости $\varkappa$ габброидов района получены данные, приведенные в таблице ниже.

Интервалы $\Delta_i = [x_{i-1}, x_i)$ в ед. CGSE · 10^6	<95	95—105	105—115	115—125	125—135	135—145	145—155	155—165	165—175	>175
количества наблюдений n_i	0	3	15	22	18	12	10	6	4	0

Проверить гипотезу о нормальном распределении $\varkappa$ при уровне значимости 0,10

Решение. Для вычисления $\bar{x}$ и $\check{\sigma}^2$ воспользуемся формулами: $\bar{x} \approx \frac{1}{n} \sum_{i=1}^{m} n_i \bar{x}_i$,

$\check{\sigma}^2 \approx \frac{1}{n} \sum_{i=1}^{m} n_i (\bar{x}_i - \bar{x})^2$, где $\bar{x}_i$ — середины интервалов Δ_i.

Общее количество наблюдений $n = \sum_{i=1}^{m} n_i = 90$, число интервалов $m = 8$.

$$\bar{x} \approx \frac{1}{90}(3 \cdot 100 + 15 \cdot 110 + 22 \cdot 120 + 18 \cdot 130 + 12 \cdot 140 + 10 \cdot 150 + 6 \cdot 160 + 4 \cdot 170) \approx 130;$$

$$\check{\sigma}^2 \approx \frac{1}{90}(3 \cdot 30^2 + 15 \cdot 20^2 + 22 \cdot 10^2 + 18 \cdot 0 + 12 \cdot 10^2 + 10 \cdot 20^2 + 6 \cdot 30^2 + 4 \cdot 40^2) \approx 310;$$

$$\check{\sigma} \approx \sqrt{310} \approx 17{,}6.$$

Взяв в качестве границ интервалов $x_0 = -\infty$, $x_1^0 = 105$, $x_2^0 = 115, \ldots, x_7^0 = 165$, $x_8^0 = +\infty$, вычисляем отклонения t_i вида (6.21) и по табл. 2 (Приложение) определяем значения $\Phi(t_i)$.

x_i^0	$-\infty$	105	115	125	135	145	155	165	∞
t_i	$-\infty$	$-1{,}42$	$-0{,}85$	$-0{,}28$	0,28	0,85	1,42	1,99	∞
$\Phi(t_i)$	0	0,08	0,20	0,39	0,61	0,80	0,92	0,98	1

Теоретические вероятности p_i оцениваются в виде (6.18′)

Δ_i	$-\infty-105$	105—115	115—125	125—135	135—145	145—155	155—165	165—∞
$\check{p}_i$	0,08	0,12	0,19	0,22	0,19	0,12	0,06	0,02
$n\check{p}_i$	7,2	10,8	17,1	19,8	17,1	10,8	5,4	1,8
$n_i - n\check{p}_i$	$-4{,}2$	4,2	4,9	$-1{,}8$	$-5{,}1$	$-0{,}8$	0,6	2,2

Величина χ^2 составит:

$$\chi^2 = \frac{4{,}2^2}{7{,}2} + \frac{4{,}2^2}{10{,}8} + \frac{4{,}9^2}{17{,}1} + \frac{1{,}8^2}{19{,}8} + \frac{5{,}1^2}{17{,}1} + \frac{0{,}8^2}{10{,}8} + \frac{0{,}6^2}{5{,}4} + \frac{2{,}2^2}{1{,}8} \approx 9{,}99.$$

Количество интервалов $m = 8$, число степеней свободы $m - 3 = 5$ (два оцениваемых параметра — математическое ожидание и среднее квадратическое отклонение). Критическую границу уровня значимости $\alpha = 0{,}10$ ($q = 1 - \alpha = 0{,}9$) определяем из табл. 3 (Приложение) как 90%-ный квантиль распределения χ^2 с пятью степенями свободы: $\chi^2_{0,9}(5) = 9{,}24$. Ввиду выполнения неравенства (6.20) гипотеза о нормальном распределении отвергается с вероятностью ошибки $\alpha \approx 0{,}10$.

Отметим, что если бы мы задали уровень значимости 0,05, то нулевая гипотеза формально была бы принята ($\chi^2_{0,95}(5) = 11{,}07$). Однако приведенный выше результат показывает, что вероятность ошибки II рода при принятии такого решения была бы велика.

Критерий согласия по методу моментов. Для проверки гипотезы о нормальном распределении существует простой метод моментов, ос-

нованный на использовании оценок коэффициентов асимметрии и эксцесса. В соответствии с (4.28) эти оценки имеют вид:

$$\check{A} = \frac{\sqrt{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}, \quad \check{E} = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3. \qquad (6.22)$$

При нормальном распределении показателя математические ожидания оценок (6.22) близки к нулю. Пределы, в которых при этой гипотезе должны находиться эти оценки с вероятностями q, имеют приближенный вид: $\pm u_{\frac{1+q}{2}}\sqrt{\mathbf{D}\check{A}}$ (для $\check{A}$); $\pm u_{\frac{1+q}{2}}\sqrt{\mathbf{D}\check{E}}$ (для $\check{E}$). $\mathbf{D}\check{A}$ и $\mathbf{D}\check{E}$ вычисляются по (4.30) либо по уточненным формулам (4.30′) с использованием оценки эксцесса (4.29′). Гипотеза о нормальном распределении показателя отвергается, если выполняется хотя бы одно из неравенств:

$$|\check{A}| \geqslant u_{\frac{1+q}{2}}\sqrt{\mathbf{D}\check{A}}; \quad |\check{E}| \geqslant u_{\frac{1+q}{2}}\sqrt{\mathbf{D}\check{E}}. \qquad (6.23)$$

При проверке гипотезы о нормальном распределении преобразованного показателя $f(\xi)$ (для логарифмически нормального закона $f(\xi) = \ln\xi$) оценки коэффициентов асимметрии и эксцесса $\check{A}_n$, $\check{E}_n$ должны вычисляться по преобразованным наблюдениям $y_i = f(x_i)$, после чего проверяется условие (6.23) для $\check{A}_n$ и $\check{E}_n$.

Следует отметить, что аппроксимация распределения $\check{E}$ нормальным законом возможна лишь при очень больших n, в отличие от коэффициента асимметрии, для которого это приближение обычно эффективно. Вместо $\check{E}$ можно рекомендовать [2] использование статистики d:

$$d = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| \Big/ \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} - \sqrt{\frac{2}{\pi}}\left(1 + \frac{2}{8n-9}\right).$$

Распределение ее близко к нормальному с математическим ожиданием, равным нулю, и дисперсией $\mathbf{D}d = \frac{1}{n}\left(0{,}045 - 0{,}08\frac{1}{n}\right)$. Гипотеза о нормальном распределении ξ отвергается, если $|\check{A}| \geqslant u_{\frac{1+q}{2}}\sqrt{\mathbf{D}\check{A}}$ либо $|d| \geqslant u_{\frac{1+q}{2}}\sqrt{\mathbf{D}d}$.

Пример 6.5. При уровне значимости 0,05 проверить гипотезу о логарифмически нормальном распределении концентраций титана по результатам его спектральных определений в 54 пробах. Оценка коэффициента асимметрии, вычисленная по логарифмам концентраций, $\check{A}_n = 0{,}39$, а коэффициента эксцесса по тем же величинам, $\check{E}_n = 0{,}79$.

Решение. Дисперсии $\check{A}_n$ и $\check{E}_n$ по формулам (4.30): $\mathbf{D}\check{A}_n \approx \frac{6}{54}$, $\mathbf{D}\check{E}_n \approx \frac{24}{54}$. Вероятности q попадания $\check{A}_n$ и $\check{E}_n$ в пределы их допустимых значений соответствующие уровню значимости 0,05, $q = 1 - \alpha = 0{,}95$; $\frac{1+q}{2} = 0{,}975$ По табл. 2 (Приложение) находим $u_{0,95} = 1{,}96$. Критические границы, используемые в критерии (6.23),

$$u_{\frac{1+q}{2}} \sqrt{\mathbf{D}\check{A}_n} = 1{,}96 \sqrt{\frac{6}{54}} \approx 0{,}65; \quad u_{\frac{1+q}{2}} \sqrt{\mathbf{D}\check{E}_n} = 1{,}96 \sqrt{\frac{24}{54}} \approx 1{,}30.$$

Так как оба условия (6.23) не выполняются, данные измерений не противоречат гипотезе логарифмически нормального закона распределения.

Критерий Колмогорова. В этом критерии используется наибольшее отклонение выборочной функции распределения $F_n(x)$ от теоретической $F(x)$: $d_n = \sup_x |F_n(x) - F(x)|$. Оно сравнивается с табличным критическим значением и нулевая гипотеза отвергается, если d_n превышает критический уровень или равно ему. Для расчета критической границы d_α можно использовать и приближенную формулу: $d_\alpha \approx \sqrt{-\frac{1}{2n} \ln \frac{\alpha}{2}}$, где α — уровень значимости критической области, n — число независимых наблюдений, по которым построена $F_n(x)$. С помощью этого правила проверяется соответствие эмпирических данных лишь вполне конкретной функции распределения (а не параметрическому семейству). Использование вместо неизвестных параметров их оценок может привести к отклонению уровня значимости критической области от того значения, которое должно ей соответствовать в действительности. В практике геолого-геофизических исследований этот критерий применяется редко. Таблица критических значений и описание ее применения даны в [2].

Критерий согласия при заданной альтернативе. Описанные критерии согласия относятся к задаче проверки гипотез без указания альтернатив. Однако специфика геологических исследований такова, что далеко не всегда можно говорить о строгом соответствии распределения показателей той или иной принятой в математической статистике модели. Если такие обоснования можно найти для описания распределений ошибок некоторых методов измерений, скажем, центральную предельную теорему при использовании нормального закона, то распределения показателей в самих геологических объектах ввиду своеобразия каждого из них могут лишь более или менее удачно аппроксимироваться (описываться приближенно) теоретическими законами распределения, опирающимися на те или иные вероятностные модели. Поэтому, если даже принята какая-либо гипотеза о теоретическом распределении показателя в объекте, в принципе всегда можно предложить более подробную схему, точнее соответствующую конкретным эмпирическим данным.

Когда же принимаются во внимание альтернативные законы распределения, задача сводится к установлению среди гипотетических законов того, который наиболее соответствует выборочным данным.

При этом можно указать естественную меру такого соответствия с помощью формулы Бейеса. Ясно, что по возможности нужно стремиться к такой постановке задачи, при которой указывается исчерпывающая группа возможных в данных условиях семейств распределений. Оценив для каждого семейства параметры распределения по возможности наиболее точным методом, с их помощью определяют оценки плотностей распределения

$$\check{p}_j(x) = p_j(x, \check{\alpha}_1, \check{\alpha}_2, \ldots, \check{\alpha}_{k_j}), \quad j = \overline{1, t} \tag{6.25}$$

($\check{\alpha}_i$ — оценки параметров). В соответствии с формулой Бейеса закон распределения, наиболее соответствующий имеющейся выборке независимых наблюдений $x_1, x_2, \ldots, x_n$, определяется по максимальной из вероятностей

$$P_j = \prod_{i=1}^{n} p_j(x_i)\Big[\sum_{j=1}^{t}\prod_{i=1}^{n} p_j(x_i)\Big]^{-1}, \quad j = \overline{1, t} \tag{6.26}$$

(P_j — вероятность того, что плотность распределения принадлежит семейству $p_j(x)$). Подставив $\check{p}_j(x_i)$ в (6.26), получим оценки вероятностей P_j, а с ними и решение задачи.

В частности, если проверяется гипотеза H_1 о нормальном распределении, а альтернативой H_2 является логарифмически нормальный закон, то оценки вероятностей (6.26) вычисляются в виде

$$\check{P}_1 = \frac{\check{p}_1(x_1, x_2, \ldots, x_n)}{\check{p}_1(x_1, x_2, \ldots x_n) + \check{p}_2(x_1, x_2, \ldots, x_n)}, \quad \check{P}_2 = 1 - \check{P}_1,$$

где

$$\check{p}_1(x_1, x_2, \ldots, x_n) = \frac{1}{(\sqrt{2\pi})^n \check{\sigma}^n} \exp\Big[-\frac{1}{2\check{\sigma}^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\Big] =$$

$$= (2\pi)^{-\frac{n}{2}} \check{\sigma}^{-n} \exp\Big(-\frac{1}{2}n\Big);$$

$$\check{p}_2(x_1, x_2, \ldots, x_n) = \frac{1}{(\sqrt{2\pi})^n \check{\delta}^n \prod_{1}^{n} x_i} \exp\Big[-\frac{1}{2\check{\delta}^2}\sum_{i=1}^{n}(\ln x_i - \check{\mu})^2\Big] =$$

$$= (2\pi)^{-\frac{n}{2}} \check{\delta}^{-n} \prod_{i=1}^{n} x_i^{-1} \exp\Big(-\frac{1}{2}n\Big).$$

Подставив в (6.26), получим

$$\check{P}_1 = \frac{\check{\sigma}^{-n}}{\check{\sigma}^{-n} + \check{\delta}^{-n}\prod_{i=1}^{n} x_i^{-1}} = \frac{\check{\sigma}^{-n}}{\check{\sigma}^{-n} + (\check{\delta}\check{\mu})^{-n}}, \quad \check{P}_2 = 1 - \check{P}_1, \tag{6.27}$$

где $\check{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$, $\check{\delta} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\ln x - \check{\mu})^2}$, $\check{\mu} =$

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

С помощью формулы Бейеса можно построить критерий и для дискретных распределений. Для расчета апостериорных вероятностей P_j применяется та же формула (6.26), с той разницей, что в качестве $p_j(x_i)$ используются теоретические вероятности получения значений, равных x_i.

Пример 6.6. Проверить гипотезу о нормальном распределении концентраций Zn в гранитах при альтернативе логнормального распределения, если оценки $\check{\sigma}$, $\check{\delta}$, $\check{M}_e = e^{\check{\mu}}$, вычисленные по данным 30 спектральных определений (10^{-3}%) составили: $\check{\sigma} = 0{,}84$, $\check{\delta} = 0{,}46$, $e^{\check{\mu}} = 1{,}56$.

Решение. По (6.27) имеем

$$\check{P}_1 = \frac{0{,}84^{-30}}{0{,}84^{-30} + (0{,}46 \quad 1{,}56)^{-30}} \approx \frac{187}{187 + 21074} = 0{,}0088; \quad \check{P}_2 = 1 - \check{P}_1 = 0{,}9912.$$

Гипотеза о нормальном распределении должна быть отвергнута в пользу логарифмически нормального закона.

§ 4. Оценка параметров и числовых характеристик распределений при анализе геолого-геофизической информации

При статистическом анализе распределений геолого-геофизических показателей ξ_j обычно используются следующие числовые характеристики: математическое ожидание (среднее) M_j, среднее квадратическое отклонение σ_j, коэффициент вариации V_j, квантили высоких и низких порядков (критические границы) B_j, медиана M_{ej}, модальное (наиболее вероятное) значение M_{0j}, коэффициенты асимметрии A_j и эксцесса E_j. Помимо параметров одномерных распределений употребляются характеристики совместных распределений признаков — коэффициенты парной, множественной и частной корреляции, коэффициенты регрессий.

Оценки числовых характеристик распределений. Точность оценок. Методы получения оценок были рассмотрены в гл. 4, где в качестве примеров приведены оценки характеристик одномерных распределений. Эти оценки с доверительными пределами, рассчитанными по (4.8) с использованием (4.22) на основании асимптотической нормальности оценок, сведены в табл. 6.1. Использованы следующие обозначения: x_{ij} — i-е наблюдение j-го показателя ξ_j из таблицы исходных данных; $u_{\frac{1+q}{2}}$ — квантиль порядка $\frac{1+q}{2}$ (0; 1)-нормального распределения, определяемый из табл. 2 (Приложение); n — количество наблюдений каж-

дого показателя; $s_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2} = \sqrt{\frac{n-1}{n}}\,\check{\sigma}_j$; p — вероятность некоторого события, ν — количество испытаний, в которых событие произошло.

При неизвестном законе распределения в качестве оценки моды обычно используется середина интервала гистограммы, которому соответствует ее наибольшее значение. Для уточнения интервал перемещают вдоль оси абсцисс и находят положение, при котором число наблюдений в нем максимально; середина интервала при таком положении дает оценку моды. Другой способ состоит в вычислении оценки моды в виде: $\check{M}_0 = \frac{n_{i-1}x^0_{i-1} + n_i x^0_i + n_{i+1}x^0_{i+1}}{n_{i-1} + n_i + n_{i+1}}$, где x^0_i — середина интервала с наибольшим числом наблюдений, n_i — число наблюдений в нем; x^0_{i-1}, x^0_{i+1} — середины соседних интервалов (одинаковых с ним по длине); n_{i-1}, n_{i+1} — количества наблюдений в них.

Оценкой медианы служит такое значение показателя, что 50% наблюдений выборки меньше его. Если n нечетное, то оценкой медианы является наблюдение с номером $\frac{n+1}{2}$ в ряду наблюдений по возрастанию; если n четное, ее оценкой служит среднее из двух наблюдений — с номерами $\frac{n}{2}$ и $\frac{n}{2}+1$.

На упомянутые оценки моды и медианы не оказывают влияния резко выделяющиеся наблюдения выборки. Это служит доводом в пользу их применения для оценки уровня местного фона показателя. В случае нормального распределения мода, медиана и математическое ожидание совпадают, так что оптимальной оценкой этих величин будет среднее арифметическое.

Дисперсии оценок коэффициентов асимметрии и эксцесса в общем случае зависят от центральных моментов высоких порядков (4.29), так что доверительные пределы для A_j и E_j могут быть оценены лишь с низкой точностью. С другой стороны, асимметрия и эксцесс употребляются, главным образом, для проверки гипотезы о нормальном распределении. Для этой цели достаточно указать пределы для оценок асимметрии и эксцесса при гипотезе о нормальном распределении показателя, какие и даны в табл. 6.1.

Приведенные в табл. 6.1 доверительные пределы построены на основании асимптотического распределения соответствующих оценок и, следовательно, применимы лишь при достаточно больших n. В некоторых случаях можно указать более точные доверительные пределы. Если, например, показатель распределен нормально, то можно доказать, что отношение

$$t = \frac{(\bar{x}_j - M_j)}{\sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}} = \frac{(\bar{x}_j - M_j)\sqrt{n}}{\check{\sigma}_j}$$

Таблица 6.1. Оценки и доверительные пределы числовых характеристик одномер

Числовая характеристика	Оценка	Среднее квадратическое отклонение оценки
Математическое ожидание (среднее) M_j	$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$	$\frac{\sigma_j}{\sqrt{n}}$
Среднее квадратическое отклонение σ_j	$\check{\sigma}_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}$	$\sim \frac{\sigma_j\sqrt{E_j+2}}{\sqrt{4n}} = \sigma(\check{\sigma}_j)$
Коэффициент вариации V_j	$\check{V}_j = \frac{\check{\sigma}_j}{\bar{x}_j}$	$\sim V\sqrt{\frac{2+E_j-4A_jV_j+4V_j^2}{4n}} = \sigma(\check{V}_j)$
Критическая граница $B_j = M_j + t\sigma_j$	$\check{B}_j = \bar{x}_j + t\check{\sigma}_j$	$\sim \frac{\sigma_j}{\sqrt{n}}\sqrt{1+tA_j+\frac{t^2(E_j+2)}{4}} = \sigma(\check{B}_j)$
Коэффициент асимметрии A_j	$\check{A}_j = \frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^3}{ns_j^3}$	$\sim\sqrt{\frac{6}{n}}$ *
Коэффициент эксцесса E_j	$\check{E}_j = \frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^4}{ns_j^4} - 3$	$\sim 2\sqrt{\frac{6}{n}}$ *
Вероятность p	$\check{p} = \frac{\nu}{n}$	$\sqrt{\frac{p(1-p)}{n}} = \sigma(\check{p})$

Примечания. * Приведены приближенные выражения при гипотезе
** Приближенные пределы для $\check{A}$ и $\check{E}$ при гипотезе нор

подчиняется распределению Стьюдента с $n-1$ степенями свободы. Обозначив $t_{\frac{1+q}{2}}$ — $\frac{1+q}{2}$-квантиль распределения Стьюдента с $n-1$ степенями свободы и приняв во внимание, что вследствие симметричности этого распределения $t_{\frac{1-q}{2}} = -t_{\frac{1+q}{2}}$, имеем:

$$\mathbf{P}\left\{-t_{\frac{1+q}{2}} \leqslant t < t_{\frac{1+q}{2}}\right\} = \mathbf{P}\left\{-t_{\frac{1+q}{2}}\frac{\check{\sigma}_j}{\sqrt{n}} + \bar{x}_j < M_j \leqslant t_{\frac{1+q}{2}}\frac{\check{\sigma}_j}{\sqrt{n}} + \bar{x}_j\right\} = q.$$

Доверительные пределы уровня q	Относительная точность (в %)
$\sim \bar{x}_j \pm u_{\frac{1+q}{2}} \frac{\sigma_j}{\sqrt{n}}$	$\pm u_{\frac{1+q}{2}} \frac{V_j}{\sqrt{n}} 100$
$\sim \check{\sigma}_j \pm u_{\frac{1+q}{2}} \sigma(\check{\sigma}_j)$	$\pm u_{\frac{1+q}{2}} \sqrt{\frac{E_j+2}{4n}} 100$
$\sim \check{V}_j \pm u_{\frac{1+q}{2}} \sigma(\check{V}_j)$	$\pm u_{\frac{1+q}{2}} \sqrt{\frac{2+E_j-4A_jV_j+4V_j^2}{4n}} 100$
$\sim \check{B}_j \pm u_{\frac{1+q}{2}} \sigma(\check{B}_j)$	$\pm u_{\frac{1+q}{2}} \frac{\sqrt{1+tA_j+\frac{1}{4}t^2(E_j+2)}}{(V_j^{-1}+t)\sqrt{n}} 100$
$\sim \pm u_{\frac{1+q}{2}} \sqrt{\frac{6}{n}}$ **	—
$\sim \pm 2u_{\frac{1+q}{2}} \sqrt{\frac{6}{n}}$ **	—
$\sim \check{p} \pm u_{\frac{1+q}{2}} \sigma(\check{p})$	$\pm u_{\frac{1+q}{2}} \sqrt{\frac{p^{-1}-1}{n}} 100$

нормального распределения показателя.
мального распределения показателя.

Таким образом, доверительные пределы уровня q для математического ожидания составят

$$\bar{x}_j - t_{\frac{1+q}{2}} \frac{\check{\sigma}_j}{\sqrt{n}}, \quad \bar{x}_j + t_{\frac{1+q}{2}} \frac{\check{\sigma}_j}{\sqrt{n}}.$$

Уже при $n \geqslant 30$ квантили распределения Стьюдента близки к соответствующим квантилям нормального распределения. Это является следствием асимптотической нормальности распределения Стьюдента (при увеличении числа степеней свободы). Так, по табл. 9 (Приложе-

ние) $t_{0,95}(30) \approx 1{,}70$, а квантиль нормального распределения $u_{0,95} \approx 1{,}645$; $t_{0,975}(30) \approx 2{,}04$, а $u_{0,975} \approx 1{,}96$. Это свидетельствует о том, что доверительные пределы, вычисленные по приближенной формуле табл. 6.1, оказываются достаточно точными уже при небольших объемах выборок.

При нормальном распределении показателя можно построить также и уточненные доверительные пределы для среднего квадратического отклонения. Их построение основывается на том, что в этом случае величина $\frac{(n-1)\check{\sigma}_j^2}{\sigma_j^2}$ подчиняется распределению χ^2 с $n-1$ степенями свободы. Очевидно,

$$\mathbf{P}\left\{\chi^2_{\frac{1-q}{2}}(n-1) \leqslant \frac{(n-1)\check{\sigma}_j^2}{\sigma_j^2} < \chi^2_{\frac{1+q}{2}}(n-1)\right\} = q, \qquad (6.28)$$

где $\chi^2_{\frac{1-q}{2}}(n-1)$, $\chi^2_{\frac{1+q}{2}}(n-1)$ — квантили распределения χ^2 с $n-1$ степенями свободы порядков $\frac{1-q}{2}$ и $\frac{1+q}{2}$. Из (6.28)

$$\mathbf{P}\left\{\frac{(n-1)\check{\sigma}_j^2}{\chi^2_{\frac{1+q}{2}}(n-1)} < \sigma_j^2 \leqslant \frac{(n-1)\check{\sigma}_j^2}{\chi^2_{\frac{1-q}{2}}(n-1)}\right\} = q, \qquad (6.29)$$

следовательно, доверительными пределами для среднего квадратического отклонения будут величины

$$\sigma_j^- = \check{\sigma}_j \sqrt{\frac{n-1}{\chi^2_{\frac{1+q}{2}}(n-1)}}, \quad \sigma_j^+ = \check{\sigma}_j \sqrt{\frac{n-1}{\chi^2_{\frac{1-q}{2}}(n-1)}}. \qquad (6.30)$$

Для сравнения с приближенными пределами, приведенными в табл. 6.1, вычислим доверительные пределы уровня $q = 0{,}95$ при $n = 50$. По формуле табл. 6.1, приняв $E_j = 0$,

$$\sigma_j^- \approx \check{\sigma}_j - 1{,}96\check{\sigma}_j \frac{1}{\sqrt{2 \cdot 50}} \approx 0{,}8\check{\sigma}_j; \quad \sigma_j^+ \approx \check{\sigma}_j + 1{,}96 \frac{1}{\sqrt{2 \cdot 50}} \approx 1{,}2\check{\sigma}_j;$$

по формуле (6.30), используя табл. 3 (Приложение),

$$\sigma_j^- = \check{\sigma}_j \sqrt{\frac{49}{\chi^2_{0,975}(49)}} \approx \check{\sigma}_j \frac{7}{\sqrt{70{,}2}} \approx 0{,}84\check{\sigma}_j;$$

$$\sigma_j^+ = \check{\sigma}_j \sqrt{\frac{49}{\chi^2_{0,025}(49)}} \approx \check{\sigma}_j \frac{7}{\sqrt{31{,}6}} \approx 1{,}24\check{\sigma}_j.$$

Сравнение показывает, что точность приближенных формул, по крайней мере при $n \geqslant 50$, удовлетворительна, хотя и наблюдается некоторое смещение.

Разумеется, для использования пределов (6.30) необходим положительный результат проверки гипотезы о нормальном распределении показателя. Отметим также, что в случае принятия такой гипотезы расчет доверительных пределов по формулам табл. 6.1 упрощается, так как в них можно положить $A_j = E_j = 0$.

Нормальный закон распределения обеспечивает возможность пользоваться простым, хотя и менее точным способом вычисления оценки среднего квадратического отклонения σ_j. Такая оценка имеет вид

$$\sigma_j^0 = \frac{R_{nj}}{d_n}, \tag{6.31}$$

где $R_{nj} = x_{\max j} - x_{\min j}$ — *размах* выборки наблюдений показателя ξ_j; $x_{\max j}$, $x_{\min j}$ — максимальное и минимальное наблюдения выборки; d_n — коэффициент, определяемый по величине n — количеству наблюдений (табл. 6, Приложение). Эта оценка применяется при малых объемах выборок ($n \leqslant 20$), так как ее использование при больших количествах наблюдений сопряжено с потерей точности. Доверительные пределы определяются в виде

$$\frac{R_{nj}}{d_n\left(\frac{1+q}{2}\right)}, \quad \frac{R_{nj}}{d_n\left(\frac{1-q}{2}\right)}, \tag{6.32}$$

где $d_n\left(\frac{1-q}{2}\right)$, $d_n\left(\frac{1+q}{2}\right)$ — коэффициенты, соответствующие уровням $\frac{1-q}{2}$, $\frac{1+q}{2}$ и определяемые из табл. 7 (Приложение). Если $n > 20$, то неупорядоченные данные выборки необходимо разбить на группы одинакового объема и вычислить среднее арифметическое из оценок σ_j^0 по каждой такой группе.

Другой способ оценки среднего квадратического отклонения нормального распределения состоит в использовании *оценки среднего абсолютного отклонения* l_j:

$$\sigma_{ja} = l_j \sqrt{\frac{\pi}{2}}, \quad l_j = \frac{1}{n}\sum_{i=1}^{n} |x_{ij} - \bar{x}_j|. \tag{6.33}$$

Здесь используется тот факт, что для нормально распределенной случайной величины ξ_j среднее абсолютное отклонение

$$\mathbf{M}|\xi_j - \mathbf{M}\xi_j| = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma_j} |x - \mathbf{M}\xi_j| \exp\left[-\frac{1}{2}\frac{(x-\mathbf{M}\xi_j)^2}{\sigma_j^2}\right] dx =$$

$$= 2\int_0^{\infty} \frac{\sigma_j u}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right) du = \sqrt{\frac{2}{\pi}}\,\sigma_j.$$

В статистическом анализе иногда используют *энтропию* распределения, служащую мерой априорной неопределенности величины ξ_j.

Энтропия в *битах* (единицах энтропии) определяется в виде $H(\xi_j) = -\mathbf{M}\log_2 p_j(\xi_j)$, где $p_j(x)$ — плотность распределения ξ_j, если ξ_j — непрерывная величина, и $p_j(x) = \mathbf{P}\{\xi_j = x\}$, если ξ_j — дискретная величина. Среди непрерывных распределений наибольшую энтропию при заданной дисперсии σ^2 имеет нормальное распределение: $H_{\max} = \log_2\sqrt{2\pi e\sigma^2}$. Для дискретной величины $H(\xi_j) \geqslant 0$, причем $H(\xi_j) = 0$, если $\xi_j = c = \text{const}$ с вероятностью, равной единице. Оценка энтропии непрерывной величины имеет вид $\check{H}(\xi_j) = -\int\limits_{-\infty}^{\infty} \log_2 \check{p}_j(x) \times \check{p}_j(x)\,dx$, где $\check{p}_j(x)$ — оценка плотности распределения. По гистограмме энтропия оценивается (при достаточно большом числе n наблюдений) в виде $\check{H}(\xi_j) = -\sum\limits_{i=1}^{m} \frac{n_i}{n}\log_2\frac{n_i}{n\delta_i}$, где δ_i — длины интервалов гистограммы, n_i — количества попавших в них наблюдений, $n = \sum\limits_{i=1}^{m} n_i$.

Пример 6.7. По выборке 26 содержаний Al_2O_3, измеренных химико-аналитическим методом, получены следующие упорядоченные по возрастанию данные (в %): 14,6; 14,8; 14,9; 15,09; 15,17; 15,68; 15,72; 15,9; 16,0; 16,06; 16,08; 16,12; 16,19; 16,20; 16,27; 16,4; 16,5; 16,6; 16,6; 16,9; 17,3; 17,3; 17,4; 17,4; 17.6; 18,2.

1) Определить оценки и доверительные пределы при уровне $q = 0{,}95$: для математического ожидания, среднего квадратического отклонения, коэффициента вариации; найти оценки асимметрии и эксцесса. 2) Проверить гипотезу о нормальном распределении методом моментов. 3) Определить оценки моды и медианы, оценку критической границы для аномально высоких содержаний при уровне значимости 0,01. 4) Оценить долю площади исследуемого объекта, на которой содержание Al_2O_3 в породе менее 14%.

Решение. 1) По формулам табл. 6.1 имеем:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \approx \frac{1}{26}\,423{,}0 \approx 16{,}27\,[\%];$$

$$\check{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \approx \frac{1}{25}\,21{,}20 \approx 0{,}85\,[\%]^2;$$

$$\check{\sigma} = \sqrt{\check{\sigma}^2} = \sqrt{0{,}85} \approx 0{,}92\,[\%];\quad \check{V} = \frac{\check{\sigma}}{\bar{x}} = \frac{0{,}92}{16{,}27} \approx 0{,}06 = 6\%;$$

$$\check{A} = \frac{1}{n\check{s}^3}\sum_{i=1}^{n}(x_i - \bar{x})^3 = \frac{\sqrt{n}}{(n-1)^{3/2}\check{\sigma}^3}\sum_{i=1}^{n}(x_i - \bar{x})^3 \approx 0{,}65;$$

$$\check{E} = \frac{1}{n\check{s}^4}\sum_{i=1}^{n}(x_i - \bar{x})^4 - 3 = \frac{n}{(n-1)^2\check{\sigma}^4}\sum_{i=1}^{n}(x_i - \bar{x})^4 - 3 \approx -0{,}81.$$

Доверительные пределы: для математического ожидания

$$\sim\bar{x} \pm u_{\frac{1+q}{2}}\frac{\check{\sigma}}{\sqrt{n}} = 16{,}27 \pm 1{,}96\frac{0{,}92}{\sqrt{26}} \approx 16{,}27 \pm 0{,}35\,[\%]$$

(при доверительной вероятности $q = 0{,}95$, $u_{\frac{1+q}{2}} = 1{,}96$);

для среднего квадратического отклонения

$$\sim\check{\sigma} \pm u_{\frac{1+q}{2}} \frac{\check{\sigma}}{2\sqrt{n}} \sqrt{\check{E}+2} = 0{,}92 \pm \frac{1{,}96 \quad 0{,}92}{2} \sqrt{\frac{-0{,}81+2}{26}} \approx 0{,}92 \pm 0{,}19\,[\%];$$

для коэффициента вариации

$$\sim \check{V} \pm u_{\frac{1+q}{2}} \frac{\check{V}}{\sqrt{4n}} \sqrt{2+\check{E}+4\check{V}^2-4\check{A}\check{V}} = 0{,}06 \pm$$

$$\pm \frac{0{,}06 \quad 1{,}96}{2 \quad 5{,}1} \sqrt{2-0{,}81+4\cdot 0{,}06^2-4\cdot 0{,}65 \quad 0{,}06} \approx 6 \pm 1{,}2\,[\%].$$

Найденные доверительные пределы заключают с вероятностью, близкой к 0,95, точные значения соответствующих числовых характеристик.

2) Для проверки гипотезы о нормальном распределении воспользуемся приближенным критерием (6.23). При $q = 0{,}95$ пределы для $\check{A}$ и $\check{E}$ составят

$$\pm u_{\frac{1+q}{2}} \sqrt{\mathbf{D}\check{A}} = \pm u_{0{,}975} \sqrt{\frac{6}{n}} = \pm 1{,}96 \sqrt{\frac{6}{26}} \approx \pm 0{,}94;$$

$$\pm u_{\frac{1+q}{2}} \sqrt{\mathbf{D}\check{E}} = \pm 2u_{\frac{1+q}{2}} \sqrt{\frac{6}{n}} \approx \pm 1{,}88.$$

Так как ни одно из условий (6.23) не выполняется, оценки $\check{A}$ и $\check{E}$ не противоречат гипотезе о нормальном распределении * содержания Al_2O_3.

3) Определим интервалы содержаний Al_2O_3 следующим образом: 14—15—16—17—18—19 (%). Наибольшее число наблюдений (12) попадает в интервал 16—17%; оценка моды $\check{M}_0 = 16{,}5\%$. Оценка медианы — среднее между значениями 16,19 и 16,20: $\check{M}_e = 16{,}195\%$.

Как видно из сопоставления этих оценок с доверительными пределами для математического ожидания, они несущественно отличаются от $\bar{x}$, что вполне согласуется с нормальным распределением содержания Al_2O_3. Ввиду этого можно, в частности, принять $\check{M}_0 = \check{M}_e = \bar{x} \approx 16{,}27\%$.

Оценкой критической границы уровня значимости 0,01 для аномально высоких содержаний служит квантиль порядка $q = 0{,}99$, который оценивается с помощью табл. 2 (Приложение) в виде:

$$\check{B}_{0{,}99} = \bar{x} + u_{0{,}99}\check{\sigma} = 16{,}27 + 2{,}33 \cdot 0{,}92 \approx 18{,}41\,[\%].$$

Доверительные пределы для критической границы

$$\sim \check{B}_{0{,}99} \pm u_{\frac{1+q}{2}} \check{\sigma} \sqrt{n}^{-\frac{1}{2}} \sqrt{1+u_{0{,}99}\check{A}+0{,}25u_{0{,}99}^2(\check{E}+2)} = 18{,}41 \pm$$

$$\pm 1{,}96 \cdot 0{,}92 \sqrt{\frac{1}{26}} \sqrt{1+2{,}33 \quad 0{,}65+0{,}25\cdot 2{,}33^2(-0{,}81+2)} \approx 18{,}41 \pm 0{,}72\,[\%].$$

4) Найденные числовые характеристики позволяют решить и последнюю из поставленных задач. Искомую долю p генеральной совокупности можно определить

---

* Применение метода моментов здесь носит несколько формальный характер, так как число наблюдений недостаточно велико.

как значение функции распределения, соответствующее квантилю 14%. Величина нормированного отклонения

$$z = \frac{B - \bar{x}}{\check{\sigma}} = \frac{14 - 16{,}27}{0{,}92} \approx -2{,}47.$$

С учетом нормальности распределения показателя по табл. 2 (Приложение) находим $\Phi(z) = \Phi(-2{,}47) = 0{,}007$ и оценкой p будет $\check{p} = 0{,}7\%$.

Оценки по сгруппированным данным. Расчет оценок по формулам табл. 6.1 вручную довольно трудоемок. Для сокращения вычислений используют простейшие преобразования наблюдений и *группирование данных*. Преобразования обычно имеют вид

$$y_{ij} = x_{ij} - a_j \tag{6.34}$$

либо

$$z_{ij} = \frac{x_{ij} - a_j}{b_j}, \tag{6.35}$$

причем величины a_j и b_j подбирают так, чтобы обеспечить по возможности получение информации в виде небольших целых положительных чисел. После получения оценок по преобразованным величинам переходят к обычным оценкам. В случае преобразования (6.34) оценки M_j, M_{0j}, M_{ej}, B_j выражаются через соответствующие оценки M_{yj}, M_{0yj}, M_{eyj}, B_{yj} в виде

$$\alpha_j = a_j + \alpha_{yj} \tag{6.36}$$

(α_{yj} — оценка по величинам y_{ij}). Оценки σ_j, A_j, E_j останутся без изменений (*инвариантны* по отношению к смещению).

Если используется преобразование (6.35), то оценки M_j, M_{0j}, M_{ej}, B_j вычисляются в виде $\alpha_j = a_j + b_j\alpha_{zj}$ (α_{zj} — оценка по преобразованным в виде (6.35) данным), оценка среднего квадратического отклонения $\check{\sigma}_j = |b_j|\check{\sigma}_{zj}$. Коэффициент эксцесса и модуль коэффициента асимметрии останутся без изменений, причем коэффициент асимметрии изменит свой знак при $b_j < 0$: $\check{A}_j = b_j|b_j|^{-1}\check{A}_{zj}$, $\check{E}_j = \check{E}_{zj}$.

Для получения оценок по сгруппированным данным область значений, принимаемых показателем, делят на интервалы $\Delta_k^{(j)}$ как при построении гистограммы. Далее, определяют середины x_{kj}^0 интервалов и количества наблюдений n_{kj}, попавших в них. Если вычисляемая оценка $\check{\alpha}_j$ представляется в виде $\check{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} f(x_{ij})$, то, полагая для наблюдений, попавших в интервалы $\Delta_k^{(j)}$ ($k = \overline{1, m}$), $\sum_{i,\ x_{ij} \in \Delta_k^{(j)}} f(x_{ij}) \approx$ $\approx n_{kj} f(x_{kj}^0)$, получим оценку по сгруппированным данным:

$$\tilde{\alpha}_j = \frac{1}{n}\sum_{k=1}^{m} n_{kj} f(x_{kj}^0) \tag{6.37}$$

(m — число интервалов).

В частности, оценка математического ожидания

$$\tilde{M}_j = \frac{1}{n}\sum_{k=1}^{m} n_{kj} x_{kj}^0, \qquad (6.38)$$

а среднего квадратического отклонения

$$\sigma_j = \sqrt{\frac{1}{n}\sum_{k=1}^{m} n_{kj}(x_{kj}^0 - \tilde{M}_j)^2}. \qquad (6.39)$$

Расхождение между оценками, получаемыми по сгруппированным и несгруппированным данным, можно уменьшить введением *поправки Шеппарда*. Ее можно применять для распределений, на концах которых частоты малы (точнее, плотности распределения имеют касание высокого порядка с осью абсцисс). Оценка дисперсии, с учетом поправки Шеппарда, составит $\check{\sigma}_j^2 = \tilde{\sigma}_j^2 - \frac{h^2}{12}$ (h — длина интервала группирования); четвертого центрального момента $\mu_{4j} = \tilde{\mu}_{4j} - \frac{1}{2}\tilde{\sigma}^2 h^2 + \frac{7h^4}{240}$; для математического ожидания и третьего центрального момента поправка равна нулю.

Оценки параметров обобщенно-логнормального распределения. Оценки табл. 6.1 могут использоваться для различных распределений, так как они состоятельны и несмещенны или асимптотически несмещенны. При нормальном распределении показателя можно гарантировать оптимальные свойства оценок $\bar{x}$ и $\check{\sigma}^2$ (эффективность). При отклонении распределения от нормального возможны более точные оценки числовых характеристик, как и более точная аппроксимация самой плотности распределения. Для однородных распределений концентраций химических элементов в горных породах, количественных показателей физических свойств пород, промыслово-геофизических данных часто дают удовлетворительную аппроксимацию логнормальный или обобщенно-логнормальный законы распределения, уже рассматривавшиеся в гл. 2 и 4.

Напомним, что ξ_j подчиняется обобщенно-логнормальному закону распределения, если $\ln(a_j + \lambda_j \xi_j)$ распределен нормально (a_j и λ_j — параметры, причем λ_j может равняться 1 или —1). Очевидно, для оценки качества аппроксимации распределения необходимо проверять гипотезу о нормальном распределении $\ln(a_j + \lambda_j \xi_j)$. Так, для критерия по методу моментов вычисляют оценки коэффициентов асимметрии и эксцесса по преобразованным наблюдениям

$$\check{A}_{nj} = \frac{1}{n\check{\sigma}_j^3}\sum_{i=1}^{n}(z_{ij} - \check{\mu}_j)^3, \quad \check{E}_{nj} = \frac{1}{n\check{\sigma}_j^4}\sum_{i=1}^{n}(z_{ij} - \check{\mu}_j)^4 - 3, \qquad (6.40)$$

где $z_{ij} = \ln(a_j + \lambda_j x_{ij})$; $\check{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} z_{ij}$, $\check{\delta}_j^2 = \frac{1}{n}\sum_{i=1}^{n}(z_{ij} - \check{\mu}_j)^2$ — оценки математического ожидания и дисперсии преобразованной величины.

Для получения наиболее правдоподобных оценок числовых характеристик необходимо в формулы (2.39) — (2.43), выражающие их через μ_j и δ_j^2, подставить оценки $\check{\mu}_j$ и $\check{\delta}_j^2$, подобно (4.46), (4.47), (4.48).

Расчет доверительных пределов для M_j, M_{0j}, M_{ej} и для критических границ B_j покажем на примере оценки математического ожидания (4.46)

$$\tilde{M}_j = \left[\exp\left(\check{\mu}_j + \frac{1}{2}\check{\delta}_j^2\right) - a_j\right]\lambda_j.$$

Так как ее вариация определяется распределением $\check{\mu}_j + \frac{1}{2}\check{\delta}_j^2$, естественно определить доверительные пределы в виде

$$\sim\left\{\exp\left[\check{\mu}_j + \frac{1}{2}\check{\delta}_j^2 \pm u_{\frac{1+\varphi}{2}}\sigma\left(\check{\mu}_j + \frac{1}{2}\check{\delta}_j^2\right)\right] - a_j\right\}\lambda_j,$$

где $\sigma\left(\check{\mu}_j + \frac{1}{2}\check{\delta}_j^2\right)$ — среднее квадратическое отклонение суммы $\check{\mu}_j + \frac{1}{2}\check{\delta}_j^2$. По (4.22) дисперсия $\mathbf{D}\left(\check{\mu}_j + \frac{1}{2}\check{\delta}_j^2\right) \approx \frac{\delta_j^2}{n} + \frac{1}{n}A_{nj}\delta_j^3 + \frac{1}{4n}(E_{nj} + 2)\delta_j^4 = \frac{1}{n}\delta_j^2 + \frac{1}{2n}\delta_j^4$, так как $\mathbf{D}\check{\mu}_j = \frac{\delta_j^2}{n}$, $\mathbf{D}\left(\frac{\check{\delta}_j^2}{2}\right) \approx \frac{1}{4n}(E_{nj}\delta_j^4 + 2\delta_j^4)$, $\mathbf{M}[(\check{\mu}_j - \mu_j)(\check{\delta}_j^2 - \delta_j^2)] \approx A_{nj}\frac{\delta_j^3}{n}$ и принимаем $A_{nj} = E_{nj} = 0$. Искомые доверительные пределы представляются в виде

$$\sim\left[\exp\left(\check{\mu}_j + \frac{1}{2}\check{\delta}_j^2 \pm u_{\frac{1+}{2}}\check{\delta}_j\sqrt{2 + \check{\delta}_j^2}\frac{1}{\sqrt{2n}}\right) - a_j\right]\lambda_j.$$

Для σ_j и V_j применение этого способа затруднительно, поэтому формула (4.22) применяется непосредственно к функциям, выражающим $\tilde{\sigma}_j$ и $\tilde{V}_j$ (табл. 6.2) через $\check{\mu}_j$ и $\check{\delta}_j^2$. Свойство асимптотически нормального закона распределения оценок $\tilde{\sigma}_j$ и $\tilde{V}_j$ позволяет оценить доверительные пределы для σ_j и V_j.

Полученные описанными способами оценки и доверительные пределы сведены в табл. 6.2.

Оценка плотности распределения, которая определяется на основе нормализующего преобразования, в соответствии с (2.38) имеет вид:

$$\check{p}_j(x) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\,\check{\delta}_j(a_j + \lambda_j x)}\exp\left\{-\dfrac{1}{2\check{\delta}_j^2}[\ln(a_j + \lambda_j x) - \check{\mu}_j]^2\right\} & \text{при } a_j + \lambda_j x > 0, \\ 0 & \text{при } a_j + \lambda_j x \leqslant 0. \end{cases} \tag{6.41}$$

График ее — одновершинная право- или левоасимметричная кривая (в зависимости от значения λ_j (2.44), $\lambda_j = 1$ или $\lambda_j = -1$), какие в большинстве случаев и встречаются на практике при обработке геолого-геофизической информации.

Параметр a_j формально является верхним пределом возможных значений показателя ξ_j при $\lambda_j = -1$ (гл. 2). В случае $\lambda_j = 1$ величина a_j, взятая с обратным знаком, служит нижним пределом возможных значений ξ_j. На практике, однако, эти пределы далеко не всегда можно указать. В этих случаях, если аппроксимация с помощью логарифмически нормального распределения (т. е. при $a_j = 0$, $\lambda_j = 1$) неудовлетворительна, возникает задача нахождения таких оценок параметров a_j и λ_j, которые обеспечивают наилучшую аппроксимацию плотности распределения и получение оценок необходимых числовых характеристик.

Оценки $\check{a}_j$, $\check{\mu}_j$, $\check{\delta}_j^2$ максимального правдоподобия параметров a_j, μ_j, δ_j^2 определяются из системы уравнений правдоподобия (4.37), которую в нашем случае можно свести к виду:

$$\begin{cases} \sum\limits_{i=1}^{n} \dfrac{\check{\delta}_j^2 - \check{\mu}_j + \ln(\check{a}_j + \lambda_j x_{ij})}{\check{a}_j + \lambda_j x_{ij}} = 0, \\ \check{\mu}_j = \dfrac{1}{n} \sum\limits_{i=1}^{n} \ln(\check{a}_j + \lambda_j x_{ij}), \\ \check{\delta}_j^2 = \dfrac{1}{n} \sum\limits_{i=1}^{n} [\ln(\check{a}_j + \lambda_j x_{ij}) - \check{\mu}_j]^2. \end{cases}$$

Решив эту систему при $\lambda_j = 1$ и $\lambda_j = -1$ одним из итерационных методов на ЭВМ (как, например, в [11]), подставим найденные оценки в выражение функции правдоподобия. Сравнив ее значения при $\lambda_j = 1$ и $\lambda_j = -1$, найдем искомые оценки $\check{a}_j$, $\check{\mu}_j$, $\check{\delta}_j$ и λ_j, а подставив их в формулы табл. 6.2, в (6.40) и (6.41), получим оценки числовых характеристик и плотности распределения. Опыт показывает, что эта аппроксимация плотности для одновершинных распределений геохимических и петрофизических показателей пород дает, как правило, удовлетворительный результат.

Оценка параметров нормального и логарифмически нормального распределений с применением вероятностной бумаги. Для получения приближенных оценок параметров нормального или логнормального распределений можно применить простой метод, основанный на использовании вероятностной бумаги (рисунки в Приложении). С этой целью на вероятностной бумаге (нормальной или логнормальной, в зависимости от того, какой закон используется для аппроксимации) строят выборочную функцию распределения. Если эмпирические данные согласуются с нормальным либо логнормальным законами рас-

пределения, нанесенные на соответствующую вероятностную бумагу точки накопленных частот будут группироваться вокруг прямой линии. По максимальному отклонению d_n выборочной функции распределения от осредняющей прямой можно с помощью критерия Колмогорова проверить гипотезу о согласии наблюдений с нормальным либо логнормальным законами распределения.

Если принята гипотеза о нормальном распределении, по абсциссе точки с ординатой 0,5 находят оценку $\tilde{M}$ математического ожидания, служащую одновременно оценкой медианы и моды. Определив абсциссу x_0 точки на прямой с ординатой 0,841, которой соответствует значение $u_0 = 1$ по оси u (см. Приложение), находят оценку среднего квадратического отклонения $\tilde{s} = x_0 - \tilde{M}$. Эта оценка может быть найдена и как разность абсцисс любых двух точек прямой, разность ординат которых по оси u равна единице. Квантиль порядка q определяют как абсциссу точки прямой, ордината которой равна q.

В случае логнормального распределения абсцисса точки с ординатой 0,5 представляет собой оценку медианы $\tilde{M}_e$. Оценка $\tilde{\delta}$ среднего квадратического отклонения логарифма анализируемого показателя определяется в виде $\tilde{\delta} = \lg \frac{x_0}{\tilde{M}_e}$, где x_0 — абсцисса точки с ординатой 0,841 ($u_0 = 1$ по оси u). Оценки математического ожидания, моды, квантилей, среднего квадратического отклонения находят по формулам табл. 6.2, приняв в них $\check{\mu} = \lg \tilde{M}_e$, $a = 0$, $\lambda = 1$. Оценкой квантиля порядка q служит также абсцисса точки прямой с ординатой q.

Оценка параметра распределения Пуассона. Иногда можно заранее указать закон распределения, которому подчиняется тот или иной показатель. Так, распределение ошибок при таких методах измерений, когда в процессе измерения происходит суммирование случайных кратковременных импульсов (например, фиксирующих радиоактивные распады) обычно следует закону Пуассона (гл. 2). Распределение Пуассона, как упоминалось в гл. 2, зависит лишь от одного параметра λ_j, равного математическому ожиданию и дисперсии ξ_j.

Если $x_{1j}, x_{2j}, \ldots, x_{nj}$ — наблюдения показателя ξ_j, следующего распределению Пуассона, то уравнение для получения оценки наибольшего правдоподобия параметра $\lambda_j = \mathbf{M}\xi_j$ примет вид

$$\frac{\partial}{\partial \lambda_j}\left[\sum_{i=1}^{n} \ln\left(e^{-\lambda_j}\lambda_j^{x_{ij}}\frac{1}{x_{ij}!}\right)\right] = \frac{\partial}{\partial \lambda_j}\left(\sum_{i=1}^{n} x_{ij}\ln\lambda_j - n\lambda_j\right) = \frac{1}{\lambda_j}\sum_{i=1}^{n} x_{ij} - n = 0,$$

откуда

$$\check{\lambda}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = \bar{x}_j. \qquad (6.42)$$

Отсюда, в частности, следует, что наиболее правдоподобной оценкой дисперсии этого распределения служит $\bar{x}_j$.

Таблица 6.2. **Оценки числовых характеристик и доверительные пределы на основе аппроксимации распределений обобщенно-логнормальным законом**

Числовая характеристика	Оценка при нормализующем преобразовании $\ln(a_j+\lambda_j\xi_j)$	Доверительные пределы уровня α
Математическое ожидание M_j	$\tilde{M}_j=\left[\exp\left(\check{\mu}_j+\frac{1}{2}\check{\delta}_j^2\right)-a_j\right]\lambda_j$	$\sim\left[\exp\left(\check{\mu}_j+\frac{1}{2}\check{\delta}_j^2\pm u_{\frac{1+q}{2}}\frac{\delta_j}{\sqrt{2n}}\sqrt{2+\delta_j^2}\right)-a_j\right]\lambda_j$
Среднее квадратическое отклонение σ_j	$\tilde{\sigma}_j=(\lambda_j\tilde{M}_j+a_j)\sqrt{\exp\check{\delta}_j^2-1}$	$\sim\tilde{\sigma}_j\pm u_{\frac{1+q}{2}}\frac{\sigma_j\delta_j}{\sqrt{n}}\sqrt{1+2\delta_j^2\left(1+\frac{1}{2}\exp(2\mu_j+\delta_j^2)\,\sigma_j^{-2}\right)^2}$
Коэффициент вариации V_j	$\tilde{V}_j=\frac{\tilde{\sigma}_j}{\tilde{M}_j}=(\lambda_j+a_j\tilde{M}_j^{-1})\sqrt{\exp\check{\delta}_j^2-1}$	$\sim\tilde{V}_j\pm u_{\frac{1+q}{2}}\frac{V_j\delta_j}{\sqrt{n}}\sqrt{1+\left(1+\frac{a_j\lambda_j}{M_j}\right)^2\left(1+\frac{\delta_j^2}{2}\right)+\frac{\delta_j^2}{2}\left[2+\left(V_j^{-1}+\frac{a_j\lambda_j}{\sigma_j}\right)^2\right]^2}$
Критическая граница B_j (квантиль порядка p)	$\tilde{B}_j=[\exp(\check{\mu}_j+\lambda_j u_p\check{\delta}_j)-a_j]\lambda_j$	$\sim\left[\exp\left(\check{\mu}_j+\lambda_j u_p\check{\delta}_j\pm u_{\frac{1+q}{2}}\delta_j\sqrt{\frac{1+\frac{1}{2}u_p^2}{n}}\right)-a_j\right]\lambda_j$
Мода M_{0j}	$\tilde{M}_{0j}=[\exp(\check{\mu}_j-\check{\delta}_j^2)-a_j]\lambda_j$	$\sim\left[\exp\left(\check{\mu}_j-\check{\delta}_j^2\pm u_{\frac{1+q}{2}}\frac{\delta_j}{\sqrt{n}}\sqrt{1+2\delta_j^2}\right)-a_j\right]\lambda_j$
Медиана M_{ej}	$\tilde{M}_{ej}=(\exp\check{\mu}_j-a_j)\lambda_j$	$\sim\left[\exp\left(\check{\mu}_j\pm u_{\frac{1+q}{2}}\frac{\delta_j}{\sqrt{n}}\right)-a_j\right]\lambda_j$

Примечание. Для логнормального распределения ($a_j=0$, $\lambda_j=1$) формулы табл. 6.2 можно записать в другой форме: $\tilde{M}_{0j}=\tilde{M}_{ej}^3\tilde{M}_j^{-2}$, $\tilde{\sigma}_j=\tilde{M}_j\sqrt{\tilde{M}_{ej}\tilde{M}_{0j}^{-1}-1}$; $\tilde{M}_{ej}=(x_{1j}x_{2j}\dots x_{nj})^{1/n}$ (оценка медианы — *среднее геометрическое* из наблюдений) и т. д.

Приближенные доверительные пределы для параметра λ_j строятся на основании асимптотически нормального распределения оценки (6.42):

$$\bar{x}_j \pm u_{\frac{1+q}{2}} \sqrt{\frac{\mathbf{D}\xi_j}{n}} \approx \bar{x}_j \pm u_{\frac{1+q}{2}} \sqrt{\frac{\bar{x}_j}{n}}. \qquad (6.43)$$

Квантили распределения Пуассона можно оценивать с помощью табл. 1 (Приложение) и по формуле (2.8). Расчет по этой формуле, впрочем, при больших λ_j довольно трудоемок. Гораздо менее трудоемкий способ расчета основан на том, что функция распределения Пуассона представляется [2] в виде

$$F(x) = \mathbf{P}\{\xi_j < x\} = 1 - F_{2x}(2\lambda_j), \quad x = 1, 2, \ldots, \qquad (6.44)$$

где $F_{2x}(u)$ — функция χ^2-распределения с $2x$ степенями свободы. Пользуясь таблицей распределения χ^2 [2] при $2x \leqslant 100$, можно оценивать функцию распределения по (6.44), подставив вместо λ_j оценку (6.42). При $2x > 100$ для расчетов можно воспользоваться аппроксимацией с помощью нормального закона распределения:

$$F(x) = \mathbf{P}\{\xi_j < x\} \approx \Phi\left(\frac{x - \lambda_j}{\sqrt{\lambda_j}}\right), \qquad (6.45)$$

где $\Phi(u)$ — функция (0; 1)-нормального распределения, λ_j — параметр распределения Пуассона, оцениваемый по (6.42).

Пример 6.8. Для изучения распределения ошибок измерения гаммаплотномером плотности пород при постоянной малой экспозиции проведены измерения одного и того же образца, по которым вычислено среднее арифметическое количеств зафиксированных импульсов, составившее $\bar{x} = 30\gamma$. 1) Считая, что распределение ошибок следует закону Пуассона, оценить погрешность измерений в виде пределов допустимых значений, в которые с вероятностью $q = 0{,}95$ должен попадать результат измерения. 2) Оценить относительную погрешность измерений при увеличении экспозиции в 20 раз и в том же уровне $q = 0{,}95$, считая $\bar{x}$ пропорциональным времени экспозиции.

Решение. 1) Искомые пределы a_q, b_q определим из условия $\mathbf{P}\{a_q \leqslant \xi < b_q\} = q = 0{,}95$. Приняв $\mathbf{P}\{\xi \geqslant b_q\} = \mathbf{P}\{\xi < a_q\} = \frac{1-q}{2}$, можно взять в качестве пределов a_q и b_q квантили порядков $\frac{1-q}{2}$ и $q + \frac{1-q}{2} = \frac{1+q}{2}$:

$$\mathbf{P}\{\xi < a_q\} = \frac{1-q}{2}, \quad \mathbf{P}\{\xi < b_q\} = \frac{1+q}{2}.$$

Из соотношения (6.44)

$$F_{2b_q}(\bar{2x}) = \frac{1-q}{2} = 0{,}025, \quad F_{2a_q}(\bar{2x}) = \frac{1+q}{2} = 0{,}975.$$

По табл. 3 (Приложение) квантилю $\chi^2_{0{,}025}(n) = 2\bar{x} = 60$ соответствует четное $n \approx 84$, а для $\chi^2_{0{,}975}(n) = 2\bar{x} = 60$ $n \approx 40$. Искомые пределы составят

$$a_q \approx \frac{40}{2} = 20, \quad b_q \approx \frac{84}{2} = 42.$$

Относительная погрешность измерения плотности оценивается в виде

$$\delta_q \approx \frac{b_q - a_q}{2\bar{x}} \cdot 100\% = \frac{22}{60} \cdot 100\% = 36{,}6\%.$$

2) При увеличении экспозиции в 20 раз среднее показаний составит $\bar{x} = 600\gamma$. Пользуясь нормальной аппроксимацией распределения ξ, имеем, по (6.45),

$$\Phi\left(\frac{a_q - \bar{x}}{\sqrt{\bar{x}}}\right) = 0{,}025, \quad \Phi\left(\frac{b_q - \bar{x}}{\sqrt{\bar{x}}}\right) = 0{,}975.$$

Таким образом, $\frac{a_q - \bar{x}}{\sqrt{\bar{x}}} = u_{0,025}$, $\frac{b_q - \bar{x}}{\sqrt{\bar{x}}} = u_{0,975}$; $a_q = \bar{x} + u_{0,025}\sqrt{\bar{x}}$, $b_q = \bar{x} + u_{0,975}\sqrt{\bar{x}}$, где $u_{0,025}$, $u_{0,975}$ — квантили (0; 1)-нормального распределения. По табл. 2 (Приложение) $u_{0,025} = -1{,}96$, $u_{0,975} = 1{,}96$ и $a_q = 600 - 1{,}96\sqrt{600} \approx 552\,[\gamma]$, $b_q = 600 + 1{,}96\sqrt{600} \approx 648\,[\gamma]$.

Относительная погрешность измерения плотности при той же вероятности $q = 0{,}95$ попадания результата измерения в задаваемый этой погрешностью интервал составит

$$\delta_q = \frac{b_q - a_q}{2\bar{x}} \cdot 100\% = \frac{u_{0,975}}{\sqrt{\bar{x}}} \approx 8\%.$$

Последняя формула показываст, в частности, что для уменьшения относительной ошибки в n раз экспозицию следует увеличить в n^2 раз.

Оценки по неравноточным данным. При статистическом анализе геолого-геофизических данных иногда возникает необходимость в вычислении оценки числовой характеристики по группе уже имеющихся состоятельных и несмещенных оценок, оценивающих эту характеристику с разной точностью. Пусть $\check{\alpha}_1, \check{\alpha}_2, \ldots, \check{\alpha}_t$ — такие оценки некоторого параметра α, вычисленные по t выборкам независимых наблюдений; математическое ожидание $\mathbf{M}\check{\alpha}_j = \alpha$ и дисперсия $\mathbf{D}\check{\alpha}_j = \frac{c_j}{n_j}$ (n_j — количества наблюдений в выборках, c_j — некоторые постоянные, $j = \overline{1, t}$).

Представим искомую оценку $\check{\alpha}$ в виде такой линейной комбинации $\check{\alpha} = \sum_{j=1}^{t} \beta_j \check{\alpha}_j$ имеющихся оценок, чтобы она была несмещенной и имела среди всех возможных линейных комбинаций минимальную дисперсию. Условие несмещенности: $\mathbf{M}\check{\alpha} = \mathbf{M}\left(\sum_{j=1}^{t} \beta_j \check{\alpha}_j\right) = \alpha \sum_{j=1}^{t} \beta_j = \alpha$, откуда $\sum_{j=1}^{t} \beta_j = 1$. Минимум дисперсии $\check{\alpha}$ при этом условии будет достигаться, по аналогии с (4.13), при

$$\beta_j = \frac{(\mathbf{D}\check{\alpha}_j)^{-1}}{\sum_{j=1}^{t} (\mathbf{D}\check{\alpha}_j)^{-1}} = \frac{\frac{n_j}{c_j}}{\sum_{j=1}^{t} \frac{n_j}{c_j}} \quad (j = \overline{1, t}). \tag{6.46}$$

В частности, если c_j во всех выборках одинаковы ($c_j = c$, $j = \overline{1, t}$), коэффициенты β_j примут вид

$$\beta_j = \frac{n_j}{\sum_{j=1}^{t} n_j}, \tag{6.46'}$$

т. е. они будут пропорциональны количествам наблюдений в соответствующих выборках. Например, если оценивается математическое ожидание M, а $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_t$ — средние арифметические, вычисленные по t выборкам с дисперсиями $D_1, D_2, \ldots, D_t$ представляемых ими величин, то средневзвешенная оценка будет иметь вид

$$\bar{x} = \sum_{j=1}^{t} \beta_j \bar{x}_j, \quad \text{где} \quad \beta_j = \frac{n_j D_j^{-1}}{\sum_{j=1}^{t} n_j D_j^{-1}}, \tag{6.47}$$

так как дисперсия оценки $\bar{x}_j$, $\mathbf{D}\bar{x}_j = \frac{D_j}{n_j}$. Этот простой вывод позволяет правильно ориентироваться в тех случаях, когда необходимо оценивать математическое ожидание по неравноточным наблюдениям: наблюдения, полученные в частях объекта с большими колебаниями показателя следует, по (6.47), брать с меньшим «весом», чем в областях малых колебаний (при том условии, что математическое ожидание в этих частях одно и то же).

Аналогично, если $\check{\sigma}_1^2, \check{\sigma}_2^2, \ldots, \check{\sigma}_t^2$ — оценки дисперсии D (предполагается, что $\mathbf{M}\check{\sigma}_i^2 = D$, $i = \overline{1, t}$), общей ее оценкой будет

$$\check{\sigma}^2 = \sum_{j=1}^{t} \beta_j \check{\sigma}_j^2, \quad \text{где} \quad \beta_j = \frac{n_j D}{\mu_{4j} - D^2} \left(\sum_{j=1}^{t} \frac{n_j D}{\mu_{4j} - D^2} \right)^{-1} = \frac{n_j}{E_j + 2} \left(\sum_{j=1}^{t} \frac{n_j}{E_j + 2} \right)^{-1}, \tag{6.48}$$

(E_j — коэффициенты эксцесса совокупностей, представленных оценками $\check{\sigma}_j^2$, μ_{4j} — четвертые центральные моменты, $j = \overline{1, t}$). Если E_j постоянны, $E_j \approx E = \text{const}$, $j = \overline{1, t}$, то

$$\check{\sigma}^2 = \sum_{j=1}^{t} \beta_j \check{\sigma}_j^2, \quad \text{где} \quad \beta_j = \frac{n_j}{\sum_{j=1}^{t} n_j}. \tag{6.48'}$$

§ 5. Оценка формы и силы связей геолого-геофизических показателей

Изучение связей геолого-геофизических показателей горных пород составляет одну из основных задач, которые решаются в процессе количественного анализа материалов геологических исследований. Результаты этого анализа вместе с оценками числовых характеристик

одномерных распределений дают общую количественную информацию об особенностях изучаемого объекта. Аппарат статистического анализа связей используется также для решения задачи косвенных измерений, например, для прогнозирования значений какого-либо параметра по комплексу значений других показателей и для оценки ошибки такого прогноза.

Задача содержит два основных вопроса: 1) оценка силы связи; 2) оценка формы, в которой она проявляется.

Воспользуемся представлениями, принятыми в гл. 3. Будем анализировать зависимость показателя ξ_i от ξ_j в форме

$$\xi_i = f(\xi_j) + \Delta_{ij};$$

$f(x)$ — условное математическое ожидание ξ_i при известном значении $\xi_j = x$, Δ_{ij} — случайная величина с нулевым математическим ожиданием, не зависящая от ξ_j.

Корреляционный анализ парных связей. Если функция $f(x)$ линейна, по (3.46) $\xi_i = \alpha_{ij}\xi_j + \beta_{ij} + \Delta_{ij}$; сила связи оценивается коэффициентом корреляции, оценка которого, как показано в гл. 4 (4.33),

$$\check{r}_{ij} = \frac{1}{ns_i s_j}\sum_{l=1}^{n}(x_{li} - \bar{x}_i)(x_{lj} - \bar{x}_j) = \frac{1}{ns_i s_j}\left(\sum_{l=1}^{n} x_{li}x_{lj} - n\bar{x}_i\bar{x}_j\right), \quad (6.49)$$

где $\bar{x}_i$, $\bar{x}_j$, s_i, s_j — оценки математических ожиданий и средних квадратических отклонений величин ξ_i, ξ_j,

$$\bar{x}_i = \frac{1}{n}\sum_{l=1}^{n} x_{li}, \quad \bar{x}_j = \frac{1}{n}\sum_{l=1}^{n} x_{lj}, \quad s_i = \sqrt{\frac{1}{n}\sum_{l=1}^{n}(x_{li} - \bar{x}_i)^2},$$
$$s_j = \sqrt{\frac{1}{n}\sum_{l=1}^{n}(x_{lj} - \bar{x}_j)^2}. \quad (6.50)$$

Доверительные пределы для коэффициента корреляции, соответствующие доверительной вероятности q, с учетом (4.34) и в предположении нормальности совместного распределения ξ_i и ξ_j,

$$r_{ij}^{-} = \check{r}_{ij} - u_{\frac{1+q}{2}}\frac{1-\check{r}_{ij}^2}{\sqrt{n}}, \quad r_{ij}^{+} = \check{r}_{ij} + u_{\frac{1+q}{2}}\frac{1-\check{r}_{ij}^2}{\sqrt{n}}. \quad (6.51)$$

Эти пределы являются приближенными, так как построены на основе асимптотического распределения выборочного коэффициента корреляции. Более точная аппроксимация достигается использованием *преобразования Фишера* вида

$$z_{ij} = \frac{1}{2}\ln\frac{1+\check{r}_{ij}}{1-\check{r}_{ij}}. \quad (6.52)$$

Математическое ожидание и дисперсия z_{ij},

$$\mathbf{M}z_{ij} = \mathbf{M}\left(\frac{1}{2}\ln\frac{1+\check{r}_{ij}}{1-\check{r}_{ij}}\right) = \frac{1}{2}\ln\frac{1+r_{ij}}{1-r_{ij}} + O\left(\frac{1}{n}\right); \quad \mathbf{D}z_{ij} = \frac{1}{n-3} + o\left(\frac{1}{n}\right).$$

Доверительные пределы для $\frac{1}{2}\ln\frac{1+r_{ij}}{1-r_{ij}}$ определяются на основании асимптотически нормального закона распределения z_{ij}:

$$z'_{ij} = \frac{1}{2}\ln\frac{1+\check{r}_{ij}}{1-\check{r}_{ij}} - u_{\frac{1+q}{2}}\sqrt{\frac{1}{n-3}}, \quad z''_{ij} = \frac{1}{2}\ln\frac{1+\check{r}_{ij}}{1-\check{r}_{ij}} + u_{\frac{1+q}{2}}\sqrt{\frac{1}{n-3}}.$$

Поскольку обратное преобразование имеет вид $r_{ij} = (e^{2z_{ij}} - 1)(e^{2z_{ij}} + 1)^{-1}$, в качестве доверительных пределов для r_{ij} можно взять

$$r_{ij}^{-} = \frac{e^{2z'_{ij}} - 1}{e^{2z'_{ij}} + 1}, \quad r_{ij}^{+} = \frac{e^{2z''_{ij}} - 1}{e^{2z''_{ij}} + 1}. \tag{6.52}$$

Перейдя к десятичным логарифмам и основаниям, получим

$$e^{2z'_{ij}} = 10^{2z'_{ij}\lg e} = \frac{1+\check{r}_{ij}}{1-\check{r}_{ij}}\frac{1}{c_n}, \quad e^{2z''_{ij}} = 10^{2z''_{ij}\lg e} = \frac{1+\check{r}_{ij}}{1-\check{r}_{ij}}c_n,$$

где $c_n = 10^{0{,}8686u_{\frac{1+q}{2}}\sqrt{\frac{1}{n-3}}}$. Таким образом,

$$r_{ij}^{-} = \frac{1 - c_n + (1+c_n)\check{r}_{ij}}{1 + c_n + (1-c_n)\check{r}_{ij}}, \quad r_{ij}^{+} = \frac{c_n - 1 + (c_n+1)\check{r}_{ij}}{c_n + 1 + (c_n-1)\check{r}_{ij}}. \tag{6.52'}$$

Хотя при отсутствии линейной связи коэффициент корреляции равен нулю, нельзя ожидать, что в этом случае выборочный коэффициент корреляции также будет равен нулю. Его колебания вокруг нуля можно учесть, указав пределы, в которых при справедливости гипотезы об отсутствии связи должен с близкой к единице вероятностью q находиться выборочный коэффициент корреляции. Эти пределы, подобно (6.51), имеют приближенный вид (см. также (5.15))

$$-u_{\frac{1+q}{2}}\sqrt{\frac{1}{n}}, \quad u_{\frac{1+q}{2}}\sqrt{\frac{1}{n}} \tag{6.51'}$$

или, более точно, с использованием преобразования Фишера, вид (6.52) при $z'_{ij} = -u_{\frac{1+q}{2}}\sqrt{\frac{1}{n-3}}$, $z''_{ij} = u_{\frac{1+q}{2}}\sqrt{\frac{1}{n-3}}$. Если $\check{r}_{ij}$ выходит из этих пределов, гипотеза об отсутствии связи отвергается с вероятностью ошибки $\sim(1-q)$. Другой более точный критерий некоррелированности основан на том, что при отсутствии связи $t = \frac{\check{r}\sqrt{n-2}}{\sqrt{1-\check{r}^2}}$ следует распределению Стьюдента с $n-2$ степенями свободы. Гипотеза об отсутствии связи отвергается, если $|t| \geqslant t_{\frac{1+q}{2}}$, где $t_{\frac{1+q}{2}}$ —

$\frac{1+q}{2}$-квантиль распределения Стьюдента с $n-2$ степенями свободы.

Форма связи (3.46) определяется коэффициентами α_{ij}, β_{ij}, оценки которых, в соответствии с (3.47), вычисляются по формулам *

$$\check{\alpha}_{ij} = \check{r}_{ij}\frac{s_i}{s_j}, \quad \check{\beta}_{ij} = \bar{x}_i - \check{\alpha}_{ij}\bar{x}_j, \tag{6.53}$$

а оценка среднего квадратического отклонения Δ_{ij} — ошибки косвенного измерения, по (3.48), в виде

$$s_{ij} = s_i\sqrt{1-\check{r}_{ij}^2}. \tag{6.54}$$

При совместном нормальном распределении коррелируемых величин или, по крайней мере, нормальном распределении Δ_{ij}, доверительные пределы для прогнозируемого с помощью функции регрессии значения ξ_i по значению ξ_j составят (приближенно)

$$\check{\alpha}_{ij}x + \check{\beta}_{ij} - u_{\frac{1+q}{2}}s_{ij}, \quad \check{\alpha}_{ij}x + \check{\beta}_{ij} + u_{\frac{1+q}{2}}s_{ij}, \tag{6.55}$$

и они обладают тем свойством, что при $\xi_j = x$ значение ξ_i, которое могло быть получено прямым измерением, с вероятностью $\sim q$ находится в этих пределах.

Пример 6.9. Проверить гипотезу о наличии связи между величинами кажущихся сопротивлений (КС) $\rho_{0,55}$ и $\rho_{1,05}$, $\rho_{0,55}$ и $\rho_{2,75}$, измеренных в нефтенасыщенных пластах терригенных отложений при различных зондах — 0,55, 1,05 и 2,75 *м*, если выборочные коэффициенты корреляции по 20 пластам составили: $r_{i_2} = 0{,}91$ (для зондов 0,55 и 1,05 *м*); $r_{i_3} = 0{,}15$ (для зондов 0,55 и 2,75 *м*).

Решение. Пределы, в которых с вероятностью $q = 0{,}95$ должен находиться выборочный коэффициент корреляции при гипотезе об отсутствии связи, составляют, по (6.51′), $r^- \approx -u_{0,975}\sqrt{\frac{1}{n}} = -1{,}96\sqrt{\frac{1}{20}} \approx -0{,}44$; $r^+ \approx u_{0,975}\sqrt{\frac{1}{n}} \approx$ $\approx 0{,}44$. Так как $r_{i_2} > r^+$, гипотеза о наличии связи между $\rho_{0,55}$ и $\rho_{1,05}$ принимается с вероятностью ошибки $\alpha < 0{,}05$. Для $\rho_{0,55}$ и $\rho_{1,05}$ эта гипотеза отвергается, так как $r^- < r_{i_3} < r^+$. Полученные результаты можно объяснить особенностями залегания исследованных пластов. Так, малая мощность и невыдержанность пластов по площади приводят к усреднению величин КС при увеличении длины зонда и к ослаблению связи с КС при малой длине зонда.

Пример 6.10. На участке проведены измерения приращения гравитационного потенциала Δg_i в пунктах бурения скважин. По скважинам определены расстояния H_i от поверхности до фундамента. По этим данным вычислены: выборочный коэффициент корреляции Δg и H, составивший $r = -0{,}90$; средние арифметические $\bar{H} = 175$ *м*, $\overline{\Delta g} = 50$ *мгл*; выборочные средние квадратические отклонения $s_H = 15$ *м*, $s_g = 9$ *мгл*. 1) Построить уравнение регрессии для расчета H по Δg и оценить ошибку такого прогноза при доверительной вероятности $q = 0{,}90$. 2) Можно ли пользоваться этим уравнением для прогноза H на другом участке, если контрольная скважина на нем дала $H = 250$ *м*, $\Delta g = 25$ *мгл*?

---

* Следует иметь в виду, что значительное отклонение распределения от нормального закона часто влечет ослабление устойчивости оценок, получаемых по формулам (6.49), (6.53), а аномальные наблюдения могут их вообще резко исказить. Оценки коэффициентов корреляции, вычисленные по выборкам, объединяющим наблюдения из нескольких совокупностей, также могут значительно отличаться от оценок, вычисленных по каждой совокупности в отдельности.

Решение. 1) По (6.53) оценки коэффициентов регрессии $H = \alpha \Delta g + \beta -$

$$\alpha = -0{,}9\frac{15}{9} = -1{,}5, \quad \beta = \bar{H} - \alpha \bar{\Delta g} = 175 + 1{,}5 \quad 50 = 250;$$

уравнение регрессии: $H = -1{,}5\,\Delta g + 250$. Ошибка прогнозирования H по Δg: $\pm u_{0{,}95}\sqrt{1 - r^2}\, s_H = \pm 1{,}645\sqrt{1 - 0{,}9^2} \quad 15 \approx \pm 10{,}8$ *м*.

2) При $\Delta g = 25$ *мгл* H должно находиться, по (6.55), с вероятностью $q = 0{,}9$ в пределах $[\tilde{H} - 10{,}8$ *м*, $\tilde{H} + 10{,}8$ *м*$]$, где $\tilde{H}$ — прогнозное значение, $\tilde{H} = -1{,}5 \times 25 + 250 = 212{,}5$ *м*. Так как это условие не выполняется, пользоваться уравнением для расчета H на другом участке нецелесообразно.

Пример 6.11. По данным измерений плотности δ и содержания кремнекислоты SiO_2 в эффузивах (52 пробы) получены следующие результаты: оценки математического ожидания (средние арифметические) $\bar{x}_1 = 2{,}68$ *г/см*^{3}, $\bar{x}_2 = 61\%$; выборочные средние квадратические отклонения $s_1 = 0{,}128$ *г/см*^{3}; $s_2 = 6{,}10\%$; выборочный коэффициент корреляции $r_{12} = -0{,}96$. 1) Определить уравнения линейной регрессии содержания SiO_2 на плотность δ и плотности на SiO_2. 2) Вычислить доверительные пределы при $q = 0{,}95$ для оцениваемых с помощью уравнений регрессии значений: содержания SiO_2 по величине плотности и плотности — по SiO_2. 3) Вычислить доверительные пределы для коэффициента корреляции с помощью преобразования Фишера и сравнить их с пределами, вычисленными по формулам (6.51), приняв доверительную вероятность $q = 0{,}95$.

Решение. 1) Представляя форму связей в виде $\delta = \alpha_{12} SiO_2 + \beta_{12} + \Delta_{12}$, $SiO_2 = \alpha_{21}\delta + \beta_{21} + \Delta_{21}$, по (6.53) имеем: $\alpha_{12} = r_{12}\frac{s_1}{s_2} = -0{,}96\,\frac{0{,}128}{6{,}10} \approx -0{,}0202$; $\beta_{12} = \bar{x}_1 - \alpha_{12}\bar{x}_2 = 2{,}68 + 0{,}0202 \quad 61 \approx 3{,}91$; $\alpha_{21} = r_{12}\frac{s_2}{s_1} \approx -45{,}7$; $\beta_{21} = x_2 - \alpha_{21}\bar{x}_1 \approx 183{,}5$.

Уравнения регрессии оцениваются в виде: $\delta = -0{,}0202\, SiO_2 + 3{,}91$; $SiO_2 = -45{,}7\delta + 183{,}5$.

2) Оценки среднего квадратического отклонения ошибки косвенного измерения: содержания кремнекислоты (в %) по плотности —

$$s_{21} = s_2\sqrt{1 - r_{12}^2} = 6{,}1\sqrt{1 - 0{,}96^2} \approx 1{,}70\ [\%];$$

плотности по содержанию SiO_2 —

$$s_{12} = s_1\sqrt{1 - r_{12}^2} = 0{,}128\sqrt{1 - 0{,}96^2} \approx 0{,}0357\ [г/см^3].$$

Доверительные пределы (6.55) для прогнозируемых с помощью уравнений регрессии значений: содержания SiO_2 —

$$-45{,}7\delta + 183{,}5 \pm 1{,}96 s_{21} \approx -45{,}7\delta + 183{,}5 \pm 3{,}33\ [\%];$$

плотности —

$$-0{,}0202\, SiO_2 + 3{,}91 \pm 1{,}96 s_{12} = -0{,}0202\, SiO_2 + 3{,}91 \pm 0{,}070\ [г/см^3].$$

3) Для вычисления доверительных пределов r воспользуемся формулами (6.52′):

$$c_n = 10^{0{,}8686 \cdot 1{,}96\sqrt{\frac{1}{49}}} \approx 1{,}751; \quad r_{12}^{-} = \frac{-0{,}751 - 2{,}751 \cdot 0{,}96}{2{,}751 + 0{,}751 \cdot 0{,}96} \approx -0{,}98;$$

$$r_{12}^{+} = \frac{0{,}751 - 2{,}751 \quad 0{,}96}{2{,}751 - 0{,}751 \cdot 0{,}96} \approx -0{,}93.$$

По упрощенным же формулам (6.51) имеем:

$$\tilde{r}^{-}_{12} = -0{,}96 - 1{,}96\frac{1-0{,}96^2}{\sqrt{52}} \approx -0{,}96 - 0{,}02 = -0{,}98;$$

$$\tilde{r}^{+}_{12} = -0{,}96 + 1{,}96\frac{1-0{,}96^2}{\sqrt{52}} \approx -0{,}94.$$

Криволинейные зависимости. Рассмотрим зависимости, которые с помощью известных преобразований переменных сводятся к линейным вида (3.46). Пусть $\varphi(x)$ и $g(x)$ некоторые функции, причем $\varphi(x)$ имеет обратную функцию $\psi(u)$ и связь двух показателей представляется в виде

$$\xi_i = \varphi(\alpha_{ij} g(\xi_j) + \beta_{ij}) + \Delta_{ij}. \tag{6.56}$$

Считая, что ξ_i принимает значения из области определения функции $\psi(u)$, а значения $|\Delta_{ij}|$ малы, применим преобразование $\psi(u)$ к обеим частям (6.56):

$$\psi(\xi_i) \approx \psi[\varphi(\alpha_{ij} g(\xi_j) + \beta_{ij})] + \psi'_{\varphi}\Delta_{ij} = \alpha_{ij} g(\xi_j) + \beta_{ij} + \delta_{ij}. \tag{6.57}$$

Введя новые переменные $\eta_i = \psi(\xi_i)$, $\eta_j = g(\xi_j)$, приходим к рассмотренному случаю линейной связи. Оценки коэффициентов α_{ij}, β_{ij} будут иметь вид: $\check{\alpha}_{ij} = \rho_{ij}\frac{s^0_i}{s^0_j}$, $\check{\beta}_{ij} = \bar{\psi}_i - \check{\alpha}_{ij}\bar{g}_j$, где ρ_{ij} — выборочный коэффициент корреляции преобразованных величин $\psi(\xi_i)$ и $g(\xi_j)$, вычисляемый по преобразованным наблюдениям $\psi(x_{ti})$, $g(x_{tj})$; $\bar{\psi}_i$, $\bar{g}_j$, s^0_i, s^0_j — оценки их математических ожиданий и средних квадратических отклонений, вычисляемые по этим наблюдениям*. Таким способом могут анализироваться регрессии вида

$$\tilde{\eta}_i = \alpha_{ij}\ln\xi_j + \beta_{ij},\quad \tilde{\eta}_i = \alpha_{ij}\xi_j^k + \beta_{ij},\quad \tilde{\eta}_i = \alpha_{ij}e^{\xi_j} + \beta_{ij},$$

где $\eta_i = \xi_i$, либо $\eta_i = \ln\xi_i$, $\eta_i = e^{\xi_i}$ и т. д.

Множественный корреляционный анализ. Рассмотрим множественную линейную связь вида (3.56) показателя ξ_{k+1} с группой показателей $\xi_1, \xi_2, \ldots, \xi_k$

$$\xi_{k+1} = \sum_{j=1}^{k}\alpha_j\xi_j + \alpha_{k+1} + \Delta_{k+1},$$

где α_j — коэффициенты, Δ_{k+1} — случайная величина с нулевым математическим ожиданием, не зависящая от $\xi_1, \xi_2, \ldots, \xi_k$. Как отмечалось в гл. 3, для коэффициентов $\alpha_1, \alpha_2, \ldots, \alpha_{k+1}$ справедливы соотношения (см. (3.61), (3.57))

$$\alpha_j = \beta_j\frac{\sigma_{k+1}}{\sigma},\ j = \overline{1,k};\quad \alpha_{k+1} = M_{k+1} - \sum_{j=1}^{k}\alpha_j M_j, \tag{6.58}$$

---

* Хотя ξ_j и δ_{ij} некоррелированы, в общем случае они зависимы. Учёт зависимости дисперсии δ_{ij} от значений ξ_j может помочь в уточнении оценок коэффициентов $\check{\alpha}_{ij}$ и $\check{\beta}_{ij}$.

где σ_j — средние квадратические отклонения показателей ξ_j ($j = \overline{1, k+1}$), M_j — математические ожидания ξ_j, $j = \overline{1, k+1}$; β_j ($j = \overline{1, k}$) — решение системы уравнений (3.59) $\mathbf{R}\beta = \mathbf{r}_{k+1}$ ($\beta = \{\beta_1, \beta_2, \ldots, \beta_k\}'$); $\mathbf{R} = \{r_{ij}\}_{i,j=1}^{k}$ — корреляционная матрица $\xi_1, \xi_2, \ldots, \xi_k$; $\mathbf{r}_{k+1} = \{r_{1k+1}, r_{2k+1}, \ldots, r_{kk+1}\}'$ — вектор коэффициентов корреляции $\xi_1, \xi_2, \ldots, \xi_k$ и ξ_{k+1}. Подставив в (6.58) вместо соответствующих величин их оценки, найденные по n наблюдениям $(k+1)$-мерной величины $\{\xi_1, \xi_2, \ldots, \xi_{k+1}\}$, получим оценки

$$\check{\alpha}_j = \check{\beta}_j \frac{s_{k+1}}{s_j} \quad (j = \overline{1, k}), \quad \check{\alpha}_{k+1} = \bar{x}_{k+1} - \sum_{j=1}^{k} \check{\alpha}_j \bar{x}_j, \tag{6.58'}$$

где s_j — оценки средних квадратических отклонений ξ_j вида (6.50) ($j = \overline{1, k+1}$); $\bar{x}_j$ — оценки математических ожиданий (6.50); $\check{\beta}_j$ ($j = \overline{1, k}$) — решение системы уравнений $\check{\mathbf{R}}\check{\beta} = \check{\mathbf{r}}_{k+1}$ ($\check{\beta} = \{\check{\beta}_1, \check{\beta}, \ldots, \check{\beta}_k\}'$); $\check{\mathbf{R}} = \{\check{r}_{ij}\}_{i,j=1}^{k}$ — матрица выборочных коэффициентов корреляции $\xi_1, \xi_2, \ldots, \xi_k$, вычисляемых по (6.49); $\check{\mathbf{r}}_{k+1}$ — вектор выборочных коэффициентов корреляции $\xi_1, \xi_2, \ldots, \xi_k$ и ξ_{k+1}: $\check{\mathbf{r}}_{k+1} = \{\check{r}_{1k+1}, \check{r}_{2k+1}, \ldots, \check{r}_{kk+1}\}'$.

Оценками $\check{\alpha}_j$ определяется оценка функции регрессии —

$$\tilde{\xi}_{k+1} = \sum_{j=1}^{k} \check{\alpha}_j \xi_j + \check{\alpha}_{k+1}. \tag{6.59}$$

Сила линейной связи характеризуется множественным коэффициентом корреляции R_{k+1}, оценка которого, в соответствии с (3.62), вычисляется в виде

$$R_{k+1} = \sqrt{\check{\beta}'\check{\mathbf{r}}_{k+1}} = \sqrt{\sum_{j=1}^{k} \check{\beta}_j \check{r}_{jk+1}}. \tag{6.60}$$

Функция распределения выборочного множественного коэффициента корреляции довольно сложна, так что расчет доверительных пределов весьма трудоемок. Мы рассмотрим тот частный случай, когда множественный коэффициент корреляции равен нулю, т. е. предполагается отсутствие линейной множественной связи. Это даст возможность, как в рассматривавшейся выше задаче парной корреляции, проверять гипотезу о наличии связи.

Доказано [1], что при $R_{k+1} = 0$ и нормальном распределении $\{\xi_1, \xi_2, \ldots, \xi_{k+1}\}$ величина

$$F = \frac{\check{R}_{k+1}^2 (n-k-1)}{(1-\check{R}_{k+1}^2)\, k} \tag{6.61}$$

подчиняется распределению Фишера с k и $n-k-1$ степенями свободы. Таким образом, если выполняется условие

$$\frac{\check{R}^2_{k+1}(n-k-1)}{(1-\check{R}^2_{k+1})k} \geqslant F_q(k, n-k-1) \tag{6.62}$$

($F_q(k, n-k-1)$ — q-квантиль распределения Фишера с k и $n-k-1$ степенями свободы), то гипотеза об отсутствии линейной связи между ξ_{k+1} и $\xi_1, \xi_2, \ldots, \xi_k$ отвергается с вероятностью ошибки $1-q$. Критерий является односторонним, так как о наличии связи могут свидетельствовать лишь «слишком высокие» значения $\check{R}_{k+1}$, а с ним и отношения $\dfrac{\check{R}^2_{k+1}}{1-\check{R}^2_{k+1}}$.

Оценка среднего квадратического отклонения Δ_{k+1}, определяющая точность косвенного измерения, в соответствии с (3.63)

$$s_\Delta = s_{k+1}\sqrt{1-\check{R}^2_{k+1}}. \tag{6.63}$$

Доверительные пределы для значений показателя ξ_{k+1} при его косвенном измерении с помощью уравнения регрессии (6.59), обладающие тем свойством, что попадание в них результата прямого измерения ξ_{k+1} гарантируется с вероятностью q, оцениваются в виде:

$$\sim \sum_{j=1}^{k} \check{\alpha}_j \xi_j + \check{\alpha}_{k+1} - u_{\frac{1+q}{2}} s_\Delta, \quad \sim \sum_{j=1}^{k} \check{\alpha}_j \xi_j + \check{\alpha}_{k+1} + u_{\frac{1+q}{2}} s_\Delta \tag{6.64}$$

(предполагая нормальным распределение случайного отклонения Δ_{k+1}).

Для оценки «долевого участия» показателей $\xi_1, \xi_2, \ldots, \xi_k$ в связи, описываемой регрессией (3.56′), очевидно, можно воспользоваться частными коэффициентами корреляции $r^{(k-1)}_{j\,k+1}$ (ξ_{k+1} и ξ_j при исключенном влиянии остальных показателей). Способы оценки частных коэффициентов корреляции мы рассмотрим ниже. Иногда употребляются и другие упрощенные характеристики, помогающие интерпретировать множественные линейные связи.

В частности, для этой цели можно использовать соответствующим образом нормированные коэффициенты $\alpha_1, \alpha_2, \ldots, \alpha_k$, учитывая, что чем больше коэффициент α_j, тем больше влияние ξ_j в зависимости (3.56). Сравнивать непосредственно величины α_j нецелесообразно, так как они являются коэффициентами при показателях, имеющих разные размерности и вариацию. Между тем показатель ξ_j влияет на ξ_{k+1} в зависимости (3.56) как раз за счет своей вариации. Если, например, в собственных единицах измерений среднее квадратическое отклонение ξ_1 равно 10^{-1}, а ξ_2 — 10, то ξ_1 и ξ_2 будут «равноправными» в том случае, если коэффициент при ξ_1 больше коэффициента при ξ_2 в 100 раз. При этом $\alpha_1\xi_1$ и $\alpha_2\xi_2$ будут иметь одинаковые дисперсии.

Нормируем величины ξ_j, введя $\xi_{jн} = \frac{1}{\sigma_j}\xi_j \quad (j = \overline{1, k})$ с тем, чтобы привести их к одной и той же дисперсии $\mathbf{D}\xi_{jн} = \frac{1}{\sigma_j^2}\sigma_j^2 = 1$ и выясним вид зависимости между ξ_{k+1} и $\xi_{1н}$, $\xi_{2н}$, ..., $\xi_{kн}$. Очевидно,

$$\xi_{k+1} = \sum_{j=1}^{k} \alpha_j \xi_j + \alpha_{k+1} + \Delta_{k+1} = \sum_{j=1}^{k} \alpha_j \sigma_j \xi_{jн} + \alpha_{k+1} + \Delta_{k+1} =$$

$$= \sum_{j=1}^{k} \beta_j \sigma_{k+1} \xi_{jн} + \alpha_{k+1} + \Delta_{k+1}$$

и коэффициенты при $\xi_{jн}$, равные $\beta_j\sigma_{k+1}$, могут служить для сравнения влияния отдельных показателей. Эти характеристики для удобства использования нормируют:

$$\gamma_j = \frac{\beta_j \sigma_{k+1}}{\sum_{j=1}^{k} |\beta_j \sigma_{k+1}|} = \frac{\beta_j}{\sum_{j=1}^{k} |\beta_j|} \quad (j = \overline{1, k}) \tag{6.65}$$

либо

$$\gamma_j' = \frac{\beta_j \sigma_{k+1}}{\sqrt{\sum_{j=1}^{k} \beta_j^2 \sigma^2_{k+1}}} = \frac{\beta_j}{\sqrt{\sum_{j=1}^{k} \beta_j^2}} \quad (j = \overline{1, k}). \tag{6.66}$$

Подставив вместо β_j их оценки $\check{\beta}_j$, получим оценки коэффициентов γ_j, γ_j'.

Этот прием дает хорошие результаты при независимых или слабо зависимых аргументах регрессии ξ_1, ξ_2, ..., ξ_k. Наличие сильной зависимости показателей приводит к неустойчивости коэффициентов γ_j, γ_j', следствием которой могут быть искаженные представления о «долевом участии» показателей в связи, так что их интерпретацию следует проводить с учетом этой особенности.

Формулы для вычисления оценок коэффициентов линейной регрессии можно использовать и при анализе некоторых криволинейных зависимостей. Например, если предполагаемая связь двух показателей η и ξ имеет вид

$$\eta = \sum_{j=1}^{k} \alpha_j f_j(\xi) + \alpha_{k+1} + \Delta_{k+1}, \tag{6.67}$$

то, положив $f_j(\xi) = \eta_j$, получим зависимость вида (3.56):

$$\eta = \sum_{j=1}^{k} \alpha_j \eta_j + \alpha_{k+1} + \Delta_{k+1}.$$

Для оценки коэффициентов α_j $(j = \overline{1, k+1})$ можно, таким образом, применить формулы (6.58′), использовав для вычисления необходимых

величин уже не сами наблюдения, а результаты их преобразований:

$$s_j = \sqrt{\frac{1}{n}\sum_{m=1}^{n}[f_j(x_m) - \bar{f}_j]^2}, \quad \bar{f}_j = \frac{1}{n}\sum_{m=1}^{n} f_j(x_m) \quad (j = \overline{1, k});$$

$$s_{k+1} = \sqrt{\frac{1}{n}\sum_{m=1}^{n}(y_m - \bar{y})^2};$$

$$\check{r}_{ij} = \frac{1}{ns_is_j}\sum_{m=1}^{n}[f_i(x_m) - \bar{f}_i][f_j(x_m) - \bar{f}_j] \quad (i, \ j = \overline{1, k}, \ i \neq j);$$

$$\check{r}_{k+1\,j} = \frac{1}{ns_{k+1}s_j}\sum_{m=1}^{n}[f_j(x_m) - \bar{f}_j][y_m - \bar{y}] \quad (j = \overline{1, k})$$

(y_m, x_m — наблюдения η и ξ соответственно, $m = \overline{1, n}$).

Такой способ можно применить и тогда, когда форма связи двух показателей существенно отличается от линейной, однако параметрическое семейство функций регрессии неизвестно. Аппроксимируя эту зависимость, например, полиномом k-го порядка, мы и получим связь вида (6.67): $\eta = \sum_{j=1}^{k}\alpha_j\xi^j + \alpha_{k+1} + \Delta_{k+1}$, полагая в (6.67) $f_j(\xi) = \xi^j$.

Для *проверки линейности связи* двух показателей ξ и η можно следующим образом использовать описанные в этой главе критерии однородности. Если x_i, y_i — значения показателей ξ и η, то в случае линейной связи $\eta = \alpha\xi + \beta + \Delta$, значения $\Delta_i = y_i - \alpha x_i - \beta$, упорядоченные по возрастанию величин x_i, должны образовывать ряд независимых однородных наблюдений с нулевыми математическими ожиданиями. Сравнением, по аналогии с (6.1′), статистики g^2, вычисленной по значениям $\tilde{\Delta}_i = y_i - \check{\alpha}x_i - \check{\beta}$ с $\check{D} = \frac{1}{n-2}\sum_{i=1}^{n}\tilde{\Delta}_i^2$ можно оценить, насколько приемлема линейная форма регрессии ($\check{\alpha}$ и $\check{\beta}$ — оценки коэффициентов регрессии вида (6.53)). Подобную оценку можно сделать и с помощью серийного критерия. Таким же образом можно действовать при анализе криволинейных зависимостей.

Оценка частного коэффициента корреляции. Частный коэффициент корреляции $r_{ij}^{(k)}$ употребляется в качестве характеристики силы связи двух показателей ξ_i и ξ_j при фиксированных значениях группы других $\xi_1, \xi_2, \ldots, \xi_k$ (т. е. при их исключенном влиянии). Для получения его оценки по n наблюдениям $(k+2)$-мерной величины $\{\xi_i, \xi_j, \xi_1, \xi_2, \ldots, \xi_k\}$ можно воспользоваться формулой (3.66), связывающей $r_{ij}^{(k)}$ с коэффициентами корреляции r_{si}, r_{sj} и коэффициентами β_s', β_s'' линей-

ных регрессий нормированных величин $\frac{\xi_i - M_i}{\sigma_i}$, $\frac{\xi_j - M_j}{\sigma_j}$ на $\frac{\xi_s - M_s}{\sigma_s}$ ($s = \overline{1, k}$). Оценкой $r_{ij}^{(k)}$ будет

$$\check{r}_{ij}^{(k)} = (\check{r}_{ij} - \sum_{s=1}^{k} \check{\beta}_s'' \check{r}_{si}) [(1 - \sum_{s=1}^{k} \check{\beta}_s' \check{r}_{si})(1 - \sum_{s=1}^{k} \check{\beta}_s'' \check{r}_{sj})]^{-1/2}, \qquad (6.68)$$

где $\check{r}_{ij}$, $\check{r}_{is}$, $\check{r}_{js}$ ($s = \overline{1, k}$) — выборочные коэффициенты корреляции вида (6.49); $\{\check{\beta}_1', \check{\beta}_2', \ldots, \check{\beta}_k'\} = \check{\beta}_i'$, $\{\check{\beta}_1'', \check{\beta}_2'', \ldots, \check{\beta}_k''\} = \check{\beta}_j'$ — решения систем уравнений $\check{\mathbf{R}}\check{\beta}_i = \check{\mathbf{r}}_i$, $\check{\mathbf{R}}\check{\beta}_j = \check{\mathbf{r}}_j$. $\check{\mathbf{R}}$ — матрица выборочных коэффициентов корреляции величин $\xi_1, \xi_2, \ldots, \xi_k$: $\check{\mathbf{R}} = \{\check{r}_{st}\}_{s,\,t=1}^{k}$; $\check{\mathbf{r}}_i = \{\check{r}_{i1}, \check{r}_{i2}, \ldots, \check{r}_{ik}\}'$, $\check{\mathbf{r}}_j = \{\check{r}_{j1}, \check{r}_{j2}, \ldots, \check{r}_{jk}\}'$. Из (6.68) можно получить другое выражение для $\check{r}_{ij}^{(k)}$:

$$\check{r}_{ij}^{(k)} = \operatorname{sign} a_i \sqrt{a_i b_j}, \qquad (6.68')$$

где a_i — оценка коэффициента при ξ_i регрессии ξ_j на $\xi_i, \xi_1, \xi_2, \ldots, \xi_k$; b_j — оценка коэффициента при ξ_j регрессии ξ_i на $\xi_j, \xi_1, \xi_2, \ldots, \xi_k$; $\operatorname{sign} a_i = a_i |a_i|^{-1}$ при $a_i \neq 0$ и $\operatorname{sign} a_i = 0$ при $a_i = 0$.

Чтобы найти доверительные пределы для частного коэффициента корреляции $r_{ij}^{(k)}$, воспользуемся важным свойством (доказательство приведено [1]): если совместное распределение $\{\xi_i, \xi_j, \xi_1, \xi_2, \ldots, \xi_k\}$ нормально, то $\check{r}_{ij}^{(k)}$ распределен так же, как выборочный коэффициент корреляции двух совместно нормально распределенных величин, вычисленный по выборке объема $n - k$. Таким образом, доверительные пределы можно определять: по приближенным формулам (6.51)

$$\sim \check{r}_{ij}^{(k)} - u_{\frac{1+q}{2}} \frac{1 - (\check{r}_{ij}^{(k)})^2}{\sqrt{n-k}}, \quad \sim \check{r}_{ij}^{(k)} + u_{\frac{1+q}{2}} \frac{1 - (\check{r}_{ij}^{(k)})^2}{\sqrt{n-k}}, \qquad (6.69)$$

либо с помощью преобразования Фишера по формулам (6.52), (6.52′), с подстановкой $\check{r}_{ij}^{(k)}$ вместо $\check{r}_{ij}$ и $n - k$ вместо n.

Пример 6.12. По результатам 69 спектральных определений содержаний кобальта, цинка и никеля в гранодиоритах массива получены следующие данные. Средние значения: Co — $\bar{x}_3 = 5{,}9 \cdot 10^{-3}\%$, Zn — $\bar{x}_1 = 3{,}8 \cdot 10^{-3}\%$; Ni — $\bar{x}_2 = 3{,}7 \cdot 10^{-4}\%$; оценки среднего квадратического отклонения: Co — $s_3 = 2{,}4 \times 10^{-3}\%$; Zn — $s_1 = 1{,}8 \cdot 10^{-3}\%$; Ni — $s_2 = 3 \cdot 10^{-4}\%$; выборочные коэффициенты корреляции: Co и Zn — $\check{r}_{13} = 0{,}69$; Co и Ni — $\check{r}_{32} = 0{,}75$; Zn и Ni — $\check{r}_{12} = 0{,}55$. 1) Вычислить оценку множественного коэффициента корреляции кобальта с цинком и никелем и проверить гипотезу о наличии множественной связи. 2) Построить уравнение множественной регрессии Co на Zn и Ni и определить точность косвенного измерения содержания Co по Zn и Ni при доверительной вероятности $q = 0{,}95$. 3) Вычислить величину ожидаемой концентрации кобальта в пробе гранодиорита того же массива и пределы, в которых с вероятностью 0,95 должен находиться результат прямого измерения, если содержания цинка и никеля в этой пробе составили: $x_1 = 3 \cdot 10^{-3}\%$ (Zn); $x_2 = 4 \cdot 10^{-4}\%$ (Ni).

Решение. 1) В соответствии с формулой (6.60) оценка множественного коэффициента корреляции $\check{R}_3 = \sqrt{\check{\beta}_1 \check{r}_{13} + \check{\beta}_2 \check{r}_{23}}$, где $\check{\beta}_1$, $\check{\beta}_2$ определяются из системы уравнений

$$\begin{cases} \check{\beta}_1 + \check{\beta}_2 \check{r}_{12} = \check{r}_{13}, \\ \check{\beta}_1 \check{r}_{12} + \check{\beta}_2 = \check{r}_{23}; \end{cases}$$

($\check{r}_{11} = \check{r}_{22} = 1$). Имеем:

$$\check{\beta}_1 = \frac{\check{r}_{13} - \check{r}_{23}\check{r}_{12}}{1 - \check{r}_{12}^2} \approx 0{,}4, \qquad \check{\beta}_2 = \frac{\check{r}_{23} - \check{r}_{13}\check{r}_{12}}{1 - \check{r}_{12}^2} \approx 0{,}53$$

и оценка множественного коэффициента корреляции $\check{R}_3 = \sqrt{0{,}69 \cdot 0{,}4 + 0{,}75 \cdot 0{,}53} \approx \approx 0{,}82$.

Для проверки гипотезы об отсутствии связи воспользуемся критерием (6.62). $\frac{\check{R}_3^2}{1 - \check{R}_3^2} \frac{n-3}{2} \approx 67{,}6$. 95%-ный квантиль распределения Фишера с $\nu_1 = k = 2$, $\nu_2 = n - k - 1 = 66$ степенями свободы, по табл. 10 (Приложение), $F_{0,95}(2, 66) \approx \approx 3{,}15$. Так как неравенство (6.62) выполняется, $67{,}6 > 3{,}15$, гипотеза об отсутствии связи отвергается с вероятностью ошибки 0,05.

2) Оценки коэффициентов множественной регрессии

$$\widetilde{\mathrm{Co}} = \alpha_1 \mathrm{Zn} + \alpha_2 \mathrm{Ni} + \alpha_3$$

вычисляем по (6.58′):

$$\check{\alpha}_1 = \check{\beta}_1 \frac{s_3}{s_1} = 0{,}4 \frac{2{,}4 \cdot 10^{-3}}{1{,}8 \cdot 10^{-3}} \approx 0{,}53; \quad \check{\alpha}_2 = \check{\beta}_2 \frac{s_3}{s_2} = 0{,}53 \frac{2{,}4 \cdot 10^{-3}}{3{,}0 \cdot 10^{-4}} = 4{,}24;$$

$$\check{\alpha}_3 = \bar{x}_3 - \check{\alpha}_1 \bar{x}_1 - \check{\alpha}_2 \bar{x}_2 = 5{,}9 \cdot 10^{-3} - 0{,}53 \cdot 3{,}8 \cdot 10^{-3} - 4{,}24 \cdot 0{,}37 \cdot 10^{-3} \approx \approx 2{,}32 \cdot 10^{-3}.$$

Итак, уравнение связи оценивается в виде

$$\mathrm{Co} = 0{,}53\,\mathrm{Zn} + 4{,}24\,\mathrm{Ni} + 2{,}32 \cdot 10^{-3} + \Delta\ [\%].$$

Оценка среднего квадратического отклонения ошибки косвенных измерений, по (6.63):

$$s_\Delta = s_3 \sqrt{1 - \check{R}_3^2} = 2{,}4 \cdot 10^{-3} \sqrt{1 - 0{,}82^2} = 1{,}37 \cdot 10^{-3}\ [\%].$$

Точность косвенного измерения содержания Co по содержаниям Zn и Ni, при доверительной вероятности $q = 0{,}95$, составляет

$$\sim \pm u_{\frac{1+q}{2}} s_\Delta = \pm 1{,}96 \cdot 1{,}37 \cdot 10^{-3} \approx 2{,}68 \cdot 10^{-3}\ [\%];$$

доверительные пределы для содержания Co, прогнозируемого по содержаниям Zn и Ni

$$0{,}53\,\mathrm{Zn} + 4{,}24\,\mathrm{Ni} + 2{,}32 \cdot 10^{-3} \pm 2{,}68 \cdot 10^{-3}\ [\%].$$

3) Ожидаемая концентрация кобальта в пробе гранодиорита данного массива

$$x_3 = \check{\alpha}_1 x_1 + \check{\alpha}_2 x_2 + \check{\alpha}_3 = 0{,}53 \cdot 3 \cdot 10^{-3} + 4{,}24 \cdot 4 \cdot 10^{-4} + 2{,}32 \cdot 10^{-3} = = 5{,}61 \cdot 10^{-3}\ [\%].$$

Доверительные пределы, в которых должен находиться результат прямого измерения содержания Co с вероятностью 0,95

$$\underline{x_3} = \check{\alpha}_1 x_1 + \check{\alpha}_2 x_2 + \check{\alpha}_3 - u_{\frac{1+q}{2}} s_\Delta = 5{,}61 \cdot 10^{-3} - 2{,}68 \cdot 10^{-3} = 2{,}93 \cdot 10^{-3}\ [\%],$$

$$\overline{x_3} = \check{\alpha}_1 x_1 + \check{\alpha}_2 x_2 + \check{\alpha}_3 + u_{\frac{1+q}{2}} s_\Delta = 5{,}61 \cdot 10^{-3} + 2{,}68 \cdot 10^{-3} = 8{,}29 \cdot 10^{-3}\ [\%].$$

Пример 6.13. По данным примера 6.12: 1) вычислить оценку частного коэффициента корреляции между содержаниями кобальта и никеля при фиксированном содержании цинка; 2) построить доверительные пределы для частного коэффициента корреляции при доверительной вероятности $q = 0{,}95$; 3) проверить гипотезу о наличии такой частной связи при уровне значимости 0,05.

Решение. 1) Обозначим $\check{r}^{(1)}_{23}$ искомую оценку, ξ_1, ξ_2, ξ_3 — содержания Zn, Ni и Co соответственно. В этой задаче $k = 1$ (k — число фиксируемых компонент); корреляционная матрица R этих компонент состоит из одного элемента, равного единице: $r_{11} = 1$. В соответствии с принятыми выше общими обозначениями векторы $\check{\beta}_2 = \check{\beta}'_1$, $\check{\beta}_3 = \check{\beta}''_1$, $\check{r}_2 = \check{r}_{12}$, $\check{r}_3 = \check{r}_{13}$ также состоят из одной компоненты. $\check{\beta}'_1$ и $\check{\beta}''_1$ имеют смысл оценок коэффициентов линейных регрессий нормированных показателей ξ_2 и ξ_3 соответственно на нормированный же показатель ξ_1. Они являются решениями уравнений $R\check{\beta}_2 = \check{r}_2$, $R\check{\beta}_3 = \check{r}_3$, т. е. $r_{11}\check{\beta}'_1 = \check{r}_{12}$, $\check{r}_{11}\check{\beta}''_1 = \check{r}_{13}$, откуда $\check{\beta}'_1 = \check{r}_{12}$, $\check{\beta}''_1 = \check{r}_{13}$.

По формуле (6.68) имеем:

$$\check{r}^{(1)}_{23} = \frac{\check{r}_{23} - \check{\beta}''_1\check{r}_{12}}{\sqrt{(1 - \check{\beta}'_1\check{r}_{12})(1 - \check{\beta}''_1\check{r}_{13})}} = \frac{\check{r}_{23} - \check{r}_{13}\check{r}_{12}}{\sqrt{(1 - \check{r}^2_{12})(1 - \check{r}^2_{13})}} =$$

$$= \frac{0{,}75 - 0{,}69 \cdot 0{,}55}{\sqrt{(1 - 0{,}55^2)(1 - 0{,}69^2)}} \approx 0{,}61.$$

2) Воспользуемся для построения доверительных пределов упрощенной формулой (6.69). В нашем примере $k = 1$, и доверительные пределы

$$r^{\pm}_{23} = \check{r}^{(1)}_{23} \pm u_{0{,}975}\frac{1 - (\check{r}^{(1)}_{12})^2}{\sqrt{n-1}} = 0{,}61 \pm 1{,}96\frac{1 - 0{,}61^2}{\sqrt{68}} \approx 0{,}61 \pm 0{,}15;$$

$$r^{-}_{23} \approx 0{,}46, \quad r^{+}_{23} \approx 0{,}76.$$

Так как нуль оказался вне этих пределов, гипотеза об отсутствии связи отвергается с вероятностью ошибки $\sim 0{,}05$.

Метод наименьших квадратов и корреляционный анализ. Рассмотрим одно важное свойство оценок коэффициентов множественной регрессии $\check{\alpha}_1, \check{\alpha}_2, \ldots, \check{\alpha}_{k+1}$. Как упоминалось в гл. 3, оцениваемые ими коэффициенты $\alpha_1, \alpha_2, \ldots, \alpha_{k+1}$ регрессии (3.56′) таковы, что математическое ожидание квадрата отклонения ξ_{k+1} от произвольной линейной комбинации $\sum_{j=1}^{k} a_j\xi_j + a_{k+1}$ величин $\xi_1, \xi_2, \ldots, \xi_k$ минимально при $a_j = \alpha_j$ ($j = \overline{1, k+1}$):

$$\min_{a_1, a_2, \ldots, a_{k+1}} \mathbf{M}\left(\sum_{j=1}^{k} a_j\xi_j + a_{k+1} - \xi_{k+1}\right)^2 = \mathbf{M}\left(\sum_{j=1}^{k}\alpha_j\xi_j + \alpha_{k+1} - \right.$$

$$\left. - \xi_{k+1}\right)^2 = \mathbf{D}\,\Delta_{k+1}. \qquad (6.70)$$

Оценкам коэффициентов $\check{\alpha}_1, \check{\alpha}_2, \ldots, \check{\alpha}_{k+1}$ присуще сходное по характеру свойство: среди всех возможных наборов коэффициентов $a_1, a_2, \ldots, a_{k+1}$ сумма квадратов отклонений

$$S = \sum_{i=1}^{n}\left(\sum_{j=1}^{k} a_j x_{ij} + a_{k+1} - x_{ik+1}\right)^2 \qquad (6.71)$$

обращается в минимум при $a_j = \check{\alpha}_j$, $j = \overline{1, k+1}$. Оно выглядит естественным, если учесть, что оценкой $D = \mathbf{M}\left(\sum_{j=1}^{k} a_j\xi_j + a_{k+1} - \xi_{k+1}\right)^2$ может служить величина $\check{D} = \frac{S}{n}$, которая отличается от S лишь постоян-

ным множителем. Чтобы удостовериться в справедливости этого, найдем значения a_j^0, при которых обращается в минимум сумма (6.71). Необходимым условием минимума S является равенство нулю частных производных по a_j, в том числе и по a_{k+1}:

$$\frac{\partial S}{\partial a_{k+1}} = 2\sum_{i=1}^{n}\left(\sum_{j=1}^{k} a_j x_{ij} + a_{k+1} - x_{ik+1}\right) = 0,$$

откуда

$$a_{k+1}^0 = \frac{1}{n}\sum_{i=1}^{n} x_{ik+1} - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} a_j x_{ij} = \bar{x}_{k+1} - \sum_{j=1}^{k} a_j \bar{x}_j.$$

Подставив $a_{k+1} = a_{k+1}^0$ в S, преобразуем (6.71) к виду

$$S = \sum_{i=1}^{n}\left[\sum_{j=1}^{k} a_j (x_{ij} - \bar{x}_j) - (x_{ik+1} - \bar{x}_{k+1})\right]^2.$$

Теперь необходимым условием минимума S является равенство нулю частных производных по $a_1, a_2, \ldots, a_k$:

$$\frac{\partial S}{\partial a_l} = 2\sum_{i=1}^{n}\left[\sum_{j=1}^{k} a_j (x_{ij} - \bar{x}_j) - (x_{ik+1} - \bar{x}_{k+1})\right](x_{il} - \bar{x}_l) = 0,\ l = \overline{1, k},$$

откуда

$$\sum_{j=1}^{k} a_j \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{il} - \bar{x}_l) = \sum_{i=1}^{n}(x_{ik+1} - \bar{x}_{k+1})(x_{il} - \bar{x}_l),\ l = \overline{1, k}.$$

Разделим обе части равенства на $ns_{k+1}s_l$, обозначив s_j выборочное среднее квадратическое отклонение величины $\xi_j\ (j = \overline{1, k+1})$ вида (6.50). Учитывая (6.49), получим: $\frac{1}{s_{k+1}}\sum_{j=1}^{k} a_j s_j \check{r}_{jl} = \check{r}_{lk+1},\ l = \overline{1, k}$. Отсюда

$$a_j^0 = \check{\beta}_j \frac{s_{k+1}}{s_j},\ j = \overline{1, k}, \qquad (6.72)$$

причем $\check{\beta}_j\ (j = \overline{1, k})$ являются решением системы уравнений $\check{\mathbf{R}}\check{\boldsymbol{\beta}} = \check{\mathbf{r}}_{k+1}$ (см. (6.58′)). Соотношения (6.72), (6.58′) и доказывают указанное свойство коэффициентов $\check{\alpha}_j$. Из него следует, что, вообще говоря, формальное применение метода наименьших квадратов для отыскания коэффициентов линейной связи одного показателя с группой других приводит к тем же оценкам, которые вычисляются при множественном корреляционном анализе. В случае равноточных наблюдений оценки $\check{\alpha}_j$ этих

коэффициентов по методу наименьших квадратов определяются условием минимума S (4.54):

$$L_0 = \sum_{i=1}^{n} p_i \Big(\sum_{j=1}^{k} a_j x_{ij} + a_{k+1} - x_{ik+1} \Big)^2 = \sum_{i=1}^{n} \Big(\sum_{j=1}^{k} a_j x_{ij} + a_{k+1} - x_{ik+1} \Big)^2 = S,$$

так как вследствие $\mathbf{D}\,\Delta_{k+1} = \text{const}$ нужно принять $p_i = 1\,(i = \overline{1, n})$.

Аналогично, если рассматривается связь двух показателей —

$$\xi_2 = a_1 \xi_1 + a_2 + \Delta \tag{6.73}$$

(Δ — случайная величина с нулевым средним, не зависящая от ξ_1), то оценки коэффициентов по методу наименьших квадратов имеют вид

$$a_1^0 = \check{r}_{12} \frac{s_2}{s_1}, \quad a_2^0 = \bar{x}_2 - a_1^0 \bar{x}_1. \tag{6.74}$$

Итак, оценки по методу наименьших квадратов, вычисляемые из условия минимума формы (6.71), будут совпадать с оценками, вычисляемыми через выборочную корреляционную матрицу, независимо от того, с какими — случайными или неслучайными — компонентами анализируется связь. Конечно, в последнем случае элементам выборочной корреляционной матрицы $\check{\mathbf{R}}$, выборочным дисперсиям и средним нельзя придавать привычный смысл — здесь они являются не более чем элементами вычислительной процедуры.

Важная особенность метода наименьших квадратов — возможность определения точности оценок в виде доверительных пределов, используя их ковариационную матрицу (4.58). Для линейной множественной зависимости (3.56) с неслучайными компонентами при равноточных наблюдениях по (4.58) эта матрица имеет вид

$$\mathbf{B} = \sigma_p^2 (\mathbf{X}'\mathbf{X})^{-1}, \tag{6.75}$$

где $\mathbf{X}$ — матрица значений $\xi_1, \xi_2, \ldots, \xi_k$: $\mathbf{X} = \{x_{ij}\}$, $i = \overline{1, n}$, $j = \overline{1, k+1}$, x_{ij} — i-е значение ξ_j при $1 \leqslant j \leqslant k$, $x_{ik+1} = 1$, $i = \overline{1, n}$; $\sigma_p^2 = \mathbf{D}\,\Delta_{k+1}$. Оценка $\mathbf{D}\,\Delta_{k+1}$ вычисляется по (4.61): $\check{\sigma}_p^2 = \frac{1}{n-k-1} \times \times \sum_{i=1}^{n} (x_{ik+1} - \sum_{j=1}^{k} \check{\alpha}_j x_{ij} - \check{\alpha}_{k+1})^2$. Нетрудно убедиться, что $\check{\sigma}_p^2 = \frac{n}{n-k-1} s_{k+1}^2 (1 - \check{R}_{k+1}^2)$, где $\check{R}_{k+1}$ вычисляется по (6.60). При малых k по сравнению с n можно принять $\check{\sigma}_p^2 \approx s_{k+1}^2 (1 - \check{R}_{k+1}^2)$.

Дисперсией оценки $\check{\alpha}_j$ будет j-й диагональный элемент матрицы $\mathbf{B}$: $b_{jj} = \sigma_p^2 c_{jj}$, где c_{jj} — j-й диагональный элемент матрицы $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$. Оценка дисперсии $\check{\alpha}_j$, $\check{b}_{jj} = \check{\sigma}_p^2 c_{jj}$. Точность оценки $\check{\alpha}_j$ определяется

в виде доверительных пределов для α_i (4.59′), в предположении нормального распределения Δ_{k+1}:

$$\check{\alpha}_i - t_{\frac{1+q}{2}} \sqrt{\check{b}_{ii}},\ \check{\alpha}_i + t_{\frac{1+q}{2}} \sqrt{\check{b}_{ii}}, \tag{6.76}$$

где $t_{\frac{1+q}{2}} - \frac{1+q}{2}$ - квантиль распределения Стьюдента с $n-k-1$ степенями свободы. Приближенные пределы с использованием квантиля $u_{\frac{1+q}{2}}$ (0; 1)-нормального распределения —

$$\check{\alpha}_i - u_{\frac{1+q}{2}} \sqrt{\check{b}_{ii}},\ \check{\alpha}_i + u_{\frac{1+q}{2}} \sqrt{\check{b}_{ii}}. \tag{6.76′}$$

Выражения (6.76′) дают удовлетворительное приближение уже при $n-k-1>30$.

В случае парной линейной связи (6.73)

$$\mathbf{X}' = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ 1 & 1 & \cdots & 1 \end{pmatrix},\ \mathbf{X}'\mathbf{X} = \begin{pmatrix} \sum\limits_{i=1}^{n} x_i^2 & \sum\limits_{i=1}^{n} x_i \\ \sum\limits_{i=1}^{n} x_i & n \end{pmatrix},\ (\mathbf{X}'\mathbf{X})^{-1} =$$

$$= \frac{1}{n s_1^2} \begin{pmatrix} 1 & -\bar{x} \\ -\bar{x} & s_1^2 + \bar{x}^2 \end{pmatrix};$$

$$s_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,\ \check{\sigma}_p^2 = s_2^2(1-\check{r}^2), \qquad \check{r} = \frac{1}{n s_1 s_2}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}),$$

$s_2^2 = \frac{1}{n}\sum\limits_{i=1}^{n}(y_i - \bar{y})^2$; x_i — значения ξ_1, y_i — наблюдения ξ_2. Оценкой ковариационной матрицы вектора $\{\check{\alpha}_1, \check{\alpha}_2\}$ (6.75) будет

$$\check{\mathbf{B}} = \frac{s_2^2(1-\check{r}^2)}{n s_1^2} \begin{pmatrix} 1 & -\bar{x} \\ -\bar{x} & s_1^2 + \bar{x}^2 \end{pmatrix}. \tag{6.75′}$$

В частности, дисперсия оценки $\check{\alpha}_1$ коэффициента при ξ_1 оценивается в виде $\frac{s_2^2(1-\check{r}^2)}{n s_1^2}$, а доверительный интервал (6.76′) для α_1

$$\sim \check{\alpha}_1 - u_{\frac{1+q}{2}} \frac{s_2}{s_1} \sqrt{\frac{1-\check{r}^2}{n}},\ \sim \check{\alpha}_1 + u_{\frac{1+q}{2}} \frac{s_2}{s_1} \sqrt{\frac{1-\check{r}^2}{n}}, \tag{6.77}$$

где q — доверительная вероятность. Из (6.77) следует, что оценка $\check{\alpha}_1$ будет тем точнее, чем больше вариация параметра ξ_1 (т. е. разброс значений по оси Ox) и чем больше количество наблюдений n. Используя (6.77), получим такой приближенный критерий для проверки

гипотезы об отсутствии влияния ξ_1 на ξ_2 (т. е. о равенстве α_1 нулю): если

$$|\check{\alpha}_1| \geqslant u_{\frac{1+q}{2}} \frac{s_2}{s_1} \sqrt{\frac{1}{n}}, \tag{6.78}$$

то эта гипотеза отвергается с вероятностью ошибки $1 - q$.

Отмеченная связь между корреляционным анализом и методом наименьших квадратов дает естественные предпосылки к использованию последнего при таких зависимостях между показателями, которые не поддаются корреляционному анализу.

Пусть общий вид зависимости ξ_{k+1} от $\xi_1, \xi_2, \ldots, \xi_k$ (последние могут быть как случайными, так и не случайными величинами)

$$\xi_{k+1} = f(\xi_1, \xi_2, \ldots, \xi_k; \alpha_1, \alpha_2, \ldots, \alpha_m) + \Delta; \tag{6.79}$$

$\alpha_1, \alpha_2, \ldots, \alpha_m$ — параметры, Δ — случайная величина с нулевым средним; функция $f(x_1, x_2, \ldots, x_k; \alpha_1, \alpha_2, \ldots, \alpha_m)$ имеет непрерывные частные производные по $\alpha_1, \alpha_2, \ldots, \alpha_m$. Имеем:

$$x_{ik+1} = f(x_{i1}, x_{i2}, \ldots, x_{ik}; \alpha_1, \alpha_2, \ldots, \alpha_m) + \Delta_i, \tag{6.79'}$$

причем дисперсия Δ_i, $D_i = \frac{D}{p_i}$ $(i = \overline{1, n})$.

По методу наименьших квадратов оценка функции регрессии определится оценками параметров $\check{\alpha}_1, \check{\alpha}_2, \ldots, \check{\alpha}_m$, приводящими к минимуму

$$S = \sum_{i=1}^{n} [x_{ik+1} - f(x_{i1}, x_{i2}, \ldots, x_{ik}; \alpha_1, \alpha_2, \ldots, \alpha_m)]^2 p_i,$$

которые определяются из системы уравнений

$$\frac{\partial S}{\partial \alpha_j} = 0, \; j = \overline{1, m}. \tag{6.80}$$

В качестве оценки D можно взять, по аналогии с (4.61),

$$\check{D} = \frac{1}{n - m} \sum_{i=1}^{n} [x_{ik+1} - f(x_{i1}, x_{i2}, \ldots, x_{ik}; \check{\alpha}_1, \check{\alpha}_2, \ldots, \check{\alpha}_m)]^2.$$

Косвенные измерения ξ_{k+1} могут производиться с помощью функции

$$\tilde{\xi}_{k+1} = f(x_1, x_2, \ldots, x_k; \check{\alpha}_1, \check{\alpha}_2, \ldots, \check{\alpha}_m), \tag{6.81}$$

где $x_1, x_2, \ldots, x_k$ — значения $\xi_1, \xi_2, \ldots, \xi_k$. Точность такого измерения оценится в виде

$$\tilde{\xi}_{k+1} - u_{\frac{1+q}{2}} \sqrt{\frac{\check{D}}{p}}, \; \tilde{\xi}_{k+1} + u_{\frac{1+q}{2}} \sqrt{\frac{\check{D}}{p}}, \tag{6.82}$$

где $\frac{D}{p} = \mathbf{D}\Delta_{k+1}$ при $\xi_j = x_j$, $j = \overline{1, k}$.

Если анализируемая зависимость не сводится с помощью преобразования или замен переменных к линейному виду, то система (6.80)

оказывается, как правило, сложной. В этом случае значения α_1, α_2, ..., α_m, приводящие S к минимуму, вычисляют итерационными методами на ЭВМ — методом наискорейшего спуска (или методом градиента), методом Ньютона и др. Если же функция регрессии представляется в виде

$$f(x_1, x_2, \ldots, x_k, \alpha_1, \alpha_2, \ldots, \alpha_m) = \varphi\left(\sum_{j=1}^{m} \alpha_j f_j(x_1, x_2, \ldots, x_m)\right) \quad (6.83)$$

(где $\varphi(x)$ — монотонная функция), то задача сводится к линейному случаю, подобно (6.67). Обозначим $g(u)$ функцию, обратную $\varphi(x)$. Считая, что $g(u)$ определена при $u = x_{ik+1}$ $(i = \overline{1, n})$, обозначим $y_{ik+1} = g(x_{ik+1})$ и перепишем (6.79′) в виде

$$y_{ik+1} = g(f(x_{i1}, x_{i2}, \ldots, x_{ik}; \alpha_1, \alpha_2, \ldots, \alpha_m) + \Delta_i) = g(f(\mathbf{x}_i, \boldsymbol{\alpha}) + \Delta_i).$$

Если связь (6.79′) достаточно тесная ($|\Delta_i|$ малы), то, полагая функцию $g(x)$ дифференцируемой, из предыдущего равенства получим

$$y_{ik+1} = g(f(\mathbf{x}_i, \alpha)) + \frac{dg}{df}\Delta_i + o(\Delta_i) \approx \sum_{j=1}^{m} \alpha_j y_{ij} + \delta_i, \quad (6.84)$$

где $y_{ij} = f_j(x_{i1}, x_{i2}, \ldots, x_{ik})$ $(j = \overline{1, m})$, $\delta_i = \Delta_i \frac{dg(f(\mathbf{x}_i, \alpha))}{df} = \Delta_i g'_{fi}$.

Так как дисперсия δ_i при условии $\xi_l = x_{il}$, $l = \overline{1, k}$ —

$$\mathbf{D}\delta_i \approx (g'_{fi})^2 Dp_i^{-1}, \quad (6.85)$$

оценки α_1, α_2, ..., α_m определяются условием минимума

$$S = \sum_{i=1}^{n} [g(x_{ik+1}) - \sum_{j=1}^{m} \alpha_j f_j(x_{i1}, x_{i2}, \ldots, x_{ik})]^2 p_i (g'_{fi})^{-2}. \quad (6.86)$$

Приняв в первом приближении $p_i (g'_{fi})^{-2} = \text{const}$, получим линейную относительно α_1, α_2, ..., α_m систему уравнений (6.80). Решив ее, получим первые приближения для $\check{\alpha}_1$, $\check{\alpha}_2$, ..., $\check{\alpha}_m$, которыми определится второе приближение для $p_i (g'_{fi})^{-2}$ и т. д.

Примеры применения метода наименьших квадратов. В ореолах рассеяния рудопроявлений зависимость средних концентраций C химических элементов от координат (x, y, z) точек пространства описывается функцией, вид которой задается из тех или иных теоретических представлений:

$$\mathbf{M}C = f(x, y, z; \alpha_1, \alpha_2, \ldots, \alpha_m) = f(x, y, z, \boldsymbol{\alpha}). \quad (6.87)$$

Влияние неоднородности вмещающей среды, ошибок измерений и иных факторов приводит к тому, что результаты измерения концентраций по данным опробования ореола отклоняются от теоретической зависимости. Представляя результаты измерений C_i в точках с координатами (x_i, y_i, z_i) $(i = \overline{1, n})$ в виде

$$C_i = f(x_i, y_i, z_i; \alpha_1, \alpha_2, \ldots, \alpha_m) + \Delta_i = f(x_i, y_i, z_i, \boldsymbol{\alpha}) + \Delta_i, \quad (6.87')$$

где Δ_i — отклонение результата измерения от значения, предписывае-

мого теоретической зависимостью, мы можем решить задачу оценки функции $f(x, y, z, \alpha)$ методом наименьших квадратов.

Отыскание оценок параметров $\check{\alpha}_1, \check{\alpha}_2, \ldots, \check{\alpha}_m$ сведением к минимуму суммы

$$S = \sum_{i=1}^{n} [C_i - f(x_i, y_i, z_i; \alpha_1, \alpha_2, \ldots, \alpha_m)]^2 \qquad (6.89)$$

целесообразно в тех случаях, когда дисперсия отклонений Δ_i в (6.87′) постоянна: $D_i = D, p_i = 1, i = \overline{1, n}$. Известно, однако, что колебания концентраций элементов, как правило, тем больше, чем выше средний уровень самих концентраций, так что тенденцию к постоянству испытывают скорее относительные колебания.

Если средние квадратические отклонения величин Δ пропорциональны значениям функции регрессии, $\sqrt{D_i} = af(x_i, y_i, z_i, \alpha) = af_i$, т. е. средняя величина относительных отклонений результатов наблюдений от функции регрессии постоянна, то целесообразно применить логарифмическое преобразование к результатам измерений и функции регрессии. Оценки $\alpha_1, \alpha_2, \ldots, \alpha_m$ будут определяться условием минимума суммы

$$S_1 = \sum_{i=1}^{n} (\ln C_i - \ln f_i)^2. \qquad (6.90)$$

Действительно, по (6.84) $\ln C_i \approx \ln f(x_i, y_i, z_i, \alpha) + \delta_i = \ln f_i + \delta_i$, причем из (6.85) $\mathbf{D}\delta_i = D_i \left(\frac{d}{df_i} \ln f_i\right)^2 = a^2 f_i^2 \frac{1}{f_i^2} = \text{const}$ и, в соответствии с (6.86), оценки $\alpha_1, \alpha_2, \ldots, \alpha_m$ должны определяться из условия минимума суммы (6.90).

В первичном ореоле рассеяния зависимость содержания элемента от расстояния x до границы рудного тела приближенно описывается регрессией вида

$$f(x) = Ae^{-\lambda x} + C_\Phi, \qquad (6.91)$$

где C_Φ — уровень фона; $\frac{1}{\lambda}$ — показатель миграционной способности элемента. По данным опробования ореола, представляющим собой содержания C_i на различных расстояниях x_i до рудного тела, можно определить оценки параметров A и λ минимизацией S_1 (при условии $\sqrt{D_i}/f_i \approx \text{const}$). Система уравнений для нахождения оценок $\check{A}$ и $\check{\lambda}$ нелинейна, так что последние приходится вычислять итерационными методами на ЭВМ из условия минимума S_1. Первое приближение $\check{A}_0$, $\check{\lambda}_0$ для оценок $\check{A}$ и $\check{\lambda}$ можно найти, используя то обстоятельство, что в области $C_i \gg C_\Phi$, $\ln f(x) \approx \ln A - \lambda x$ — функция, линейная относительно x. Взяв наибольший интервал значений x_i, на котором сохраняется линейность зависимости $\ln C_i$ от x_i, определим $\check{A}_0$, $\check{\lambda}_0$ в виде: $\check{A}_0 = e^{\check{\beta}}$, $\check{\lambda}_0 = -\check{\alpha}$, где $\check{\alpha}$ и $\check{\beta}$ вычисляются по (6.53), (6.49), (6.50), с подстановкой $\ln C_i$ вместо x_{1i} и x_i вместо x_{1i}. Линейность зависимости проверяется так же, как было показано ранее при анализе связи двух показателей.

Распределение содержаний химических элементов во вторичном элювиальном ореоле рассеяния для маломощных рудных тел жильной формы, согласно формуле А. П. Соловова, описывается регрессией

$$f(x) = Ae^{-\beta(x-m)^2} + C_\Phi, \tag{6.92}$$

где x — абсцисса на оси, направленной вкрест простирания рудного тела (рис. 29). Как и в предыдущем примере, при $\sqrt{D_i}/f_i \approx \text{const}$ ($i = \overline{1, n}$), оценки $\check{A}$, $\check{m}$, $\check{\beta}$ можно определить из условия минимума (6.90), применив итерационные методы. Первое приближение для них можно получить по наибольшему участку опробования, где $\ln C_i$ ($i = \overline{n_0, n_1}$) следуют параболической зависимости от x. Это приближение определяют из условия минимума суммы

$$S_1 = \sum_{i=n_0}^{n_1} (\ln C_i - \ln A + \beta m^2 + \beta x_i^2 - 2\beta m x_i)^2, \tag{6.93}$$

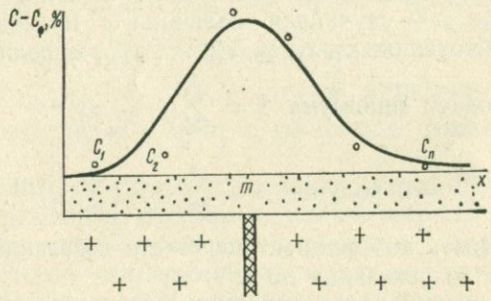


Рис. 29.

которое приводит к линейной относительно ($\ln A - \beta m^2$), $-\beta$, $2\beta m$ системе уравнений. Величина $4\check{\sigma}_0 = 4\sqrt{\frac{1}{2\check{\beta}}}$ будет характеризовать ширину ореола (точнее интервала, в котором $f(x) - C_\Phi \geqslant 0{,}135\,(C_{\max} - C_\Phi) = 0{,}135\,(f(0) - C_\Phi)$); $\check{m}$ оценивает абсциссу точки с максимальной средней концентрацией, $\check{A} + C_\Phi$ — величину этой концентрации.

Аналогично решается задача с использованием несимметричных относительно m функций $f(x)$. При изучении площадных вторичных ореолов от точечных линзообразных источников используется регрессия распределения по площади вида

$$C(x, y) = A \exp[-a_{11}(x - m_1)^2 - a_{12}(x - m_1)(y - m_2) - a_{22}(x - m_2)^2] + C_\Phi. \tag{6.94}$$

Метод наименьших квадратов иногда полезен и при решении задачи об оценке расстояния до рудного тела по комплексу показателей, получаемых в результате опробования ореола рассеяния химических элементов. Построив по данным опробования участков со сходным геологическим строением (однотипных по характеру залегания рудных

тел, вмещающим породам и т. д.) функцию регрессии, описывающую зависимость между расстояниями и этими показателями, можно в дальнейшем ее использовать для такого прогноза по принципу косвенных измерений. Среднее квадратическое отклонение ошибок этих измерений будет характеризовать точность прогноза.

При решении обратных задач геофизики метод наименьших квадратов дает одну из возможностей реализации метода моделирования на ЭВМ. Суммы квадратов отклонений, подобные (6.89), или их интегральные аналоги характеризуют степень соответствия теоретического поля $f(\mathbf{x}, \boldsymbol{\alpha})$, определяемого заданной моделью геологического тела и окружающей его среды, значениям наблюденного поля. Варьируя параметры так, чтобы S свести к минимуму, методом последовательных приближений определяют оптимальную модель или класс эквивалентных ей моделей.

Рассмотрим другие примеры применения метода наименьших квадратов.

Если предполагаемая зависимость двух величин η и ξ имеет вид $\eta = \alpha\xi + \Delta$, где Δ — случайная величина с нулевым средним, то оценку α по наблюдениям $y_1, x_1; y_2, x_2; \ldots, y_n, x_n$ величин η и ξ можно определить условием минимума $S = \sum_{i=1}^{n} (y_i - \alpha x_i)^2$.

Из условия $\frac{\partial S}{\partial \alpha} = 0$ получим $\alpha = \sum_{i=1}^{n} y_i x_i \left(\sum_{i=1}^{n} x_i^2\right)^{-1}$. Например, так можно рассчитывать коэффициент пересчета показаний по одному методу измерений на показания по другому.

Метод наименьших квадратов можно использовать и при анализе своеобразных зависимостей, в которых отдельные коэффициенты регрессии связаны функциональными соотношениями. Например, если предполагаемая зависимость

$$\eta = f_0(\alpha) + f_1(\alpha)\, g(\xi) + \Delta, \tag{6.95}$$

то оценкой α будет такое значение, при котором обращается в минимум

$$S = \sum_{i=1}^{n} [y_i - f_0(\alpha) - f_1(\alpha)\, g(x_i)]^2.$$

Необходимое условие минимума $-\frac{dS}{d\alpha} = 0$, следовательно, искомая оценка является одним из решений уравнения

$$\sum_{i=1}^{n} [y_i - f_0(\alpha) - f_1(\alpha)\, g(x_i)] \left[\frac{df_0(\alpha)}{d\alpha} + \frac{df_1(\alpha)}{d\alpha} g(x_i)\right] = 0. \tag{6.96}$$

В практике обработки данных специальных геолого-геофизических исследований иногда возникает задача анализа зависимости некоторого показателя ξ от одного или группы неслучайных факторов X_1,

$X_2, \ldots, X_k$, определяющих те или иные условия опыта. Примером такой задачи может служить зависимость отдельных физических свойств образцов от давления либо от давления и температуры. В широких диапазонах изменения такие зависимости, как правило, бывают криволинейными. Для получения более точного результата нередко делают по нескольку измерений при каждой комбинации значений величин $X_1, X_2, \ldots, X_m$. В этом случае также целесообразно применить метод наименьших квадратов.

Действительно, пусть при каждом наборе значений $x_{i1}, x_{i2}, \ldots, x_{im}$ величин $X_1, X_2, \ldots, X_m$ получены n_i результатов измерений ξ, по которым вычислено среднее $\bar{\xi}_i$ $(i = \overline{1, N})$. Вид зависимости: $\bar{\xi} = \sum_{j=1}^{k} \alpha_j Y_j + \Delta$, где $Y_j = f_j(X_1, X_2, \ldots, X_m)$ — некоторые функции (в случае линейной зависимости $Y_j = X_j$, $j \leqslant k - 1 = m$; $Y_k = 1$), Δ — случайная величина с нулевым средним и неизвестной дисперсией D. Очевидно, $\bar{\xi}_i = \sum_{j=1}^{k} \alpha_j y_{ij} + \bar{\Delta}_i$, где $\bar{\Delta}_i$ — среднее арифметическое, составленное из n_i наблюдений величины Δ, имеющее дисперсию $\frac{D}{n_i}$; $y_{ij} = f_j(x_{i1}, x_{i2}, \ldots, x_{im})$, $j = \overline{1, k}$. Положив $p_i = n_i$, приходим к задаче вычисления оценок α_j по методу наименьших квадратов.

В практике геолого-геофизических исследований нередко возникает необходимость сравнивать две или несколько регрессий, используемых для аппроксимации одной и той же зависимости, с целью определения той из них, которая точнее соответствует эмпирическим данным. Естественной мерой для такого сравнения служит дисперсия $\mathbf{D}\Delta$ отклонения Δ показателя от регрессии, оцениваемая по одной из приводившихся выше формул. Минимальной величине $\mathbf{D}\Delta$, очевидно, будет соответствовать наилучшая аппроксимация.

Если параметрический вид регрессии неизвестен, обычно используют аппроксимирующие функции — большей частью, полиномы различных порядков. Мерой качества такой аппроксимации также служит $\mathbf{D}\Delta$. Этот способ используют для оценки регионального фона, а также трендов физических и геохимических полей, для описания поверхностей отдельных складчатых структур, зависимостей глубин залегания фундамента от координат на плоскости и т. п., обычно с целью автоматизации последующей обработки данных (изображение полей в изолиниях, трансформация полей с учетом фона, оценка характерных точек — максимумов, минимумов и т. д.). Порядок полинома l определяют сравнением при различных l величин $\mathbf{D}\Delta$ между собой, а также с минимальным значением $\mathbf{D}_0\Delta$. При обработке наблюдений полей, в зависимости от характера задачи, масштаба работ, $\mathbf{D}_0\Delta$ может изменяться от минимального значения — дисперсии ошибок измерений, до дисперсии колебаний поля относительно регионального фона. Значение последней оценивают по известному разрезу либо по участкам с заведомо пренебрежимым изменением регионального фона в их пределах. Предварительная оценка описанными в этой главе

методами статистической однородности наблюдений, линейно упорядоченных по их координатам, позволяет заранее выяснить, насколько выражен тренд изучаемого поля. Величина g^2 в (6.1) дает ориентировочную оценку $\mathbf{D}_0\Delta$ при построении тренда, а отношение g^2 к D (6.1) показывает, в какой степени проявлен тренд на фоне случайных флюктуаций.

Следует подчеркнуть, что в подобных задачах особое значение имеют такие семейства трендов, которые учитывают природу изучаемых объектов. Если поле от отдельного «источника» (отдельная аномалия) описывается функцией $z = f(x, y, \boldsymbol{\alpha})$, которая зависит от параметров $\{\alpha_1, \alpha_2, \ldots, \alpha_m\}' = \boldsymbol{\alpha}$, определяющих положение, интенсивность источника, характер окружающей среды, то вид тренда можно задать линейной комбинацией $z = \sum_{i=1}^{k} f(x, y, \boldsymbol{\alpha}_i) + a_\Phi$ и свести задачу к вычислению оценок $\boldsymbol{\alpha}_i$, a_Φ. Такой подход допускает простую интерпретацию получаемых результатов, хотя его практическая реализация часто бывает далеко не простой задачей.

Метод скользящего среднего. При анализе тренда часто полезен простейший метод — сглаживание с помощью скользящего среднего. Этот способ легко реализуется в одномерном варианте, когда наблюдения ξ_i $(i = \overline{1, n})$ упорядочены в ряд по какому-либо признаку, например по профилю, по глубине скважины, либо по величине какого-либо показателя пород. Сглаживание производится следующим образом: по первым k точкам ряда вычисляют два средних: по значениям показателя $\bar{\xi}_1 = \frac{1}{k}\sum_{i=1}^{k} \xi_i$ и по значениям признака упорядочения $\bar{x}_1 = \frac{1}{k}\sum_{i=1}^{k} x_i$. Точку с абсциссой $\bar{x}_1$ и ординатой $\bar{\xi}_1$ наносят на график (рис. 30). Далее, сдвигают «окно» на одну точку и повторяют процедуру: вычисляют средние $\bar{\xi}_2 = \frac{1}{k}\sum_{i=2}^{k+1} \xi_i$, $\bar{x}_2 = \frac{1}{k}\sum_{i=2}^{k+1} x_i$ по k точкам, начиная со второй; получившуюся точку $(\bar{x}_2, \bar{\xi}_2)$ снова наносят на график и так до конца ряда. В результате получим $n - k + 1$ точек сглаженного ряда. Соединив их, построим сглаживающую кривую. При недостаточной степени сглаживания можно повторить процедуру по уже полученным осредненным точкам либо по исходным с бо́льшим k. Интервал первичного сглаживания обычно колеблется от 4 до 20 точек в зависимости от имеющегося материала. Колебания значений $\bar{\xi}_i$ в интервале осреднения, при условии пренебрежимой вариации тренда в этом интервале и независимости отклонений ξ_i от него будут уменьшены по сравнению с исходными данными в 2—4,5 раза (дисперсия средних уменьшается в k раз при уровне осреднения k, а среднее квадратическое отклонение — в $\sqrt{k}$ раз). Метод требует определенной осмотрительности: при больших уровнях осреднения могут

искажаться особенности тренда — функциональной составляющей ряда. Так, если эта составляющая имеет разрыв («уступ»), например при переходе через контакт пород, то при сглаживании случайных колебаний произойдет и некоторое сглаживание уступа, тем большее, чем больше интервал осреднения (рис. 31, *а*). В областях резких локальных повышений или понижений тренда, не подтвержденных окружающими точками или подтвержденных лишь их малым числом, также может произойти искажение, выражающееся в понижении амплитуды соответствующих «всплесков» (рис. 31, *б*).

Способ скользящего среднего используется для оценки тренда и в двумерном варианте, особенно при равномерном распределении пунктов наблюдений, когда размер перемещаемого при сглаживании «окна» и количество точек в нем постоянны. В этом случае «окно» имеет вид прямоугольника, сдвигаемого после каждого расчета, производимого

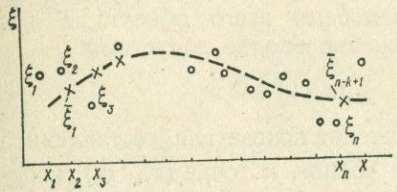


Рис. 30

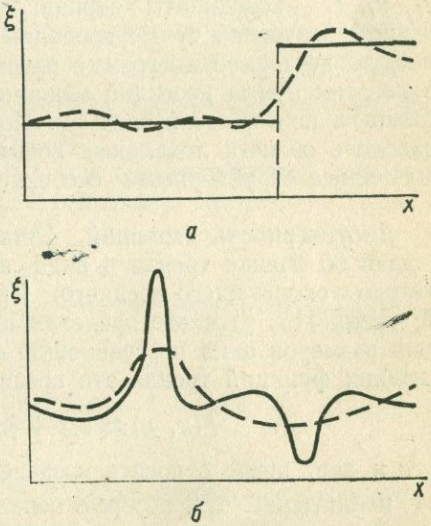


Рис. 31.

по точкам, попадающим в «окно». При начальном его положении вычисляют среднее $\bar{\xi}_1$ из значений показателя в точках, попавших в окно, средние из абсцисс и ординат пунктов наблюдений (опробования) $\bar{x}_1$, $\bar{y}_1$. Величину $\bar{\xi}_1$ относят к точке с координатами $\bar{x}_1$, $\bar{y}_1$, сдвигают окно на один шаг вдоль профилей и повторяют описанное построение. Результаты осреднения, как правило, более четко отражают функциональную составляющую поля, чем облегчают построение его карты в изолиниях.

Эту схему можно использовать и при автоматизации графического представления геофизической информации (в виде карт полей в изолиниях, векторных полей) с применением ЭВМ. Пример такого применения метода — построение поля усреднённых векторов проекций нормали тренда на горизонтальную плоскость. Измеренная функция $z = f(x, y)$ аппроксимируется плоскостями треугольников с вершинами в трёх точках, проекции которых попадают в «окно» и наименее удалены друг от друга. Проекция единичной нормали к плоскости треугольника определяется её величиной и азимутальным углом: $n_p =$

$= \sqrt{a^2 + b^2}$, $\varphi = -\frac{b}{|b|} \arccos \frac{-a}{\sqrt{a^2 + b^2}} + \pi$, где $a = A/C$, $b = B/C$, $C = \sqrt{A^2 + B^2 + 1}$ при $b \neq 0$; если $b = 0$, $a \neq 0$, то $\varphi = \arccos a \, |a|^{-1}$;

$$A = \frac{(y_j - y_i)(z_k - z_i) - (y_k - y_i)(z_j - z_i)}{(x_j - x_i)(y_k - y_i) - (x_k - x_i)(y_j - y_i)},$$

$$B = \frac{(x_k - x_i)(z_j - z_i) - (x_j - x_i)(z_k - z_i)}{(x_j - x_i)(y_k - y_i) - (x_k - x_i)(y_j - y_i)},$$

x_l, y_l, z_l — координаты вершин треугольника ($l = i, j, k$). Вектор, равный векторной сумме проекций в пределах «окна» либо вычисленный по векторной диаграмме проекций в интервале углов диаграммы, в котором сумма проекций максимальна, наносят на карту, сдвигают «окно» и повторяют процедуру. Полученная схема дает возможность находить области локальных повышений или понижений поля, обусловленные структурными особенностями, изменением состава пород и т. п.

Достоверность изолиний. Остановимся на вероятностной стороне задачи об оценке тренда в виде изолиний по данным, осредненным методом скользящего среднего. При произвольном положении окна Ω_r тренд $f(x, y)$ в его пределах предполагается линейным. Ввиду малых размеров окна по сравнению с размерами всего объекта T для гладких функций тренда это предположение вполне допустимо:

$$f(x, y) \approx \alpha_r x + \beta_r y + \gamma_r, \quad (x, y) \in \Omega_r.$$

Это и дает право относить среднее значение показателя, составленное из наблюдений $\xi_i^{(r)}$, которые попали в «окно», к точке $(\bar{x}_r, \bar{y}_r)$, координаты которой получают осреднением соответствующих координат $x_i^{(r)}$, $y_i^{(r)}$ пунктов наблюдений в «окне»:

$$\bar{\xi}_r = \frac{1}{k} \sum_{i=1}^{k} \xi_i^{(r)} = \frac{1}{k} \sum_{i=1}^{k} [f(x_i^{(r)}, y_i^{(r)}) + \Delta_i^{(r)}] \approx \frac{1}{k} \sum_{i=1}^{k} (\alpha_r x_i^{(r)} +$$

$$+ \beta_r y_i^{(r)} + \gamma_r + \Delta_i^{(r)}) = \alpha_r \bar{x}_r + \beta_r \bar{y}_r + \gamma_r + \bar{\Delta}_r,$$

где $\Delta_i^{(r)}$ — отклонение результата $\xi_i^{(r)}$ измерения от значения тренда $f(x_i^{(r)}, y_i^{(r)})$ в пункте наблюдения $(x_i^{(r)}, y_i^{(r)})$, $\bar{\Delta}_r = \sum_{i=1}^{k} \frac{\Delta_i^{(r)}}{k}$. Дисперсия усредненного отклонения $\bar{\Delta}_r$ от тренда уменьшается в k раз по сравнению с дисперсией $D = \mathbf{D}\,\Delta_i^{(r)} : \mathbf{D}\bar{\Delta}_r = \frac{D}{k}$. Обозначим u_1, u_2 $(u_1 < u_2)$ уровни двух соседних изолиний и $\xi_r^{(T)}$ — случайную величину, представляющую собой значение $\bar{\xi}_r$ в произвольной точке объекта T; $\mathbf{M}(\xi_r^{(T)}/x, y) = f(x, y)$ — условное математическое ожидание $\xi_r^{(T)}$ в точке (x, y). Будем исходить из того, что распределение значений функции

$f(x, y)$ в интервале (u_1, u_2) равномерно (длина $d = u_2 - u_1$ интервала (u_1, u_2) мала по сравнению с общей вариацией функции $f(x, y)$). Разбив отрезок $[u_1, u_2]$ точками $m_0 = u_1$, $m_1 = u_1 + \delta$, $m_2 = u_1 + 2\delta, \ldots,$ $m_N = u_2$ на малые интервалы длиной δ, по формуле полной вероятности имеем:

$$P = \mathbf{P}\{u_1 \leqslant \xi_r^{(T)} < u_2 / u_1 \leqslant f(x, y) < u_2\} = \sum_{i=0}^{N-1} \mathbf{P}\{u_1 \leqslant \xi_r^{(T)} < u_2 / m_i \leqslant f(x, y) < m_{i+1}\} \frac{\delta}{u_2 - u_1}, \tag{6.97}$$

где P — условная вероятность попадания в интервал $[u_1, u_2)$ значения $\xi_r^{(T)}$ при условии, что $f(x, y)$ находится в интервале $[u_1, u_2)$; $\mathbf{P}\{u_1 \leqslant \xi_r^{(T)} < u_r / m_i \leqslant f(x, y) < m_{i+1}\}$ — условная вероятность того же события при условии $\{m_i \leqslant f(x, y) < m_{i+1}\}$.

Считая условное распределение $\xi_r^{(T)}$ при фиксированном значении $f(x, y)$ нормальным и переходя в (6.97) к пределу при $N \to \infty$, $\delta \to 0$, получим

$$P = \lim_{N \to \infty} \sum_{i=0}^{N-1} \left[\Phi\left(\frac{u_2 - m_i}{\sqrt{Dk^{-1}}}\right) - \Phi\left(\frac{u_1 - m_i}{\sqrt{Dk^{-1}}}\right)\right] \frac{\delta}{u_2 - u_1} = \frac{1}{u_2 - u_1} \int_{u_1}^{u_2} \left[\Phi\left(\frac{u_2 - m}{\sqrt{Dk^{-1}}}\right) - \Phi\left(\frac{u_1 - m}{\sqrt{Dk^{-1}}}\right)\right] dm = \frac{1}{t}\left[\int_0^t \Phi(z)\, dz - \int_{-t}^0 \Phi(z)\, dz\right], \tag{6.97'}$$

где $t = \frac{u_2 - u_1}{\sqrt{Dk^{-1}}}$, $\Phi(z)$ — функция (0; 1)-нормального распределения. Очевидно, чем ближе величина P к единице, тем более надежными следует считать изолинии. P является монотонно возрастающей функцией t, поэтому задаваясь надежностью P, можно определить соотношение между шагом изолиний $u_2 - u_1$, уровнем осреднения и дисперсией D. Поскольку увеличение шага сказывается на детальности карты (уменьшение числа изолиний), следует останавливаться на минимально возможных значениях P. Как показывает непосредственный расчет по (6.97′), надежность P для $t = 10$ составит примерно 0,9. Иными словами, при уровне осреднения $k = 10$ для проведения изолиний с надежностью $P = 0{,}9$ их шаг следует брать примерно равным утроенному локальному среднему квадратическому отклонению: $u_2 - u_1 = 3\sqrt{D}$. Приближенную оценку величины D можно получить в процессе обработки данных способом скользящего среднего. Для этого нужно вычислить выборочные дисперсии внутри отдельных «окон» и найти среднее из полученных оценок.

Ранговая корреляция. Рассмотрим приемы изучения связей таких признаков, которые не поддаются точному измерению, но по каждому из них исследуемые объекты можно сравнивать друг с другом. Это дает возможность «ранжировать» группу из n объектов по таким признакам, приписав им номера от 1 до n в соответствии с порядком

следования объектов, упорядоченных по каждому признаку. Запишем результаты ранжирования n объектов по двум признакам ξ_1, ξ_2 в виде

$$\begin{pmatrix} 1 & 2 & \ldots & n \\ r_1 & r_2 & \ldots & r_n \end{pmatrix}, \tag{6.98}$$

упорядочив их по признаку ξ_1. В (6.98) r_i — порядковый номер по признаку ξ_2 того объекта, который по признаку ξ_1 имеет номер i. Эти данные несут информацию о связи признаков ξ_1 и ξ_2: при прямой зависимости значения $r_1, r_2, \ldots, r_k$ будут более или менее выраженно следовать нумерации по признаку ξ_1 $(1, 2, \ldots, n)$, при обратной связи будет отмечаться тенденция к обратному порядку $(n, n-1, \ldots, 2, 1)$. Отсутствие связи повлечет за собой отсутствие закономерности в ряду $r_1, r_2, \ldots, r_n$. Мерой подобных связей служит *ранговый коэффициент корреляции* ρ *Спирмена* [2]:

$$\rho = 1 - \frac{6s_\rho}{n^3 - n} \quad \left(s_\rho = \sum_{i=1}^{n} (r_i - i)^2\right). \tag{6.99}$$

Другой коэффициент ранговой корреляции — коэффициент *Кендалла* [2] имеет вид:

$$\tau = \frac{4N}{n(n-1)} - 1 = \frac{2s_\tau}{n(n-1)} \quad \left(N = N_1 + N_2 + \right.$$
$$\left. + \cdots + N_{n-1},\ s_\tau = 2N - \frac{n(n-1)}{2}\right); \tag{6.100}$$

величина N_i равна количеству тех r_j в (6.98), для которых $j > i$ и одновременно $r_j > r_i$ $(i = \overline{1, n-1})$.

Ранговые коэффициенты корреляции (6.99), (6.100), как и обычные коэффициенты парной корреляции, заключены в пределах от -1 до 1. Близость их к нулю свидетельствует о слабой связи или вообще об отсутствии ее; знаки ρ и τ указывают на характер связи (прямая или обратная). Абсолютная величина служит характеристикой силы связи: чем ближе $|\rho|$ и $|\tau|$ к единице, тем теснее зависимость между исследуемыми признаками.

Описанные коэффициенты являются выборочными характеристиками, т. е. они лишь оценивают те величины, которые, например, могли бы быть получены по очень большому числу наблюдений. Для проверки гипотезы об отсутствии связи необходимо сравнивать коэффициенты ранговой корреляции с пределами, вычисленными для них в предположении отсутствия связи. Если оценки окажутся вне этих пределов, следует принимать гипотезу о наличии связи.

Используя то свойство, что при отсутствии связи ρ и τ распределены приближенно нормально [2] с математическими ожиданиями и дисперсиями, соответственно

$$\mathbf{M}\rho = 0,\ \mathbf{D}\rho = \frac{1}{n-1},\ \mathbf{M}\tau = 0,\ \mathbf{D}\tau = \frac{2(2n+5)}{9n(n-1)}, \tag{6.101}$$

получим упомянутые пределы:

для ρ

$$\rho_q^- = -u_{\frac{1+q}{2}}\sqrt{\frac{1}{n-1}},\quad \rho_q^+ = u_{\frac{1+q}{2}}\sqrt{\frac{1}{n-1}}; \tag{6.102}$$

для τ

$$\tau_q^- = -u_{\frac{1+q}{2}}\sqrt{\frac{2n(2n+5)}{9n(n-1)}},\quad \tau_q^+ = u_{\frac{1+q}{2}}\sqrt{\frac{2n(2n+5)}{9n(n-1)}} \tag{6.103}$$

(q — вероятность, с которой эти пределы заключают ρ и τ при отсутствии связи). Вероятность ошибки I рода — при принятии гипотезы об отсутствии связи — будет близка к $1-q$.

Как указывается в [2], эти пределы можно использовать уже при $n > 10$. При $n < 10$ следует обратиться к соответствующим таблицам, содержащим величины критических значений s_ρ и s_τ, рассчитанные на основе их точного распределения [2].

В качестве меры связи нескольких (m) признаков, допускающих ранжирование, можно использовать *коэффициент согласованности W Кендалла и Смита*:

$$W = \frac{12 s_W}{m^2(n^3-n)}\quad \left(s_W = \sum_{i=1}^{n}\left[\sum_{i=1}^{m} r_{ij} - \frac{m(n+1)}{2}\right]^2\right), \tag{6.104}$$

где r_{ij} — элементы матрицы результатов ранжирования $\mathbf{R} = \{r_{ij}\}$, $i = \overline{1, m}$, $j = \overline{1, n}$ (i-я строка матрицы содержит результаты ранжирования по ξ_i объектов, следующих в одном и том же порядке). Коэффициент W принимает значения из отрезка [0, 1], причем в случае отсутствия связи $\mathbf{M}W = \frac{1}{m}$. Ввиду узкой специфичности коэффициента согласованности мы не будем здесь более подробно на нем останавливаться. Более подробные сведения о нем приведены в [2, 4].

§ 6. Понятие о факторном анализе. Главные компоненты

Различные геолого-геофизические показатели, характеризующие тот или иной объект, связаны более или менее выраженными статистическими зависимостями. Влияние совокупности всех таких связей нередко оказывается настолько значительным, что отдельные показатели можно вообще не рассматривать, будучи уверенным в небольшой потере информации при их исключении. В самом деле, пусть показатель ξ_m среди рассматриваемых $\{\xi_1, \xi_2, \ldots, \xi_m\}' = \xi$ оказался настолько связанным с остальными $\xi_1, \xi_2, \ldots, \xi_{m-1}$, что дисперсия ошибки его косвенного измерения по величинам $\xi_1, \xi_2, \ldots, \xi_{m-1}$ с помощью функции регрессии близка к дисперсии ошибки прямого измерения. В этом случае значения ξ_m могут быть восстановлены по значениям $\xi_1, \xi_2, \ldots, \xi_{m-1}$ с той же точностью, с которой он измерялся независимо от них. Прямые измерения такого показателя ξ_m не несут добавочной информации по сравнению с той, которую дают измерения величин $\xi_1, \xi_2, \ldots, \xi_{m-1}$.

Это можно объяснить тем, что поведение показателей $\xi_1, \xi_2, \ldots, \xi_m$ определяется некоторой группой независимых факторов, число k которых меньше количества показателей: $k < m$. Такую схему можно представить в виде общей статистической модели [15]:

$$\xi_i = M_i + \sum_{j=1}^{k} a_{ij} f_j + e_i \quad (i = \overline{1, m},\ k \leqslant m), \tag{6.105}$$

где M_i — математические ожидания ξ_i; $e_1, e_2, \ldots, e_m$ — случайные величины с нулевыми математическими ожиданиями, независимые между собой, интерпретирующие ошибки измерений и, возможно, неточность модели. Дисперсии e_i $(i = \overline{1, m})$ принимаются одинаковыми и равными σ^2. Если это не так, $\mathbf{D}e_i = \sigma_i^2$, то простейшим преобразованием $\tilde{\xi}_i = \sigma \dfrac{\xi_i}{\sigma_i}$ задача сводится к случаю $\mathbf{D}e_i = \sigma^2$. f_j — случайные величины с нулевыми математическими ожиданиями и дисперсиями, равными единице, независимые между собой и не зависящие от e_i. Величины f_j именуются факторами.

Оценка количества факторов. В векторно-матричной записи соотношение (6.105) выглядит так:

$$\boldsymbol{\xi} = \mathbf{m} + \mathbf{A}\mathbf{f} + \mathbf{e}, \tag{6.106}$$

где $\boldsymbol{\xi}$, $\mathbf{m}$, $\mathbf{e}$ — векторы-столбцы, составленные из компонент ξ_i, M_i, e_i $(i = \overline{1, m})$ соответственно, $\mathbf{A} = \{a_{ij}\}$ — $(m \times k)$-матрица, составленная из коэффициентов a_{ij}, $\mathbf{f} = \{f_1, f_2, \ldots, f_k\}'$.

Одной из задач факторного анализа является определение ранга k матрицы $\mathbf{A}$, который равен числу независимых факторов распределения $\boldsymbol{\xi}$. Ковариационная матрица $\boldsymbol{\xi}$

$$\mathbf{B} = \mathbf{M}[(\boldsymbol{\xi} - \mathbf{m})(\boldsymbol{\xi} - \mathbf{m})'] = \mathbf{M}[(\mathbf{A}\mathbf{f} + \mathbf{e})(\mathbf{A}\mathbf{f} + \mathbf{e})'] = \mathbf{A}\mathbf{A}' + \sigma^2\mathbf{I} \tag{6.107}$$

($\mathbf{I}$ — единичная матрица).

В предположении малой величины σ^2 для оценки количества независимых факторов можно использовать, как будет показано ниже, характеристические корни (собственные значения) λ_i корреляционной матрицы $\mathbf{R}$ вектора $\boldsymbol{\xi}$. Корни λ_i определяются из уравнения

$$|\mathbf{R} - \lambda\mathbf{I}| = 0 \tag{6.108}$$

($|\mathbf{R} - \lambda\mathbf{I}|$ — определитель матрицы $\mathbf{R} - \lambda\mathbf{I}$). Расположив корни в порядке убывания, отбрасывают те из них, которые незначительно отличаются от нуля (с учетом того, что сумма всех корней равна порядку m корреляционной матрицы). Количество оставшихся корней дает оценку количества независимых факторов.

Метод главных компонент. Оценка матрицы $\mathbf{A}$, участвующей в модели (6.106), на практике оказывается довольно сложной задачей. Кроме того, геологическая интерпретация самих факторов часто затруднительна. Мы рассмотрим задачу отыскания *главных компонент*, т. е. новой системы независимых величин, образующихся линейным

преобразованием исходных. Такая система может служить опорной по отношению к исходной группе показателей, а число компонент системы определяет количество независимых факторов. При этом каждая главная компонента допускает интерпретацию на основе свойств, характеризуемых исходными показателями $\xi_1, \xi_2, \ldots, \xi_m$. Приведенное ниже решение описано Т. Андерсоном [1].

Первая главная компонента определяется как линейная комбинация исходных величин $\xi_1, \xi_2, \ldots, \xi_m$

$$\eta = \sum_{i=1}^{m} \beta_i \xi_i = \beta' \xi \quad (\beta = \{\beta_1, \beta_2, \ldots, \beta_m\}'), \tag{6.109}$$

обладающая максимальной дисперсией среди всех возможных линейных комбинаций вида (6.109) при условии

$$\sum_{i=1}^{m} \beta_i^2 = \beta' \beta = 1. \tag{6.110}$$

Обозначим $\mathbf{B} = \{b_{ij}\}_{i,j=1}^{m}$ ковариационную матрицу ξ, $\mathbf{m} = \mathbf{M}\xi$ — математическое ожидание ξ. Дисперсия η

$$\mathbf{D}\eta = \mathbf{M}(\eta - \mathbf{M}\eta)^2 = \mathbf{M}[(\eta - \mathbf{M}\eta)(\eta - \mathbf{M}\eta)'] = \mathbf{M}[(\beta'\xi - \beta'\mathbf{m})(\beta'\xi -$$

$$- \beta'\mathbf{m})'] = \mathbf{M}[\beta'(\xi - \mathbf{m})(\xi - \mathbf{m})'\beta] = \beta'\mathbf{B}\beta = \sum_{i,\,j=1}^{m} \beta_i b_{ij} \beta_j, \tag{6.111}$$

учитывая, что $\eta' = \eta$ (η — скалярная величина) и $[\beta'(\xi - \mathbf{m})]' = (\xi - \mathbf{m})'\beta$ по правилу транспонирования произведения векторов и матриц. Для определения η необходимо отыскать вектор β, дающий максимум дисперсии (6.111) при условии $\sum_{i=1}^{m} \beta_i^2 = 1$.

По методу неопределенных множителей Лагранжа, необходимым условием максимума (6.111) будет

$$\frac{\partial}{\partial \beta_i}[\beta'\mathbf{B}\beta - \lambda(\beta'\beta - 1)] = 2\sum_{j=1}^{m} \beta_j b_{ij} - 2\lambda\beta_i = 0, \quad i = \overline{1, m},$$

или в векторно-матричной записи

$$\mathbf{B}\beta - \lambda\beta = (\mathbf{B} - \lambda\mathbf{I})\beta = 0. \tag{6.112}$$

Необходимым условием существования решения этой системы уравнений является вырожденность матрицы ее коэффициентов, т. е. равенство нулю определителя матрицы: $|\mathbf{B} - \lambda\mathbf{I}| = 0$. Таким образом, λ является характеристическим корнем ковариационной матрицы. Подставив λ в (6.112), находим β — нормированное по (6.110) решение системы (6.112).

Дисперсия главной компоненты η будет

$$\mathbf{D}\eta = \beta'\mathbf{B}\beta = \beta'\lambda\beta = \lambda\beta'\beta = \lambda \tag{6.113}$$

(из (6.112) $\mathbf{B}\beta = \lambda\beta$).

Таким образом, для получения первой главной компоненты η, имеющей максимальную дисперсию, необходимо взять максимальное собственное значение — наибольший корень λ_1 характеристического уравнения и определить ее в виде $\eta_1 = \beta_1'\xi$, где β_1 — собственный вектор, соответствующий λ_1 и определяемый из (6.112) при $\lambda = \lambda_1$ с учетом (6.110). Первая главная компонента имеет простую геометрическую интерпретацию: в m-мерном пространстве значений ξ ось наибольшей протяженности эллипсоида рассеяния ξ определяется направляющими косинусами, равными компонентам вектора β_1. Проекция вектора ξ на это направление $\beta_1'\xi$ имеет наибольшую дисперсию по сравнению с проекциями на другие направления. Максимальный среди оставшихся корней λ_2 определяет второй собственный вектор β_2, а с ним и вторую главную компоненту $\eta_2 = \beta_2'\xi$. Перебрав все собственные значения, не равные между собой и упорядоченные по убыванию, $\lambda_1 > \lambda_2 > \cdots > \lambda_k$, получим искомую систему главных компонент $\eta_i = \beta_i'\xi$ $(i = \overline{1, k})$. Каждый вектор β_i определяет ось эллипсоида рассеяния ξ, причем векторы β_i будут ортогональны между собой: $\beta_i'\beta_j = 0$, $i \neq j$. В случае кратных характеристических корней собственный вектор определяется неоднозначно: в этом случае следует выбирать группу t взаимно ортогональных решений (6.112) при значении λ, равном кратному корню (t — кратность). Однако в практике такая ситуация нетипична.

Из ортогональности векторов β_i между собой следует некоррелированность главных компонент:

$$\mathbf{M}\,(\eta_i - \mathbf{M}\eta_i)(\eta_j - \mathbf{M}\eta_j)'] = \mathbf{M}\,[\beta_i'\,(\xi - \mathbf{m})(\xi - \mathbf{m})'\,\beta_j] = \\ = \beta_i'\,\mathbf{B}\beta_j = \lambda_j\beta_i'\,\beta_j = 0.$$

Дисперсии этих компонент, $\mathbf{D}\eta_i = \lambda_i$, будут убывать от η_1 к η_k, причем сумма корней λ_i (с учетом кратности каждого) равна сумме дисперсий ξ_i. Как показано в [1], η_i будет иметь максимальную дисперсию среди всех линейных комбинаций вида (6.109), некоррелированных с $\eta_1, \eta_2, \ldots, \eta_{i-1}$. Если ξ_i предварительно нормируются, $\xi_i^0 = \frac{\xi_i - \mathbf{M}\xi_i}{\sqrt{\mathbf{D}\xi_i}}$, то уравнение $|\mathbf{B} - \lambda\mathbf{I}| = 0$ приобретает форму (6.108), и так как $\mathbf{D}\xi_i^0 = 1$, сумма собственных чисел окажется равной m. Если, начиная с некоторого номера r, собственные значения окажутся малыми, соответствующими главными компонентами можно пренебречь (по (6.113) такие компоненты будут иметь незначительное влияние) и оценкой количества действующих факторов будет число оставшихся собственных значений $(r - 1)$.

Вычислительные операции, используемые в описанном методе, — нахождение характеристических корней и матричные процедуры — оказываются весьма трудоемкими уже при небольших объемах данных, так что реализация их вручную обычно невыполнима. Поэтому метод можно рекомендовать лишь при налаженном автоматизированном процессе вычислений на ЭВМ.

Может оказаться полезным и такой способ преобразования исходной системы величин $\xi_1, \xi_2, \ldots, \xi_m$ в базис взаимно независимых величин. Будем считать, что ξ_i нормированы ($\mathbf{D}\xi_i = 1$, $\mathbf{M}\xi_i = 0$, $i = \overline{1, m}$) и пронумерованы так, что среди всех пар $\xi_i, \xi_j (1 \leqslant i < j \leqslant m)$ величины ξ_1, ξ_2 наименее коррелированы между собой (модуль коэффициента корреляции минимален); среди $\xi_3, \xi_4, \ldots, \xi_m$ величина ξ_3 наименее коррелирована с ξ_1, ξ_2, т. е. имеет наименьший множественный коэффициент корреляции; среди $\xi_4, \xi_5, \ldots, \xi_m$ величина ξ_4 наименее коррелирована с ξ_1, ξ_2, ξ_3 и т. д. Обозначим r_{12} коэффициент корреляции ξ_1 и ξ_2, $R_{k \cdot 1, 2, \ldots, k-1}$ — множественный коэффициент корреляции ξ_k и $\xi_1, \xi_2, \ldots, \xi_{k-1}$ ($m \geqslant k \geqslant 3$). Очевидно, $R_{3 \cdot 1,2} \geqslant |r_{31}| \geqslant |r_{12}|$; $R_{k \cdot 1,2, \ldots, k-1} \geqslant R_{k \cdot 1,2, \ldots, k-2} \geqslant R_{k-1 \cdot 1,2, \ldots, k-2}$ ($k \geqslant 4$), т. е. $R_{k \cdot 1, 2, \ldots, k-1}$ не убывают с ростом k.

Возьмем в качестве первой компоненты базиса $\eta_1 = \xi_1$, в качестве второй, в соответствии с (3.50), $\eta_2 = \xi_2 - r_{12}\xi_1$, третьей, с учетом (3.56‴), $\eta_3 = \xi_3 - \beta_1^{(3)}\xi_1 - \beta_2^{(3)}\xi_2$ и вообще $\eta_{k+1} = \xi_{k+1} - \sum_{i=1}^{k} \beta_i^{(k+1)}\xi_i$ ($\beta_i^{(k+1)}$ — коэффициенты множественной связи ξ_{k+1} с $\xi_1, \xi_2, \ldots, \xi_k$, определяемые из (3.59), $k = \overline{2, m-1}$). Полученные величины будут некоррелированы между собой, причем их дисперсии $\mathbf{D}\eta_{1k} = 1 - R^2_{k \cdot 1, 2, \ldots, k-1}$ не возрастают с ростом k. Если, начиная с некоторого номера, дисперсии $\mathbf{D}\eta_k$ окажутся малыми настолько, что соответствующими компонентами можно пренебречь, получим базис, число некоррелированных компонент которого меньше числа исходных величин.

Глава 7

МЕТОДЫ СОПОСТАВЛЕНИЯ И КЛАССИФИКАЦИИ ГЕОЛОГО-ГЕОФИЗИЧЕСКИХ ДАННЫХ

Методы сопоставления геологических объектов по геолого-геофизическим показателям пород часто используются в практике геологических исследований. Это связано с тем, что в изучении различных геологических объектов значительное место занимает сравнительный анализ. Понимаемый в широком смысле, такой анализ должен включать, помимо сравнения качественных особенностей и средних значений количественных показателей пород, сопоставление других числовых характеристик распределений, а также функций и плотностей распределения. Кроме сопоставления различных геологических объектов может предусматриваться сравнение распределений показателей на одном и том же объекте, имеющих одну и ту же физическую природу, но определенных разными способами (например, кажущихся сопротивлений при различных длинах зонда). Сопоставление отдельных признаков по величине информации, которую несет каждый

о различиях геологических объектов или их частей, дает возможность выделить наиболее информативные («контрастные») с этой точки зрения признаки. К задачам сравнительного анализа относится сопоставление форм связей и характеристик силы связей показателей, например, при установлении генетического родства геологических образований по распределениям в них минералов и химических элементов.

Широкую область применения получают в настоящее время методы статистической классификации геологических объектов по комплексу количественных показателей. В эту область входят: задачи геологического картирования, выделения петрографических разностей пород по геофизическим и геохимическим данным, комплексирование методов при геологическом картировании; выделение зон, перспективных в отношении рудоносности по комплексу геолого-геофизических показателей; задачи промысловой геофизики, в частности, классификация пластов осадочных пород по промыслово-геофизическим данным в соответствии с характером их насыщения и т. д.

Перечисленные и подобные им задачи, как правило, решаются на основе статистических методов проверки гипотез и классификации. Общая постановка и принципы решения этих вопросов рассматривались в гл. 5. Теперь мы переходим к изучению способов их практического решения.

§ 1. Сравнение числовых характеристик и функций распределения

Пусть $x_{11}, x_{21}, \ldots, x_{n_1 1}$ — независимые наблюдения показателя ξ, принадлежащие некоторой генеральной совокупности Q_1; $x_{12}, x_{22} \ldots, x_{n_2 2}$ — независимые наблюдения того же показателя, принадлежащие другой генеральной совокупности Q_2. Совокупности Q_1 и Q_2 обычно связываются с определенными геологическими объектами или их частями, наличием или отсутствием характерного геологического признака, районами, участками и т. д.

Сравнение математических ожиданий. Пусть нулевая гипотеза состоит в равенстве математических ожиданий ξ: $M_1 = M_2$. Оценками величин M_1 и M_2 будут $\bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{i1}$ и $\bar{x}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{i2}$. В простейшей форме сравнение этих оценок можно осуществить уже по доверительным интервалам для M_1 и M_2:

$$\sim \bar{x}_1 \pm u_{\frac{1+q}{2}} \sqrt{\frac{\check{D}_1}{n_1}}, \sim \bar{x}_2 \pm u_{\frac{1+q}{2}} \sqrt{\frac{\check{D}_2}{n_2}}, \tag{7.1}$$

где D_1 и $\check{D}_2$ — оценки дисперсий ξ в Q_1 и Q_2,

$$\check{D}_1 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_{i1} - \bar{x}_1)^2 = \check{\sigma}_1^2, \quad \check{D}_2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(x_{i2} - \bar{x}_2)^2 = \check{\sigma}_2^2, \tag{7.2}$$

q — доверительная вероятность. При массовой обработке информации и нескольких совокупностях Q_i такое сопоставление удобно производить с помощью столбчатых диаграмм (рис. 32). Столбцы (1, 2, 3, ..., h), соответствующие сравниваемым объектам (Q_1, Q_2, ..., Q_h), строят по величинам $\bar{x}_j + u_{\frac{1+q}{2}} \sqrt{\frac{\check{D}_j}{n_j}}$ ($j = \overline{1, h}$). Части столбцов, соответствующие доверительным интервалам — от $\bar{x}_j - u_{\frac{1+q}{2}} \sqrt{\frac{\check{D}_j}{n_j}}$ до $\bar{x}_j + u_{\frac{1+q}{2}} \sqrt{\frac{\check{D}_j}{n_j}}$ — заштриховывают. Эти диаграммы, построенные для каждого показателя (ξ_1, ξ_2, ..., ξ_m), дают возможность достаточно уверенно ориентироваться в планировании дальнейшего анализа.

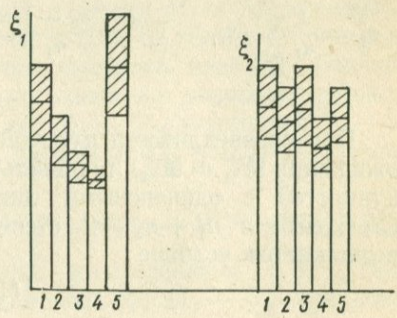


Рис. 32.

Пределы (7.1) обладают тем свойством, что они заключают величины математических ожиданий (M_1 и M_2 соответственно) с вероятностями, приближенно равными q. Если определяемые ими интервалы не пересекаются, то можно утверждать с вероятностью ошибки, не превышающей $1 - q$, что $M_1 \neq M_2$, так как вероятность этого события при $M_1 = M_2$ меньше $1 - q$. Это, однако, не означает, что пересечение доверительных интервалов всегда свидетельствует о справедливости нулевой гипотезы. Более точное решение получают с помощью соответствующих статистических критериев.

Принимая во внимание асимптотически нормальное распределение оценок $\bar{x}_1$ и $\bar{x}_2$, можно воспользоваться критерием, описанным в гл. 5. По условию (5.13) нулевая гипотеза должна отвергаться с вероятностью ошибки $\alpha \approx 1 - q$, если

$$|\bar{x}_1 - \bar{x}_2| \geqslant u_{\frac{1+q}{2}} \sqrt{\mathbf{D}\bar{x}_1 + \mathbf{D}\bar{x}_2}, \qquad (7.3)$$

где $\mathbf{D}\bar{x}_1$, $\mathbf{D}\bar{x}_2$ — дисперсии оценок $\bar{x}_1$ и $\bar{x}_2$, $\mathbf{D}\bar{x}_1 = \frac{D_1}{n_1}$, $\mathbf{D}\bar{x}_2 = \frac{D_2}{n_2}$. Подстановкой в (7.3) вместо обычно неизвестных дисперсий D_1 и D_2 их оценок (7.2) неравенство преобразуется к виду

$$|\bar{x}_1 - \bar{x}_2| \geqslant u_{\frac{1+q}{2}} \sqrt{\frac{\check{D}_1}{n_1} + \frac{\check{D}_2}{n_2}}. \qquad (7.4)$$

Поскольку такой критерий основан на асимптотическом распределении оценок $\bar{x}_1$ и $\bar{x}_2$, а вместо дисперсий $\mathbf{D}\bar{x}_1$ и $\mathbf{D}\bar{x}_2$ используются их

оценки, критическую границу для разности $|\bar{x}_1 - \bar{x}_2|$ в (7.4) надо расценивать как приближенную, уточняющуюся с увеличением объемов n_1 и n_2 обеих выборок. Удовлетворительная для практических расчетов степень приближения достигается при $n_1 \geqslant 30$, $n_2 \geqslant 30$. Более точный *критерий Стьюдента* можно применить в том случае, когда распределения сравниваемых генеральных совокупностей предполагаются нормальными, а дисперсии — одинаковыми ($D_1 = D_2$).

Критерий Стьюдента. Обозначим

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2 n_2^{-1} + s_2^2 n_1^{-1}}} \sqrt{\frac{n_1 + n_2 - 2}{n_1 + n_2}}, \tag{7.5}$$

где s_1^2, s_2^2 — оценки дисперсий вида

$$s_1^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} (x_{i1} - \bar{x}_1)^2 = \frac{n_1 - 1}{n_1} \check{D}_1, \quad s_2^2 = \frac{1}{n_2} \sum_{i=1}^{n_2} (x_{i2} - \bar{x}_2)^2 = \frac{n_2 - 1}{n_2} \check{D}_2. \tag{7.2'}$$

При справедливости нулевой гипотезы (о равенстве математических ожиданий, $\mathbf{M}\bar{x}_1 = \mathbf{M}\bar{x}_2$) и нормальном распределении сравниваемых совокупностей с одинаковыми дисперсиями, t следует распределению Стьюдента с $n_1 + n_2 - 2$ степенями свободы. Таким образом, если выполняется условие

$$|t| \geqslant t_{\frac{1+q}{2}}, \tag{7.6}$$

где $t_{\frac{1+q}{2}} - \frac{1+q}{2}$-квантиль распределения Стьюдента с $n_1 + n_2 - 2$ степенями свободы, то нулевая гипотеза должна быть отвергнута с вероятностью ошибки $1 - q$. В математической статистике доказано, что этот критерий равномерно наиболее мощный, т. е. обладает наибольшей чувствительностью при сравнении математических ожиданий двух нормально распределенных случайных величин. Отметим, что если $D_1 = D_2$, то в (7.4) $\check{D}_1$ и $\check{D}_2$ можно переставить местами, после чего с учетом $\sqrt{\frac{n_1 + n_2 - 2}{n_1 + n_2}} \approx 1$, получим критерий, подобный (7.6), с тем отличием, что вместо $u_{\frac{1+q}{2}}$ в (7.6) используется $t_{\frac{1+q}{2}}$. Сравнение квантилей распределения Стьюдента и (0; 1)-нормального распределения показывает, что таким приближенным критерием можно ограничиться уже при $n_1 + n_2 \geqslant 40$.

Критерий (7.6) двухсторонний, т. е. альтернативами для него являются условия $M_1 > M_2$ и $M_2 > M_1$. С помощью статистики t легко построить, в случае необходимости, и односторонний критерий, например для проверки гипотезы $M_1 \leqslant M_2$ либо $M_1 = M_2$ при альтернативе $M_1 > M_2$. Пусть при этой альтернативе нулевая гипотеза $M_1 = M_2$. Очевидно, чем больше величина t при фиксированных n_1, n_2 и D ($D = D_1 = D_2$), тем вероятнее альтернатива, так что критическая граница должна определять допустимые значения t как «не слишком большие», т. е. ограничивать t сверху так, чтобы вероят-

ность получения величиной t значения, большего или равного этой границе при справедливости нулевой гипотезы была мала. Такой границей при заданном уровне значимости $\alpha = 1 - q$ будет t_q — q-квантиль распределения Стьюдента с $n_1 + n_2 - 2$ степенями свободы. При выполнении условия $t \geqslant t_q$ нулевая гипотеза будет отвергаться (с вероятностью ошибки $\alpha = 1 - q$), иными словами, принимается альтернатива. Это правило используется и для проверки нулевой гипотезы $M_1 < M_2$ при альтернативе $M_1 \geqslant M_2$ либо $M_1 \leqslant M_2$ при альтернативе $M_1 > M_2$. В первом случае вероятность ошибки I рода $\alpha < 1 - q$, во втором $\alpha \leqslant 1 - q$.

Разность выборочных средних как характеристика различия. Используемая в критерии Стьюдента разность $\bar{x}_1 - \bar{x}_2$ характеризует различие сравниваемых совокупностей по математическому ожиданию ξ, так как является несмещенной оценкой разности $M_1 - M_2$ математических ожиданий M_1 и M_2 показателя (ξ) в Q_1 и Q_2. Это дает возможность из групп геологических объектов выделять наиболее близкие и наиболее отличающиеся объекты по среднему значению данного показателя. Точность оценки разности $M_1 - M_2$ величиной $\bar{x}_1 - \bar{x}_2$ можно рассчитать с помощью оценок дисперсий $\mathbf{D}\bar{x}_1$ и $\mathbf{D}\bar{x}_2$. Дисперсия разности $\bar{x}_1 - \bar{x}_2$

$$\mathbf{M}[\bar{x}_1 - \bar{x}_2 - (M_1 - M_2)]^2 = \mathbf{M}(\bar{x}_1 - M_1)^2 + \mathbf{M}(\bar{x}_2 - M_2)^2 = \mathbf{D}\bar{x}_1 + \mathbf{D}\bar{x}_2$$

и с учетом асимптотически нормального распределения $\bar{x}_1$ и $\bar{x}_2$ приближенные доверительные пределы для разности $M_1 - M_2$ при доверительной вероятности q будут:

$$\bar{x}_1 - \bar{x}_2 - u_{\frac{1+q}{2}} \sqrt{\frac{D_1}{n_1} + \frac{D_2}{n_2}},\ \bar{x}_1 - \bar{x}_2 + u_{\frac{1+q}{2}} \sqrt{\frac{D_1}{n_1} + \frac{D_2}{n_2}}. \qquad (7.7)$$

Подставив вместо D_1 и D_2 оценки (7.2), получим оценки доверительных пределов. Если ξ в Q_1 и Q_2 распределяется по нормальному закону с одной и той же дисперсией, эти оценки можно уточнить. Так как

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2 - M_1 + M_2}{\sqrt{n_1 s_1^2 + n_2 s_2^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}} \qquad (7.8)$$

подобно (7.5) подчиняется распределению Стьюдента с $n_1 + n_2 - 2$ степенями свободы,

$$\mathbf{P}\{-t_{\frac{1+q}{2}} < t_0 < t_{\frac{1+q}{2}}\} = q,$$

откуда, положив $c = \sqrt{(s_1^2 n_2^{-1} + s_2^2 n_1^{-1}) \frac{n_1 + n_2}{n_1 + n_2 - 2}}$, получим

$$\mathbf{P}\{\bar{x}_1 - \bar{x}_2 - t_{\frac{1+q}{2}} c < M_1 - M_2 < \bar{x}_1 - \bar{x}_2 + t_{\frac{1+q}{2}} c\} = q.$$

Доверительные пределы для разности $M_1 - M_2$

$$A_j^{+,-} = \bar{x}_1 - \bar{x}_2 \pm t_{\frac{1+q}{2}} c. \qquad (7.7')$$

Рассмотренная постановка задачи вполне естественна. Используя статистический критерий, мы выясняем значимость различия на фоне случайных колебаний. В сущности, для любых двух геологических объектов математические ожидания непрерывно распределенного показателя различны, но в одних случаях это различие слабо выражено, а в других — сопоставимо со случайными колебаниями показателя или превосходит их. За счет большого увеличения числа независимых наблюдений в каждой выборке почти наверняка можно обеспечить положительный результат проверки гипотезы о различии средних двух распределений. Доверительные же пределы (7.7), (7.7′) определяют интервал, в котором находится значение разности $M_1 - M_2$, которое могло бы быть получено при очень большом числе наблюдений в обеих выборках.

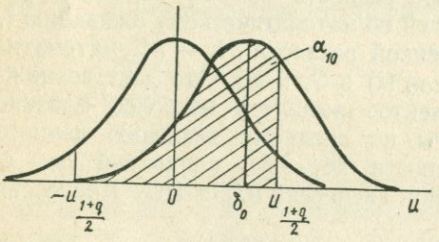


Рис. 33.

При конечных объемах выборок, малая вероятность ошибки II рода обусловливается достаточной величиной разности $M_1 - M_2$, тем большей, чем меньшим числом наблюдений мы располагаем. Вероятность выполнения неравенства (7.3) при нулевой гипотезе $M_1 = M_2$ составит $\alpha_{01} \approx 1 - q$, следовательно, вероятность противоположного события при этой же гипотезе

$$\mathbf{P}_{H_0}\left\{-u_{\frac{1+q}{2}} < \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{D_1 n_1^{-1} + D_2 n_2^{-1}}} < u_{\frac{1+q}{2}}\right\} \approx q. \qquad (7.9)$$

Пусть альтернативой H_1 является условие $M_1 - M_2 = \Delta$ при дисперсии первой совокупности D_{11}, дисперсии второй D_{21}. Вероятность ошибки II рода критерия (7.3), аналогично (5.6)

$$\alpha_{10} = \mathbf{P}_{H_1}\left\{-u_{\frac{1+q}{2}} < \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{D_1 n_1^{-1} + D_2 n_2^{-1}}} < u_{\frac{1+q}{2}}\right\} =$$

$$= \mathbf{P}_{H_1}\left\{\frac{-u_{\frac{1+q}{2}}\sqrt{D_1 n_1^{-1} + D_2 n_2^{-1}} - \Delta}{\sqrt{D_{11} n_1^{-1} + D_{21} n_2^{-1}}} < \frac{\bar{x}_1 - \bar{x}_2 - \Delta}{\sqrt{D_{11} n_1^{-1} + D_{21} n_2^{-1}}} < \right.$$

$$\left. < \frac{u_{\frac{1+q}{2}}\sqrt{D_1 n_1^{-1} + D_2 n_2^{-1}} - \Delta}{\sqrt{D_{11} n_1^{-1} + D_{21} n_2^{-1}}}\right\} \approx \Phi(u - \delta) - \Phi(-u - \delta), \qquad (7.10)$$

где $\delta = \dfrac{\Delta}{\sqrt{D_{11} n_1^{-1} + D_{21} n_2^{-1}}}$, $u = u_{\frac{1+q}{2}} \sqrt{\dfrac{n_2 D_1 + n_1 D_2}{n_2 D_{11} + n_1 D_{21}}}$; $\Phi(z)$—функция (0; 1)-нормального распределения.

Если $D_{11} = D_1$, $D_{21} = D_2$, вероятность ошибки II рода примет вид (рис. 33):

$$\alpha_{10} = \Phi\left(u_{\frac{1+q}{2}} - \delta_0\right) - \Phi\left(-u_{\frac{1+q}{2}} - \delta_0\right), \quad \delta_0 = \frac{\Delta}{\sqrt{D_1 n_1^{-1} + D_2 n_2^{-1}}}. \tag{7.11}$$

Частный случай (5.7) этой формулы (при $n_1 = n_2$, $D_1 = D_2 = D$) получен в примере, рассматривавшемся в гл. 5. Таким образом, при относительно малой величине альтернативного различия (δ_0 по сравнению с $u_{\frac{1+q}{2}}$) получим вероятность ошибки II рода, близкую к $1 - \alpha_{01} = q$:

$$\Phi\left(u_{\frac{1+q}{2}} - \delta_0\right) - \Phi\left(-u_{\frac{1+q}{2}} - \delta_0\right) \approx \Phi\left(u_{\frac{1+q}{2}}\right) - \Phi\left(-u_{\frac{1+q}{2}}\right) = \frac{1+q}{2} - \frac{1-q}{2} = q.$$

Для обеспечения же вероятности ошибки II рода, близкой к нулю — $\alpha_{10} = 1 - q_1$ $(0{,}9 < q_1 < 1)$, абсолютная величина δ_0 должна быть близкой к $u_{\frac{1+q}{2}} + u_{q_1}$. Это следует из того, что уже при $|\delta_0| > 2$

$$\alpha_{10} = \Phi\left(u_{\frac{1+q}{2}} - \delta_0\right) - \Phi\left(-u_{\frac{1+q}{2}} - \delta_0\right) \approx \begin{cases} \Phi\left(u_{\frac{1+q}{2}} - \delta_0\right) & (\text{при } \delta_0 > 2), \\ 1 - \Phi\left(-u_{\frac{1+q}{2}} - \delta_0\right) & (\text{при } \delta_0 < -2), \end{cases} \tag{7.12}$$

так как для $q \geqslant 0{,}9$ $u_{\frac{1+q}{2}} \geqslant 1{,}645$ (табл. 2, Приложение) и для $\delta_0 > 2$ $\Phi\left(-u_{\frac{1+q}{2}} - \delta_0\right) \leqslant \Phi(-3{,}645) \approx 0$; для $\delta_0 < -2$ и $q \geqslant 0{,}9$ $\Phi\left(u_{\frac{1+q}{2}} - \delta_0\right) \geqslant \Phi(3{,}645) \approx 1$. Из условия $\alpha_{10} = 1 - q_1$ и с учетом (7.12):

в области $\delta_0 > 2$

$$\Phi\left(u_{\frac{1+q}{2}} - \delta_0\right) = 1 - q_1, \quad u_{\frac{1+q}{2}} - \delta_0 = u_{1-q_1} = -u_{q_1}; \quad \delta_0 = u_{\frac{1+q}{2}} + u_{q_1}, \tag{7.13}$$

а при $\delta_0 < -2$

$$1 - \Phi\left(-u_{\frac{1+q}{2}} - \delta_0\right) = 1 - q_1, \quad -u_{\frac{1+q}{2}} - \delta_0 = u_{q_1};$$

$$\delta_0 = -u_{\frac{1+q}{2}} - u_{q_1}. \tag{7.13'}$$

Итак, заданную вероятность ошибки II рода, равную $1 - q_1$, обеспечивает следующее соотношение между разностью математических ожиданий $\Delta = M_1 - M_2$ и количествами наблюдений n_1 и n_2:

$$|\Delta| = \left(u_{\frac{1+q}{2}} + u_{q_1}\right) \sqrt{\frac{D_1}{n_1} + \frac{D_2}{n_2}}. \tag{7.14}$$

Следует отметить, что свойство оптимальности (наибольшей мощности) критерия Стьюдента для проверки гипотезы о равенстве математических ожиданий нормально распределенных случайных величин, имеющих одинаковые дисперсии, дает возможность на практике оценивать качество других критериев — применяя их в тех случаях, когда применим и критерий Стьюдента, и сравнивая результаты с получаемыми по нему выводами.

Рассмотренные методы предполагают независимость наблюдений. Если это условие не выполняется, при расчете оценок дисперсий $\mathbf{D}\bar{x}_1$ и $\mathbf{D}\bar{x}_2$ нужно учитывать автокорреляцию наблюдений. В частности, если пункты наблюдений обеих выборок линейно упорядочены и размещены с шагом h_1 и h_2 соответственно, то при достаточно больших объемах выборок и малых в сравнении с общими интервалами наблюдений радиусами автокорреляции ρ_{01} и ρ_{02} можно использовать оценки вида

$$\check{\mathbf{D}}\bar{x}_1 = \frac{c_1(0)}{n_1} + \frac{2}{n_1^2}\sum_{i=1}^{k_1}(n_1 - i)\,c_1(i), \quad \check{\mathbf{D}}\bar{x}_2 = \frac{c_2(0)}{n_2} + \frac{2}{n_2^2}\sum_{i=1}^{k_2}(n_2 - i)\,c_2(i),$$

где $c_j(i) = \dfrac{1}{n_j - i}\displaystyle\sum_{l=1}^{n_j - i}(x_{lj} - \bar{x}_j)(x_{l+i,j} - \bar{x}_j)$, $i = \overline{0,\ k_j}$, $k_j = [\rho_{0j}h_j^{-1}] + 1$, $j = 1, 2$.

Пример 7.1. По данным измерений общей радиоактивности автомагматических брекчий двух массивов получены следующие оценки: математического ожидания — $\bar{x}_1 = 18{,}8\gamma$, $\bar{x}_2 = 20{,}9\gamma$; средних квадратических отклонений — $s_1 = \sqrt{\dfrac{1}{n_1}\displaystyle\sum_{i=1}^{n_1}(x_{i1} - \bar{x}_1)^2} = 1{,}45\gamma$; $s_2 = \sqrt{\dfrac{1}{n_2}\displaystyle\sum_{i=1}^{n_1}(x_{i2} - \bar{x}_2)^2} = 1{,}81\gamma$. Количество наблюдений в первой выборке $n_1 = 22$, во второй $n_2 = 20$. 1) Проверить критерием Стьюдента гипотезу $M_1 = M_2$ о равенстве математических ожиданий при уровне значимости $\alpha_{01} = 0{,}05$. 2) Определить доверительные пределы для разности математических ожиданий при доверительной вероятности $q = 0{,}95$. 3) Оценить объемы выборок независимых наблюдений, обеспечивающие вероятность ошибки II рода не более 0,05 при альтернативном различии математических ожиданий $M_2 - M_1 \geqslant 1\gamma$, считая эти объемы одинаковыми, а дисперсии такими же, как и при нулевой гипотезе, если для ее проверки предполагается использовать приближенный критерий (7.4).

Решение. 1) По формуле (7.5) $t = \dfrac{18{,}8 - 20{,}9}{\sqrt{1{,}45^2\dfrac{1}{20} + 1{,}81^2\dfrac{1}{22}}}\sqrt{\dfrac{40}{42}} = \dfrac{-2{,}1}{0{,}516} \approx$ $\approx -4{,}07$. Табличный квантиль распределения Стьюдента с $n_1 + n_2 - 2 = 40$ степенями свободы порядка $\dfrac{1+q}{2} = \dfrac{1 + (1 - \alpha_{01})}{2} = 1 - \dfrac{\alpha_{01}}{2} = 0{,}975$, $t_{0{,}975} = 2{,}021$ (Приложение, табл. 9). Так как $|t| > t_{0{,}975}$, гипотеза о тождественности математических ожиданий отвергается с вероятностью ошибки (I рода) 0,05.

2) Доверительные пределы для разности $M_1 - M_2$ математических ожиданий по (7.7′):

$$\bar{x}_1 - \bar{x}_2 \pm ct_{0{,}975} = -2{,}1 \pm 0{,}516 \cdot 2{,}021 = \begin{cases} -1{,}06, \\ -3{,}14. \end{cases}$$

Таким образом, M_2 превышает M_1 не более чем на $3{,}14\gamma$ и не менее чем на $1{,}06\gamma$ (с вероятностью ошибки 0,05).

3) Искомые количества наблюдений, обеспечивающие вероятность ошибки II рода не более 0,05, наверняка будут значительно превышать имеющиеся объемы выборок; об этом свидетельствует приведенный расчет доверительных пределов для $M_1 - M_2$. Поэтому для проверки гипотезы можно ограничиться применением критерия (7.4), а для расчета оценок количеств наблюдений использовать соотношение (7.14).

Рассмотрим наименее благоприятный случай условия $M_2 - M_1 \geqslant 1\gamma : M_2 - M_1 = 1\gamma$. При количестве наблюдений N в каждой выборке по (7.14) имеем

$$|\Delta| = (u_{0,975} + u_{0,95})\sqrt{D_1 N^{-1} + D_2 N^{-1}} = 1$$

$\left(\text{считая } q = 1 - \alpha_{01} = 0{,}95,\ \frac{1+q}{2} = 0{,}975,\ q_1 = 1 - \alpha_{10} = 0{,}95\right)$. Из этого равенства

$$N = \frac{1}{\Delta^2}(u_{0,975} + u_{0,95})^2 (D_1 + D_2).$$

По табл. 2 (Приложение) находим $u_{0,975} = 1{,}96$, $u_{0,95} = 1{,}645$ и, использовав вместо D_1 и D_2 оценки s_1^2 и s_2^2, получим

$$\check{N} = (1{,}96 + 1{,}645)^2 (1{,}45^2 + 1{,}81^2) \approx 70.$$

Критерий Вилкоксона. Этот критерий относится к *непараметрическим*, так как для его применения не нужно заранее устанавливать законы распределения сравниваемых генеральных совокупностей (точнее, неизвестны семейства функций, к которым принадлежат плотности распределения). В соответствии с критерием Вилкоксона независимые наблюдения двух выборок x_i, y_k $(i = \overline{1, n_1},\ k = \overline{1, n_2})$ располагаются в общую последовательность в порядке их возрастания. В полученной последовательности подсчитывают *число инверсий* U — сумму количеств всех y_k, стоящих впереди каждого x_i. В случае справедливости нулевой гипотезы (о тождественности распределений) U распределяется асимптотически нормально (при возрастании объемов выборок n_1 и n_2) с математическим ожиданием и дисперсией, соответственно

$$\mathbf{M}U = \frac{n_1 n_2}{2},\ \mathbf{D}U = \frac{n_1 n_2}{12}(n_1 + n_2 + 1). \tag{7.15}$$

Благодаря этому критические значения для U можно оценивать с помощью таблицы значений функции (0; 1)-нормального распределения, приняв для двухстороннего критерия в качестве нижней и верхней границ, соответственно,

$$U_q^- = \mathbf{M}U - u_{\frac{1+q}{2}}\sqrt{\mathbf{D}U},\ U_q^+ = \mathbf{M}U + u_{\frac{1+q}{2}}\sqrt{\mathbf{D}U}, \tag{7.16}$$

где $q = 1 - \alpha$, $\alpha = \alpha_{01}$ — уровень значимости критерия; $u_{\frac{1+q}{2}}$ — квантиль (0; 1)-нормального распределения порядка $\frac{1+q}{2}$. При соблюдении условия $U \geqslant U_q^+$ либо $U \leqslant U_q^-$ нулевая гипотеза должна быть отвергнута и вероятность ошибки при этом близка к $1 - q$.

Критические границы (7.16) являются приближенными, так как основаны на асимптотическом распределении U. Однако приближение оказывается вполне удовлетворительным уже при $n_1 > 10$, $n_2 > 10$. При $n_1 \leqslant 10$, $n_2 \leqslant 10$ критические значения для U можно определить по специальным таблицам [4]. Критерий Вилкоксона имеет достаточно большую мощность при сравнении математических ожиданий двух величин, но в отличие от критерия Стьюдента он не требует соответствия обоих распределений нормальному закону и равенства их дисперсий между собой. В случае выполнения этих условий критерий Вилкоксона обеспечивает ту же мощность, что и равномерно наиболее мощный критерий Стьюдента при увеличении объемов выборок всего в $\frac{\pi}{3} \approx 1{,}05$ раза [4].

Критерий Вилкоксона не фиксирует различий распределений по степени рассеяния (дисперсии) и центральным моментам более высоких порядков, так что его можно рекомендовать лишь для проверки систематического смещения между двумя выборками по средним или наиболее вероятным значениям.

Аналогично описанному критерию построен *критерий Ван дер Вардена* [4]. Он характеризуется несколько большей мощностью по сравнению с критерием Вилкоксона. При его использовании наблюдения x_i и y_k двух выборок располагают в виде последовательности в порядке возрастания и вычисляют суммы

$$X = \sum_{i=1}^{n_1} \Phi^{-1}\left(\frac{r_i}{n_1 + n_2 + 1}\right), \quad Y = \sum_{k=1}^{n_2} \Phi^{-1}\left(\frac{s_k}{n_1 + n_2 + 1}\right), \qquad (7.17)$$

где r_i, s_k — порядковые номера x_i, y_k в последовательности ($i = \overline{1, n_1}$, $k = \overline{1, n_2}$; $\Phi^{-1}(z)$ — функция, обратная функции (0; 1)-нормального распределения, значения которой определяются из табл. 2 (Приложение) как квантили порядка z: $\Phi^{-1}(z) = u_z$, $\Phi(u_z) = z$. Критическую область двухстороннего критерия уровня значимости $\alpha = 1 - q$ составляют значения X и Y, для которых выполняется одно из неравенств $X \geqslant X_\alpha$ или $Y \geqslant X_\alpha$.

Для вычисления X_α используется свойство асимптотически нормального распределения величин X и Y при нулевой гипотезе с математическими ожиданиями $\mathbf{M}X = \mathbf{M}Y = 0$ и дисперсиями

$$\sigma_X^2 = \sigma_Y^2 = \frac{n_1 n_2}{(n_1 + n_2 - 1)(n_1 + n_2)} \sum_{i=1}^{n_1+n_2} \left[\Phi^{-1}\left(\frac{i}{n_1 + n_2 + 1}\right)\right]^2. \qquad (7.18)$$

Нормальное приближение рекомендуется использовать при $n_1 + n_2 > 50$, однако хорошие результаты получаются на практике уже при $n_1 \geqslant 10$, $n_2 \geqslant 10$. Критическая граница X_α получает вид $X_\alpha = u_{\frac{1+q}{2}} \sigma_X = u_{1-\frac{\alpha}{2}} \sigma_X$. При $n_1 + n_2 \leqslant 50$ для нахождения X_α можно воспользо-

ваться таблицами, приведенными в [4]. Если некоторые наблюдения x_i и y_k, занимающие в выписанной последовательности места с номерами $r+1, r+2, \ldots, r+c$ образуют группу равных между собой значений, для них необходимо вычислить

$$S_c = \Phi^{-1}\left(\frac{r+1}{n_1+n_2+1}\right) + \Phi^{-1}\left(\frac{r+2}{n_1+n_2+1}\right) + \ldots + \Phi^{-1}\left(\frac{r+c}{n_1+n_2+1}\right)$$

и к X и Y, найденным по остальным x_i и y_k, добавить, соответственно, $\frac{a}{a+b}S_c$ и $\frac{b}{a+b}S_c$, где a — количество равных x_i, b — количество равных y_k в указанной группе.

Пример 7.2. В пробах неизмененных гранодиоритов двух массивов измерены содержания титана, результаты x_i и y_k (в 10^{-1} %) приведены ниже. Проверить критерием Вилкоксона гипотезу о равенстве средних содержаний титана в неизмененных гранодиоритах массивов при уровне значимости критической области $\alpha=0{,}05$.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}
0,68	0,85	1,0	1,0	1,0	1,0	1,0	1,1	1,2	1,2	1,4	1,4	1,8	1,9	2,1	2,1	2,2	2,3	2,3	2,4

y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}	y_{13}	y_{14}	y_{15}	y_{16}	y_{17}	y_{18}	y_{19}	y_{20}
0,9	0,9	1,0	1,1	1,2	1,2	1,4	1,5	1,7	1,7	1,7	1,7	1,8	2,0	2,0	2,0	2,2	2,3	2,5	2,7

Решение. Составим общую последовательность наблюдений по возрастанию: $x_1x_2y_1y_2x_3y_3x_4x_5x_6x_7x_8y_4x_9y_5x_{10}y_6x_{11}y_7x_{12}y_8y_9y_{10}y_{11}y_{12}x_{13}y_{13}x_{14}y_{14}y_{15}y_{16}x_{15}x_{16}x_{17}y_{17}x_{18}\rightarrow y_{18}x_{19}x_{20}y_{19}y_{20}$. Количество U инверсий y на x равно сумме произведений r_ik_i, где r_i — номера тех y_k, которые сменяются в последовательности наблюдениями x_j; k_i — количество наблюдений в серии наблюдений x_j, следующих за y_{r_i}.

$$U = 2+3\cdot 5+4+5+6+7+12+13+16\cdot 3+17+18\cdot 2 = 165.$$

Математическое ожидание U при нулевой гипотезе (о тождественности распределений)

$$\mathrm{M}U = \frac{n_1n_2}{2} = \frac{20\cdot 20}{2} = 200;$$

среднее квадратическое отклонение

$$\sigma_U = \sqrt{\mathrm{D}U} = \sqrt{\frac{20\cdot 20}{12}(20+20+1)} \approx 37.$$

По (7.16) критические границы уровня значимости $\alpha = 0{,}05$, с учетом того, что $q = 1-\alpha = 0{,}95$, $\frac{1+q}{2} = 0{,}975$: $U^-_{0,95} = 200 - 1{,}96\cdot 37 \approx 127{,}5$; $U^+_{0,95} = 200 + 1{,}96\cdot 37 \approx 272{,}5$. Так как выполняется условие $127{,}5 < U < 272{,}5$, результат применения критерия Вилкоксона не противоречит нулевой гипотезе.

Сравнение дисперсий. Рассмотренные критерии используются для сопоставления двух распределений по средним значениям. Аналогичен принцип сопоставления и других числовых характеристик: как правило, оно основывается на сравнении оценок с учетом их дисперсий.

Простейший способ состоит в сопоставлении доверительных интервалов параметров распределений — как это делалось для математических ожиданий. Если доверительные интервалы не пересекаются, гипотеза о равенстве соответствующих числовых характеристик отвергается с вероятностью ошибки, не превышающей $1-q$ (q — доверительная вероятность интервалов). Если распределения этих оценок аппроксимируются нормальным законом, для получения более точного результата можно воспользоваться правилом, приведенным в гл. 5 (5.13). Если $\check{a}_1$, $\check{a}_2$ — несмещенные оценки параметра a распределений показателя в Q_1 и Q_2, то критическая граница для $|\check{a}_1 - \check{a}_2|$ уровня значимости $\alpha = 1-q$ по (5.13) оценивается в виде: $a_\alpha = u_{\frac{1+q}{2}} \times$ $\times \sqrt{\mathbf{D}\check{a}_1 + \mathbf{D}\check{a}_2}$ ($\mathbf{D}\check{a}_1$, $\mathbf{D}\check{a}_2$ — дисперсии оценок $\check{a}_1$ и $\check{a}_2$ соответственно при гипотезе $a_1 = a_2$; a_1 и a_2 — значения параметра в Q_1 и Q_2).

Используя выражения для средних квадратических отклонений оценок различных параметров, приводившиеся в гл. 6 (табл. 6.1), нетрудно оценить критическую границу в каждом конкретном случае. Понятно, что критерий, в котором используется эта граница, будет приближенным в том смысле, что истинный уровень значимости будет несколько отличаться от гарантируемого ($1-q$), причем отличие будет тем меньше, чем больше количество наблюдений n_1 и n_2 в каждой выборке.

Иногда удобно пользоваться другой формой сравнения оценок. Рассмотрим отношение $\check{f} = \frac{\check{a}_1}{\check{a}_2}$ как оценку $f = \frac{a_1}{a_2}$, полагая $\check{a}_2 \geqslant c > 0$. Обозначив $\Delta_1 = \check{a}_1 - a_1$, $\Delta_2 = \check{a}_2 - a_2$ и считая Δ_1, Δ_2 малыми, по формуле для приращения функции двух переменных получим: $\frac{\check{a}_1}{\check{a}_2} - \frac{a_1}{a_2} \approx \frac{\partial f}{\partial a_1}\Delta_1 + \frac{\partial f}{\partial a_2}\Delta_2 = \frac{1}{a_2}\Delta_1 - \frac{a_1}{a_2^2}\Delta_2$, откуда $\mathbf{M}\frac{\check{a}_1}{\check{a}_2} \approx \frac{a_1}{a_2}$ и $\mathbf{M}\left(\frac{\check{a}_1}{\check{a}_2} - \frac{a_1}{a_2}\right)^2 \approx$ $\approx \frac{\mathbf{D}\check{a}_1}{a_2^2} + \frac{a_1^2}{a_2^4}\mathbf{D}\check{a}_2$. При $a_1 = a_2$, $\mathbf{M}\left(\frac{\check{a}_1}{\check{a}_2} - 1\right)^2 \approx V_{1a}^2 + V_{2a}^2$, где V_{1a}, V_{2a} — коэффициенты вариации оценок $\check{a}_1$, $\check{a}_2$. Нулевая гипотеза $a_1 = a_2$ будет отвергаться с вероятностью ошибки, близкой к $1-q$, если

$$\left|\frac{\check{a}_1}{\check{a}_2} - 1\right| \geqslant u_{\frac{1+q}{2}} \sqrt{V_{1a}^2 + V_{2a}^2}. \qquad (7.19)$$

Рассмотрим задачу о сравнении средних квадратических отклонений. Результат этого сравнения полностью относится к дисперсиям. По (7.19) гипотеза $\sigma_1 = \sigma_2$ о равенстве средних квадратических отклонений будет отвергаться с вероятностью ошибки $\alpha_{01} \approx 1-q$, если

$$\left|\frac{s_1}{s_2} - 1\right| \geqslant u_{\frac{1+q}{2}} \sqrt{\frac{E_1+2}{4n_1} + \frac{E_2+2}{4n_2}}, \qquad (7.20)$$

где E_1, E_2 — коэффициенты эксцесса сравниваемых распределений, оцениваемые по независимым наблюдениям x_{i1} и x_{i2} в виде $\check{E}_1 = \frac{1}{n_1 s^4}\sum_{i=1}^{n_1}(x_{i1}-\bar{x}_1)^4 - 3$, $\check{E}_2 = \frac{1}{n_2 s^4}\sum_{i=1}^{n_2}(x_{i2}-\bar{x}_2)^4 - 3$ (предполагается, что n_1 и n_2 достаточно велики); s^2 — оценка дисперсии, вычисляемая ввиду условия $\sigma_1 = \sigma_2$ по формуле: $s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}$, $s_1^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(x_{i1} - \bar{x}_1)^2$, $s_2^2 = \frac{1}{n_2}\sum_{i=1}^{n_2}(x_{i2}-\bar{x}_2)^2$. Предполагая при нулевой гипотезе ($\sigma_1^2 = \sigma_2^2$) близость и центральных моментов более высоких порядков, можно в (7.20) вместо $\check{E}_1$, $\check{E}_2$ использовать их общую оценку: $\check{E} = \frac{n_1\check{E}_1 + n_2\check{E}_2}{n_1 + n_2}$.

Расчет критической границы упрощается в предположении нормального распределения сравниваемых совокупностей: в формуле (7.20) следует принять $E_1 = E_2 = 0$. Справедливость такого предположения, впрочем, обеспечивает возможность применения критерия, построенного с учетом не асимптотического, а точно известного закона распределения оценок дисперсий σ_1^2 и σ_2^2 — *критерия Фишера*. В этом критерии используется отношение вида

$$F = \frac{\check{\sigma}_1^2}{\check{\sigma}_2^2} = \frac{n_1(n_2-1)}{n_2(n_1-1)}\frac{s_1^2}{s_2^2}, \tag{7.21}$$

где $\check{\sigma}_1^2$, $\check{\sigma}_2^2$ — несмещенные оценки дисперсий σ_1^2, σ_2^2 (7.2).

Величина

$$F_1 = \frac{\check{\sigma}_1^2\sigma_2^2}{\check{\sigma}_2^2\sigma_1^2} \tag{7.22}$$

подчиняется распределению Фишера с $n_1 - 1$ и $n_2 - 1$ степенями свободы. Это следует из (2.59) и того, что $(n_1-1)\check{\sigma}_1^2/\sigma_1^2$ и $(n_2-1)\check{\sigma}_2^2/\sigma_2^2$ следуют распределению χ^2 с $n_1 - 1$ и $n_2 - 1$ степенями свободы. При нулевой гипотезе ($\sigma_1 = \sigma_2$) F_1 совпадает с F, и построив пределы допустимых значений для отношения (7.21), получим критические границы, определяющие критерий. Если выполняется одно из условий

$$F \leqslant F_{\frac{1-q}{2}}(n_1-1,\ n_2-1),\quad F \geqslant F_{\frac{1+q}{2}}(n_1-1,\ n_2-1), \tag{7.23}$$

то гипотеза о равенстве средних квадратических отклонений отвергается с вероятностью ошибки $1-q$. В (7.23) $F_\alpha(n_1-1,\ n_2-1)$ — α-квантиль распределения Фишера с $n_1 - 1$ и $n_2 - 1$ степенями свободы. Верхние критические границы $F_{\frac{1+q}{2}}(n_1-1,\ n_2-1)$ при

$q = 0{,}9$ и $q = 0{,}95$ $\left(\frac{1+q}{2} = 0{,}95$ и $\frac{1+q}{2} = 0{,}975$ соответственно$\right)$ можно определить непосредственно из табл. 10, 11 (Приложение), полагая $\nu_1 = n_1 - 1$, $\nu_2 = n_2 - 1$. Нижние критические границы $F_{\frac{1-q}{2}}(n_1 - 1, n_2 - 1)$ определяются на основании очевидного свойства: так как

$$\mathbf{P}\{F < F_\alpha(n_1 - 1, n_2 - 1)\} = \alpha = \mathbf{P}\left\{\frac{1}{F} > \frac{1}{F_\alpha(n_1 - 1, n_2 - 1)}\right\},$$

$$\mathbf{P}\left\{\frac{1}{F} \leqslant \frac{1}{F_\alpha(n_1 - 1, n_2 - 1)}\right\} = 1 - \alpha,$$

значение $[F_\alpha(n_1 - 1, n_2 - 1)]^{-1}$ служит квантилем порядка $1 - \alpha$ для величины $\frac{1}{F}$, которая подчиняется также распределению Фишера, но с $n_2 - 1$ и $n_1 - 1$ степенями свободы. Поэтому всегда справедливо равенство

$$F_\alpha(n_1 - 1, n_2 - 1) = \frac{1}{F_{1-\alpha}(n_2 - 1, n_1 - 1)}. \tag{7.24}$$

По этой формуле нижние критические границы определяются в виде

$$F_{\frac{1-q}{2}}(n_1 - 1, n_2 - 1) = \frac{1}{F_{\frac{1+q}{2}}(n_2 - 1, n_1 - 1)}. \tag{7.25}$$

Таким образом, допустимые значения F при уровне значимости $\alpha_{01} = 1 - q$ определяются условием

$$\frac{1}{F_{\frac{1+q}{2}}(n_2 - 1, n_1 - 1)} < F < F_{\frac{1+q}{2}}(n_1 - 1, n_2 - 1). \tag{7.26}$$

Допустимые значения отношения оценок $\check{\sigma}_1$ и $\check{\sigma}_2$ средних квадратических отклонений, при этом же уровне значимости, будут определяться условием

$$\sqrt{\frac{1}{F_{\frac{1+q}{2}}(n_2 - 1, n_1 - 1)}} < \frac{\check{\sigma}_1}{\check{\sigma}_2} < \sqrt{F_{\frac{1+q}{2}}(n_1 - 1, n_2 - 1)}, \tag{7.27}$$

а допустимые значения отношения выборочных средних квадратических отклонений s_1 и s_2, вычисляемых по (7.2′), — условием

$$\sqrt{\frac{n_2(n_1 - 1)}{n_1(n_2 - 1)} \frac{1}{F_{\frac{1+q}{2}}(n_2 - 1, n_1 - 1)}} < \frac{s_1}{s_2} <$$

$$< \sqrt{\frac{n_2(n_1 - 1)}{n_1(n_2 - 1)} F_{\frac{1+q}{2}}(n_1 - 1, n_2 - 1)}. \tag{7.27′}$$

Сравнением полученных пределов с пределами, вычисленными для приближенного критерия (7.20), можно оценить минимальные количества

наблюдений, при которых приближение остается удовлетворительным. Если сравниваемые распределения нормальны ($E_1 = E_2 = 0$), критические границы для $\frac{s_1}{s_2}$ по приближенной формуле (7.20) приобретут вид

$$1 \pm u_{\frac{1+q}{2}} \sqrt{\frac{2}{4n_1} + \frac{2}{4n_2}} = 1 \pm u_{\frac{1+q}{2}} \sqrt{\frac{1}{2n_1} + \frac{1}{2n_2}}.$$

Например, для $n_1 = n_2 = 25$ и $q = 0{,}90$ (уровень значимости критической области $\alpha_{01} = 1 - q = 0{,}10$) эти границы равны

$$1 - u_{0{,}95} \sqrt{\frac{1}{25}} = 1 - 1{,}645 \cdot \frac{1}{5} \approx 0{,}67; \quad 1 + u_{0{,}95} \sqrt{\frac{1}{25}} \approx 1{,}33,$$

а по формуле (7.27′) при тех же n_1, n_2 и q: $\sqrt{\frac{24 \cdot 25}{F_{0{,}95}(24{,}24) \cdot 24 \cdot 25}} = \sqrt{\frac{1}{1{,}98}} \approx 0{,}71$ (нижняя граница); $\sqrt{\frac{24 \cdot 25}{25 \cdot 24} F_{0{,}95}(24{,}24)} = \sqrt{1{,}98} \approx 1{,}41$ (верхняя граница). Квантиль $F_{0{,}95}(24{,}24) = 1{,}98$ находим из табл. 10 (Приложение). Значения критических границ, вычисленные по обеим формулам для различных $n = n_1 = n_2$, приведены в табл. 7.1 ниже. Как видно из сравнения этих данных, уже при $n > 60$ формула (7.20) дает вполне удовлетворительное приближение.

Таблица 7.1. Значения критических границ (u^-, u^+), (f^-, f^+) для отношения выборочных средних квадратических отклонений, вычисленные при $q = 0{,}9$ по приближенной формуле (7.20) и по формуле (7.27′) с использованием квантилей распределения Фишера.

	$n_1 = n_2 = 25$		$n_1 = n_2 = 31$		$n_1 = n_2 = 41$		$n_1 = n_2 = 61$		$n_1 = n_2 = 121$	
u^-, u^+	0,67	1,33	0,70	1,30	0,74	1,26	0,79	1,21	0,85	1,15
f^-, f^+	0,71	1,41	0,74	1,36	0,77	1,30	0,81	1,24	0,86	1,16

Подобно тому, как определялась точность оценки разности математических ожиданий с помощью критических границ критериев различия, вычисляются и доверительные пределы для отношения средних квадратических отклонений $\frac{\sigma_1}{\sigma_2}$, оцениваемого отношением $\frac{\check{\sigma}_1}{\check{\sigma}_2}$. Вероятность соблюдения условия

$$\frac{1}{F_{\frac{1+q}{2}}(n_2 - 1,\ n_1 - 1)} < \frac{\check{\sigma}_1^2 \sigma_2^2}{\check{\sigma}_2^2 \sigma_1^2} < F_{\frac{1+q}{2}}(n_1 - 1,\ n_2 - 1)$$

равна q и доверительными пределами уровня q для отношения $\frac{\sigma_2}{\sigma_1}$ будут

$$\frac{\check{\sigma}_2}{\check{\sigma}_1} \sqrt{\frac{1}{F_{\frac{1+q}{2}}(n_2 - 1,\ n_1 - 1)}}, \quad \frac{\check{\sigma}_2}{\check{\sigma}_1} \sqrt{F_{\frac{1+q}{2}}(n_1 - 1,\ n_2 - 1)} \qquad (7.28)$$

либо, при использовании оценок s_1 и s_2 (7.2′), пределы, рассчитанные по этим же формулам с подстановкой $\check{\sigma}_1 = s_1\sqrt{\frac{n_1}{n_1-1}}$, $\check{\sigma}_2 = s_2\sqrt{\frac{n_2}{n_2-1}}$.

Аналогично, при использовании приближенного критерия (7.20) оценки доверительных пределов для отношения $\frac{\sigma_2}{\sigma_1}$

$$\frac{s_2}{s_1}\left(1 \mp u_{\frac{1+q}{2}}\sqrt{\frac{E_1+2}{4n_1}+\frac{E_2+2}{4n_2}}\right). \qquad (7.29)$$

Пример 7.3. По данным опробования терригенных отложений получены оценки средних квадратических отклонений показаний нейтронного гамма-метода (НГМ), составившие: для неколлекторов $s_1 = 0{,}22$; для нефтенасыщенных пластов $s_2 = 0{,}10$ (в относительных единицах). Выборочные коэффициенты асимметрии и эксцесса составили, соответственно: по нефтенасыщенным пластам $\check{A}_1 = -0{,}53$, $\check{E}_1 = 0{,}86$; по неколлекторам $\check{A}_2 = 0{,}11$, $\check{E}_2 = -0{,}99$; количества наблюдений в обеих выборках одинаковы: $n_1 = n_2 = 20$. Проверить гипотезу: 1) о нормальности обоих распределений; 2) о различии средних квадратических отклонений (при уровне значимости 0,05). 3) Оценить доверительные пределы для отношения средних квадратических отклонений при доверительной вероятности $q = 0{,}95$.

Решение. 1) По формулам табл. 6.1 приближенные пределы, в которых с вероятностью $q = 0{,}95$ должна находиться оценка коэффициента асимметрии при гипотезе о нормальном распределении показателя, составят $\pm 1{,}96\sqrt{\frac{6}{20}} \approx \pm 1{,}07$;

такие же пределы для оценки коэффициента эксцесса — $\pm 2 \cdot 1{,}96\sqrt{\frac{6}{20}} \approx \pm 2{,}14$.

Сравнение их с $\check{A}_i$, $\check{E}_i$ показывает, что выборочные данные не противоречат гипотезе о нормальности обоих распределений*.

2) Используя полученный результат, применим для сопоставления средних квадратических отклонений критерий Фишера. Отношение F по (7.21)

$$F = \frac{s_1^2 n_1 (n_2-1)}{s_2^2 n_2 (n_1-1)} = \frac{0{,}22^2 \cdot 19 \cdot 20}{0{,}10^2 \cdot 19 \cdot 20} = 4{,}84.$$

Критические границы для отношения F при уровне значимости $\alpha_{01} = 0{,}05$ по табл. 11 (Приложение)

$$F_{0{,}975}(19{,}19) \approx 2{,}53,\ F_{0{,}025}(19{,}19) = \frac{1}{F_{0{,}975}(19{,}19)} \approx 0{,}40.$$

Гипотеза о равенстве дисперсий отвергается с вероятностью ошибки 0,05: в неколлекторах рассеяние показаний НГМ значимо больше, чем в нефтенасыщенных пластах, что является, очевидно, следствием бо́льшей неоднородности пластов, объединенных в класс «неколлекторы».

3) Доверительные пределы для отношения $\frac{\sigma_2}{\sigma_1}$, по (7.28),—

$$\frac{0{,}10}{0{,}22}\sqrt{\frac{19 \cdot 19}{19 \cdot 19}[F_{0{,}975}(19{,}19)]^{-1}} \approx 0{,}29,\quad \frac{0{,}10}{0{,}22}\sqrt{F_{0{,}975}(19{,}19)} \approx 0{,}72,$$

а отношение $\frac{\sigma_1}{\sigma_2}$ с вероятностью $q = 0{,}95$ заключено в пределах $\frac{1}{0{,}72} \approx 1{,}39$ $\frac{1}{0{,}29} \approx 3{,}44$.

* Сравнение коэффициента эксцесса с пределами допустимых значений в этой задаче носит чисто иллюстративный характер — объемы обеих выборок слишком малы.

Сравнение плотностей распределения. Мы рассмотрели способы сопоставления основных числовых характеристик одномерных распределений — математических ожиданий и средних квадратических отклонений. Они характеризуют существенно различные стороны показателей: различие математических ожиданий свидетельствует о систематическом смещении между сравниваемыми совокупностями, а средних квадратических отклонений — о различной степени рассеяния значений показателя вокруг своего среднего. Тождественность обеих основных числовых характеристик свидетельствует о близости плотностей распределения. Впрочем, иногда представляет интерес сопоставление и других числовых характеристик. На рис. 34 изображен случай, когда распределения различаются по коэффициенту ассиметрии, а по математическим ожиданиям и дисперсиям близки.

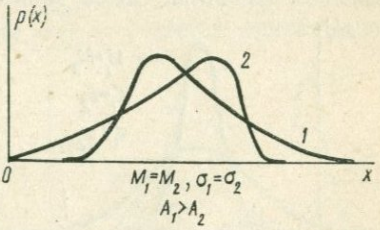


Рис. 34.

Задача сопоставления других числовых характеристик носит, как правило, специфический характер. Обычно вполне достаточно воспользоваться приближенным критерием в форме (5.13) или (7.19), оценив необходимые величины (дисперсии $D\check{a}_i$ или коэффициенты вариации V_{ia}) с учетом условий, налагаемых на эти величины нулевой гипотезой и другими правдоподобными предположениями. Такой учет, в частности, предусматривает выведение общих оценок для тех величин, которые по нулевой гипотезе должны быть равны. Если исследуемые распределения нормальны, то коэффициенты асимметрии и эксцесса в соответствующих расчетных формулах принимаются равными нулю.

Этот способ можно применить и для сравнения оценок, полученных с помощью нормализующих преобразований, например, в случае аппроксимации распределений обобщенно-логнормальным законом. Так, для сравнения медиан двух логарифмически нормальных распределений достаточно применить один из рассматривавшихся критериев сравнения параметров μ_1 и μ_2, представляющих собой математические ожидания логарифмов показателей и оцениваемых в виде средних арифметических, составленных из логарифмов наблюдений. Это обусловливается тем, что медианы по (2.43) являются экспоненциальными функциями этих параметров. Для сравнения математических ожиданий логнормальных распределений достаточно проверить гипотезу о равенстве величин, которыми они определяются. По табл. 6.2 (гл. 6) оценки математических ожиданий двух распределений

$$\check{M}_1 = \exp\left(\check{\mu}_1 + \frac{1}{2}\check{\delta}_1^2\right), \quad \check{M}_2 = \exp\left(\check{\mu}_2 + \frac{1}{2}\check{\delta}_2^2\right),$$

где $\check{\mu}_1$, μ_2 — средние арифметические, составленные из логарифмов наблюдений; $\check{\delta}_1^2$, $\check{\delta}_2^2$ — выборочные дисперсии по этим же данным. Используя для вычисления дисперсий сумм $\check{\mu}_1 + \frac{1}{2}\check{\delta}_1^2$ и $\check{\mu}_2 +$

$+\frac{1}{2}\delta_2^2$ формулы (4.22), не трудно построить по (5.13) приближенный критерий сравнения этих величин, который и даст возможность проверить гипотезу о тождественности математических ожиданий самих показателей.

Сопоставление отдельных числовых характеристик в известном смысле является односторонним, так как распределения сравниваются по каким-либо отдельным свойствам. Поэтому наряду со сравнением числовых характеристик возникает задача сравнения самих функций или плотностей распределения, которая охватывает возможные различия по отдельным характеристикам или их группам. Особое значение она приобретает в тех случаях, когда необходимо из группы показателей выделить такие, по распределениям которых сравниваемые объекты наиболее разнятся. Здесь может оказаться недостаточным сравнение

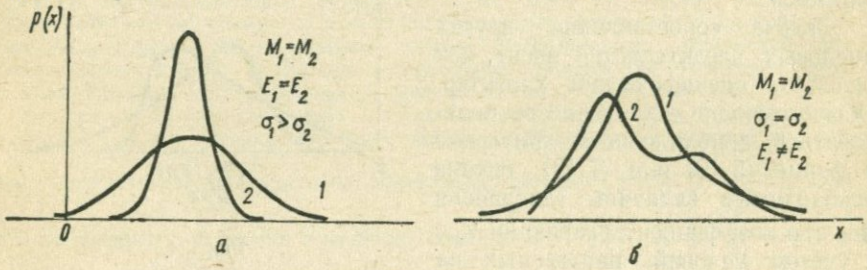


Рис. 35

по какой-либо одной числовой характеристике, так как значения последней могут быть близкими, а распределения в целом значительно отличаться друг от друга (рис. 35). Простой и удобной количественной характеристикой различия двух плотностей распределения $p_i(x)$ и $p_j(x)$ может служить величина $k_{ij} = 1 - s_{ij}$, где s_{ij} — общая часть площадей, ограничиваемых кривыми $p_i(x)$ и $p_j(x)$ над осью Ox (рис. 36). Поскольку площадь, ограничиваемая плотностью распределения и осью абсцисс, равна единице, s_{ij} и k_{ij} всегда ограничены сверху единицей: $0 \leqslant k_{ij} \leqslant 1$, $0 \leqslant s_{ij} \leqslant 1$. Если плотности распределения полностью тождественны, $s_{ij} = 1$ и $k_{ij} = 0$ — характеристика различия принимает минимальное значение, равное нулю. Если плотности распределения отделены друг от друга так, что $s_{ij} = 0$, то k_{ij} принимает максимальное значение, равное единице. Очевидно, по одному показателю, для которого $k_{ij} = 1$, можно безошибочно определить, к какому из двух распределений и соответствующих им совокупностей (Q_i или Q_j) принадлежит наблюдение, если заранее не известно, к какой из них он относится. Один такой показатель дает вполне достаточную информацию для решения задач классификации по принадлежности наблюдений к совокупностям Q_i и Q_j.

Величина k_{ij} имеет простой вероятностный смысл —

$$k_{ij} = 1 - (q_{ij} + q_{ji}), \qquad (7.30)$$

где q_{ji} — вероятность ошибочного принятия гипотезы о соответствии

испытуемого наблюдения плотности распределения $p_i(x)$ при альтернативе $p_j(x)$ по оптимальному критерию Неймана – Пирсона (5.8) с критической величиной отношения $\frac{p_i(x)}{p_j(x)}$, равной единице (в (5.8) $c(\alpha_{01}) = 1$) и, одним испытуемым наблюдением; q_{ij} — вероятность ошибочного принятия гипотезы о плотности $p_j(x)$ по тому же правилу. Равенство величины k_{ij} единице означает, что $q_{ij} = q_{ji} = 0$, т. е. критерий будет безошибочным. Другое выражение величины k_{ij} —

$$k_{ij} = q_{ii} - q_{ji} = q_{jj} - q_{ij}, \qquad (7.30')$$

где q_{ii} и q_{jj} — мощности того же критерия при принятии в качестве нулевой гипотезы соответственно $p_i(x)$ и $p_j(x)$.

В справедливости формул (7.30), (7.30') не трудно убедиться. Для простоты будем считать, что уравнение $\frac{p_i(x)}{p_j(x)} = 1$ имеет единственное решение. В этом случае области принятия гипотез о плотностях $p_i(x)$ и $p_j(x)$ разделяются значением показателя x_0, при котором $p_i(x_0) = p_j(x_0)$, т. е. x_0 — абсцисса точки пересечения кривых $p_i(x)$ и $p_j(x)$. На рис. 36 часть заштрихованной площади слева от x_0 соответствует вероятности q_{ji} ошибочного принятия гипотезы о плотности распределения $p_i(x)$, справа — вероятности q_{ij} ошибочного принятия $p_j(x)$; q_{ii} — площадь над осью Ox, ограниченная сверху кривой $p_i(x)$ и справа — вертикальным отрезком от точки x_0. Незаштрихованная часть площади под кривой $p_i(x)$ равна k_{ij} и представляется в виде разности площади под всей кривой и заштрихованной ее части: $k_{ij} = 1 - (q_{ij} + q_{ji})$. С другой стороны, k_{ij} равно разности q_{ii} и q_{ji}.

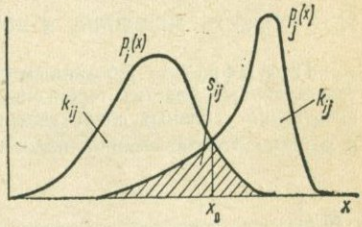


Рис. 36.

Вычислять величину k_{ij} на практике проще всего геометрическим способом. Для этого следует из точки P пересечения кривых плотностей распределения провести наклонные так, чтобы получить треугольник, равновеликий общей части площадей, ограничиваемых этими кривыми и осью абсцисс. Кривые, ограничивающие эту площадь, как правило, вогнутые. Поэтому наклонные проводят так, чтобы часть B, остающаяся вне треугольника (рис. 37), оказалась близкой по площади к сегменту A, включенному в треугольник. Величина k_{ij} вычисляется в виде

$$k_{ij} = 1 - \frac{x_2 - x_1}{2} h, \qquad (7.31)$$

где h — ордината точки P, $x_2 - x_1$ — длина основания треугольника. Способ сведения к простейшим геометрическим фигурам используется и при вычислении k_{ij} по более сложным пересечениям площадей, ограничиваемых кривыми плотностей распределения $p_i(x)$ и $p_j(x)$.

Для определения оценок плотностей $p_i(x)$ и $p_j(x)$ при больших объемах выборок можно ограничиться использованием гистограмм, сгладив получаемые кусочно-постоянные функции непрерывными кривыми, с соблюдением того условия, чтобы ограничиваемые ими площади равнялись единице. При удовлетворительной аппроксимации распределений способом нормализующих преобразований, например с помощью обобщенно-логнормального закона, оценки плотностей распределения получаются в аналитическом виде по (2.38). Величину k_{ij} можно вычислять по формуле

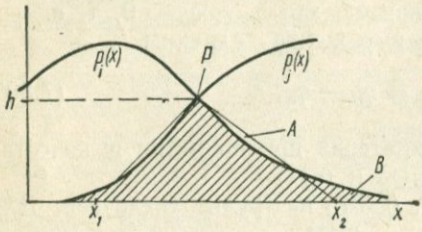


Рис. 37.

$$k_{ij} = 1 - \int_{\Omega_i} p_i(x)\,dx - \int_{\Omega_j} p_j(x)\,dx, \tag{7.32}$$

где Ω_i — область значений показателя, в которой $p_i(x) \leqslant p(x)$, Ω_j — область значений, в которой $p_i(x) > p_j(x)$.

Пример 7.4. По данным скважинных измерений в водонасыщенных и нефтенасыщенных пластах терригенных отложений получены оценки плотностей распределения удельных сопротивлений $\rho_п$, графики которых приведены на рис. 38,*а*, и потенциалов собственной поляризации $\Delta E_{сп}$ в относительных единицах (рис. 38,б).

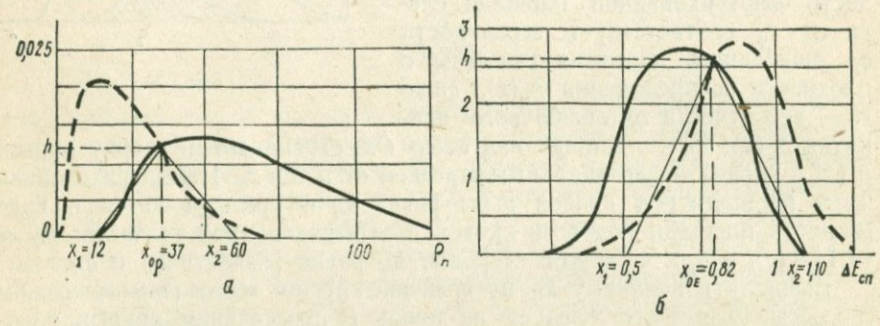


Рис. 38

Пунктирной линией обозначены плотности распределения в водонасыщенных пластах, сплошной — в нефтенасыщенных. Сравнить с помощью характеристики k_{ij} параметры $\rho_п$ и $\Delta E_{сп}$ как признаки различия пластов по характеру их насыщения.

Решение. Проведя наклонные из точки пересечения кривых плотностей распределения для получения треугольника, равновеликого площади пересечения ограничиваемых этими кривыми фигур, имеем по (7.31):

$$\text{для } \rho_п\ x_1 = 12,\ x_2 = 60,\ h = 0{,}012,\ k'_{12} = 1 - \frac{0{,}012\,(60 - 12)}{2} \approx 0{,}71;$$

$$\text{для } \Delta E_{сп}\ x_1 = 0{,}50,\ x_2 = 1{,}10,\ h = 2{,}6,\ k''_{12} = 1 - \frac{2{,}6\,(1{,}10 - 0{,}50)}{2} \approx 0{,}22.$$

Таким образом, удельное сопротивление как признак характера насыщения пласта несет значительно большую информацию по сравнению с потенциалом собственной поляризации.

Проведем вероятностную интерпретацию полученных значений k_{ij}. Величина x_0, соответствующая отношению плотностей, равному единице (т. е. $p_1(x_0) = p_2(x_0)$) составляет: для $\rho_{\text{п}}$ $x_{0\rho} = 37$ *ом м*, для $\Delta E_{\text{сп}}$ $x_{0e} = 0{,}82$. В соответствии с критерием Неймана — Пирсона по результату x измерения $\rho_{\text{п}}$ пласт должен признаваться нефтенасыщенным, если $\frac{p_1(x)}{p_2(x)} > 1$, т. е. $x > x_{0\rho} = 37$ *ом м* ($p_1(x)$ и $p_2(x)$ — плотности распределения в нефтенасыщенных и водонасыщенных пластах соответственно). В противном случае, $x \leqslant x_{0\rho} = 37$ *ом м*, пласт должен признаваться водонасыщенным. Вероятность q'_{12} ошибочного признания пласта водонасыщенным равна площади, ограниченной плотностью $p_1(x)$ слева от критического значения $x_{0\rho}$:

$$q'_{12} = \frac{h(x_{0\rho} - x_1)}{2} = \frac{0{,}012(37 - 12)}{2} = 0{,}15.$$

Вероятность ошибочного признания пласта нефтенасыщенным (по значению $\rho_{\text{п}}$, находящемуся справа от $x_{0\rho}$) —

$$q'_{21} = \frac{h(x_2 - x_{0\rho})}{2} = \frac{0{,}012(60 - 37)}{2} \approx 0{,}14,$$

и $s'_{12} = 1 - k'_{12} \approx 0{,}15 + 0{,}14 = 0{,}29$.

Для потенциала собственной поляризации критическое значение $x_{0e} = 0{,}84$; вероятности ошибок

$$q''_{12} = \frac{2{,}6(1{,}1 - 0{,}84)}{2} \approx 0{,}34, \quad q''_{21} = \frac{2{,}6(0{,}84 - 0{,}5)}{2} \approx 0{,}44$$

и $s''_{12} = 1 - k''_{12} \approx 0{,}34 + 0{,}44 = 0{,}78$.

Вероятности q''_{12} и q''_{21} оказались довольно близкими к 0,5 — величине, получаемой в том случае, когда характер насыщения указывается наугад с одинаковыми, равными по 0,5 вероятностями.

Описанный способ сопоставления плотностей распределения дает возможность получить количественную характеристику их отличия (k_{ij}). Задача допускает и другую постановку — в виде проверки гипотез об однородности распределений, т. е. о тождественности функций распределения. Рассмотрим некоторые критерии.

Критерий Смирнова. Пусть $F_1(x)$, $F_2(x)$ — выборочные функции распределения некоторого показателя, построенные по выборкам независимых наблюдений объемов n_1 и n_2; D — наибольшее расхождение между $F_1(x)$ и $F_2(x)$, $D = \sup_x |F_1(x) - F_2(x)|$. Нулевая гипотеза состоит в совпадении функций распределения во всем интервале значений аргумента. Критическую область уровня значимости α составляют значения D, удовлетворяющие условию $D \geqslant D_\alpha$, где D_α — критическая граница, вычисляемая по приближенной формуле

$$D_\alpha \approx \sqrt{-\frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \ln \frac{\alpha}{2}}. \tag{7.33}$$

Эта граница обладает тем свойством, что при нулевой гипотезе $\mathbf{P}\{D \geqslant D_\alpha\} \approx \alpha$, поэтому при $D \geqslant D_\alpha$ гипотеза о тождественности распределений должна быть отвергнута с вероятностью ошибки, близкой к α.

Пример 7.5. Графики выборочных функций распределения содержаний CaO в неизмененных породах двух последовательных интрузий представлены на рис. 39. Количества наблюдений в выборках $n_1 = n_2 = 44$. Проверить гипотезу о тождественности распределений содержаний CaO при уровне значимости $\alpha = 0{,}05$.

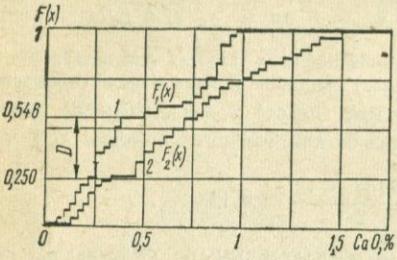


Рис. 39.

Решение. На рис. 39 наибольшее расхождение между выборочными функциями распределения $D = 0{,}546 - 0{,}250 = 0{,}296$. Критическая граница уровня значимости $\alpha = 0{,}05$, по формуле (7.33) —

$$D_{0,05} = \sqrt{\frac{1}{2}\left(\frac{1}{44} + \frac{1}{44}\right) \left| \ln 0{,}025 \right|} \approx$$

$$\approx 0{,}29.$$

Нулевая гипотеза отвергается, т. е. распределения содержаний CaO в сравниваемых интрузиях различны.

Критерии χ^2. Этот критерий также является приближенным и употребляется для проверки гипотезы о совпадении двух функций распределения при достаточно больших объемах выборок.

Область значений, принимаемых показателем в обеих выборках, разбивается на k интервалов так, чтобы в каждый из них попадало не менее пяти — семи наблюдений какой-либо из двух выборок. Далее, вычисляется статистика, характеризующая отличие распределений:

$$\chi^2_k = \sum_{i=1}^{k} \frac{(n_i' - n_1\tilde{p}_i)^2 (n_1 + n_2)}{n_1 n_2 \tilde{p}_i} = n_1 n_2 \sum_{i=1}^{k} \frac{1}{n_i' + n_i''} \left(\frac{n_i'}{n_1} - \frac{n_i''}{n_2}\right)^2, \qquad (7.34)$$

где $\tilde{p}_i = \dfrac{n_i' + n_i''}{n_1 + n_2}$ — оценка вероятности попадания значения показателя в i-й интервал выборок при нулевой гипотезе (об однородности сравниваемых распределений); n_i' — количество наблюдений первой выборки, попавших в i-й интервал; n_i'' — количество наблюдений второй выборки, попавших в тот же интервал; n_1, n_2 — объемы первой и второй выборок.

Величина χ^2_k при справедливости нулевой гипотезы следует, приближенно, распределению χ^2 с $k - 1$ степенями свободы. Поэтому критическую область уровня значимости α составляют значения χ^2_k, удовлетворяющие неравенству $\chi^2_k \geqslant \chi^2_{1-\alpha}(k-1)$, где $\chi^2_{1-\alpha}(k-1)$ — $(1-\alpha)$-квантиль распределения χ^2 с $k - 1$ степенями свободы, определяемый из табл. 3 (Приложение). В случае выполнения неравенства нулевая гипотеза отвергается с вероятностью ошибки, приближенно равной α.

Так как критерии Смирнова и χ^2 являются приближенными, может наблюдаться некоторое расхождение получаемых по ним результатов. Кроме того, к недостаткам критерия χ^2 следует отнести зависимость величины χ^2_k от положения границ и количества выделенных интер-

валов, относительно выбора которых нет строго определенных рекомендаций.

В тех случаях, когда закон сравниваемых распределений известен, задача допускает иное решение. Так, если обе совокупности распределены нормально, для проверки гипотезы об однородности двух распределений достаточно проверить совпадение числовых характеристик, определяющих функции распределения: математических ожиданий и дисперсий. Такой критерий будет описан ниже. Аналогично можно сравнивать плотности логарифмически нормальных распределений — проверкой гипотезы о равенстве математических ожиданий и дисперсий, оцениваемых по логарифмам наблюдений.

Рассмотрим теперь задачи сопоставления, которые решаются при многомерном статистическом анализе. Методы решения этих задач приобретают в настоящее время большое значение. Это обусловливается, с одной стороны, использованием в практике геологических исследований все большего числа разнообразных характеристик горных пород и необходимостью их комплексного, совместного сопоставления; с другой стороны, внедрением в эту практику электронно-вычислительной техники, без которой вообще широкое использование методов многомерного анализа было бы невозможным.

Критерий для сравнения векторов средних значений. Пусть объекты Q_1 и Q_2 охарактеризованы каждый m показателями $\{\xi_1, \xi_2, \ldots, \xi_m\}' = \boldsymbol{\xi}$; $\mathbf{x}_{11}, \mathbf{x}_{21}, \ldots, \mathbf{x}_{n_1 1}, \mathbf{x}_{12}, \mathbf{x}_{22}, \ldots, \mathbf{x}_{n_2 2}$ — независимые наблюдения-векторы этих показателей: $\mathbf{x}_{i1}$ — в Q_1, $\mathbf{x}_{i2}$ — в Q_2; $\mathbf{x}_{i1} = \{x'_{i1}, x'_{i2}, \ldots, x'_{im}\}'$, $\mathbf{x}_{k2} = \{x''_{k1}, x''_{k2}, \ldots, x''_{km}\}'$; x'_{ij}, x''_{kj} — наблюдения ξ_j в Q_1 и Q_2 соответственно ($i = \overline{1, n_1}$, $k = \overline{1, n_2}$); $\mathbf{m}_1 = \{M'_1, M'_2, \ldots, M'_m\}'$, $\mathbf{m}_2 = \{M''_1, M''_2, \ldots, M''_m\}'$ — математические ожидания $\boldsymbol{\xi}$ в Q_1 и Q_2. Задача состоит в проверке гипотезы о равенстве векторов $\mathbf{m}_1$ и $\mathbf{m}_2$, т. е. о равенстве математических ожиданий всех показателей: $M'_j = M''_j$, $j = \overline{1, m}$. Рассмотрим простейший случай, когда все компоненты ξ_j в обеих совокупностях независимы.

Пусть для каждого показателя ξ_j построено решающее правило для проверки гипотезы $M'_j = M''_j$, т. е. указан способ вычисления статистики τ_j, характеризующей различие математических ожиданий, и критическая граница $\tau_{\alpha j}$ для τ_j, соответствующая уровню значимости α_j. Вероятность получения значения τ_j, которое не противоречит гипотезе ($\tau_j < \tau_{\alpha j}$), будет $1 - \alpha_j$.

Построим критерий гипотезы $\mathbf{m}_1 = \mathbf{m}_2$ следующим образом. Будем ее принимать всякий раз, когда $\tau_j < \tau_{\alpha j}$ для всех компонент ξ_j ($j = \overline{1, m}$) и отвергать, если хотя бы одно из этих условий не выполняется. При нулевой гипотезе вероятность совместного выполнения этих неравенств ввиду независимости распределений ξ_j будет равна произведению вероятностей событий $\{\tau_j < \tau_{\alpha j}\}$ ($j = \overline{1, m}$):

$$\mathbf{P}\{\tau_j < \tau_{\alpha j},\ j = \overline{1, m}\} = \prod_{j=1}^{m} \mathbf{P}\{\tau_j < \tau_{\alpha j}\} = \prod_{j=1}^{m} (1 - \alpha_j). \qquad (7.35)$$

Вероятность ошибки I рода такого критерия составит $\alpha = 1 -$

$-\prod_{j=1}^{m}(1-\alpha_j)$. Если принять α_j одинаковыми, $\alpha_j=\alpha_0$, то $\alpha=1-$ $-(1-\alpha_0)^m$, откуда

$$\alpha_0=1-\sqrt[m]{1-\alpha}. \tag{7.36}$$

Полученная формула показывает, что для обеспечения вероятности ошибки I рода, равной α, уровень значимости α_0 критических областей критериев сравнения отдельных компонент следует брать существенно меньшим α. Например, если $m=4$, а $\alpha=0{,}05$, то по формуле (7.36) $\alpha_0=1-\sqrt[4]{0{,}95}\approx 0{,}013$. Если хотя бы одна из величин $\tau_j=$ $=|\bar{x}_j'-\bar{x}_j''|\left(\bar{x}_j'=\frac{1}{n_1}\sum_{i=1}^{n_1}x_{ij}',\ \bar{x}_j''=\frac{1}{n_2}\sum_{i=1}^{n_2}x_{ij}''\right)$ превзойдет критическую границу уровня значимости $\alpha_0=0{,}013$, равную по (7.3) $u_{0,994}\times$ $\times\sqrt{(\sigma_j')^2n_1^{-1}+(\sigma_j'')^2n_2^{-1}}$, гипотеза тождественности векторов средних $\mathbf{m}_1=\{M_1', M_2', M_3', M_4'\}'$ и $\mathbf{m}_2=\{M_1'', M_2'', M_3'', M_4''\}'$ будет отвергаться с вероятностью ошибки (I рода) $\alpha\approx 0{,}05$.

Подобные выводы, конечно, справедливы и для оценок других числовых характеристик, если эти оценки независимы между собой: при совместном сравнении двух групп параметров на основе критериев для каждой пары из них уровень значимости существенно выше, чем у каждой критической области в отдельности.

Рассмотрим некоторые критерии для проверки гипотез о векторах параметров.

Пусть $\{\check{a}_1', \check{a}_2', \ldots, \check{a}_k'\}=\check{\mathbf{a}}_1$ — несмещенные состоятельные оценки параметров $a_1, a_2, \ldots, a_k$ по наблюдениям одной совокупности Q_1; $\{\check{a}_1'', \check{a}_2'', \ldots, \check{a}_k''\}=\check{\mathbf{a}}_2$ — оценки тех же параметров по наблюдениям другой совокупности Q_2, причем распределение каждой оценки аппроксимируется нормальным законом распределения. Обозначим дисперсии оценок D_j', D_j'': $\mathbf{D}\check{a}_j'=D_j'$, $\mathbf{D}\check{a}_j''=D_j''$ $(j=\overline{1,k})$. Нулевая гипотеза состоит в попарном равенстве параметров a_j в Q_1 и Q_2: $\mathbf{M}\check{a}_j'=$ $=\mathbf{M}\check{a}_j''$ $(j=\overline{1,k})$. Будем считать также, что $\check{a}_1', \check{a}_2', \ldots, \check{a}_k', \check{a}_1'', \check{a}_2'', \ldots,$ $\check{a}_k''$ независимы.

Характеристикой отличия векторов $\check{\mathbf{a}}_1$, $\check{\mathbf{a}}_2$ может, например, служить величина $S=\sum_{j=1}^{k}(\check{a}_j'-\check{a}_j'')^2$. В такой характеристике, однако, не учитываются различия оценок по их вариации. Слагаемые в сумме S можно сделать в этом смысле равноценными, умножив каждое на коэффициент β_j, обратно пропорциональный дисперсии $D_j=D_j'+D_j''$ разности оценок $\check{a}_j'-\check{a}_j''$: $\beta_j=(D_j'+D_j'')^{-1}$. В характеристике

$$\rho=\sum_{j=1}^{k}\left(\frac{\check{a}_j'-\check{a}_j''}{\sqrt{D_j'+D_j''}}\right)^2=\sum_{j=1}^{k}t_j^2 \tag{7.37}$$

нормированные величины

$$t_j = \frac{\check{a}_j' - \check{a}_j''}{\sqrt{D_j' + D_j''}} \tag{7.38}$$

имеют одну и ту же дисперсию, равную единице; дисперсии их квадратов также примерно одинаковы и близки каждая к трем, как четвертый центральный момент нормированной случайной величины, распределение которой близко к нормальному (2.21).

Выяснив закон распределения статистики ρ при нулевой гипотезе, можно построить для ρ критическую границу, а с нею и нужный критерий. Так как нормированные разности (7.38) подчиняются при нулевой гипотезе нормальному закону распределения с нулевым математическим ожиданием и дисперсией, равной единице, сумма ρ их квадратов имеет распределение χ^2 с k степенями свободы. Поэтому критическая граница ρ_q уровня значимости $1-q$ равна $\chi_q^2(k)$ — квантилю порядка q распределения χ^2 с k степенями свободы. При выполнении условия $\rho \geqslant \chi_q^2(k)$ нулевая гипотеза отвергается с вероятностью ошибки $\alpha \approx 1-q$.

Сравнивая слагаемые t_j^2 в (7.37), можно выделить те оценки $\check{a}_j'$, $\check{a}_j''$, по которым различия проявляются наиболее отчетливо. Каждое слагаемое имеет распределение χ^2 с одной степенью свободы. Поэтому критическая граница для проверки гипотезы $a_j' = a_j''$ будет для каждой величины t_j^2 одной и той же и равной q-квантилю распределения χ^2 с одной степенью свободы. Большему значению t_j^2 будет соответствовать бóльшая вероятность того, что значения a_j в Q_1 и Q_2 различны. Опираясь на эту особенность, можно подобрать группу m параметров a_j $(m < k)$ так, чтобы им соответствовала среди прочих групп наибольшая сумма значений t_j^2. Тем самым найдем группу наиболее «контрастных» параметров среди всех k рассматривавшихся.

Примером применения критерия может служить сравнение одновременно оценок средних $\bar{x}_1$, $\bar{x}_2$ и средних квадратических отклонений $\check{\sigma}_1$, $\check{\sigma}_2$ двух нормально распределенных величин. Обозначим $\check{a}_1' = \bar{x}_1$, $\check{a}_1'' = \bar{x}_2$, $\check{a}_2' = \ln \check{\sigma}_1$, $\check{a}_2'' = \ln \check{\sigma}_2$. Используя выражения $\mathbf{D}\bar{x}_i$, $\mathbf{D}\check{\sigma}_i$ из табл. 6.1, имеем:

$$\mathbf{D}\bar{x}_i = \frac{\sigma_i^2}{n_i}, \quad \mathbf{D} \ln \check{\sigma}_i \approx \frac{1}{\sigma_i^2} \mathbf{D}\check{\sigma}_i \approx \frac{1}{2n_i} \quad (i = 1, 2),$$

где n_i — объемы выборок, σ_i — средние квадратические отклонения сравниваемых величин. Так как $\mathbf{M}[(\bar{x}_i - \mathbf{M}\bar{x}_i)(\ln \check{\sigma}_i - \ln \sigma_i)] \approx \frac{A_i \sigma_i}{2n_i} = 0$ (A_i — асимметрия, $i = 1, 2$), зависимостью между $\bar{x}$ и $\ln \check{\sigma}_i$ можно пренебречь. Так как при нулевой гипотезе $\sigma_1 = \sigma_2$, в качестве оценки σ_i^2 используем $\check{\sigma}^2 = \frac{n_1 \check{\sigma}_1^2 + n_2 \check{\sigma}_2^2}{n_1 + n_2}$. По (7.37) получим:

$$\rho \approx \frac{n_1 n_2}{n_1 + n_2} \left[\frac{(\bar{x}_1 - \bar{x}_2)^2}{\check{\sigma}^2} + 2 \left(\ln \frac{\check{\sigma}_1}{\check{\sigma}_2} \right)^2 \right]. \tag{7.39}$$

Гипотеза $\{\mathbf{M}\bar{x}_1 = \mathbf{M}\bar{x}_2,\ \mathbf{M}\check{\sigma}_1 = \mathbf{M}\check{\sigma}_2\}$, которая в данном случае означает тождественность обоих распределений, отвергается с вероятностью ошибки равной, приближенно, $1 - q$, если $\rho \geqslant \chi_q^2(2)$.

Рассмотрим векторно-матричную форму записи ρ в (7.37). Ковариационные матрицы векторов $\check{\mathbf{a}}_1$ и $\check{\mathbf{a}}_2$, $\mathbf{B}_i = \mathbf{M}\,[(\check{\mathbf{a}}_i - \mathbf{a}_i)(\check{\mathbf{a}}_i - \mathbf{a}_i)']$ $(i = 1,2)$, ввиду независимости их компонент, диагональны: ковариационная матрица разности $\check{\mathbf{a}}_1 - \check{\mathbf{a}}_2$, по (3.40), $\mathbf{B} = \mathbf{B}_1 + \mathbf{B}_2$. Характеристика ρ записывается в виде

$$\rho = (\check{\mathbf{a}}_1 - \check{\mathbf{a}}_2)'(\mathbf{B}_1 + \mathbf{B}_2)^{-1}(\check{\mathbf{a}}_1 - \check{\mathbf{a}}_2). \tag{7.40}$$

Нетрудно заметить аналогию между правилом (5.13) и описанным критерием сравнения векторов $\check{\mathbf{a}}_1$ и $\check{\mathbf{a}}_2$ с помощью характеристики ρ. В (5.13) в качестве характеристики различия используется величина $t = \dfrac{\check{a}_1 - \check{a}_2}{\sqrt{\mathrm{D}\check{a}_1 + \mathrm{D}\check{a}_2}}$, которая после возведения в квадрат записывается в форме (7.40), если под $\mathbf{B}_1$ и $\mathbf{B}_2$ понимать дисперсии соответствующих оценок.

Пусть теперь компоненты векторов $\check{\mathbf{a}}_1$ и $\check{\mathbf{a}}_2$ зависимы (ковариационные матрицы $\mathbf{B}_1$ и $\mathbf{B}_2$ не диагональны). Обозначим $\mathbf{C}$ симметрическую матрицу, обладающую свойством $\mathbf{C}(\mathbf{B}_1 + \mathbf{B}_2)\mathbf{C} = \mathbf{I}$ ($\mathbf{I}$ — единичная матрица), т. е. $(\mathbf{B}_1 + \mathbf{B}_2)^{-1} = \mathbf{C}\mathbf{C}' = \mathbf{C}\mathbf{C}$ ($\mathbf{B}_1$ и $\mathbf{B}_2$ предполагаются невырожденными). Из теории матриц известно, что для положительно определенной матрицы, какой является $\mathbf{B}_1 + \mathbf{B}_2$, такая матрица $\mathbf{C}$ существует [1]. Преобразуем вектор $\check{\mathbf{a}}_1 - \check{\mathbf{a}}_2$ в виде $\mathbf{C}(\check{\mathbf{a}}_1 - \check{\mathbf{a}}_2) = \check{\mathbf{b}}$. Математическое ожидание $\check{\mathbf{b}}$ при нулевой гипотезе, $\mathbf{M}\check{\mathbf{b}} = \mathbf{C}\mathbf{M}(\check{\mathbf{a}}_1 - \check{\mathbf{a}}_2) = 0$. Ковариационная матрица —

$$\mathbf{M}\check{\mathbf{b}}\check{\mathbf{b}}' = \mathbf{M}\,[\mathbf{C}(\check{\mathbf{a}}_1 - \check{\mathbf{a}}_2)(\check{\mathbf{a}}_1 - \check{\mathbf{a}}_2)'\mathbf{C}'] = \mathbf{C}(\mathbf{B}_1 + \mathbf{B}_2)\mathbf{C} = \mathbf{I}.$$

Отсюда следует, что $\check{b}_j$ — компоненты $\check{\mathbf{b}}$ — не коррелированы, причем каждая имеет дисперсию, равную единице. Поэтому величина

$$\sum_{j=1}^{k} \check{b}_j^2 = \check{\mathbf{b}}'\check{\mathbf{b}} = (\check{\mathbf{a}}_1 - \check{\mathbf{a}}_2)'\,\mathbf{C}\mathbf{C}\,(\check{\mathbf{a}}_1 - \check{\mathbf{a}}_2) =$$
$$= (\check{\mathbf{a}}_1 - \check{\mathbf{a}}_2)'(\mathbf{B}_1 + \mathbf{B}_2)^{-1}(\check{\mathbf{a}}_1 - \check{\mathbf{a}}_2) = \rho$$

имеет распределение χ^2 с k степенями свободы.

Итак, подобно рассмотренному выше критерию для независимых оценок $\check{a}_j'$, $\check{a}_j''$ $(j = \overline{1, k})$, нулевая гипотеза должна отвергаться с вероятностью ошибки $1 - q$, если будет выполнено условие

$$(\check{\mathbf{a}}_1 - \check{\mathbf{a}}_2)'(\mathbf{B}_1 + \mathbf{B}_2)^{-1}(\check{\mathbf{a}}_1 - \check{\mathbf{a}}_2) \geqslant \chi_q^2(k). \tag{7.41}$$

Рассмотрим критерий сравнения математических ожиданий многомерных величин. Пусть $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$ — оценки этих математических ожиданий,

составленные из средних арифметических, вычисленных по результатам измерений m показателей в Q_1 и Q_2:

$$\bar{\mathbf{x}}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{x}_{i1} = \{\bar{x}'_1, \bar{x}'_2, \ldots, \bar{x}'_m\}', \quad \bar{\mathbf{x}}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2}\mathbf{x}_{i2} = \{\bar{x}''_1, \bar{x}''_2, \ldots, \bar{x}''_m\}',$$

где $\bar{x}'_j = \frac{1}{n_1}\sum_{i=1}^{n_1} x'_{ij}$, $\bar{x}''_j = \frac{1}{n_2}\sum_{i=1}^{n_2} x''_{ij}$; x'_{ij}, x''_{ij} — i-е наблюдения j-го показателя (ξ_j) в совокупностях Q_1 и Q_2 соответственно ($j = \overline{1, m}$); n_1, n_2 — количества векторов наблюдений $\mathbf{x}_{i1}$, $\mathbf{x}_{i2}$ в них. Пусть, далее, $\mathbf{W}_1$, $\mathbf{W}_2$ — ковариационные матрицы исследуемых показателей в Q_1 и Q_2. Нулевая гипотеза состоит в равенстве математических ожиданий: $\mathbf{m}_1 = \mathbf{m}_2$.

Ковариационная матрица оценки $\bar{\mathbf{x}}_1$ будет

$$\mathbf{B}_1 = \mathbf{M}\left[\left(\frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{x}_{i1} - \mathbf{m}_1\right)\left(\frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{x}_{i1} - \mathbf{m}_1\right)'\right] =$$

$$= \frac{1}{n_1^2}\mathbf{M}\sum_{i=1}^{n_1}(\mathbf{x}_{i1} - \mathbf{m}_1)(\mathbf{x}_{i1} - \mathbf{m}_1)' = \frac{\mathbf{W}_1}{n_1}.$$

Аналогично, ковариационная матрица $\bar{\mathbf{x}}_2$, $\mathbf{B}_2 = \frac{\mathbf{W}_2}{n_2}$ и по (7.41) гипотеза о равенстве математических ожиданий будет отвергаться с вероятностью ошибки $1 - q$, если

$$\rho = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\left(\frac{\mathbf{W}_1}{n_1} + \frac{\mathbf{W}_2}{n_2}\right)^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \geqslant \chi^2_q(m). \qquad (7.42)$$

Применение этого правила в практике обработки геолого-геофизических данных осложняется тем, что ковариационные матрицы $\mathbf{W}_1$ и $\mathbf{W}_2$ обычно неизвестны. Если исследуемые показатели независимы, можно воспользоваться приближенным критерием, подставив вместо дисперсий D'_j, D''_j в (7.37) их оценки $\check{D}'_j = \frac{(\check{\sigma}'_j)^2}{n_1}$, $\check{D}''_j = \frac{(\check{\sigma}''_j)^2}{n_2}$, где $(\check{\sigma}'_j)^2$, $(\check{\sigma}''_j)^2$ — оценки дисперсий показателя ξ_j,

$$(\check{\sigma}'_j)^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x'_{ij} - \bar{x}'_j)^2, \quad (\check{\sigma}''_j)^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(x''_{ij} - \bar{x}''_j)^2 \quad (j = \overline{1, m}).$$

Тогда характеристика различия средних ρ примет вид:

$$\rho = \sum_{j=1}^{m}\tilde{t}_j^2, \quad \text{где } \tilde{t}_j = (\bar{x}'_j - \bar{x}''_j)\left[\frac{(\check{\sigma}'_j)^2}{n_1} + \frac{(\check{\sigma}''_j)^2}{n_2}\right]^{-1/2}. \qquad (7.43)$$

Если величины ξ_j распределены в Q_1 и Q_2 нормально, причем $\mathbf{M}(\check{\sigma}'_j)^2 = \mathbf{M}(\check{\sigma}''_j)^2$, $n_1 = n_2 = n$, то, как следует из сравнения величины $\tilde{t}_j$ с (7.5), ее распределение при нулевой гипотезе близко к распределению Стьюдента с $n_1 + n_2 - 2$ степенями свободы. Из сравнения

квантилей (0; 1)-нормального распределения и распределения Стьюдента (Приложение, табл. 2, 9) видно, что они почти не отличаются при числе степеней свободы большем 80. Это дает возможность аппроксимировать распределение характеристики ρ при $n \geqslant 40$ распределением χ^2 с m степенями свободы*.

Критерий T^2. Рассмотрим случай, когда ковариационные матрицы неизвестны и нет априорных сведений о независимости показателей в Q_1 и Q_2. Подстановка в (7.42) вместо ковариационных матриц их оценок приводит к тому, что левая часть неравенства уже не будет следовать распределению χ^2, так что критическая граница в (7.42) $\chi^2_q(m)$ будет лишь некоторым приближением. Оно может оказаться грубым, так как ковариационная матрица обычно содержит довольно большое число элементов, вследствие чего совокупное влияние погрешностей их оценок оказывается значительным.

Построение более точных критериев основывается на широко употребляемой в многомерном статистическом анализе *обобщенной T^2-статистике*. Если $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ — независимые векторы-наблюдения m-мерной нормально распределенной случайной величины ξ, то T^2-статистика определяется в виде

$$T^2 = (\bar{\mathbf{x}} - \mathbf{m})' \left(\frac{\check{\mathbf{W}}}{N}\right)^{-1} (\bar{\mathbf{x}} - \mathbf{m}), \tag{7.44}$$

где $\bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$, $\check{\mathbf{W}}$ — оценка ковариационной матрицы,

$$\check{\mathbf{W}} = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'.$$

T^2 является многомерным аналогом квадрата статистики t Стьюдента. Доказано, что если $\mathbf{m}$ — математическое ожидание ξ, то статистика $\frac{T^2 (N-m)}{(N-1)\, m}$ имеет распределение Фишера с m и $N-m$ степенями свободы. Так, определенное распределение величины T^2 называют *T^2-распределением* с $N-1$ степенями свободы [1].

Это свойство используется в критерии для проверки гипотезы о том, что вектор математического ожидания равен данному вектору $(\mathbf{m})$. Гипотеза отвергается с вероятностью ошибки α, если

$$T^2 \geqslant \frac{(N-1)\, m}{N-m} F_{1-\alpha}(m,\ N-m), \tag{7.45}$$

где $F_{1-\alpha}(m,\ N-m)$ — квантиль порядка $1-\alpha$ распределения Фишера с m и $N-m$ степенями свободы.

---

* При описании критерия Стьюдента в этой главе упоминалось, что возможность использования нормального приближения обеспечивается меньшим числом степеней свободы: $n_1 + n_2 \geqslant 40$. В рассматриваемом случае, однако, требования к аппроксимации распределений должны быть более жесткими, так как в сумме (7.43) ошибки аппроксимации могут накапливаться.

На этом свойстве основан и критерий T^2, применяемый для сопоставления двух объектов по комплексу геолого-геофизических показателей, распределенных нормально или приведенных некоторым преобразованием к нормальному закону. Пусть количества наблюдений в обеих совокупностях Q_1 и Q_2 одинаковы, $n_1 = n_2 = n$, и проверяется гипотеза о равенстве векторов математических ожиданий, $\mathbf{m}_1 = \mathbf{m}_2$. Введем наблюдения $\mathbf{y}_i = \mathbf{x}_{i1} - \mathbf{x}_{i2}$ $(i = \overline{1, n})$, которые при нулевой гипотезе имеют своим математическим ожиданием нулевой вектор и ковариационную матрицу $\mathbf{W} = \mathbf{W}_1 + \mathbf{W}_2$. Задача сводится к проверке гипотезы о равенстве математического ожидания вектора $\bar{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i$ нулевому вектору. T^2-статистика, вычисленная по наблюдениям $\mathbf{y}_i$, составит

$$T^2 = \bar{\mathbf{y}}' \left(\frac{\check{\mathbf{W}}}{n}\right)^{-1} \bar{\mathbf{y}} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \left(\frac{\check{\mathbf{W}}}{n}\right)^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2),$$

где $\check{\mathbf{W}} = \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})' = \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{x}_{i1} - \mathbf{x}_{i2} - \bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)(\mathbf{x}_{i1} - \mathbf{x}_{i2} - \bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)'$. Нулевая гипотеза должна отвергаться с вероятностью ошибки α, если

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \left(\frac{\check{\mathbf{W}}}{n}\right)^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \geqslant \frac{(n-1)\,m}{n-m} F_{1-\alpha}(m,\ n-m). \qquad (7.46)$$

Сравним это правило с приближенным критерием, который можно построить по (7.42) с подстановкой вместо ковариационной матрицы разности $\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$ ее оценки. В данном случае в качестве оценки $\frac{\mathbf{W}_1}{n} + \frac{\mathbf{W}_2}{n}$ можно взять $\frac{\check{\mathbf{W}}}{n}$ и приближенной критической границей для $(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \times \left(\frac{\check{\mathbf{W}}}{n}\right)^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ по (7.42) будет $\chi^2_{1-\alpha}(m)$. Для сравнения ее с критической границей в (7.46) ниже приведены значения последней при $m = 10$, $\alpha = 0{,}05$ и различных n, вычисленные с помощью табл. 10 (Приложение).

n	20	30	40	50	70	130	∞
$\frac{(n-1)\,10}{n-10} F_{0,95}(10,\ n-10)$	56,6	34,1	28,1	25,5	22,9	20,5	18,3

Критическая граница $\chi^2_{0,95}(10) \approx 18{,}3$ (Приложение, табл. 3). Таким образом, приближенный критерий можно применять лишь при значительных объемах наблюдений в обеих выборках ($n > 100$). При его использовании вероятность ошибки I рода будет несколько больше, чем тот уровень значимости, для которого определена критическая граница. С другой стороны, этот критерий более чувствителен к различиям математических ожиданий.

Если $n_1 \neq n_2$, но ковариационные матрицы предполагаются одинаковыми ($\mathbf{W}_1 = \mathbf{W}_2$), то T^2 вычисляется в виде [1]

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \check{\mathbf{W}}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2),$$

где $\check{\mathbf{W}} = \frac{1}{n_1+n_2-2}\left[\sum_{i=1}^{n_1} (\mathbf{x}_{i1} - \bar{\mathbf{x}}_1)(\mathbf{x}_{i1} - \bar{\mathbf{x}}_1)' + \sum_{i=1}^{n_2} (\mathbf{x}_{i2} - \bar{\mathbf{x}}_2)(\mathbf{x}_{i2} - \bar{\mathbf{x}}_2)'\right]$. Нулевая гипотеза отвергается с вероятностью ошибки α, если $T^2 \geqslant \frac{(n_1 + n_2 - 2)m}{n_1 + n_2 - m - 1} F_{1-\alpha}(m,\ n_1 + n_2 - m - 1)$.

Рассмотрим случай, когда $\mathbf{W}_1 \neq \mathbf{W}_2$, $n_1 \neq n_2$. Пусть для определенности $n_1 < n_2$. Обозначим для $i = \overline{1, n_1}$

$$\mathbf{y}_i = \mathbf{x}_{i1} - \sqrt{\frac{n_1}{n_2}}\,\mathbf{x}_{i2} + \frac{1}{\sqrt{n_1 n_2}} \sum_{j=1}^{n_1} \mathbf{x}_{j2} - \frac{1}{n_2} \sum_{j=1}^{n_2} \mathbf{x}_{j2};$$

$$\bar{\mathbf{y}} = \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{y}_i = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2; \quad \check{\mathbf{W}} = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'.$$

Можно показать, что при справедливости нулевой гипотезы ($\mathbf{m}_1 = \mathbf{m}_2$) статистика $T^2 = \bar{\mathbf{y}}' \left(\frac{\check{\mathbf{W}}}{n_1}\right)^{-1} \bar{\mathbf{y}}$ имеет T^2-распределение с $n_1 - 1$ степенями свободы. Поэтому нулевая гипотеза должна отвергаться с вероятностью ошибки α, если

$$\bar{\mathbf{y}}' \left(\frac{\check{\mathbf{W}}}{n_1}\right)^{-1} \bar{\mathbf{y}} \geqslant \frac{(n_1 - 1)\, m}{n_1 - m} F_{1-\alpha}(m,\ n_1 - m). \qquad (7.46')$$

В описанных выше решающих правилах (7.42), (7.45), (7.46) содержатся характеристики, в которых участвуют обратные матрицы. При расчетах, однако, можно обойтись без трудоемкой операции непосредственного обращения матриц. Так как упомянутые характеристики имеют вид $\mathbf{z}'\mathbf{G}^{-1}\mathbf{z}$, ($\mathbf{z} = \{z_1, z_2, \ldots, z_m\}'$ — вектор, $\mathbf{G} = \{g_{ij}\}_{i,j=1}^{m}$ — матрица), их можно вычислять в виде

$$\mathbf{z}'\mathbf{G}^{-1}\mathbf{z} = \mathbf{z}'\mathbf{a} = \sum_{i=1}^{m} z_i a_i,$$

где вектор $\mathbf{a} = \{a_1, a_2, \ldots, a_m\}'$ — решение системы линейных уравнений $\mathbf{G}\mathbf{a} = \mathbf{z}$, т. е. $\sum_{j=1}^{m} g_{ij} a_j = z_i$, $i = \overline{1, m}$.

Следует подчеркнуть, тем не менее, что широкое применение на практике описанных критериев, особенно при большом числе компонент сопоставляемых многомерных величин, как и подавляющего большинства других методов многомерного статистического анализа, возможно лишь с привлечением современной электронно-вычислительной техники.

В многомерном статистическом анализе разработаны методы проверки и таких гипотез, как равенство ковариационных матриц, а также одновременно математических ожиданий и ковариационных матриц нескольких многомерных нормальных распределений, т. е. о тождественности последних [1]. Эти методы, однако, сложны для практической реализации и к тому же задачи, решаемые с их помощью, достаточно специфичны. Поэтому мы на них не останавливаемся, рекомендуя интересующемуся читателю ознакомиться с ними по специальной литературе [1].

Пример 7.6. По результатам химического анализа определены содержания породообразующих окислов в образцах двух пород массива: 1) андезито-дацитах; 2) андезитах. Количество образцов в первой выборке $n_1 = 40$, во второй — $n_2 = 42$. Средние содержания по первой выборке: SiO_2 — $\bar{x}'_1 = 59{,}3\%$, Al_2O_3 — $\bar{x}'_2 = 16{,}36\%$; средние содержания по второй выборке: SiO_2 — $\bar{x}''_1 = 57{,}39\%$, Al_2O_3 — $\bar{x}''_2 = 16{,}13\%$.

Оценки средних квадратических отклонений по первой выборке: SiO_2 — $\check{\sigma}'_1 = 3{,}29\%$; Al_2O_3 — $\check{\sigma}'_2 = 0{,}74\%$; по второй выборке: SiO_2 — $\check{\sigma}''_1 = 2{,}76\%$, Al_2O_3 — $\check{\sigma}''_2 = 1{,}12\%$. Оценки коэффициентов корреляции между содержаниями SiO_2 и Al_2O_3: по первой выборке $\check{r}_1 = -0{,}10$, по второй $\check{r}_2 = -0{,}08$. 1) Сопоставить математические ожидания содержаний SiO_2 и Al_2O_3 в андезито-дацитах и андезитах, проверив при уровне значимости $\alpha = 0{,}05$ гипотезу о равенстве векторов математических ожиданий содержаний SiO_2 и Al_2O_3 в предположении нормальности сравниваемых распределений. 2) Определить, по какому из этих двух окислов различие более значительно.

Решение. 1) Выясним возможность применения правила с использованием (7.43), основанного на независимости компонент ξ_j в Q_1 и Q_2. Для этого проверим гипотезу о равенстве нулю коэффициентов корреляции между содержаниями SiO_2 и Al_2O_3. В соответствии с формулой (6.51′) при справедливости этой гипотезы выборочные коэффициенты корреляции $\check{r}_1$ и $\check{r}_2$ с вероятностями $q \approx 0{,}95$ должны были бы находиться в пределах, соответственно:

$$r_1^{\pm} = \pm u_{0{,}975}\sqrt{\frac{1}{n_1}} = \pm 1{,}96\sqrt{\frac{1}{40}} \approx \pm 0{,}31;\quad r_2^{\pm} = \pm 1{,}96\sqrt{\frac{1}{42}} \approx \pm 0{,}30.$$

Оценки $\check{r}_1$ и $\check{r}_2$, оказавшиеся в допустимых пределах, не противоречат гипотезе о некоррелированности содержаний SiO_2 и Al_2O_3. Характеристика различия векторов математических ожиданий (7.43)

$$\rho = \sum_{j=1}^{2} \tilde{t}_j^2 = \sum_{j=1}^{2} \frac{(\bar{x}'_j - \bar{x}''_j)^2}{(\check{\sigma}'_j)^2 n_1^{-1} + (\check{\sigma}''_j)^2 n_2^{-1}} = \frac{(59{,}3 - 57{,}39)^2}{\frac{3{,}29^2}{40} + \frac{2{,}76^2}{42}} +$$

$$+ \frac{(16{,}36 - 16{,}13)^2}{\frac{0{,}74^2}{40} + \frac{1{,}12^2}{42}} \approx 8{,}07 + 1{,}21 = 9{,}28.$$

Критическая граница для ρ определяется по табл. 3 (Приложение) как квантиль порядка 0,95 распределения χ^2 с двумя степенями свободы: $\chi^2_{0{,}95}(2) = 5{,}99$.

Так как $\rho > 5{,}99$, нулевая гипотеза отвергается с вероятностью ошибки $\alpha \approx 0{,}05$.

2) Сравнение слагаемых, составляющих в сумме ρ ($\tilde{t}_1^2 = 8{,}07$ и $\tilde{t}_2^2 = 1{,}21$) позволяет выделить показатель, по которому отличие сравниваемых объектов видно наиболее отчетливо. Как и следовало ожидать, им оказалось содержание SiO_2.

§ 2. Сопоставление характеристик силы и формы связей

Сравнение коэффициентов корреляции. Описанные в предыдущем разделе методы можно применять для сопоставления количественных характеристик связей геолого-геофизических показателей. Рассмотрим задачу сравнения коэффициентов корреляции, служащих характеристиками силы линейной статистической связи двух показателей. Пусть $\check{r}_1$ и $\check{r}_2$ — оценки коэффициентов корреляции вида (6.49) величин ξ и η, вычисленные по независимым наблюдениям из генеральных совокупностей Q_1 и Q_2; n_1 и n_2 — количества пар наблюдений, по которым вычислены эти оценки. Для построения критерия в форме (5.13), применимого для нормально или, по крайней мере, приближенно нормально распределенных оценок, воспользуемся преобразованием Фишера. Как упоминалось в гл. 6, распределение преобразованного с его помощью выборочного коэффициента корреляции аппроксимируется нормальным законом распределения. Нулевой гипотезе соответствует тождественность математических ожиданий преобразованных величин: $\mathbf{M}z_1 = \mathbf{M}z_2$, где $z_1 = \frac{1}{2}\ln\frac{1+\check{r}_1}{1-\check{r}_1}$, $z_2 = \frac{1}{2}\ln\frac{1+\check{r}_2}{1-\check{r}_2}$. В предположении нормальности распределений $\zeta = \{\xi, \eta\}$ в Q_1 и Q_2 дисперсиями z_1 и z_2 будут $\mathbf{D}z_1 \approx \frac{1}{n_1-3}$ и $\mathbf{D}z_2 \approx \frac{1}{n_2-3}$. По правилу (5.13) нулевая гипотеза должна отвергаться с вероятностью ошибки $\alpha \approx 1-q$, если

$$|z_1 - z_2| \geqslant u_{\frac{1+q}{2}} \sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}. \qquad (7.47)$$

Аналогично строится критерий сравнения частных коэффициентов корреляции. Пусть $\check{r}_1^{(k)}$, $\check{r}_2^{(m)}$ — оценки частных коэффициентов корреляции двух показателей ξ и η, полученные по двум выборкам наблюдений из совокупностей Q_1 и Q_2 соответственно; первая $\check{r}_1^{(k)}$ — при k фиксированных показателях $\xi_{j_1}, \xi_{j_2}, \ldots, \xi_{j_k}$, вторая $\check{r}_2^{(m)}$ — при m фиксированных показателях $\xi_{i_1}, \xi_{i_2}, \ldots, \xi_{i_m}$. Эти две группы могут иметь отдельные показатели общими или вообще совпадать — в зависимости от условий поставленной задачи. Пусть, далее, n_1 — число независимых наблюдений вектора $\{\xi, \eta, \xi_{j_1}, \xi_{j_2}, \ldots, \xi_{j_k}\}$ из Q_1, по которым вычислялась оценка $\check{r}_1^{(k)}$; n_2 — число независимых наблюдений вектора $\{\xi, \eta, \xi_{i_1}, \xi_{i_2}, \ldots, \xi_{i_m}\}$ из Q_2, по которым вычислялась оценка $\check{r}_2^{(m)}$. Распределения обоих векторов предполагаются нормальными.

В соответствии с упоминавшимся в гл. 6 выводом о распределении выборочного частного коэффициента корреляции, $\check{r}_1^{(k)}$ и $\check{r}_2^{(m)}$ распределены так же, как обычные выборочные коэффициенты корреляции, вычисляемые по $n_1 - k$ и $n_2 - m$ парам независимых наблюдений. Положим

$$z_1^{(k)} = \frac{1}{2}\ln\frac{1+\check{r}_1^{(k)}}{1-\check{r}_1^{(k)}} \text{ и } z_2^{(m)} = \frac{1}{2}\ln\frac{1+\check{r}_2^{(m)}}{1-\check{r}_2^{(m)}}.$$

Дисперсиями этих величин будут $\mathbf{D}z_1^{(k)} \approx \frac{1}{n_1-k-3}$ и $\mathbf{D}z_2^{(m)} \approx \frac{1}{n_2-m-3}$. Нулевая гипотеза о равенстве частных коэффициентов корреляции между собой будет отвергаться с вероятностью ошибки $\alpha \approx 1-q$, если

$$|z_1^{(k)} - z_2^{(m)}| \geqslant u_{\frac{1+q}{2}} \sqrt{\frac{1}{n_1-k-3} + \frac{1}{n_2-m-3}}. \qquad (7.48)$$

Пример 7.7 Выборочные частные коэффициенты корреляции между содержаниями кобальта и цинка при исключенном влиянии магния и марганца в гранодиоритах и жильных гранитах массива составили, соответственно, $\check{r}_1^{(2)} = 0{,}63$; $\check{r}_2^{(2)} = 0{,}15$. Количество проб в первой выборке $n_1 = 38$, во второй — $n_2 = 33$. Сопоставить полученные оценки, проверив гипотезу о тождественности частных коэффициентов корреляции между содержаниями кобальта и цинка при уровне значимости 0,05.

Решение. Применим преобразование Фишера:

$$z_1^{(2)} = \frac{1}{2} \ln \frac{1+0{,}63}{1-0{,}63} \approx 0{,}741, \quad z_2^{(2)} = \frac{1}{2} \ln \frac{1+0{,}15}{1-0{,}15} \approx 0{,}151;$$

$$z_1^{(2)} - z_2^{(2)} \approx -0{,}59.$$

Критическая граница для $|z_1^{(2)} - z_2^{(2)}|$

$$u_{0{,}975} \sqrt{\frac{1}{n_1-5} + \frac{1}{n_2-5}} = 1{,}96 \sqrt{\frac{1}{33} + \frac{1}{28}} \approx 0{,}504.$$

Нулевая гипотеза отвергается с вероятностью ошибки $\alpha \approx 0{,}05$ ввиду выполнения (7.48).

Сопоставление форм связей. Рассмотрим задачу проверки гипотез о тождестве линейных регрессий, оцениваемых по методу наименьших квадратов. Такая проверка дает вполне достаточные основания дальнейшей интерпретации геолого-геофизических данных. Получение более общего решения и его практическая реализация являются сложной задачей, требующей применения специального математического аппарата.

Пусть Q_t и Q_s — сравниваемые совокупности, $\boldsymbol{\alpha}_t$ и $\boldsymbol{\alpha}_s$ — $(k+1)$-мерные векторы коэффициентов связи между показателем ξ_{k+1} и величинами $\xi_1, \xi_2, \ldots, \xi_k$: $\boldsymbol{\alpha}_t = \{\alpha_{1t}, \alpha_{2t}, \ldots, \alpha_{k+1,\,t}\}'$, $\boldsymbol{\alpha}_s = \{\alpha_{1s}, \alpha_{2s}, \ldots, \alpha_{k+1,\,s}\}'$, т. е. в Q_t и Q_s, соответственно, $\xi_{k+1} = \sum_{j=1}^{k} \alpha_{jt}\xi_j + \alpha_{k+1,\,t} + \Delta_t$ и $\xi_{k+1} = \sum_{j=1}^{k} \alpha_{js}\xi_j + \alpha_{k+1,\,s} + \Delta_s$, где Δ_t и Δ_s — нормально распределенные случайные величины с нулевыми средними и дисперсиями σ_t^2 и σ_s^2 соответственно. Далее, пусть $\check{\boldsymbol{\alpha}}_t$, $\check{\boldsymbol{\alpha}}_s$ — оценки векторов $\boldsymbol{\alpha}_t$, $\boldsymbol{\alpha}_s$ по методу наименьших квадратов; $\mathbf{X}_t$ и $\mathbf{X}_s$ — матрицы вида:

$$\mathbf{X}_t = \{x_{ij}^{(t)}\}, \; i = \overline{1, n_t}, \; j = \overline{1, k+1};$$

$$\mathbf{X}_s = \{x_{ij}^{(s)}\}, \; i = \overline{1, n_s}, \; j = \overline{1, k+1};$$

$x_{ij}^{(t)}$ и $x_{ij}^{(s)}$ — i-е значения ξ_j в Q_t и Q_s соответственно при $j = \overline{1, k}$; $x_{ik+1}^{(t)} = 1$, $x_{ik+1}^{(s)} = 1$. Обозначим $\mathbf{C}_t = \mathbf{X}_t'\mathbf{X}_t$, $\mathbf{C}_s = \mathbf{X}_s'\mathbf{X}_s$. По (4.56),

$$\check{\boldsymbol{\alpha}}_t = \mathbf{C}_t^{-1}\mathbf{X}_t'\mathbf{y}_{k+1}^{(t)}, \quad \check{\boldsymbol{\alpha}}_s = \mathbf{C}_s^{-1}\mathbf{X}_s'\mathbf{y}_{k+1}^{(s)}, \tag{7.49}$$

где $\mathbf{y}_{k+1}^{(t)} = \{y_{ik+1}^{(t)}\}_{i=1}^{n_t}$, $\mathbf{y}_{k+1}^{(s)} = \{y_{ik+1}^{(s)}\}_{i=1}^{n_s}$ — векторы наблюдений ξ_{k+1} ($x_{ij}^{(t)}$, $x_{ij}^{(s)}$ считаем неслучайными величинами). Ковариационными матрицами векторов $\check{\boldsymbol{\alpha}}_t$, $\check{\boldsymbol{\alpha}}_s$, по (4.58), будут $\mathbf{B}_t = \sigma_t^2\mathbf{C}_t^{-1}$, $\mathbf{B}_s = \sigma_s^2\mathbf{C}_s^{-1}$, а ковариационная матрица вектора $\check{\boldsymbol{\alpha}}_t - \check{\boldsymbol{\alpha}}_s$, в соответствии с (3.40),

$$\mathbf{B}_{ts} = \mathbf{B}_t + \mathbf{B}_s = \sigma_t^2\mathbf{C}_t^{-1} + \sigma_s^2\mathbf{C}_s^{-1}. \tag{7.50}$$

Задача сравнения функций регрессии рассматривается как проверка гипотезы $\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_s$. По (7.41), эта гипотеза должна отвергаться с вероятностью ошибки $1 - q$, если

$$(\check{\boldsymbol{\alpha}}_t - \check{\boldsymbol{\alpha}}_s)'\,\mathbf{B}_{ts}^{-1}(\check{\boldsymbol{\alpha}}_t - \check{\boldsymbol{\alpha}}_s) \geqslant \chi_q^2(k+1). \tag{7.51}$$

Если σ_t^2 и σ_s^2 неизвестны, используются их оценки $\check{\sigma}_t^2$ и $\check{\sigma}_s^2$ вида (4.61) при $p_i = 1$ с подстановкой в (4.61) $k + 1$ вместо k.

Такая процедура позволяет сопоставлять оценки регрессий и лишь по коэффициентам при $\xi_1, \xi_2, \ldots, \xi_k$, пренебрегая отличием свободных членов α_{k+1t} и α_{k+1s}. Для этого в ковариационной матрице (7.50) следует отбросить последние строку и столбец и проверять условие (7.51) для векторов $\overset{\circ}{\boldsymbol{\alpha}}_t = \{\check{\alpha}_{1t}, \check{\alpha}_{2t}, \ldots, \check{\alpha}_{kt}\}'$ и $\overset{\circ}{\boldsymbol{\alpha}}_s = \{\check{\alpha}_{1s}, \check{\alpha}_{2s}, \ldots, \check{\alpha}_{ks}\}'$, заменив $\chi_q^2(k+1)$ на $\chi_q^2(k)$.

При анализе множественных связей может быть полезным сравнение оценок $\check{\sigma}_t^2$ и $\check{\sigma}_s^2$. Равенство $\sigma_t^2 = \sigma_s^2$ свидетельствует об одинаковой степени рассеяния значений ξ_{k+1} вокруг функции регрессии в обеих совокупностях. Критерий для проверки этой гипотезы строится точно так же, как рассматривавшиеся выше критерии для сравнения дисперсий показателей с учетом того, что $(n_t - k - 1)\check{\sigma}_t^2$ и $(n_s - k - 1)\check{\sigma}_s^2$ следуют распределению χ^2 с $n_t - k - 1$ и $n_s - k - 1$ степенями свободы.

Для иллюстрации расчета рассмотрим сравнение регрессий при $k = 1$:

$$\xi_2 = \alpha_{1t}\xi_1 + \alpha_{2t} + \Delta_t \text{ (в } Q_t), \quad \xi_2 = \alpha_{1s}\xi_1 + \alpha_{2s} + \Delta_s \text{ (в } Q_s).$$

Оценками коэффициентов будут, по (6.74),

$$\check{\alpha}_{1u} = \check{r}_u\frac{s_{2u}}{s_{1u}}, \quad \check{\alpha}_{2u} = \bar{x}_{2u} - \check{\alpha}_{1u}\bar{x}_{1u} \quad (u = t, s), \tag{7.52}$$

где s_{1u}, s_{2u}, $\bar{x}_{1u}$, $\bar{x}_{2u}$ вычисляются в виде (6.50) по значениям ξ_1 и ξ_2 в Q_t и Q_s; $\check{r}_t$ и $\check{r}_s$ вычисляются по формуле (6.49) выборочного коэффициента корреляции.

Оценки ковариационных матриц $\mathbf{B}_t$ и $\mathbf{B}_s$, по (6.75′), имеют вид

$$\check{\mathbf{B}}_u = \frac{\check{\sigma}_u^2}{n_u s_{1u}^2}\begin{pmatrix} 1 & -\bar{x}_{1u} \\ -\bar{x}_{1u} & s_{1u}^2 + \bar{x}_{1u}^2 \end{pmatrix},$$

где $\bar{x}_{1u} = \frac{1}{n_u}\sum_{i=1}^{n_u} x_{i1}^{(u)}$, $\check{\sigma}_u^2 = s_{2u}^2(1-\check{r}_u^2)$, $u = t,\ s$ ($x_{i1}^{(u)}$ считаем неслучайными величинами). Подставив найденное выражение в (7.50), получим оценку матрицы $\mathbf{B}_{ts}$:

$$\check{\mathbf{B}}_{ts} = \begin{pmatrix} \check{b}_{11} & \check{b}_{12} \\ \check{b}_{12} & \check{b}_{22} \end{pmatrix},\quad \check{b}_{11} = \frac{s_{2t}^2(1-\check{r}_t^2)}{n_t s_{1t}^2} + \frac{s_{2s}^2(1-\check{r}_s^2)}{n_s s_{1s}^2},$$

$$\check{b}_{12} = -\frac{s_{2t}^2(1-\check{r}_t^2)}{n_t s_{1t}^2}\bar{x}_{1t} - \frac{s_{2s}^2(1-\check{r}_s^2)}{n_s s_{1s}^2}\bar{x}_{1s},$$

$$\check{b}_{22} = \frac{(s_{1t}^2 + \bar{x}_{1t}^2)s_{2t}^2(1-\check{r}_t^2)}{n_t s_{1t}^2} + \frac{(s_{1s}^2 + \bar{x}_{1s}^2)s_{2s}^2(1-\check{r}_s^2)}{n_s s_{1s}^2}. \qquad (7.53)$$

Функция, используемая в решающем правиле (7.51), примет вид

$$(\check{\boldsymbol{\alpha}}_t - \check{\boldsymbol{\alpha}}_s)'\,\mathbf{B}_{ts}^{-1}(\check{\boldsymbol{\alpha}}_t - \check{\boldsymbol{\alpha}}_s) = \frac{1}{b_{11}b_{22} - b_{12}^2}[b_{22}(\check{\alpha}_{1t} - \check{\alpha}_{1s})^2 -$$

$$- 2b_{12}(\check{\alpha}_{1t} - \check{\alpha}_{1s})(\check{\alpha}_{2t} - \check{\alpha}_{2s}) + b_{11}(\check{\alpha}_{2t} - \check{\alpha}_{2s})^2].$$

Дисперсия разности $\check{\alpha}_{1t} - \check{\alpha}_{1s}$ оценок коэффициентов при ξ_1 будет равна первому диагональному элементу матрицы $\mathbf{B}_{ts}$: $\mathbf{D}(\check{\alpha}_{1t} - \check{\alpha}_{1s}) = b_{11}$. Если выполняется условие

$$(\check{\alpha}_{1t} - \check{\alpha}_{1s})^2 b_{11}^{-1} \geqslant \chi_q^2(1) \qquad (7.54)$$

или эквивалентное ему

$$\frac{|\check{\alpha}_{1t} - \check{\alpha}_{1s}|}{\sqrt{b_{11}}} \geqslant u_{\frac{1+q}{2}}; \qquad (7.54')$$

гипотеза $\alpha_{1t} = \alpha_{1s}$ отвергается с вероятностью ошибки $1 - q$.

Пример 7.8. По результатам измерений содержаний породообразующих окислов в образцах двух генетических связанных петрографических разновидностей пород A и B получены следующие данные. В разновидности A: выборочная дисперсия содержания кремнекислоты $s_{2a}^2 = \frac{1}{n_a}\sum_{t=1}^{n_a}(x_{t2}^{(a)} - \bar{x}_2^{(a)})^2 = 9\,(\%)^2$; выборочная дисперсия суммы содержаний окислов $G = Al_2O_3 + TiO_2$, $s_{1a}^2 = 2{,}25(\%)^2$; выборочный коэффициент корреляции содержаний SiO_2 и G, $\check{r}_a = -0{,}76$. В разновидности B: выборочная дисперсия содержания SiO_2, $s_{2b}^2 = 8{,}25\,(\%)^2$; выборочная дисперсия суммы G, $s_{1b}^2 = 2{,}9\,(\%)^2$; выборочный коэффициент корреляции $\check{r}_b = -0{,}88$. Число образцов в выборке А $n_a = 50$, в выборке B — $n_b = 60$. Сравнить регрессии SiO_2 на G по коэффициентам при G, проверив при уровне значимости 0,05 гипотезу об их равенстве между собой.

Решение. По формуле (7.52) для разновидности A

$$\check{\alpha}_{1a} = \check{r}_a \frac{s_{2a}}{s_{1a}} = -0{,}76 \sqrt{\frac{9}{2{,}25}} = -1{,}52;$$

для разновидности B —

$$\check{\alpha}_{1b} = \check{r}_b \frac{s_{2b}}{s_{1b}} = -0{,}88 \sqrt{\frac{8{,}25}{2{,}9}} \approx -1{,}48.$$

По (7.53) имеем

$$\check{b}_{11} = \frac{s_{2a}^2 (1 - \check{r}_a^2)}{n_a s_{1a}^2} + \frac{s_{2b}^2 (1 - \check{r}_b^2)}{n_b s_{1b}^2} = \frac{9\,(1 - 0{,}76^2)}{50 \cdot 2{,}25} + \frac{8{,}25\,(1 - 0{,}88^2)}{60 \cdot 2{,}9} \approx 0{,}0445;$$

$$\frac{|\check{\alpha}_{1a} - \check{\alpha}_{1b}|}{\sqrt{\check{b}_{11}}} = \frac{0{,}04}{0{,}211} \approx 0{,}19.$$

Так как эта величина не превышает $u_{\frac{1+q}{2}} = u_{0{,}975} = 1{,}96$ в (7.54′), данные измерений не противоречат гипотезе о равенстве коэффициентов α_{1a}, α_{1b} между собой, т. е. о тождественности собственно функциональных частей регрессий.

§ 3. Дисперсионный анализ

При обработке геолого-геофизических данных нередко возникает задача оценки влияния тех или иных факторов на распределения отдельных показателей. Рассмотрим такой случай, когда влияние факторов сказывается на математическом ожидании, приводя к его изменению, но оставляя неизменной характеристику рассеяния — дисперсию. Установление такого влияния можно отождествить с отрицательным результатом проверки гипотезы о совпадении математических ожиданий по выборкам, сформированным каждая при каком-либо фиксированном состоянии фактора или группы факторов. Случай двух таких состояний и соответственно двух выборок фактически уже рассматривался — тогда задача решается одним из критериев проверки гипотез о равенстве математических ожиданий. Так, если необходимо убедиться в отсутствии систематического смещения в результатах измерений показателя, получаемых на двух приборах, достаточно проверить гипотезу о равенстве математических ожиданий результатов его измерения в одном и том же образце на этих приборах. Однако, если нас интересует вообще влияние прибора как фактора, такой опыт может оказаться недостаточным. Для получения обоснованного вывода необходимо провести измерения при нескольких состояниях фактора, в данном случае — на нескольких приборах, и убедиться в отсутствии существенных расхождений в получаемых результатах.

Описанный пример относится к задаче однофакторного дисперсионного анализа, когда исследуется влияние одного фактора. Этот анализ мы и рассмотрим в настоящем разделе. Разработан соответствующий аппарат для двух и более факторов (двух-, трехфакторный дисперсионный анализ). Анализ влияния нескольких факторов подобен однофакторному анализу, однако содержит более сложные вычислительные

процедуры. Читатель может ознакомиться с ними по дополнительной литературе [27].

Однофакторный дисперсионный анализ. Постановка задачи предполагает наличие такой исходной информации. Последовательно фиксируется I состояний φ_i некоторого фактора φ ($i = \overline{1, I}$). При каждом состоянии имеется J_i наблюдений показателя ξ. Обозначим эти наблюдения следующим образом:

$$
\begin{aligned}
&x_{11}, x_{12}, \ldots, x_{1J_1} \text{ (при состоянии } \varphi_1);\\
&x_{21}, x_{22}, \ldots, x_{2J_2} \text{ (при состоянии } \varphi_2);\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&x_{I1}, x_{I2}, \ldots, x_{IJ_I} \text{ (при состоянии } \varphi_I).
\end{aligned}
$$

Наблюдения представляются в виде

$$x_{ij} = \beta_i + \Delta_{ij}, \tag{7.55}$$

где x_{ij} — j-е наблюдение ξ при состоянии фактора φ_i ($j = \overline{1, J_i}$, $i = \overline{1, I}$), β_i — математическое ожидание ξ при фиксированном i; Δ_{ij} — отклонение x_{ij} от своего математического ожидания при фиксированном i, представляющее собой наблюдение случайной величины с нулевым математическим ожиданием, $\mathbf{M}\Delta_{ij} = 0$.

Если фактор влияет на распределение ξ, приводя к смещению среднего, величины β_i будут отличаться друг от друга. Для того, чтобы убедиться в таком влиянии, достаточно проверить гипотезу о существовании расхождений между β_i. Решение этой задачи и составляет содержание однофакторного дисперсионного анализа.

Предположим дополнительно, что распределение ξ при каждом φ_i нормально и имеет одну и ту же дисперсию $\mathbf{D}\xi = D$, не зависящую от φ_i. Дисперсионный анализ основывается на сравнении двух оценок, которые при нулевой гипотезе ($\beta_i = \beta_k$, $i, k = \overline{1, I}$, $i \neq k$) оценивают одну и ту же величину D, но имеют различные математические ожидания, если эта гипотеза не верна.

Каковы бы ни были β_i, несмещенными оценками дисперсии D будут

$$s_i^2 = \frac{1}{J_i - 1} \sum_{j=1}^{J_i} (x_{ij} - \bar{x}_i)^2 \quad (i = \overline{1, I}), \tag{7.56}$$

где $\bar{x}_i = \frac{1}{J_i} \sum_{j=1}^{J_i} x_{ij}$ — оценка математического ожидания при состоянии φ_i. Можно составить средневзвешенную оценку из величин s_i^2:

$$s_e^2 = \frac{1}{N - I} \sum_{i=1}^{I} (J_i - 1) s_i^2 = \frac{1}{N - I} \sum_{i=1}^{I} \sum_{j=1}^{J_i} (x_{ij} - \bar{x}_i)^2, \tag{7.57}$$

где $N = \sum_{i=1}^{I} J_i$ — общее количество наблюдений. Ясно, что

$$\mathbf{M}s_e^2 = \frac{1}{N-I}\sum_{i=1}^{I}(J_i - 1)\,\mathbf{M}s_i^2 = \frac{D}{N-I}\sum_{i=1}^{I}(J_i - 1) = D,$$

так что s_e^2 — несмещенная оценка D независимо от β_i. Рассмотрим теперь статистику

$$s_H^2 = \frac{1}{I-1}\sum_{i=1}^{I} J_i(\bar{x}_i - \bar{x})^2, \tag{7.58}$$

где $\bar{x} = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{J_i} x_{ij} = \frac{1}{N}\sum_{i=1}^{I} J_i\bar{x}_i$ — среднее арифметическое, составленное из всех наблюдений. В соответствии с (7.55),

$$\bar{x}_i = \beta_i + \frac{1}{J_i}\sum_{j=1}^{J_i}\Delta_{ij} = \beta_i + \bar{\Delta}_i, \tag{7.59}$$

причем математическое ожидание $\bar{\Delta}_i$ — среднего арифметического из величин Δ_{ij} при фиксированном i — равно нулю, а дисперсия

$$\mathbf{D}\bar{\Delta}_i = \mathbf{M}\bar{\Delta}_i^2 = \frac{1}{J_i}D. \tag{7.60}$$

Обозначим $\tilde{\beta} = \frac{1}{N}\sum_{i=1}^{I} J_i\beta_i$, $\tilde{\Delta} = \frac{1}{N}\sum_{i=1}^{I} J_i\bar{\Delta}_i$; тогда $\bar{x} = \tilde{\beta} + \tilde{\Delta}$. Вычислим математическое ожидание s_H^2:

$$\mathbf{M}s_H^2 = \mathbf{M}\left[\frac{1}{I-1}\sum_{i=1}^{I} J_i(\beta_i + \bar{\Delta}_i - \tilde{\beta} - \tilde{\Delta})^2\right] = \frac{1}{I-1}\mathbf{M}\left\{\sum_{i=1}^{I} J_i[(\beta_i - \tilde{\beta})^2 +\right.$$

$$\left. + 2(\beta_i - \tilde{\beta})(\bar{\Delta}_i - \tilde{\Delta}) + (\bar{\Delta}_i - \tilde{\Delta})^2]\right\} = \frac{1}{I-1}\sum_{i=1}^{I} J_i(\beta_i - \tilde{\beta})^2 +$$

$$+ \frac{1}{I-1}\mathbf{M}\sum_{i=1}^{I} J_i(\bar{\Delta}_i - \tilde{\Delta})^2.$$

Так как

$$\frac{1}{I-1}\mathbf{M}\sum_{i=1}^{I} J_i(\bar{\Delta}_i - \tilde{\Delta})^2 = \frac{1}{I-1}\mathbf{M}\sum_{i=1}^{I}\left[J_i\bar{\Delta}_i^2 - 2\bar{\Delta}_i J_i\frac{1}{N}\sum_{i=1}^{I} J_i\bar{\Delta}_i +\right.$$

$$+ J_i \frac{1}{N^2}\left(\sum_{i=1}^{I} J_i \bar{\Delta}_i\right)^2\Big] = \frac{1}{I-1}\left(DI - 2\sum_{i=1}^{I} DJ_i \frac{1}{N} + \frac{1}{N}\sum_{i=1}^{I} J_i D\right) =$$

$$= \frac{1}{I-1}(DI - 2D + D) = D,$$

получим

$$\mathbf{M}s_H^2 = \frac{1}{I-1}\sum_{i=1}^{I} J_i(\beta_i - \bar{\beta})^2 + D. \tag{7.61}$$

Таким образом, если различий между β_i нет, то статистика s_H^2 также будет несмещенной оценкой D, в противном случае она в среднем будет превышать D. На этом свойстве и основано решающее правило дисперсионного анализа. Гипотеза об отсутствии влияния фактора отвергается с вероятностью ошибки $\alpha = 1 - q$, если выполняется условие

$$\frac{s_H^2}{s_e^2} \geqslant F_q(I-1,\ N-I), \tag{7.62}$$

где $F_q(I-1,\ N-I)$ — квантиль порядка q распределения Фишера с $I-1$ и $N-I$ степенями свободы, который определяется из табл. 10, 11 (Приложение).

Количественная характеристика влияния фактора. В тех случаях, когда условие (7.62) выполняется, свидетельствуя о влиянии фактора, нередко возникает необходимость в количественной характеристике этого влияния. Если вариация состояний фактора имеет случайный характер, такой характеристикой, помимо отношения $\frac{s_H^2}{s_e^2}$, может служить дисперсия D_β значений β_i, которая как раз и обусловлена влиянием фактора. Оценкой этой дисперсии служит

$$\bar{D}_\beta = \frac{1}{I-1}\sum_{i=1}^{I}(\beta_i - \bar{\beta})^2, \quad \text{где } \bar{\beta} = \frac{1}{I}\sum_{i=1}^{I}\beta_i.$$

Для большей общности будем теперь считать дисперсии величины ξ при различных фиксированных состояниях фактора неодинаковыми: $\mathbf{D}x_{ij} = \mathbf{D}\Delta_{ij} = D_i$, $j = \overline{1, J_i}$, $i = \overline{1, I}$. Имея в виду получение оценки D_β по наблюдениям x_{ij}, рассмотрим величину

$$S^2 = \frac{1}{I-1}\sum_{i=1}^{I}(\bar{x}_i - \bar{x}_0)^2, \quad \text{где } \bar{x}_0 = \frac{1}{I}\sum_{i=1}^{I}\bar{x}_i. \tag{7.63}$$

Аналогично (7.60), $\mathbf{D}\bar{\Delta}_i = \frac{D_i}{J_i}$. Положим $\bar{\Delta} = \frac{1}{I}\sum_{i=1}^{I}\bar{\Delta}_i$. Математическое ожидание S^2 при фиксированных β_i

$$\mathbf{M}_\beta S^2 = \mathbf{M}_\beta\left[\frac{1}{I-1}\sum_{i=1}^{I}(\beta_i + \bar{\Delta}_i - \bar{\beta} - \bar{\Delta})^2\right] = \frac{1}{I-1}\mathbf{M}_\beta\left\{\sum_{i=1}^{I}[(\beta_i - \bar{\beta})^2 + \right.$$

$$\left. + (\bar{\Delta}_i - \bar{\Delta})^2]\right\} = \frac{1}{I-1}\sum_{i=1}^{I}(\beta_i - \bar{\beta})^2 + \mathbf{M}\left[\frac{1}{I-1}\sum_{i=1}^{I}(\bar{\Delta}_i - \bar{\Delta})^2\right].$$

Второе слагаемое

$$\mathbf{M}\left[\frac{1}{I-1}\left(\sum_{i=1}^{I}\bar{\Delta}_i^2 - I\bar{\Delta}^2\right)\right] = \frac{1}{I-1}\left(\sum_{i=1}^{I}\frac{D_i}{J_i} - I\mathbf{M}\bar{\Delta}^2\right) =$$

$$= \frac{1}{I-1}\left(\sum_{i=1}^{I}\frac{D_i}{J_i} - \frac{1}{I}\sum_{i=1}^{I}\frac{D_i}{J_i}\right) = \frac{1}{I}\sum_{i=1}^{I}\frac{D_i}{J_i}.$$

Таким образом, $\mathbf{M}S^2 = D_\beta + \frac{1}{I}\sum_{i=1}^{I}\frac{D_i}{J_i}$, и поскольку s_i^2 (7.56) — несмещенная оценка D_i, в качестве оценки D_β можно взять

$$\check{D}_\beta = \begin{cases} S^2 - \frac{1}{I}\sum_{i=1}^{I}\frac{s_i^2}{J_i}, & \text{если } S^2 > \frac{1}{I}\sum_{i=1}^{I}\frac{s_i^2}{J_i}, \\ 0, & \text{если } S^2 \leqslant \frac{1}{I}\sum_{i=1}^{I}\frac{s_i^2}{J_i}. \end{cases} \tag{7.64}$$

Величина $\check{D}_\beta$ оценивает часть общей дисперсии колебаний ξ, обусловленную лишь влиянием фактора, с которым связана вариация значений β_i. Оценку средней величины общей дисперсии можно определить в виде

$$\check{D}_0 = \check{D}_\beta + \frac{1}{I}\sum_{i=1}^{I}s_i^2. \tag{7.65}$$

Таким образом, отношение

$$f = \frac{\check{D}_\beta}{\check{D}_0}, \tag{7.66}$$

принимающее значения из отрезка [0, 1] ($0 \leqslant f \leqslant 1$), служит относительной характеристикой степени влияния рассматриваемого фактора: чем оно ближе к единице, тем сильнее влияние последнего. В примере с измерительными приборами D_β получает простую интерпрета-

цию «межприборной» дисперсии. Величина $\frac{1}{I}\sum_{i=1} s_i^2$ является оценкой средней дисперсии воспроизводимости отдельного прибора. Так же можно анализировать работу различных лабораторий, влияние тех или иных параметров режима измерений; сопоставлять выборки образцов горных пород, составленные на основе определенных геологических предпосылок. В последнем случае результат дисперсионного анализа может в известной мере служить подтверждением или опровержением этих предпосылок.

Отклонения распределений показателя ξ от нормального, как показывает практика, не столь сильно искажают результат описанного анализа, как это можно было бы ожидать, если только выборки не содержат резко выделяющихся наблюдений. Тем не менее иногда полезно несколько изменить метод расчета. Так, если при каждом состоянии фактора наблюдения подчиняются логнормальному закону распределения (например, ошибки спектральных определений концентраций химических элементов), целесообразно оперировать не самими наблюдениями, а их логарифмами.

Отметим еще, что наиболее рациональное распределение общего количества наблюдений N — поровну между различными состояниями фактора: $J_i = J$, $i = \overline{1, I}$. При этом упрощается и расчет по формуле (7.64). Увеличение числа I состояний фактора при сохранении J дает возможность более точно оценивать D_β, а увеличение J при сохранении I — дисперсию показателя при фиксированном состоянии фактора.

Пример 7.9. Для проверки отсутствия систематического смещения во времени в результатах спектральных определений химического элемента в лаборатории проведен следующий опыт. Одна и та же контрольная проба измерялась четырьмя сериями по 25 измерений, причем серии отделены друг от друга продолжительными промежутками времени. Ввиду логарифмически нормального закона распределения ошибок измерений результаты прологарифмированы — $x_{ij} = \lg z_{ij}$ и по ним вычислены необходимые данные для дисперсионного анализа: средние арифметические и оценки дисперсии по каждой серии вида (7.56): $\bar{x}_1 = -1{,}7$, $\bar{x}_2 = -1{,}4$, $\bar{x}_3 = -2$, $\bar{x}_4 = -1{,}5$; $s_1^2 = 0{,}04$, $s_2^2 = 0{,}05$, $s_3^2 = 0{,}03$, $s_4^2 = 0{,}01$. Проверить гипотезу об отсутствии смещения во времени методом дисперсионного анализа при уровне значимости 0,05.

Решение. Имеем $I = 4$, $J_i = J = 25$, $N = JI = 100$. По формулам (7.57) и (7.58)

$$s_e^2 = \frac{24}{96}\sum_{i=1}^{4} s_i^2 = \frac{24}{96}(0{,}04 + 0{,}05 + 0{,}03 + 0{,}01) \approx 0{,}033; \quad \bar{x} = \frac{1}{4}\sum_{i=1}^{4} \bar{x}_i =$$

$$= \frac{1}{4}(-1{,}7 - 1{,}4 - 2 - 1{,}5) = -1{,}65; \quad s_H^2 = \frac{25}{3}\sum_{i=1}^{4} (\bar{x}_i - \bar{x})^2 =$$

$$= \frac{25}{3}(0{,}05^2 + 0{,}25^2 + 0{,}35^2 + 0{,}15^2) = 1{,}75.$$

Так как отношение $\frac{s_H^2}{s_i^2} \approx 53$ превышает критическую границу 5%-го уровня значимости $F_{0,95}(3,96) \approx 2,71$ (табл. 10, Приложение), гипотеза об отсутствии сдвига во времени отвергается с вероятностью ошибки 0,05.

Пример 7.10. На исследуемом объекте предполагается переменная функциональная составляющая поля значений намагниченности насыщения, проявляющаяся на фоне случайных колебаний. Для проверки этой гипотезы отобрано по 12 проб с 10 различных участков, удаленных друг от друга; в каждой измерена намагниченность насыщения I_s и результаты прологарифмированы: $x_{ij} = \lg I_{sij}$ (с учетом близости к логнормальному закону распределений I_s на участках). По этим данным вычислены средние $\bar{x}_i$ и выборочные дисперсии (7.56), результаты вычислений приведены ниже.

Номер участка (i)	1	2	3	4	5	6	7	8	9	10
$\bar{x}_i = \frac{1}{12}\sum_{j=1}^{12} x_{ij}$	2,51	3,09	2,68	1,16	2,00	3,55	3,01	2,19	2,08	1,41
$s_i^2 = \frac{1}{11}\sum_{j=1}^{12}(x_{ij} - \bar{x}_i)^2$	0,25	0,37	0,40	0,23	0,21	0,60	0,49	0,20	0,12	0,18

1) Проверить гипотезу об отсутствии вариации функциональной составляющей I_s при уровне значимости 0,05. 2) Вычислить характеристику f (7.66), выражающую степень проявления этой составляющей на фоне случайных колебаний I_s.

Решение. В этом примере $I = 10$, $J_i = J = 12$; $N = IJ = 120$.

1) По (7.57) и (7.58) имеем $s_e^2 = \sum_{i=1}^{10} 11\frac{s_i^2}{110} = \frac{1}{10}\,3,05 = 0,305$; $\bar{x} = \frac{1}{120}\sum_{i=1}^{10} 12\bar{x}_i = \frac{1}{10}\sum_{i=1}^{10}\bar{x}_i \approx 2,37$; $s_H^2 = \frac{1}{9}\sum_{i=1}^{10} 12(\bar{x}_i - \bar{x})^2 \approx 6,06$. По табл. 10 (Приложение) находим критическую границу $F_{0,95}(9, 110) \approx 1,96$. Так как $\frac{s_H^2}{s_e^2} \approx 19,9 > 1,96$, гипотеза о постоянстве функциональной составляющей отвергается с вероятностью ошибки 0,05.

2) Для вычисления f найдем S^2 по формуле (7.63):

$$S^2 = \frac{1}{9}\sum_{i=1}^{10}(\bar{x}_i - \bar{x})^2 \approx 0,505.\quad \frac{1}{I}\sum_{i=1}^{I}\frac{s_i^2}{J} = \frac{3,05}{120} \approx 0,0254.$$

По (7.64) и (7.65) $\check{D}_\beta = 0,505 - 0,0254 \approx 0,480$; $\check{D}_0 = \check{D}_\beta + \frac{1}{10}\sum_{i=1}^{10} s_i^2 \approx 0,785$. Коэффициент $f = \frac{0,480}{0,785} \approx 0,611$. Таким образом, бо́льшую часть $\check{D}_0$ — оценки общей дисперсии $\lg I_s$ — составляет $\check{D}_\beta$. Средняя величина выборочных «локальных» диспер-

сий, вычисленных по отдельным участкам $\frac{1}{I}\sum_{i=1}^{I} s_i^2$, составляет часть $\check{D}_0$, равную $\check{D}_0(1-f) = 0{,}389\check{D}_0$.

Пример 7.9 иллюстрирует нетрадиционную для дисперсионного анализа постановку задачи. Аналогично можно решать и другие подобные задачи: анализировать распределения химических элементов и характеристик физических свойств в различных направлениях на поверхности и по глубине; проводить сопоставление различных показателей по «контрастности» их как признаков тех или иных особенностей, например по степени изменчивости, обусловленной той же функциональной составляющей, с помощью коэффициента f. В практике геологических исследований достаточно часто встречаются задачи, для решения которых с успехом можно применять дисперсионный анализ.

Критерии Бартлетта и Кочрена. Решающее правило дисперсионного анализа применяется при условии, что анализируемый показатель имеет одну и ту же дисперсию при всех состояниях фактора. Такое предположение иногда нуждается в проверке, которую можно произвести с помощью *критерия Бартлетта* или *критерия Кочрена*.

Пусть имеется I оценок s_i^2 дисперсий D_i вида (7.56), вычисленных по I выборкам независимых наблюдений показателя ξ, причем i-я выборка содержит J_i наблюдений ($J_i > 4$, $i = \overline{1, I}$); распределения генеральных совокупностей, соответствующих выборкам, при нулевой гипотезе нормальны.

Критерий Бартлетта основан на использовании статистики

$$M = N_0 \ln\left(\frac{1}{N_0}\sum_{i=1}^{I}(J_i - 1)s_i^2\right) - \sum_{i=1}^{I}(J_i - 1)\ln s_i^2 \quad \left(N_0 = \sum_{i=1}^{I} J_i - I\right). \tag{7.67}$$

Если нулевая гипотеза верна, т. е. $D_1 = D_2 = \cdots = D_I$, то

$$m = \frac{M}{1 + \frac{1}{3(I-1)}\left(\sum_{i=1}^{I}\frac{1}{J_i - 1} - \frac{1}{N_0}\right)} \tag{7.68}$$

приближенно следует распределению χ^2 с $I - 1$ степенями свободы. Поэтому по критерию Бартлетта нулевая гипотеза отвергается с вероятностью ошибки $\alpha \approx 1 - q$, если

$$m \geqslant \chi_q^2(I-1). \tag{7.69}$$

Критерий Кочрена, более простой по сравнению с критерием Бартлетта, но несколько менее мощный, можно употреблять в тех случаях, когда все J_i одинаковы: $J_i = J$ ($i = \overline{1, I}$). Этот критерий основан на использовании статистики

$$G = \frac{s_{\max}^2}{s_1^2 + s_2^2 + \cdots + s_I^2} \quad \left(s_{\max}^2 = \max_i s_i^2\right). \tag{7.70}$$

Нулевая гипотеза $D_1 = D_2 = \cdots = D_l$ отвергается с вероятностью ошибки, не превышающей α, если G больше или равно критической границе $g(\alpha, l, J-1)$, определяемой из табл. 14 (Приложение). При использовании обоих критериев необходимо учитывать, что на результат их применения существенно влияют отклонения распределений ξ от нормального закона.

Пример 7.11. По данным примера 7.9 проверить гипотезу о равенстве дисперсий воспроизводимости по оценкам s_i^2 с помощью критерия Кочрена при уровне значимости $\alpha = 0{,}05$.

Решение. По формуле (7.70) $G = \dfrac{0{,}05}{0{,}04 + 0{,}05 + 0{,}03 + 0{,}01} \approx 0{,}385$. По табл. 14 (Приложение) критическая граница для G, $g(0{,}05,\ 4,\ 24) = 0{,}411$. Так как $G < 0{,}411$, различия оценок s_i^2 незначимы.

§ 4. Методы статистической классификации

Применение вероятностно-статистических методов для решения задач классификации представляет собой одно из наиболее перспективных направлений внедрения математических методов в геологические и геофизические исследования. В настоящее время это направление находит все более широкую область применения и является одним из наиболее разработанных. Этому способствует, в первую очередь, специфика геологических исследований — в форме задачи классификации формулируется широкий круг вопросов этих исследований. Примерами задач классификации могут служить: классификация образцов пород по принадлежности их к тем или иным разновидностям на основе химических и физических показателей; классификация пластов по характеру их насыщения на основе промыслово-геофизических и геологических данных; распознавание рудовмещающих и безрудных объектов на основе комплекса их количественных характеристик; разделение пород по их фациальной и формационной принадлежности и т. д. Рассматриваемую проблему нередко именуют задачей *распознавания образов*, заимствуя терминологию из кибернетики. Учитывая общую вероятностно-статистическую базу изучаемых методов, мы будем придерживаться терминологии, принятой в математической статистике. По этой же причине лишь вкратце будут рассмотрены основные принципы построения многочисленных эмпирических алгоритмов решения указанной задачи.

С основными теоретическими принципами мы уже ознакомились в гл. 5. Напомним общую постановку задачи. Имеется t классов $Q_1, Q_2, \ldots, Q_t$ и группа наблюдений $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N$ m-мерной случайной величины $\boldsymbol{\xi} = \{\xi_1, \xi_2, \ldots, \xi_m\}'$. Каждое m-мерное наблюдение $\mathbf{y}$ может принадлежать одному из классов $Q_1, Q_2, \ldots, Q_t$. Задача заключается в определении принадлежности наблюдений $\mathbf{y}_i$. Как упоминалось в гл. 5, в случае известных распределений $\boldsymbol{\xi}$ в каждом Q_l задача решается с помощью критерия Бейеса: для каждого $\mathbf{y}$ вычисляются вероятности принадлежности (5.22) или (5.23) и определяется максимальная из них.

В дальнейшем при изучении методов классификации мы и будем придерживаться бейесовской трактовки задачи классификации. При этом мы руководствуемся свойствами критериев Бейеса — прежде всего, оптимальностью их среди прочих, как об этом упоминалось в гл. 5. Кроме того, правила, построенные по бейесовскому принципу, позволяют сразу оценивать достоверность классификации в виде вероятностей принадлежности (5.22):

$$\mathbf{P}\{\mathbf{y} \in Q_j\} = \frac{q_j p_j(\mathbf{y})}{\sum_{i=1}^{t} q_i p_i(\mathbf{y})} \quad (j = \overline{1, t}), \tag{7.71}$$

где q_j — априорные вероятности, $p_j(\mathbf{x})$ — плотности распределения ξ в Q_j.

В практике применения методов классификации нередко возникает ситуация, когда несколько классов, $Q_{j_1}, Q_{j_2}, \ldots, Q_{j_k}$, настолько близки по распределениям ξ, что достоверное отнесение к одному из группы таких близких классов заведомо невозможно. При этом, если анализируемый вектор $\mathbf{y}$ действительно принадлежит к такой группе, каждая из вероятностей $\mathbf{P}\{\mathbf{y} \in Q_{j_i}\}$ $(i = \overline{1, k})$ может быть в отдельности невелика, но в сумме они составляют настолько большую величину, что практически достоверно мы можем отнести его к группе классов Q_{j_i}. Таким образом, критерий Бейеса позволяет классифицировать как по отдельным классам, так и по их различным объединениям, причем на основании одних и тех же данных — вероятностей (7.71) суммированием последних с указанием достоверности отнесения при таких объединениях:

$$\mathbf{P}\{\mathbf{y} \in Q_{j_1} \cup Q_{j_2} \cup \ldots \cup Q_{j_k}\} = \sum_{i=1}^{k} \mathbf{P}\{\mathbf{y} \in Q_{j_i}\}. \tag{7.72}$$

На практике, однако, применить критерий Бейеса непосредственно в форме (7.71) обычно не удается, так как точные выражения плотностей распределения $p_j(\mathbf{x})$ неизвестны. Как правило, можно указать лишь выборки $\tilde{Q}_j$ наблюдений ξ, заведомо принадлежащих Q_j: $\tilde{Q}_j = \{\mathbf{x}_1^{(j)}, \mathbf{x}_2^{(j)}, \ldots, \mathbf{x}_{n_j}^{(j)}\}$, $\mathbf{x}_i^{(j)} \in Q_j$, $i = \overline{1, n_j}$, $j = \overline{1, t}$ (n_j—количество таких наблюдений из $\tilde{Q}_j$). Каждый вектор $\mathbf{x}_i^{(j)}$ составлен из наблюдений компонент $\xi_1, \xi_2, \ldots, \xi_m$: $\mathbf{x}_i^{(j)} = \{x_{i1}^{(j)}, x_{i2}^{(j)}, \ldots, x_{im}^{(j)}\}'$. Если пользоваться геометрической интерпретацией, то наблюдения $\mathbf{x}_i^{(j)}$ можно изобразить в виде точек m-мерного пространства, причем классам Q_j будут соответствовать определенные, возможно, пересекающиеся, области этого пространства (рис. 40). Точки выборок $\tilde{Q}_j$ сгущаются в местах, соответствующих максимальным значениям неизвестных плотностей распределения $p_j(\mathbf{x})$, определяя наиболее характерные для классов значения ξ.

Наблюдения $\mathbf{x}_i^{(j)}$ предстоит использовать как своего рода эталоны, на основании которых и строится критерий для определения класса,

к которому принадлежит $\mathbf{y}$. Отсутствие точных выражений плотностей $p_j(\mathbf{x})$ усложняет построение критерия, обладающего гарантированными оптимальными свойствами, предоставляя, впрочем, широкий простор для работы над различными эмпирическими правилами. В лучшем случае можно указать лишь параметрические семейства плотностей распределения ξ в классах Q_j:

$$p_j(x_1, x_2, \ldots, x_m; \alpha_{1j}, \alpha_{2j}, \ldots, \alpha_{k_j j}) = p_j(\mathbf{x}, \boldsymbol{\alpha}_j) \quad (j = \overline{1, t}). \tag{7.73}$$

Если имеется возможность достаточно точно оценить параметры α_{1j}, α_{2j}, $\ldots$, $\alpha_{k_j j}$, задача допускает решение в виде параметрической классификации, основанной на использовании выражений плотностей (7.73). Рассмотрим способы такого решения.

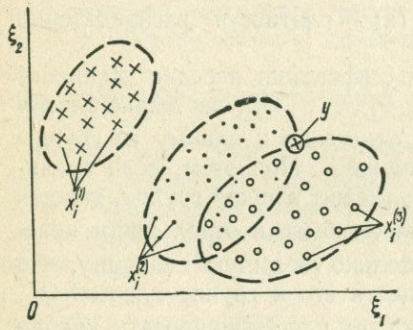


Рис. 40

Параметрическая классификация. Если $\{\check{\alpha}_{1j}, \check{\alpha}_{2j}, \ldots, \check{\alpha}_{k_j j}\} = \check{\boldsymbol{\alpha}}_j$ — оценки параметров $\{\alpha_{1j}, \alpha_{2j}, \ldots, \alpha_{k_j}\} = \boldsymbol{\alpha}_j$, $q_1, q_2, \ldots, q_t$ — априорные вероятности принадлежности $\mathbf{y} = \{y_1, y_2, \ldots, y_m\}'$ классам $Q_1, Q_2, \ldots, Q_t$ соответственно, то в качестве оценок апостериорных вероятностей можно взять, по (7.71),

$$\check{\mathbf{P}}\{\mathbf{y} \in Q_j\} = \frac{q_j p_j(\mathbf{y}, \check{\boldsymbol{\alpha}}_j)}{\sum_{j=1}^{t} q_j p_j(\mathbf{y}, \check{\boldsymbol{\alpha}}_j)} \quad (j = \overline{1, t}). \tag{7.74}$$

В дальнейшем мы будем рассматривать случай, когда априорной информации нет, т. е. будем считать q_j одинаковыми и равными по $\frac{1}{t}$, так как на практике обычно имеют дело именно с такими условиями. Приводимые ниже правила можно легко трансформировать для случая неодинаковых q_j. Таким образом, при $q_j = \frac{1}{t}$ $(j = \overline{1, t})$

$$P_j(\mathbf{y}) = \mathbf{P}\{\mathbf{y} \in Q_j\} \approx \frac{p_j(\mathbf{y}, \check{\boldsymbol{\alpha}}_j)}{\sum_{j=1}^{t} p_j(\mathbf{y}, \check{\boldsymbol{\alpha}}_j)}. \tag{7.74'}$$

Пусть, например, распределения $\xi = \{\xi_1, \xi_2, \ldots, \xi_m\}'$ нормальны, т. е. плотности (7.73) имеют вид (3.53):

$$p_j(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{m}{2}} |\mathbf{B}_j|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{x} - \mathbf{m}_j)' \mathbf{B}_j^{-1} (\mathbf{x} - \mathbf{m}_j)\right], \tag{7.75}$$

где $\mathbf{m}_j$ — математическое ожидание ξ в Q_j, $\mathbf{m}_j = \{M_1^{(j)}, M_2^{(j)}, \ldots,$

$M_m^{(j)}\}'$; $\mathbf{B}_j$ — ковариационная матрица; $|\mathbf{B}_j|$ — ее определитель. Оценки математических ожиданий и ковариационных матриц имеют вид:

$$\check{\mathbf{m}}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} \mathbf{x}_i^{(j)} - \tag{7.76}$$

вектор средних арифметических, составленных по наблюдениям из $\tilde{Q}_j$,

$$\check{\mathbf{m}}_j = \{\check{M}_1^{(j)}, \check{M}_2^{(j)}, \ldots, \check{M}_m^{(j)}\}', \quad \check{M}_k^{(j)} = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ik}^{(j)} \quad (k = \overline{1, m});$$

$$\check{\mathbf{B}}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} (\mathbf{x}_i^{(j)} - \check{\mathbf{m}}_j)(\mathbf{x}_i^{(j)} - \check{\mathbf{m}}_j)' \text{ — матрица выборочных ковариаций,}$$

$$\check{\mathbf{B}}_j = \{\check{b}_{gs}^{(j)}\}_{g,\ s=1}^{m}, \quad \check{b}_{gs}^{(j)} = \frac{1}{n_j}\sum_{i=1}^{n_j} (x_{ig}^{(j)} - \check{M}_g^{(j)})(x_{is}^{(j)} - \check{M}_s^{(j)}). \tag{7.77}$$

Подставив эти оценки в (7.75), получим оценки плотностей распределения $p_j(\mathbf{x})$. По (7.74′) оценки апостериорных вероятностей

$$\check{P}_j(\mathbf{y}) = \frac{|\check{\mathbf{B}}_j|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{y} - \check{\mathbf{m}}_j)'\,\check{\mathbf{B}}_j^{-1}(\mathbf{y} - \check{\mathbf{m}}_j)\right]}{\sum_{j=1}^{t} |\check{\mathbf{B}}_j|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{y} - \check{\mathbf{m}}_j)'\,\check{\mathbf{B}}_j^{-1}(\mathbf{y} - \check{\mathbf{m}}_j)\right]} \quad (j = \overline{1, t}). \tag{7.78}$$

Надо сказать, что этот способ громоздкий, а главное, требует значительного числа наблюдений, особенно при большой размерности m из-за необходимости получения устойчивых оценок обратных матриц $\check{\mathbf{B}}_j^{-1}$. Поэтому обычно приходится делать правдоподобные допущения о тех или иных особенностях распределений, позволяющие облегчить решение и обеспечить устойчивость.

Пусть, например, компоненты $\xi_1, \xi_2, \ldots, \xi_m$ взаимно независимы или слабо зависимы во всех Q_j. Это предположение обычно проверяют по выборочным коэффициентам корреляции соответствующими критериями. Тогда плотности совместных распределений $\xi_1, \xi_2, \ldots, \xi_m$ в каждом классе Q_j представляются в виде произведений одномерных плотностей ξ_l $(l = \overline{1, m})$:

$$p_j(\mathbf{x}) = p_j(x_1, x_2, \ldots, x_m) = p_1^{(j)}(x_1)\, p_2^{(j)}(x_2) \ldots p_m^{(j)}(x_m) \tag{7.79}$$

($p_l^{(j)}(x_l)$ — плотность распределения ξ_l в Q_j). При нормальном распределении ξ_l

$$p_l^{(j)}(x_l) = \frac{1}{\sqrt{2\pi}\,\sigma_l^{(j)}} \exp\left[-\frac{1}{2}\,\frac{(x_l - M_l^{(j)})^2}{(\sigma_l^{(j)})^2}\right], \tag{7.80}$$

где $M_l^{(j)}$ и $\sigma_l^{(j)}$ — математическое ожидание и среднее квадратическое отклонение ξ_l в Q_j. Вычислив их оценки по наблюдениям $x_{il}^{(j)}$ из выборок $\tilde{Q}_j$ — $\check{M}_l^{(j)}$ по (7.76) и $\check{\sigma}_l^{(j)} = \sqrt{\check{b}_{ll}^{(j)}}$ по (7.77),— получим возможность оценить одномерные плотности (7.80). Таким образом, при независимых ξ_l оценками вероятностей принадлежности (7.74′) будут

$$\check{\mathbf{P}}\{\mathbf{y} \in Q_j\} = \check{P}_j(\mathbf{y}) = \frac{\prod_{l=1}^{m} \check{p}_l^{(j)}(y_l)}{\sum_{j=1}^{t} \prod_{l=1}^{m} \check{p}_l^{(j)}(y_l)} \quad (j = \overline{1, t}), \tag{7.81}$$

где $\check{p}_l^{(j)}(y_l)$ определяются по (7.80) при $M_l^{(j)} = \check{M}_l^{(j)}$ и $\sigma_l^{(j)} = \check{\sigma}_l^{(j)}$.

В частности, если $t = 2$ (задача *дихотомии*, т. е. классификации на два класса), правило максимума $P_j(\mathbf{y})$ эквивалентно правилу (5.21) с использованием отношения плотностей распределения. Если

$$\frac{p_1(\mathbf{y})}{p_2(\mathbf{y})} = \prod_{l=1}^{m} p_l^{(1)}(y_l) \left[\prod_{l=1}^{m} p_l^{(2)}(y_l)\right]^{-1} > 1, \tag{7.82}$$

принимается гипотеза о принадлежности $\mathbf{y}$ классу Q_1. Взяв логарифм от обеих частей неравенства, получим, с учетом (7.80), эквивалентное условие: если

$$U = -\sum_{l=1}^{m} \ln \sigma_l^{(1)} - \frac{1}{2} \sum_{l=1}^{m} \frac{(y_l - M_l^{(1)})^2}{(\sigma_l^{(1)})^2} + \sum_{l=1}^{m} \ln \sigma_l^{(2)} + \frac{1}{2} \sum_{l=1}^{m} \frac{(y_l - M_l^{(2)})^2}{(\sigma_l^{(2)})^2} > 0, \tag{7.83}$$

принимается гипотеза о принадлежности $\mathbf{y}$ классу Q_1. В случае выполнения противоположного неравенства более вероятен класс Q_2.

Рассмотрим другую возможность: ξ_l зависимы, но ковариационные матрицы одинаковы, $\mathbf{B}_1 = \mathbf{B}_2 = \ldots = \mathbf{B}_m = \mathbf{B}$ или мало отличаются друг от друга. В этом случае можно вместо оценок матриц $\mathbf{B}_j$, участвующих в (7.78), использовать одну оценку $\check{\mathbf{B}}$, вычислив ее как средневзвешенную из $\check{\mathbf{B}}_j$:

$$\check{\mathbf{B}} = \frac{1}{N} \sum_{j=1}^{t} n_j \check{\mathbf{B}}_j \quad \left(N = \sum_{j=1}^{t} n_j\right). \tag{7.84}$$

Оценки вероятностей (7.78) запишутся в виде

$$\check{P}_j(\mathbf{y}) = \frac{\exp\left[-\frac{1}{2}(\mathbf{y} - \check{\mathbf{m}}_j)' \check{\mathbf{B}}^{-1} (\mathbf{y} - \check{\mathbf{m}}_j)\right]}{\sum_{j=1}^{t} \exp\left[-\frac{1}{2}(\mathbf{y} - \check{\mathbf{m}}_j)' \check{\mathbf{B}}^{-1} (\mathbf{y} - \check{\mathbf{m}}_j)\right]} \quad (j = \overline{1, t}). \tag{7.85}$$

В частности, решение задачи дихотомии, аналогично (7.82), будет заключаться в проверке условия

$$\frac{\check{p}_1(y)}{\check{p}_2(y)} = \frac{\exp\left[-\frac{1}{2}(y-\check{m}_1)'\check{B}^{-1}(y-\check{m}_1)\right]}{\exp\left[-\frac{1}{2}(y-\check{m}_2)'\check{B}^{-1}(y-\check{m}_2)\right]} > 1. \qquad (7.86)$$

При его выполнении наблюдение y классифицируется как принадлежащее Q_1, при выполнении противоположного неравенства его относят в Q_2. Взяв логарифм от обеих частей (7.86), получим эквивалентное условие:

$$-\frac{1}{2}(y-\check{m}_1)'\check{B}^{-1}(y-\check{m}_1)+\frac{1}{2}(y-\check{m}_2)'\check{B}^{-1}(y-\check{m}_2)>0, \qquad (7.86')$$

которое приводится к виду

$$\left(y-\frac{\check{m}_1+\check{m}_2}{2}\right)'\check{B}^{-1}(\check{m}_1-\check{m}_2)>0. \qquad (7.86'')$$

Выражение $y'B^{-1}(m_1-m_2)$, оценка которого участвует в (7.86″), называется *дискриминантной функцией*, а метод дихотомии по этому правилу — методом дискриминантных функций.

Так как y входит в (7.86″) линейно, уравнение

$$\left(y-\frac{1}{2}(\check{m}_1+\check{m}_2)\right)'\check{B}^{-1}(\check{m}_1-\check{m}_2)=0 \qquad (7.87)$$

описывает гиперплоскость в m-мерном пространстве, разделяющую его на области принятия гипотез $\{y\in Q_1\}$ и $\{y\in Q_2\}$. Математическое ожидание величины $\left(y-\frac{1}{2}(m_1+m_2)\right)'B^{-1}(m_1-m_2)$ составит $\frac{1}{2}(m_1-m_2)'B^{-}(m_1-m_2)=\frac{1}{2}d$, если $y\in Q_1$ и $\left(-\frac{1}{2}d\right)$, если $y\in Q_2$; дисперсия в обоих случаях равна d. Отсюда следует, что с увеличением d классификация производится с большей достоверностью. Величина d, именуемая *обобщенным расстоянием* (по Махаланобису), служит естественной мерой различия распределений в Q_1 и Q_2 при равенстве ковариационных матриц.

Описанный метод можно применить и для классификации на несколько классов $Q_1, Q_2, \ldots, Q_t$ $(t>2)$, определив области принятия гипотез $\{y\in Q_j\}$ $(j=\overline{1,t})$ по классификационному правилу (7.86″) для всех возможных пар классов Q_i, Q_j. Метод дискриминантных функций хорошо зарекомендовал себя, особенно при решении задач классификации пород по петрохимическим данным.

Тем не менее следует предостеречь от возможных ошибок при использовании этого метода, которые могут быть связаны с несоблюдением условий его применимости. Особенно это касается условия равенства ковариационных матриц. Различия между ними приводят к увеличению числа ошибок, а можно привести примеры, когда этим методом задача вообще не решается. На рис. 41 изображен такой пример взаимного расположения областей значений двумерной вели-

чины ξ для двух классов, Q_1 и Q_2. Распределения ξ значительно отличаются по коэффициентам корреляции компонент ξ_1 и ξ_2 ($r_{12}^{(1)} \neq r_{12}^{(2)}$), а математические ожидания совпадают: $M_1^{(1)} = M_1^{(2)}$, $M_2^{(1)} = M_2^{(2)}$. Ясно, что никакая линейная разделяющая функция не обеспечит классификации хотя бы с небольшой достоверностью.

Формулы (7.85) и (7.86) выведены в предположении нормального распределения ξ в классах $Q_1, Q_2, \ldots, Q_l$. Отклонения от нормального закона также оказывают влияние на качество классификации, приводя к увеличению количества ошибок. Если распределения используемых показателей $\xi_1, \xi_2, \ldots, \xi_m$ в классах Q_j описываются логнормальным законом, то решение достигается очевидной модификацией бейесовского правила (7.78), (7.85), которая состоит в оценке необходимых числовых характеристик ($\mathbf{m}_j$, $\mathbf{B}_j$, $\mathbf{B}$) не по самим наблюдениям $x_{il}^{(j)}$, а по их логарифмам $z_{il}^{(j)} = \lg x_{il}^{(j)}$. Соответственно преобразуется и наблюдение $\mathbf{y} = \{y_1, y_2, \ldots, y_m\}'$, подлежащее классификации — вместо него в решающем правиле используется $\mathbf{u} = \{\lg y_1, \lg y_2, \ldots, \lg y_m\}'$.

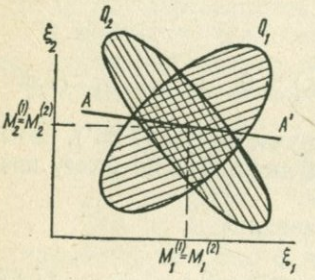


Рис. 41.

Как упоминалось ранее, однородные распределения количественных показателей пород часто аппроксимируются обобщенно-логнормальным законом. Эту аппроксимацию также можно использовать для построения параметрического критерия классификации. Обозначим $a_l^{(j)}$, $\lambda_l^{(j)}$ — параметры нормализующего преобразования ξ_l в Q_j:

$$\eta_l = \begin{cases} \ln(a_l^{(j)} + \lambda_l^{(j)} \xi_l) & \text{при обобщенно-логнормальном распределении } \xi_l, \\ \xi_l & \text{при нормальном распределении } \xi_l. \end{cases}$$

Плотность многомерного распределения ξ в Q_j согласно (3.32) будет иметь вид

$$p_j(\mathbf{x}) = (2\pi)^{-\frac{m}{2}} |\mathbf{B}_j|^{-\frac{1}{2}} \prod_{l=1}^{m} (a_l^{(j)} + \lambda_l^{(j)} x_l)^{s_l} \exp\left(-\frac{1}{2} \mathbf{z}' \mathbf{B}_j^{-1} \mathbf{z}\right), \quad (7.88)$$

где $\mathbf{z} = \{z_1, z_2, \ldots, z_m\}'$; $z_l = x_l - M_l^{(j)}$, $s_l = 0$, если принята гипотеза о нормальном распределении ξ_l в Q_j; $z_l = \ln(a_l^{(j)} + \lambda_l^{(j)} x_l) - \mu_l^{(j)}$, $\mu_l^{(j)} = \mathbf{M} \ln(a_l^{(j)} + \lambda_l^{(j)} \xi_l)$, $s_l = -1$, если принята гипотеза о нормальном распределении $\ln(a_l^{(j)} + \lambda_l^{(j)} \xi_l)$ в Q_j; $\mathbf{B}_j$ — ковариационная матрица преобразованных величин $\eta_1, \eta_2, \ldots, \eta_m$.

Вычислив оценки параметров, например, по методу максимума правдоподобия, можно получить оценки плотностей (7.88), а с ними и оценки вероятностей $P_j(\mathbf{y})$ по (7.74′). Разумеется, эффективность применения этого метода, ввиду необходимости оценки ковариационных матриц $\mathbf{B}_j$, обеспечивается большим количеством наблюдений. В случае независимости ξ_l во всех Q_j эта необходимость отпадает,

так как вероятности $P_j(\mathrm{y})$ определяются в виде (7.81) через одномерные плотности распределения ξ_l в Q_j:

$$p_l^j(x_l) = (2\pi)^{-\frac{1}{2}} (\sigma_l^{(j)})^{-1} (a_l^{(j)} + \lambda_l^{(j)} x_l)^{s_l} \exp\left\{-\frac{1}{2}\frac{z_l^2}{(\sigma_l^{(j)})^2}\right\},$$

где $(\sigma_l^{(j)})^2$ — дисперсии преобразованных величин η_l.

Пример 7.12. По данным каротажа скважин двух классов: 1) давших притоки воды и 2) давших притоки нефти,— проведен анализ распределений удельных пластовых сопротивлений $\rho_\text{п}$ (в ом · метрах) и показаний нейтронного гамма-метода $I_{n\gamma}$ (в относительных единицах). Результаты статистического анализа показали, что распределение $\rho_\text{п}$ в обоих классах описывается логнормальным законом, а $I_{n\gamma}$ — нормальным. Оценки математических ожиданий и средних квадратических отклонений логарифма $\rho_\text{п}$, $R = \lg \rho_\text{п}$, составили: для водонасыщенных пластов $\check{M}_R^{(1)} = 1{,}02$; $\check{\sigma}_R^{(1)} = 0{,}38$; для нефтенасыщенных $\check{M}_R^{(2)} = 1{,}80$; $\check{\sigma}_R^{(2)} = 0{,}27$. Оценки математических ожиданий и средних квадратических отклонений $I_{n\gamma}$: для водонасыщенных пластов $\check{M}_I^{(1)} = 0{,}40$; $\check{\sigma}_I^{(1)} = 0{,}18$; для нефтенасыщенных $\check{M}_I^{(2)} = 0{,}43$; $\check{\sigma}_I^{(2)} = 0{,}09$. 1) Считая связь между $I_{n\gamma}$ и $\rho_\text{п}$ (R) незначимой, построить правило классификации пластов по $\rho_\text{п}$ и $I_{n\gamma}$. 2) Пользуясь этим правилом, классифицировать пласт, для которого $\rho_\text{п} = 54$ *ом · м*, $I_{n\gamma} = 0{,}45$. 3) Оценить вероятность того, что этот пласт нефтенасыщенный.

Решение. 1) Учитывая, что $\rho_\text{п}$ и $I_{n\gamma}$ предполагаются независимыми и распределения $\lg \rho_\text{п}$, $I_{n\gamma}$ в обоих классах описываются нормальным законом, воспользуемся правилом (7.83), полагая $y_1 = \lg \rho_\text{п}$, $y_2 = I_{n\gamma}$; $\mathrm{y} = \{y_1, y_2\}'$ — вектор, подлежащий классификации, Q_1, Q_2 — классы водонасыщенных и нефтенасыщенных пластов соответственно. По (7.83) имеем

$$U = -\ln \check{\sigma}_R^{(1)} - \ln \check{\sigma}_I^{(1)} - \frac{1}{2}\frac{(\lg \rho_\text{п} - \check{M}_R^{(1)})^2}{(\check{\sigma}_R^{(1)})^2} - \frac{1}{2}\frac{(I_{n\gamma} - \check{M}_I^{(1)})^2}{(\check{\sigma}_I^{(1)})^2} + \ln \check{\sigma}_R^{(2)} +$$

$$+ \ln \check{\sigma}_I^{(2)} + \frac{1}{2}\frac{(\lg \rho_\text{п} - \check{M}_R^{(2)})^2}{(\check{\sigma}_R^{(2)})^2} + \frac{1}{2}\frac{(I_{n\gamma} - \check{M}_I^{(2)})^2}{(\check{\sigma}_I^{(2)})^2} = -\ln 0{,}38 - \ln 0{,}18 -$$

$$- \frac{1}{2}\frac{(\lg \rho_\text{п} - 1{,}02)^2}{0{,}38^2} - \frac{1}{2}\frac{(I_{n\gamma} - 0{,}40)^2}{0{,}18^2} + \ln 0{,}27 + \ln 0{,}09 + \frac{1}{2}\frac{(\lg \rho_\text{п} - 1{,}8)^2}{0{,}27^2} +$$

$$+ \frac{1}{2}\frac{(I_{n\gamma} - 0{,}43)^2}{0{,}09^2} \approx -1{,}03 - 3{,}46\,(\lg \rho_\text{п} - 1{,}02)^2 - 15{,}4\,(I_{n\gamma} - 0{,}40)^2 +$$

$$+ 6{,}86\,(\lg \rho_\text{п} - 1{,}8)^2 + 61{,}7\,(I_{n\gamma} - 0{,}43)^2.$$

Если полученное выражение будет положительным, пласт следует признать водонасыщенным, если отрицательным — нефтенасыщенным.

2) Подставив $\rho_\text{п} = 54$, $I_{n\gamma} = 0{,}45$ в найденное выражение U, получим $U \approx \approx -2{,}75 < 0$. Пласт классифицируется как нефтенасыщенный.

3) Для вычисления оценки искомой вероятности воспользуемся формулой (7.74′). Оценка плотности совместного распределения $\lg \rho_\text{п}$ и $I_{n\gamma}$ в Q_2, по (7.79) и (7.80)

$$\check{p}_2(\mathrm{y}) = \check{p}_2(\lg \rho_\text{п},\ I_{n\gamma}) = \frac{1}{2\pi \check{\sigma}_R^{(2)} \check{\sigma}_I^{(2)}} \exp\left[-\frac{1}{2}\frac{(\lg \rho_\text{п} - \check{M}_R^{(2)})^2}{(\check{\sigma}_R^{(2)})^2} -\right.$$

$$\left. - \frac{1}{2}\frac{(I_{n\gamma} - \check{M}_I^{(2)})^2}{(\check{\sigma}_I^{(2)})^2}\right] \approx \frac{1}{2\pi} 41{,}2 \exp\left[-6{,}86\,(\lg \rho_\text{п} - 1{,}8)^2 - 61{,}7\,(I_{n\gamma} - 0{,}43)^2\right];$$

аналогично

$$p_1(\mathrm{y}) = p_1(\lg \rho_{\mathrm{п}},\ I_{n\gamma}) \approx \frac{1}{2\pi} 14{,}6 \exp\left[-3{,}46\,(\lg \rho_{\mathrm{п}} - 1{,}02)^2 - 15{,}4\,(I_{n\gamma} - 0{,}40)^2\right].$$

Подставив $\rho_{\mathrm{п}} = 54$, $I_{n\gamma} = 0{,}45$, получим:

$$\check{p}_2(\lg 54,\ 0{,}45) \approx \frac{41{,}2}{2\pi} \exp(-0{,}06) \approx \frac{1}{2\pi} 38{,}8;$$

$$\check{p}_1(\lg 54,\ 0{,}45) \approx \frac{14{,}6}{2\pi} \exp(-1{,}78) \approx \frac{1}{2\pi} 2{,}46.$$

Искомая оценка, по (7.74′),

$$\check{P}_2(\mathrm{y}) = \frac{38{,}8}{2{,}46 + 38{,}8} \approx 0{,}92.$$

Непараметрическая классификация. Препятствием к применению параметрических правил классификации, помимо серьезных вычислительных трудностей, бывает необходимость привлечения большого числа наблюдений для обеспечения устойчивого решения, особенно при увеличении количества используемых параметров и зависимости компонент. К ошибкам приводят и отклонения распределений показателей от используемых для аппроксимации законов. Непараметрическая классификация реализуется непосредственно по эталонным выборкам, без использования каких-либо параметрических приближений для плотностей распределения. Рассмотрим способ построения такого решающего правила.

Предположим, плотности многомерного распределения показателей $\xi_1, \xi_2, \ldots, \xi_m$ существуют, однако вид их неизвестен. Образуем вокруг каждой координаты y_l $(l = \overline{1, m})$ интервалы малой длины $\Delta_l(y_l)$ с центрами y_l, которые будут проекциями на координатные оси m-мерного параллелепипеда Ω вида (5.18) с центром у. Как показывает (5.19), можно приближенно считать, по-прежнему, полагая априорные вероятности одинаковыми,

$$P_j(\mathrm{y}) = \mathbf{P}\{\mathrm{y} \in Q_j\} = \frac{P_j(\Omega)}{\sum_{j=1}^{t} P_j(\Omega)} \quad (j = \overline{1, t}), \tag{7.89}$$

где $P_j(\Omega) = \mathbf{P}\{\xi \in \Omega / \xi \in Q_j\}$ — вероятность того, что значения величин $\xi_1, \xi_2, \ldots, \xi_m$ при наблюдении их в классе Q_j образуют точку m-мерного пространства, попадающую в Ω. Воспользуемся оценками величин $P_j(\Omega)$ $(j = \overline{1, t})$. В качестве таких оценок возьмем

$$\check{P}_j(\Omega) = \frac{1}{n_j} n_j(\Omega) \quad (j = \overline{1, t}), \tag{7.90}$$

где $n_j(\Omega)$ — количество векторов-наблюдений $\mathrm{x}_i^{(j)}$ из Q_j, попавших в Ω.

Простейшая реализация описываемого принципа состоит в определении групп наблюдений $\mathrm{x}_l^{(j)}$ в каждой выборке $\tilde{Q}_j$, наиболее близких к у по некоторому «расстоянию» между у и $\mathrm{x}_l^{(j)}$ заранее сфор-

мулированной структуры. В качестве такого расстояния можно взять

$$\rho(\mathbf{x}_i^{(j)}, \mathbf{y}) = \sqrt{\sum_{l=1}^{m} \alpha_l^2 (x_{il}^{(j)} - y_l)^2}, \qquad (7.91)$$

где α_l — коэффициенты масштаба, выбираемые в соответствии с вариацией каждого показателя в классах Q_j и, возможно, степенью их информативности. В частности, хорошие результаты обеспечивает выбор α_l^2 обратно пропорциональным средней дисперсии l-го показателя:

$$\alpha_l^2 = \frac{1}{\bar{\sigma}_l^2}\left(\sum_{l=1}^{m}\frac{1}{\bar{\sigma}_l^2}\right)^{-1}, \quad \bar{\sigma}_l^2 = \frac{1}{t}\sum_{j=1}^{t}(\check{\sigma}_l^{(j)})^2, \qquad (7.92)$$

или в виде $\alpha_l = \frac{1}{\bar{\sigma}_l}\left(\sum_{l=1}^{m}\frac{1}{\bar{\sigma}_l}\right)^{-1}$. Задаваясь некоторым фиксированным уровнем $\beta < 1$, вычисляем f_j — округленные до целого значения βn_j. Далее, в каждой выборке $\tilde{Q}_j$ определяем f_j ближайших к $\mathbf{y}$ по расстоянию (7.91) векторов $\mathbf{x}_i^{(j)}$, а среди них наиболее удаленный $\mathbf{x}_d^{(j)}$. Обозначив $\rho^* = \min_j \rho(\mathbf{x}_d^{(j)}, \mathbf{y})$, находим в каждой выборке $\tilde{Q}_j$ количество наблюдений n_j^*, для которых $\rho(\mathbf{x}_i^{(j)}, \mathbf{y}) < \rho^*$. Вероятности (7.89) оцениваются в виде

$$\check{P}_j(\mathbf{y}) = \frac{n_j^*}{n_j}\left(\sum_{j=1}^{t}\frac{n_j^*}{n_j}\right)^{-1}, \qquad (7.89')$$

а класс, к которому относят $\mathbf{y}$, определяется максимальной среди этих оценок. Группа наиболее вероятных классов определяется наибольшими значениями оценок (7.89′), сумма которых по (7.72) оценивает вероятность того, что $\mathbf{y}$ принадлежит этой группе.

Рассмотренный непараметрический метод эффективен и в случае частично пересекающихся совокупностей, когда метод линейных дискриминантных функций может оказаться несостоятельным (рис. 41). Без особых затруднений он реализуется и в тех случаях, когда часть используемых показателей отождествляется с качественными признаками. Такие показатели кодируются обычным способом: $\xi_l = 0$ при наличии признака и $\xi_l = 1$ в противном случае. Аналогично кодирование нескольких качественных состояний, например: $\xi_l = 0$ — «нет признака», $\xi_l = 1$ — «признак проявлен частично», $\xi_l = 2$ — «есть признак». Пример применения непараметрического метода приведен в [12]— классификация проб по петрографическим разновидностям на основании геохимических данных.

Метод требует значительного объема вычислений, поэтому его применение практически невозможно без ЭВМ. Можно указать способ непараметрической классификации, свободный от этого недостатка — в условиях независимости или слабой зависимости компонент ξ_l

в классах Q_l. В этом случае вероятности $P_l(\mathrm{y})$ оцениваются по (7.81) через оценки одномерных плотностей распределения $p_l^{(j)}(x)$ [19]. В качестве этих оценок можно использовать гистограммы. Сняв с графиков гистограмм (6.15) значения $\check{p}_l^{(j)}(y_l)$ и подставив их в (7.81), получим искомые оценки $\check{P}_l(\mathrm{y})$. При $t=2$ по (7.82) класс, к которому следует отнести y, определяется оценкой произведения $\prod_{l=1}^{m} k_l$, где $k_l = \frac{p_l^{(1)}(y_l)}{p_l^{(2)}(y_l)}$ — отношение значений плотностей распределения показателя ξ_l в классах Q_1 и Q_2. При $\prod_{l=1}^{m} \check{k}_l > 1$ или $\sum_{l=1}^{m} \lg \check{k}_l > 0$ принимается гипотеза о принадлежности к классу Q_1; в случае противоположного неравенства y относят в Q_2.

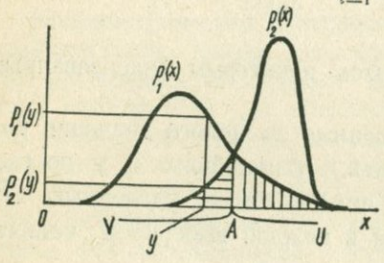


Рис. 42.

Оценка ошибок классификации. Анализ ошибок составляет необходимый этап в применении любого метода классификации. Приступая к реализации метода, необходимо знать: достаточны ли метод и исходная информация для решения задачи; какая достоверность в имеющихся условиях вообще возможна; в какой мере и какие именно результаты следует считать достоверными.

Рассмотрим простейший случай, когда имеется всего два класса и классификация производится по одному показателю ξ (рис. 42). Согласно критерию (7.82), область принятия решения $\{y \in Q_1\}$ определяется условием $\frac{p_1(y)}{p_2(y)} > 1$. На рис. 42 эту область образуют значения слева от точки A, в которой $\frac{p_1(A)}{p_2(A)} = 1$. Справа от A — область принятия решения $\{y \in Q_2\}$. Если считать нулевой гипотезой $\{y \in Q_1\}$, вероятность ошибки I рода будет

$$q_{12} = \mathbf{P}\{y \in U / y \in Q_1\} = \int_U p_1(x)\,dx \qquad (7.93)$$

$\left(U$ — область значений x, для которых $\frac{p_1(x)}{p_2(x)} < 1\right)$. На рис. 42 q_{12} равна площади вертикально заштрихованной фигуры. Вероятность ошибки II рода

$$q_{21} = \int_V p_2(x)\,dx = 1 - \int_U p_2(x)\,dx - \qquad (7.93')$$

площадь, заштрихованная горизонтально $\left(V$ — область значений x, для которых $\frac{p_1(x)}{p_2(x)} > 1\right)$. Среднее количество ошибочно классифициру-

мых наблюдений из общего их числа $n = n_1 + n_2$ составит $n_1 q_{12} + n_2 q_{21}$ (n_1 — число наблюдений из Q_1, n_2 — из Q_2; $p_1(x)$ и $p_2(x)$ считаем непрерывными).

При многомерной классификации (по комплексу показателей) провести подобный расчет довольно сложно. Поэтому для характеристики критерия вместе с имеющимися исходными данными используют обычно группы контрольных объектов из каждого класса. Классифицировав каждый контрольный объект, вычисляют оценки вероятностей ошибок $\check{q}_{ij} = \frac{N_{ij}}{N_i}$, где N_i — количество контрольных объектов из Q_i, N_{ij} — число тех из них, которые были отнесены к классу Q_j. Заодно эти оценки позволяют выделить наиболее близкие классы, которые можно разделить лишь с малой достоверностью. Это дает основание объединять отдельные классы как практически неразличимые. Часто в качестве характеристик критерия используют лишь величины $\check{q}_{ii} = \frac{N_{ii}}{N_i}$, имеющие смысл оценок вероятностей отнесения наблюдения из Q_i к своему классу. Ясно, что $\check{q}_{ii} = 1 - \sum_{j,\ j \neq i}^{t} \check{q}_{ij}$.

Нередко, однако, нельзя указать представительные выборки контрольных объектов, как того требует этот способ. Тогда приходится ограничиться использованием в качестве контрольных эталонных выборок $\tilde{Q}_j$.

Для уменьшения ошибок иногда поступают следующим образом. При классификации на два класса Q_1 и Q_2 определяют два предела c_1 и c_2 такие, что $c_1 > c_2$ и при $\frac{p_1(y)}{p_2(y)} > c_1$ у относят к Q_1, а при $\frac{p_1(y)}{p_2(y)} < c_2$ — к Q_2. За счет выбора c_1 и c_2 можно обеспечить малые вероятности ошибок q_{12} и q_{21}. Те наблюдения, для которых $c_2 \leqslant \frac{p_1(y)}{p_2(y)} \leqslant c_1$, по такому критерию не классифицируются, т. е. для уверенной классификации имеющуюся информацию считают недостаточной. В одномерном случае — при классификации по одному показателю — часто достаточно определить два предела для его значений A_1 и A_2 такие, что при $y < A_1$ наблюдение относят в один из двух классов, Q_1 или Q_2, а при $y > A_2$ — в другой.

Многомерная классификация в задаче сопоставления. Результаты классификации исходных выборок $\tilde{Q}_j$, составленных из представителей классов Q_j, можно использовать для решения самостоятельной задачи — сопоставления классов по комплексу показателей $\xi_1, \xi_2, \ldots, \xi_m$. Величины q_{ij} и q_{ji} служат при этом характеристиками близости классов Q_i и Q_j по многомерным распределениям этих показателей. Близость q_{ii} к единице указывает на обособленное положение класса Q_i по отношению к остальным, возможность его достоверного выделения по $\xi_1, \xi_2, \ldots, \xi_m$.

Информативные показатели. Показатели $\xi_1, \xi_2, \ldots, \xi_m$ могут по-разному проявлять себя в задаче классификации: некоторые несут

малую информацию о различиях классов Q_j, по другим наблюдаются отличия только отдельных классов. Неинформативные показатели, ведущие себя одинаково во всех классах, могут ухудшить результат классификации и стать источником дополнительных ошибок. Поэтому приобретает значение оценка *информативности* отдельных показателей и их групп, т. е. характеристики их по той информации, которую они несут о различиях классов Q_j.

Простейшими такими характеристиками отдельных показателей служат значения $k_{ij}^{(l)}$ величины k_{ij} (7.32), которые определяются при сопоставлении классов Q_i и Q_j по плотностям распределения ξ_l — $p_i^{(l)}(x)$ и $p_j^{(l)}(x)$. Эту характеристику отличия распределений мы подробно рассматривали в § 1 настоящей главы. Если значение $k_{ij}^{(l)}$ равно или близко к единице, ξ_l — исчерпывающе информативный показатель для разделения Q_i и Q_j; если $k_{ij}^{(l)} \approx 0$, взятый в отдельности, он не несет информации о различиях Q_i и Q_j. В непараметрическом правиле с использованием (7.91) можно учесть различия в информативности отдельных показателей, придав им бóльшие по сравнению с (7.92) коэффициенты a_l.

В задаче классификации можно использовать большое количество показателей, причем нередко их число можно увеличить за счет дополнительных признаков. Такие признаки можно получить, например, при трансформациях полей, заданных в узлах равномерной сети наблюдений, вводя специальные пересчеты, вычисляя характеристики локальной неоднородности, анизотропии, градиент в пределах скользящего окна. Однако увеличение числа признаков, особенно если они зависимы и без соответствующего увеличения эталонных выборок, далеко не всегда повышает эффективность классификации. Возникает задача выделения наименьшего набора признаков, при котором вероятности ошибок остаются на минимальном уровне. Такой набор определяют последовательным отбрасыванием от полного набора по одному показателю так, чтобы эффективность классификации не ухудшалась (например, не уменьшалось расстояние по Махаланобису). Другой способ состоит в последовательном присоединении к наиболее информативному показателю таких, которые в совокупности с уже отобранными в наибольшей мере улучшают классификацию.

Следует упомянуть еще об одной возможности повышения эффективности решающих правил — *иерархической классификации*. Сущность этого принципа состоит в классификации не сразу на классы $Q_1, Q_2, \ldots, Q_t$, а на определенные их объединения $Q_1^0, Q_2^0, \ldots, Q_k^0$ $(k < t)$; Q_j^0 $(j = \overline{1, k})$ состоят каждое из одного или нескольких классов Q_i. При этом объединяются наиболее близкие классы так, чтобы получить обособленные группы классов. Для классификации на $Q_1^0, Q_2^0, \ldots, Q_k^0$ выбирается система показателей ξ_l с соответствующими им коэффициентами масштаба, если метод предусматривает возможность учета информативности показателей с их помощью. После этого этапа классификации, если вектор y отнесен к объединению Q_j^0, определяется его принадлежность к одному из более мел-

ких объединений или непосредственно классов Q_l, слагающих Q_j^0. Для этого снова подбираются информативные показатели и система учета их информативности. Ясно, что в определенных ситуациях такой подход может повысить эффективность применения непараметрических методов классификации.

Информативные сочетания. При сопоставлении распределений отдельных компонент ξ_l с целью выяснения их информативности не учитываются взаимоотношения показателей, особенности их совместного поведения в разных классах. Между тем эти особенности могут нести значительную информацию. На рис. 41 изображен такой случай: математические ожидания и дисперсии двух компонент ξ_1 и ξ_2 в обоих классах не отличаются, так что соответствующие одномерные плотности их распределений близки или вообще совпадают. Тем не менее классы различаются по совместному распределению ξ_1 и ξ_2 — коэффициенты корреляции этих величин существенно различны.

Таким образом, представляет интерес оценка информативности сочетаний различных показателей, которую проще всего произвести следующим способом. Разобьем шкалу значений каждого показателя ξ_l на некоторое число F_l интервалов $\Delta_k^{(l)}$ $(k = \overline{1, F_l})$ и введем целочисленные величины ξ_l^0: $\xi_l^0 = k$, если ξ_l принимает значение, попадающее в интервал $\Delta_k^{(l)}$. Если по этому правилу преобразовать наблюдения эталонных выборок $\tilde{Q}_j$, получим их коды — целочисленные векторы наблюдений $\mathbf{z}_i^{(j)} = \{z_{i1}^{(j)}, z_{i2}^{(j)}, \ldots, z_{im}^{(j)}\}'$, где $z_{il}^{(j)} = k$, если $x_{il}^{(j)} \in \Delta_k^{(l)}$. Очевидно, при достаточно широких интервалах значения $z_{il}^{(j)}$ и отдельные их сочетания будут повторяться в кодах $\tilde{Q}_j$, причем тем чаще, чем более характерны эти значения и сочетания для данного класса.

Пусть $\mathbf{z} = \{z_1, z_2, \ldots, z_u\}$ — сочетание u целых чисел, представляющих собой номера некоторых интервалов для группы показателей $\xi_{l_1}, \xi_{l_2}, \ldots, \xi_{l_u}$. Мерой информативности может служить вероятность появления такого сочетания в классе Q_j, оцениваемая как $p_j(\mathbf{z}) = \frac{n_j(\mathbf{z})}{n_j}$, где $n_j(\mathbf{z})$ — число наблюдений $\mathbf{x}_i^{(j)}$ из $\tilde{Q}_j$, в кодах которых есть сочетание $\mathbf{z}$. Подобна ей функция $I_j(\mathbf{z}) = \log p_j(\mathbf{z})$*. Ясно, что непосредственно на меры $p_j(\mathbf{z})$ или $I_j(\mathbf{z})$ опираться нельзя, так как сочетание $\mathbf{z}$ может встречаться в других классах, причем, возможно, не реже, чем в данном. С точки зрения задачи классификации такие сочетания не будут представлять интереса. Особые черты класса Q_j будут выражать те сочетания, которые встречаются в нем с высокой вероятностью, а в остальных — с низкой. *Информативными* относительно Q_j уровня $L > 0$ будут такие сочетания $\mathbf{z}$, для которых $\min\limits_{i,\ i \neq j} \frac{p_j(\mathbf{z})}{p_i(\mathbf{z})} > L$ или

$$I_0(\mathbf{z}) = \min_{i,\ i \neq j} [\log p_j(\mathbf{z}) - \log p_i(\mathbf{z})] = \min_{i,\ i \neq j} [I_j(\mathbf{z}) - I_i(\mathbf{z})] > \log L. \quad (7.94)$$

* Основание логарифма обычно берут равным двум или десяти

Наиболее информативными будут сочетания, для которых значения $I_0(z)$ максимальны. Аналогичные меры можно ввести и для групп классов, например при иерархическом способе классификации.

Число всех возможных сочетаний даже при сравнительно малом числе градаций каждого показателя составляет довольно значительную величину. Так, если для всех показателей число градаций одинаково и равно F, то количество всех возможных сочетаний составит $C_m^u F^u$. Это ограничивает возможности рассмотрения сочетаний из большого числа показателей даже с применением ЭВМ — обычно используются сочетания по три признака ($u = 3$).

Многомерная классификация при прогнозировании месторождений. Применение методов классификации для прогнозирования месторождений предусматривает следующие этапы.

1. Описание системы признаков. Исследуемая площадь разбивается на элементарные ячейки, по каждой рассчитывается набор признаков. В этот набор могут входить: количественные характеристики геофизических, геохимических полей (изменчивость, средний уровень, градиент, параметры аномалий), оцениваемые по каждой ячейке; геологические признаки (минералого-петрографический состав пород, характер и интенсивность постмагматических процессов, тектоники, уровень эрозионного среза), геоморфологические, гидрохимические и др.

2. Выделение классов Q_j (в их числе должны быть представлены различные типы оруденения, возможные на данной площади, а также заведомо непродуктивные объекты) и формирование эталонных выборок, характеризующих каждый класс.

3. Оценка информативности признаков и их сочетаний с последующим выделением опорной системы признаков.

4. Проверка эффективности используемого метода классификации и оценка ошибок классификации.

5. Классификация основного массива данных. Вероятности принадлежности к различным типам оруденения выносят на карту и изображают в изолиниях; области наибольших вероятностей соответствуют, в случае небольших ошибок классификации, наиболее перспективным площадям.

Принципы построения алгоритмов распознавания. Обзор всех существующих алгоритмов и программ распознавания образов занял бы здесь слишком много места. К тому же статистическое решение проблемы классификации, основные аспекты которого мы рассмотрели, представляется достаточно полным и обоснованным. Поэтому ограничимся тем, что приведем основные принципы, используемые в эмпирических методах классификации.

1. Классификация осуществляется по такой схеме: а) имеющаяся количественная информация кодируется в целочисленных кодах; б) выделяется набор информативных сочетаний («обучение»); в) подсчитываются суммы значений $I_j(z)$ по информативным относительно классов Q_j сочетаниям z или значения какой-либо иной подобной меры по коду вектора y, представленного для классификации, и сравниваются между собой («распознавание»).

2. По целочисленному коду вектора $\mathbf{y}$ отыскиваются тождественные ему коды в эталонных выборках $\tilde{Q}_j$. Если таких нет, код $\mathbf{y}$ трансформируется так, чтобы обеспечить бо́льшую вероятность его встречи среди кодов $\tilde{Q}_j$, например, отбрасывается некоторая компонента или расширяются интервалы кодирования. После этого снова отыскиваются тождественные коды и т. д. Количества найденных аналогов сравниваются между собой.

3. По данным выборок $\tilde{Q}_j$ рассчитываются функции $f_{ij}(\xi) = f_{ij}(\xi_1, \xi_2, \ldots, \xi_m)$ показателей $\xi_1, \xi_2, \ldots, \xi_m$. Этими функциями определяются уравнения вида $f_{ij}(\mathbf{x}) = c_{ij}$ поверхностей, разделяющих области $\tilde{Q}_i, \tilde{Q}_j\ (i \neq j)$ в m-мерном пространстве значений ξ. Поверхности должны быть такими, чтобы по возможности наибольшее число наблюдений попадало в «свои» области классификации.

4. Вводится метрика, характеризующая «расстояние» классифицируемого вектора $\mathbf{y}$ до каждого из классов Q_j по показателям $\xi_1, \xi_2, \ldots, \xi_m$. Примером может служить обычное эвклидово расстояние до «центров» каждой Q_j в m-мерном пространстве значений ξ. Роль центра Q_j выполняет вектор с компонентами, равными математическим ожиданиям $\mathbf{M}\xi_l\ (l = \overline{1, m})$ в Q_j. Результат классификации определяется минимумом расстояния.

5. Подобно этому вводится мера сходства вектора $\mathbf{y}$ с произвольной группой векторов. Классификация производится по наилучшему сходству с классами Q_j. Мерой сходства двух объектов может служить, например, убывающая функция расстояния ρ (7.91) либо мера, равная числу совпадающих компонент в кодах обоих объектов. В методе *потенциальных функций* мера сходства данного вектора $\mathbf{y}$ с группой $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ — выражающийся через потенциальную функцию суммарный потенциал, отнесенный к числу векторов k. Пример потенциальной функции — $\dfrac{1}{1 + \alpha\rho^2}$.

Классификация и задача косвенных измерений. Пусть m-е компоненты векторов $\mathbf{x}_i^{(j)}$, $x_{im}^{(j)}$ — значения показателя ξ_m, относительно которого в дальнейшем ставится задача косвенных измерений по остальным $\xi_1, \xi_2, \ldots, \xi_{m-1}$. Такая ситуация возникает в тех случаях, когда прямые измерения ξ_m в большом количестве затруднительны, причем форма связи ξ_m с $\xi_1, \xi_2, \ldots, \xi_{m-1}$ существенно различна в разных Q_j.

По наблюдениям каждой выборки $\tilde{Q}_j$ можно построить уравнения множественной связи ξ_m с остальными компонентами. Задача состоит в оценке y_m по наблюдениям первых $m - 1$ компонент $\{y_1, y_2, \ldots, y_{m-1}\}' = \mathbf{y}_{m-1}$. Для определения нужного уравнения регрессии достаточно классифицировать вектор $\mathbf{y}_{m-1}$ и взять уравнение той совокупности Q_j, к которой он будет отнесен *. Подставив $y_1, y_2, \ldots, y_{m-1}$ в это уравнение, получим оценку y_m.

Нетрудно оценить точность такого косвенного измерения. Вокруг значения функции регрессии в каждом классе Q_j можно построить

---

* Наборы показателей для классификации и для косвенных измерений, конечно, не обязательно должны совпадать.

доверительные интервалы $\Delta_j^{(m)}$ для y_m. Точность определяется распределением y_m по интервалам $\Delta_j^{(m)}$: $\mathbf{P}\{y_m \in \Delta_j^{(m)}\} = q_j \mathbf{P}\{y_{m-1} \in Q_j\}$, где q_j — доверительная вероятность интервала $\Delta_j^{(m)}$. Выделив из $\bigcup_{j=1}^{t} \Delta_j^{(m)}$ множество Δ такое, что вероятность попадания в него истинного значения y_m составляет q, получим доверительную область для y_m. Если $\mathbf{P}\{y_{m-1} \in Q_{j_0}\}$ — максимальная из вероятностей $\mathbf{P}\{y_{m-1} \in Q_j\}$ $(j = \overline{1, t})$ — близка к единице и $q_{j_0} = q$, в качестве этой области можно взять $\Delta_{j_0}^{(m)}$.

Случай классификации группы наблюдений. Достоверность классификации иногда можно повысить, используя не единичные наблюдения y, а целые группы $\{y_1, y_2, \ldots, y_k\}$, о которых заведомо известно, что все они принадлежат к какому-либо одному классу. В этом случае проще всего использовать результаты классификации отдельных наблюдений, вычислив средние арифметические из соответствующих оценок вероятностей принадлежности

$$\bar{P}_j(y_1, y_2, \ldots, y_k) = \frac{1}{k} \sum_{i=1}^{k} P_j(y_i) \quad (j = \overline{1, t}) \tag{7.95}$$

и по этим оценкам определить наиболее вероятный класс для группы $\{y_i\}$.

Другой способ состоит в использовании формулы (5.25):

$$\check{\mathbf{P}}\{(y_1, y_2, \ldots, y_k) \in Q_j\} = \frac{\prod_{i=1}^{k} \check{p}_j(y_i)}{\sum_{j=1}^{t} \prod_{i=1}^{k} \check{p}_j(y_i)}, \tag{7.96}$$

где $\check{p}_j(x)$ — оценка плотности многомерного распределения ξ в Q_j. Такой подход удобен при параметрической классификации.

Наконец, можно использовать формулу (7.74′) для вектора $\bar{y}$, полученного усреднением $y_1, y_2, \ldots, y_k$ $\left(\bar{y} = \frac{1}{k} \sum_{i=1}^{k} y_i\right)$ с соответствующим преобразованием плотностей распределения. При нормальном распределении ξ плотности рассчитываются по (7.75) с подстановкой $\frac{1}{k}\mathbf{B}_j$ вместо $\mathbf{B}_j$ — ковариационных матриц вектора ξ в классах Q_j.

§ 5. Способы выделения аномальных наблюдений

Пусть $y_1, y_2, \ldots, y_n$ — выборка наблюдений показателя ξ, среди которых требуется выделить аномальные. Как упоминалось в § 3 гл. 5, задача состоит в построении критерия, позволяющего разделить выборку на фоновые и аномальные наблюдения. Ясно, что нельзя говорить об аномальных наблюдениях вообще — таковыми мы можем их считать лишь по отношению к какому-либо распределению, под которым и подразумевается фоновое. Решение возможно, если имеется информация о законе распределения фоновой совокупности.

Критерии аномальности с использованием эталонной выборки. Наиболее доступный способ получить представление о фоновом рас-

пределении ξ состоит в привлечении репрезентативной эталонной выборки — наблюдений $x_1, x_2, \ldots, x_k$, заведомо ему принадлежащих, по которой можно достаточно точно оценить плотность фонового распределения $u(x)$. Если альтернатива (аномальное распределение) неизвестна, задача сводится к проверке гипотезы о принадлежности каждого y_i к фоновому распределению с плотностью $u(x)$, которую проще всего произвести, сравнивая y_i с квантилями высоких или низких порядков. Квантили оцениваются с помощью какой-либо аппроксимации распределения x_i — нормальным, логнормальным, обобщенно-логнормальным (2.38) законами или иными схемами. Следует отметить, что оценки критических границ, вычисляемые по выборке самих проверяемых наблюдений $y_1, y_2, \ldots, y_n$, подвержены значительным неучитываемым колебаниям из-за невыявленных аномальных наблюдений.

Пусть A_q — квантиль порядка q: для фонового распределения $\mathbf{P}\{\xi < A_q\} = q$, $\mathbf{P}\{\xi \geqslant A_q\} = 1 - q$. Если $q = q^+$ близко к единице (обычно $q^+ > 0{,}95$), то событие $\xi > A_q^+$ маловероятно и наблюдения y_i, превышающие A_q^+, следует признать аномально высокими. При малом $q = q^-$ ($q^- < 0{,}05$) наблюдения y_i признают аномально низкими, если $y_i < A_q^-$ (A_q^- и A_q^+ — квантили порядков q^- и q^+ соответственно).

При выборе значения q необходимо иметь в виду следующее. Если среди наблюдений $y_1, y_2, \ldots, y_n$ заведомо нет аномальных, математическое ожидание числа тех из них, которые окажутся меньше A_q^-, составит по (2.3) nq^-, а бо́льших A_q^+ — $n(1 - q^+)$. Поэтому наличие аномально высоких наблюдений в выборке $y_1, y_2, \ldots, y_n$ следует признать лишь тогда, когда n^+ — число превышающих критический уровень A_q^+ наблюдений y_i — значимо больше $n(1 - q^+)$. Аналогичный вывод делают об аномально низких наблюдениях, если n^- — количество наблюдений y_i, меньших A_q^- — значимо больше nq^-.

Распределения n^+ и n^- приближенно следуют закону Пуассона с параметрами: для n^+ — $\lambda^+ = (1 - q^+)n$; для n^- — $\lambda^- = q^-n$. Вычислив квантиль m_α распределения Пуассона порядка α, можно определить с вероятностью ошибки $1 - \alpha$ наличие в выборке аномально низких или аномально высоких наблюдений, если, соответственно, $n^- > m_\alpha$ или $n^+ > m_\alpha$. Они будут среди тех y_i, которые удовлетворяют условию $y_i < A_q^-$ или $y_i > A_q^+$ соответственно.

В практике геохимических поисков нередко пользуются упрощенным правилом, принимая в качестве критической границы для аномально высоких содержаний химических элементов $A_q^+ = \bar{x} + 3\bar{\sigma}$ ($\bar{x}$ и σ — оценки математического ожидания и стандарта фонового распределения). При нормальном распределении ξ этой границе соответствует вероятность $q \approx 0{,}9987$, так что наблюдение, превышающее уровень $\bar{x} + 3\bar{\sigma}$, следует считать аномальным (для фонового распределения вероятность такого события $\sim 0{,}0013$). При логнормальном распределении используется критическая граница вида $e^{\breve{\mu}+3\breve{\delta}} = \breve{M}_e e^{3\breve{\delta}}$, где $\breve{\mu}$ — среднее арифметическое, $\breve{\delta}$ — оценка среднего квадратического отклонения, вычисленные по логарифмам содержаний ($\ln x_i$). Повышения эффективности подобных критериев можно достичь предварительным осред-

нением методом скользящего окна наблюдений, вынесенных на профиль или карту, опираясь на возможность подтверждения аномалии несколькими рядом расположенными пробами. Критической границей уровня значимости $1-q$ для средних арифметических, составляемых из k наблюдений, будет $A_q^{+} = \bar{x} + u_q \check{\sigma} \frac{1}{\sqrt{k}}$, где u_q — q-квантиль $(0; 1)$-нормального распределения. При обработке данных, вынесенных на карту, иногда производят осреднение с использованием окон нескольких различных форм (параллелограммов с разными углами при вершине). Подбором оптимальной формы окна удается повысить надежность выделения аномалий за счет корреляции наблюдений по профилям.

Рассмотрим задачу выделения аномалий по комплексу показателей. Пусть $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ — наблюдения m-мерной величины, составленные каждое из результатов измерений m показателей $\{\xi_1, \xi_2, \ldots, \xi_m\}' = \boldsymbol{\xi}$: $\mathbf{y}_i = \{y_{i1}, y_{i2}, \ldots, y_{im}\}'$ $(i = \overline{1, n})$; $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ — наблюдения той же величины, среди которых заведомо нет аномальных. Необходимо в выборке $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ указать аномальные наблюдения, выделяющиеся по комплексу показателей $\xi_1, \xi_2, \ldots, \xi_m$ или по каким-либо из них в отдельности.

Обозначим $\check{p}(\mathbf{x})$ — оценку плотности m-мерного распределения, полученную по эталонной выборке $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$. В соответствии с (5.10) критерий будет таким: если

$$\check{p}(\mathbf{y}_i) \leqslant C(\alpha), \quad \text{или} \quad \ln \check{p}(\mathbf{y}_i) \leqslant \ln C(\alpha), \tag{7.97}$$

то $\mathbf{y}_i$ следует признать аномальным с вероятностью ошибки α. $C(\alpha)$ — критическая граница, выбираемая в соответствии с величиной α. Например, если $p(\mathbf{x})$ — плотность нормального распределения, то оценка ее определяется по (3.53) с помощью оценок математического ожидания $\check{\mathbf{m}}$ и ковариационной матрицы $\check{\mathbf{B}}$. Имеем

$$\ln \check{p}(\mathbf{y}_i) = -\frac{m}{2} \ln (2\pi) - \frac{1}{2} \ln |\mathbf{B}| - \frac{1}{2} (\mathbf{y}_i - \check{\mathbf{m}})' \check{\mathbf{B}}^{-1} (\mathbf{y}_i - \check{\mathbf{m}}).$$

После подстановки этого выражения в (7.97) константы, не зависящие от $\mathbf{y}_i$, перенесем в правую часть неравенства. Величина $u = (\mathbf{y}_i - \check{\mathbf{m}})' \check{\mathbf{B}}^{-1} (\mathbf{y}_i - \check{\mathbf{m}})$ при фоновом распределении $\mathbf{y}_i$ распределена приближенно* по закону χ^2 с m степенями свободы, так что гипотеза об аномальности $\mathbf{y}_i$ будет приниматься с вероятностью ошибки $\alpha \approx 1 - q$, если

$$(\mathbf{y}_i - \check{\mathbf{m}})' \check{\mathbf{B}}^{-1} (\mathbf{y}_i - \check{\mathbf{m}}) \geqslant \chi_q^2(m), \tag{7.98}$$

где $\chi_q^2(m)$ — квантиль распределения χ^2 с m степенями свободы порядка q. Если при фоновом распределении ξ_j независимы, то условие (7.98) приобретает простой вид: $\sum_{j=1}^{m} \frac{(y_{ij} - \bar{x}_j)^2}{\check{\sigma}_j^2} \geqslant \chi_q^2(m)$, где $\bar{x}_j$ и $\check{\sigma}_j^2$ — оценки математических ожиданий и дисперсий ξ_j, вычисляемые по эталонной выборке.

* Приближение эффективно при больших k, причем требования к объему эталонной выборки возрастают с увеличением m.

Для величины ξ, компоненты которой распределены логнормально, критерий строится подобно (7.98) с использованием при вычислении $\check{\mathbf{m}}$ и $\check{\mathbf{B}}$, а также в качестве компонент $\mathbf{y}_i$ вместо значений показателей их логарифмов.

Способ оценки аномалий по комплексу показателей, при котором не нужно задавать вид фонового распределения, состоит в следующем. Для вектора $\mathbf{y}_i$ определяется число $n(\mathbf{y}_i)$ m-мерных наблюдений $\mathbf{x}_j$, расстояние (7.91) до которых не больше фиксированного значения ρ_0. Для $n(\mathbf{y}_i)$ определяют критическое значение n_0 так, чтобы для всех наблюдений $\mathbf{x}_j$ эталонной выборки $n(\mathbf{x}_j) \geqslant n_0$ $(j = \overline{1, k})$. Если $n(\mathbf{y}_i) < n_0$, вектор $\mathbf{y}_i$ считают аномальным. При малых ρ_0 выделяются слабые аномалии, с увеличением ρ_0 будут фиксироваться те $\mathbf{y}_i$, аномальность которых обозначается все более резко. В простейшем варианте этого способа $n_0 = 1$, $\rho_0 = \max\limits_{j} \min\limits_{i,\, i \neq j} \rho(\mathbf{x}_i, \mathbf{x}_j)$ и аномальность вектора $\mathbf{y}_i$ характеризуется величиной $\min\limits_{j} \rho(\mathbf{y}_i, \mathbf{x}_j)$ по сравнению с ρ_0. Этот метод можно применить и в задаче классификации, когда только один класс представлен эталонами; в этом случае выясняют, принадлежит или не принадлежит $\mathbf{y}_i$ к этому классу.

При обработке данных геохимического опробования нередко оказывается эффективным использование простейших функций, определяемых так, чтобы учесть отклонения от фона нескольких химических элементов в одних и тех же пробах. При этом элементы $C_1, C_2, \ldots, C_t$, имеющие тенденцию к накоплению в объектах поиска, объединяются в одну группу и для каждого пункта опробования по их концентрациям вычисляется величина $T = \lg(C_1 C_2 \ldots C_t)$. Если можно указать еще и группу элементов $C^{(1)}, C^{(2)}, \ldots, C^{(s)}$ с предполагаемой тенденцией к выносу, то используется функция $T_1 = \lg \frac{C_1 C_2 \ldots C_t}{C^{(1)} C^{(2)} \ldots C^{(s)}}$. Полученные данные выносятся на карту, после чего обрабатываются обычным способом (как наблюдения одномерной величины). Выделяемые при этом аномалии именуют «мультипликативными».

Часто достаточно суммировать концентрации элементов $C_1, C_2, \ldots, C_t$, предварительно их нормировав: $S = \alpha_1 C_1 + \alpha_2 C_2 + \cdots + \alpha_t C_t$. Получаемые при этом аномалии называют «аддитивными». Если есть возможность оценить для каждого элемента C_i среднее квадратическое отклонение σ_i фонового распределения (для данного типа пород), полагают $\alpha_i = \frac{1}{\sigma_i}$ $(i = \overline{1, t})$. Если же эталонной выборки нет, ограничиваются использованием оценок модальных значений M_{0i}, получаемых непосредственно по выборке испытуемых наблюдений: $\alpha_i = \frac{1}{M_{0i}}$ $(i = \overline{1, t})$.

Критерии исключения резко выделяющихся наблюдений нормального распределения. Существует группа методов, основанных на предположении о нормальном законе распределения показателя при отсутствии аномалий. В этом случае аномальными фактически считаются те отдельные наблюдения, которые не согласуются с нормальным распределением основной массы членов выборки. Такая постановка

задачи позволяет избежать предварительной оценки фонового распределения — решение задачи осуществляется непосредственно по наблюдениям $y_1, y_2, \ldots, y_n$, среди которых нужно выявить аномальные.

Для проверки гипотезы о наличии одного аномального наблюдения, состоящей в том, что $\mathbf{M}y_1 = \mathbf{M}y_2 = \cdots = \mathbf{M}y_{r-1} = \mathbf{M}y_{r+1} = \cdots = \mathbf{M}y_n = a$, $\mathbf{M}y_r = a + d$, $d \neq 0$, удобно использовать статистику

$$\zeta_n = \max_i \frac{|y_i - \bar{y}|}{s}, \text{ где } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}. \quad (7.99)$$

Если выполняется условие $\zeta_n \geqslant \zeta_n(q)$ ($\zeta_n(q)$ — квантиль распределения ζ_n порядка q при нулевой гипотезе), то наблюдение y_r такое, что $\zeta_n = \frac{|y_r - \bar{y}|}{s}$, признают аномальным с вероятностью ошибки $\alpha = 1 - q$. Значения $\zeta_n(q)$ для $n \leqslant 50$ даны в табл. 12 (Приложение). При $n > 50$ хорошее приближение для $\zeta_n(q)$ дает формула

$$\zeta_n(q) = u_z \sqrt{\frac{2(n-1)}{(2n-5) + u_z^2}},$$

где u_z — квантиль (0; 1)-нормального распределения порядка $z = 1 - \frac{\alpha}{2n}$.

Можно применять и более простые критерии, основанные на использовании отношений $\frac{y_n - y_{n-1}}{y_n - y_1}$, $\frac{y_n - y_{n-1}}{y_n - y_2}$ и $\frac{y_n - y_{n-2}}{y_n - y_1}$ (y_i пронумерованы в порядке возрастания своих величин, $y_1 \leqslant y_2 \leqslant \cdots \leqslant y_n$). Верхние критические границы для этих отношений даны в табл. 13 (Приложение). Для выделения аномально низких наблюдений эти статистики вычисляют по значениям y_i, занумерованным в порядке убывания.

Отметим, что оценки параметров фонового распределения можно получить, опираясь на модель усеченного нормального распределения. Задав точку усечения a так, чтобы среди наблюдений $y_i \leqslant a$ ($i = \overline{1, n_a}$) наверняка не было аномальных, определяют по ним оценки математического ожидания m и среднего квадратического отклонения σ фонового распределения. Оценки имеют вид

$$\check{\sigma} = \frac{1}{n_a}\sum_{i=1}^{n_a}(a - y_i)\, g(z), \quad \check{m} = z\check{\sigma} + a, \quad (7.100)$$

где $z = f(y)$, $y = n_a \sum_{i=1}^{n_a}(a - y_i)^2 \Big/ \{2[\sum_{i=1}^{n_a}(a - y_i)]^2\}$; $f(y)$ и $g(z)$ определяются по табл. 15 (Приложение); n_a — число наблюдений, не превышающих a. Оценкой степени усечения служит $\alpha = \Phi(z)$. При усечении слева в приведенные формулы вместо $a - y_i$ подставляют $y_i - a$. Оценки $\check{m}$ и $\check{\sigma}$ асимптотически нормальны, причем их дисперсии оцениваются в виде

$$\mathbf{D}\check{m} \approx \frac{\check{\sigma}^2}{n_a}\mu_1(z), \quad \mathbf{D}\check{\sigma} \approx \frac{\check{\sigma}^2}{n_a}\mu_2(z);$$

$\mu_1(z)$ и $\mu_2(z)$ находят по табл. 15 (Приложение).

Описанные критерии применимы и для логнормального распределения, в этом случае необходимые статистики рассчитывают по логарифмам наблюдений.

Пример 7.13. По выборке 30 содержаний кремнекислоты y_i ($i = \overline{1, 30}$) получены следующие данные: $\bar{y} = \frac{1}{30}\sum_{i=1}^{30} y_i = 60\%$; $s = \sqrt{\frac{1}{30}\sum_{i=1}^{30}(y_i - \bar{y})^2} = 2{,}5\%$. Максимальное наблюдение $y_n = 68\%$; при $y_i = y_n$ модуль разности $|y_i - \bar{y}|$ также максимален. Проверить гипотезу об аномальности y_n при уровне значимости 0,05, пользуясь: 1) статистикой (7.99); 2) статистикой $\frac{y_n - y_{n-1}}{y_n - y_1}$, если наибольшее после y_n значение $y_{n-1} = 64{,}05\%$, а наименьшее — $y_1 = 56{,}35\%$.

Решение: 1) Имеем: $\zeta_n = \max_i \left(\frac{1}{s}|y_i - \bar{y}|\right) = \frac{1}{s}|y_n - \bar{y}| = \frac{8}{2{,}5} = 3{,}2$. Так как критическое значение $\zeta_n(0{,}95) = 2{,}96$ (по табл. 12, Приложение) и $\zeta_n = 3{,}2 > 2{,}96$, наблюдение $y_n = 68\%$ следует признать аномальным.

2) Так как $\frac{y_n - y_{n-1}}{y_n - y_1} = \frac{3{,}95}{11{,}65} \approx 0{,}339$ превышает критическую границу 0,26 (табл. 13, Приложение), и по этому критерию значение $y_n = 68\%$ следует признать аномальным.

Глава 8

СТАТИСТИЧЕСКИЕ МОДЕЛИ В СПЕЦИАЛЬНЫХ ЗАДАЧАХ ГЕОЛОГО-ГЕОФИЗИЧЕСКИХ ИССЛЕДОВАНИЙ

Методы, с которыми мы познакомились в предыдущих главах, базируются на основной статистической модели, используемой при анализе распределений геолого-геофизических показателей — случайной величине. Они охватывают широкий круг типовых задач статистического анализа геолого-геофизических данных. Вместе с тем, условия получения информации, специфика постановки отдельных задач могут потребовать конкретизации такой модели, учитывающей определенные факторы, либо применения других, более сложных специальных моделей.

Примером подобной конкретизации служит представление результатов измерений как наблюдений величины вида

$$\eta = \xi + \Delta, \tag{8.1}$$

где $\xi = \xi_р + \xi_л + \xi_ф$ — истинное значение показателя; $\xi_р$ — региональная составляющая, обусловленная условиями формирования, залегания и особенностями строения геологического тела, колебания которой проявляются на больших площадях или объемах; $\xi_л$ — локальная составляющая, обусловленная поздне-, послемагматическими и другими

процессами, носящими локальный характер; ξ_ϕ — величина флюктуации показателя, вызванная структурно-текстурными неоднородностями пород; Δ — взятая со знаком ошибка измерения.

Другая модель — *случайное поле* описывает поведение показателя с его функциональной (зависящей от координат в пространстве) и случайной составляющими. Ниже мы рассмотрим некоторые аспекты использования этой модели.

§ 1. Анализ ошибок измерений по контрольным пробам

В практике геологоразведочных работ распространены методы измерений, обеспечивающие высокую производительность, но и дающие немалые погрешности. Ошибки измерений могут значительно искажать функции природных распределений и отдельные статистические характеристики, что обусловливает необходимость учета этих ошибок при интерпретации данных измерений и результатов проводимого по ним статистического анализа.

Оценка распределения ошибок измерений по контрольным пробам. Будем считать, что в (8.1) Δ не зависит от ξ. Точность измерения можно выразить с помощью доверительных пределов, заключающих с заданной, близкой к единице вероятностью неизвестное истинное значение показателя ξ. Пусть A_q^-, A_q^+ — пределы допустимых с вероятностью q значений величины Δ. Так как

$$q = \mathbf{P}\{A_q^- \leqslant \Delta \leqslant A_q^+\} = \mathbf{P}\{\xi + \Delta - A_q^- \geqslant \xi \geqslant \xi + \Delta - A_q^+\} = \\ = \mathbf{P}\{\eta - A_q^- \geqslant \xi \geqslant \eta - A_q^+\}, \qquad (8.2)$$

доверительными пределами, заключающими с вероятностью q истинное значение показателя будут: нижний $\eta - A_q^+$, верхний $\eta - A_q^-$, где η — измеренное значение. Если распределение Δ нормально, то эти пределы приобретут вид: $\eta \pm u_{\frac{1+q}{2}} \sqrt{\mathbf{D}\Delta}$.

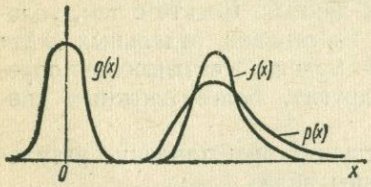


Рис. 43.

Плотность распределения величины η, наблюдениями которой являются результаты измерений, по (1.28) будет

$$p(x) = \int_{-\infty}^{\infty} f(t)\, g(x-t)\, dt, \qquad (8.3)$$

где $f(x)$ и $g(x)$ — плотности распределения истинных значений (ξ) и ошибок измерений (рис. 43). Таким образом, оценка плотности распределения $\check{p}(x)$, найденная по результатам рядовых измерений показателя, будет отклоняться от $f(x)$, причем тем сильнее, чем больше влияние ошибок. Используя оценки плотностей $\check{p}(x)$ и $\check{g}(x)$, можно поставить задачу об оценке плотности

природного распределения $f(x)$. Эта оценка $\check{f}(x)$ будет решением интегрального уравнения

$$\check{p}(x) = \int_{-\infty}^{\infty} \check{f}(t)\, \check{g}(x-t)\, dt. \qquad (8.3')$$

Уравнение такого типа часто встречается в математической статистике случайных процессов (с ними мы познакомимся ниже). Один из способов его решения основан на использовании преобразования Фурье: $\bar{p}(z) = \sqrt{2\pi}\, \bar{f}(z)\, \bar{g}(z)$, где $\bar{p}(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \check{p}(x)\, e^{-izx}\, dx$, $\bar{g}(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \check{g}(x)\, e^{-izx}\, dx$. Оценку плотности $\check{f}(x)$ находят с помощью обратного преобразования Фурье:

$$\check{f}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \bar{f}(z)\, e^{izx}\, dz = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\bar{p}(z)}{\bar{g}(z)}\, e^{izx}\, dz.$$

Другой, приближенный метод заключается в решении системы линейных уравнений по значениям функций $\check{p}(x)$ и $\check{g}(x)$ в k дискретных точках x_i^0 (с учетом того, что $\check{f}(x) \approx 0$ в области, где $\check{p}(x) \approx 0$)

$$\check{p}(x_j^0) = \sum_{i=1}^{k} \check{f}(x_i^0)\, \check{g}(x_j^0 - x_i^0)\, \Delta x, \quad j = \overline{1, k}, \qquad (8.4)$$

получаемой заменой интеграла в (8.3′) суммой Дарбу. Точки x_i^0 располагают через равные интервалы Δx в области значений, где $\check{p}(x) > 0$. Если η и Δ распределены нормально, то решение упрощается, так как тогда ξ тоже подчиняется нормальному закону, причем при $\mathbf{M}\Delta = 0$ математическое ожидание ξ оценивается средним из наблюдений η, т. е. результатов рядовых измерений $x_1, x_2, \ldots, x_n$. Дисперсия $\mathbf{D}\xi = \mathbf{D}\eta - D_0$ (1.51); $\mathbf{D}\eta$ оценивается выборочной дисперсией по $x_1, x_2, \ldots, x_n$, $D_0 = \mathbf{D}\Delta$ — дисперсия ошибок, называемая еще *дисперсией воспроизводимости* метода, которая и характеризует влияние вариации, обусловленной методикой измерений: при $2\mathbf{D}\Delta > \mathbf{D}\eta$ дисперсия ошибок будет больше, чем дисперсия самого природного распределения.

Случай ошибок, действующих в виде подверженного вариации коэффициента ζ,

$$\eta = \xi\zeta \qquad (8.5)$$

(спектральный анализ концентраций химических элементов) сводится к рассмотренному, так как $\ln\eta = \ln\xi + \ln\zeta$. Величины ξ и ζ в дальнейшем будем считать независимыми. Если $\mathbf{M}\ln\zeta = 0$, то по (8.2) доверительными пределами для $\ln\xi$ будут $\ln\eta - a_q^+$, $\ln\eta - a_q^-$, где

a_q^-, a_q^+ — пределы для $\ln\zeta$ такие, что $\mathbf{P}\{a_q^- \leqslant \ln\zeta \leqslant a_q^+\} = q$. Доверительными пределами для ξ уровня q будут $\eta e^{-a_q^+}$, $\eta e^{-a_q^-}$.

В рассматриваемом случае ошибки измерений удобно характеризовать относительными величинами $1 - e^{-a_q^+}$, $e^{-a_q^-} - 1$. Будучи умноженными на результат измерения η, они дают возможные отклонения в меньшую и большую сторону от него.

Если $\check{f}_L(x)$ — оценка плотности распределения $\ln\xi$, то плотность распределения ξ оценивается по (1.25) в виде $\check{f}(x) = \frac{1}{x}\check{f}_L(\ln x)$ $(x > 0)$. В частности, если распределения η и ζ логнормальны, величина ξ будет распределена также логнормально с плотностью (2.38), причем в (2.38)

$$\mu = \mathbf{M}\ln\xi = \mathbf{M}\ln\eta, \quad \delta^2 = \mathbf{D}\ln\xi = \mathbf{D}\ln\eta - \mathbf{D}\ln\zeta, \quad a = 0, \quad \lambda = 1.$$

Рассмотрим способы анализа распределения ошибок. При схеме (8.1) $(\eta = \xi + \Delta)$ для построения оценки плотности этого распределения используют результаты кратных независимых измерений контрольной пробы, выбираемой так, чтобы значение показателя в ней было близко к средней величине в планируемой серии измерений. Независимость кратных измерений обеспечивается одними и теми же условиями для каждого измерения, не зависящими от предыдущего результата; достаточными промежутками времени между наблюдениями, если режим работы прибора подвержен колебаниям во времени и т. д.

Оценкой дисперсии воспроизводимости метода D_0 будет

$$\check{D}_0 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2, \tag{8.6}$$

где y_i — результаты N независимых кратных измерений контрольной пробы. Если контрольная проба является эталоном, т. е. для нее известно истинное значение показателя m, то оценкой величины систематического смещения будет $\check{m}_c = \bar{y} - m$. Гипотезу $\mathbf{M}\check{m}_c = 0$ об отсутствии смещения проверяют сравнением $\check{m}_c$ с приближенными пределами $\pm u_{\frac{1+q}{2}}\sqrt{\check{D}_0/N}$, где $u_{\frac{1+q}{2}}$ — $\frac{1+q}{2}$-квантиль (0; 1)-нормального распределения порядка $\frac{1+q}{2}$. Если $|\check{m}_c| \geqslant u_{\frac{1+q}{2}}\sqrt{\check{D}_0/N}$, гипотеза об отсутствии смещения отвергается с вероятностью ошибки, близкой к $1 - q$. При отсутствии систематического смещения в качестве оценки D_0 можно взять

$$\tilde{D}_0 = \frac{1}{N}\sum_{i=1}^{N}(y_i - m)^2. \tag{8.7}$$

Если имеется t контрольных проб, то по данным их кратных измерений находят общую оценку

$$\overline{D}_0 = \frac{1}{N-t}\sum_{i=1}^{t}(k_i - 1)\check{D}_{0i}, \tag{8.8}$$

где k_i — число измерений i-й пробы, $\check{D}_{0i}$ — вычисленная по ним выборочная дисперсия вида (8.6) ($i = \overline{1, t}$); $N = k_1 + k_2 + \ldots + k_t$.

В дальнейшем будем считать метод измерений несмещенным, полагая, что в рядовые результаты внесена необходимая поправка — из наблюдений вычтена величина систематического смещения, если таковая отличается от нуля. При статистическом анализе данных без такой поправки следует иметь в виду, что его результаты относятся к показателям, выраженным в условных единицах — с изменённым началом отсчета.

При нормальном распределении ошибок доверительные пределы для истинного значения показателя ξ, результат измерения которого равен η, определяются в виде

$$\eta - u_{\frac{1+q}{2}}\sqrt{\overline{D}_0}, \quad \eta + u_{\frac{1+q}{2}}\sqrt{\overline{D}_0}, \tag{8.9}$$

где $u_{\frac{1+q}{2}}$ — квантиль (0; 1)-нормального распределения порядка $\frac{1+q}{2}$. При доверительной вероятности q погрешность составит $\pm u_{\frac{1+q}{2}}\sqrt{\overline{D}_0}$.

Подставив оценку плотности распределения ошибок измерений $\check{g}(x) = \frac{1}{\sqrt{2\pi\check{D}_0}}\exp\left(-\frac{1}{2}\frac{x^2}{\check{D}_0}\right)$ в (8.3′), получим уравнение для определения оценки плотности природного распределения. При нормальном распределении результатов измерений эта оценка имеет вид

$$\check{f}(x) = \frac{1}{\sqrt{2\pi(\check{D} - \check{D}_0)}}\exp\left[-\frac{1}{2}\frac{(x-\overline{x})^2}{\check{D} - \check{D}_0}\right], \tag{8.10}$$

где $\check{D}$ — оценка дисперсии по выборке $x_1, x_2, \ldots, x_n$ ($\check{D} > \check{D}_0$).

При схеме (8.5) с логнормальным распределением ошибок, при $\mu_0 = \mathbf{M}\ln\zeta = 0$, доверительные пределы для ξ составят

$$\eta a_q^{-1}, \quad \eta a_q \quad \left(a_q = \exp\left(u_{\frac{1+q}{2}}\delta_0\right)\right), \tag{8.11}$$

где δ_0^2 — дисперсия $\ln\zeta$, оцениваемая по логарифмам результатов контрольных измерений. Относительная погрешность, соответствующая доверительной вероятности q: вверх от η — $(a_q - 1)\,100\%$, вниз от η — $(1 - a_q^{-1})\,100\%$; средняя относительная погрешность — $0{,}5\,(a_q - a_q^{-1})\,100\%$.

Логарифмически нормальный закон распределения ошибок не приводит к систематическому смещению лишь при условии $\mu_0 + \frac{1}{2}\delta_0^2 = 0$. При фиксированном значении ξ, согласно (2.39), математическое ожидание результата измерения

$$\mathbf{M}\eta = \xi \mathbf{M}\zeta = \xi \exp\left(\mu_0 + \frac{1}{2}\delta_0^2\right). \tag{8.12}$$

Наиболее вероятное значение результата измерения будет, по (2.41),

$$M_{0\eta} = \xi M_{0\zeta} = \xi \exp(\mu_0 - \delta_0^2). \tag{8.13}$$

Если $\mu_0 + \frac{1}{2}\delta_0^2 \neq 0$, систематическое смещение проявляется в изменении масштаба измерительной шкалы. Для учета этого смещения результат измерения надо умножить на коэффициент $k_0 = \exp\left(-\mu_0 - \frac{1}{2}\delta_0^2\right)$, оценив μ_0 и δ_0^2 по контрольным измерениям эталона. Если такое смещение не учитывается, результаты измерений будут выражены в условных единицах — с измененным масштабом измерительной шкалы.

В рассмотренной схеме, для данного метода и интервала измеряемых значений, δ_0 предполагается постоянным; при этом относительная погрешность также постоянна (не зависит от измеряемых значений). При схеме (8.1) в предположении $D_0 = \text{const}$ постоянна абсолютная погрешность. Эта особенность может служить критерием для выбора одной из схем.

Нередко бывает, что погрешность не подчиняется ни той, ни другой схеме, хотя обычно среднее квадратическое отклонение ошибок измерений в достаточно широких интервалах линейно зависит от измеряемых значений. Чтобы получить оценку этой зависимости, поступают следующим образом. Выбирают группу контрольных образцов со значениями показателя, охватывающими интервал возможных значений в планируемой серии измерений. Каждый образец измеряется несколько раз, и по каждой группе наблюдений вычисляется оценка дисперсии. Эти данные позволяют построить зависимость среднего квадратического отклонения ошибок от измеряемых значений, с помощью которой и устанавливаются величины погрешности для различных ξ.

Если вариация показателя на изучаемом объекте невелика, вполне можно ограничиться схемой (8.1), так как обычно зависимость дисперсии ошибок от измеряемых значений проявляется лишь на больших интервалах. Конечно, используемые контрольные пробы должны быть характерными для этого объекта.

Пример 8.1. С целью изучения распределения ошибок спектральных определений концентраций Cr сделано 40 независимых (разделенных длительным промежутком времени) контрольных измерений y_i ($i = \overline{1, 40}$) пробы гранита. По

логарифмам результатов контрольных измерений вычислены выборочные среднее квадратическое отклонение коэффициенты асимметрии и эксцесса:

$$\check{\delta}_0 = \sqrt{\frac{1}{39}\sum_{i=1}^{40}(\ln y_i - \check{\mu}_0)^2} = 0{,}15, \quad \check{A}_n = 0{,}29, \quad \check{E}_n = 0{,}40.$$

1) Проверить гипотезу о логнормальном распределении ошибок при уровне значимости 0,05. 2) Выяснить вид доверительных пределов для истинных концентраций при доверительной вероятности $q = 0{,}90$, полагая метод несмещенным по логарифмам результатов измерений: $\mu_0 = \mathbf{M} \ln \zeta = 0$.

Решение. 1) Приближенные критические границы при гипотезе о нормальном распределении $\ln y_i$ составят, по (4.30): для $\check{A}_n - \pm A_q = \pm u_{0{,}975}\sqrt{\frac{6}{40}} = \pm 1{,}96\sqrt{\frac{6}{40}} \approx \pm 0{,}76$; для $\check{E}_n - \pm E_q = \pm 2A_q \approx \pm 1{,}52$. Так как $|\check{A}_n| < A_q$, $|\check{E}_n| < E_q$, гипотеза о логнормальном распределении ошибок принимается.

2) Воспользуемся формулой (8.11): $a_{0,9} = \exp\left(u_{\frac{1+q}{2}}\check{\delta}_0\right) = e^{1{,}645 \cdot 0{,}15} \approx 1{,}28$. Доверительные пределы уровня 0,9 для истинной концентрации ξ при результате измерения, равном η, составят: $\frac{\eta}{1{,}28}$, $1{,}28\eta$.

Оценка точности расчетных параметров. Пусть некоторый параметр ξ является дифференцируемой функцией m измеряемых показателей $\xi_1, \xi_2, \ldots, \xi_m$, так что его значение η вычисляется в виде $\eta = f(\eta_1, \eta_2, \ldots, \eta_m)$, где η_i — результат измерения ξ_i $(i = \overline{1, m})$. Обозначим $\Delta_i = \eta_i - \xi_i$; $D_{0i} = \mathbf{M}\Delta_i^2$ — дисперсия отклонения η_i от измеряемой величины. Считая $\Delta_1, \Delta_2, \ldots, \Delta_m$ малыми, имеем приближенно:

$$\eta = f(\xi_1 + \Delta_1, \xi_2 + \Delta_2, \ldots, \xi_m + \Delta_m) \approx f(\xi_1, \xi_2, \ldots, \xi_m) + \sum_{i=1}^{m}\frac{\partial f}{\partial \xi_i}\Delta_i =$$

$$= \xi + \sum_{i=1}^{m}\frac{\partial f}{\partial \xi_i}\Delta_i.$$

Дисперсия η при фиксированном ξ, с учетом $\mathbf{M}\Delta_i = 0$ $(i = \overline{1, t})$,

$$\mathbf{D}\eta \approx \mathbf{M}(\eta - \xi)^2 \approx \mathbf{M}\left(\sum_{i=1}^{m}\frac{\partial f}{\partial \xi_i}\Delta_i\right)^2 = \sum_{i,\ j=1}^{m}\frac{\partial f}{\partial \xi_i}\frac{\partial f}{\partial \xi_j}\mathbf{M}(\Delta_i\Delta_j) =$$

$$= \sum_{i,\ j=1}^{m}\frac{\partial f}{\partial \xi_i}\frac{\partial f}{\partial \xi_j} r_{ij}^0 \sqrt{D_{0i}D_{0j}}, \qquad (8.14)$$

где r_{ij}^0 — коэффициент корреляции Δ_i и Δ_j. Точность определения ξ характеризуют доверительные пределы, которые при нормальном распределении $\Delta_1, \Delta_2, \ldots, \Delta_m$ оцениваются в виде

$$\eta - u_{\frac{1+q}{2}}\sqrt{\check{\mathbf{D}}\eta}, \quad \eta + u_{\frac{1+q}{2}}\sqrt{\check{\mathbf{D}}\eta}, \text{ где } \check{\mathbf{D}}\eta = \sum_{i,\ j=1}^{m}\frac{\partial f}{\partial \eta_i}\frac{\partial f}{\partial \eta_j} r_{ij}^0 \sqrt{D_{0i}D_{0j}}. \qquad (8.15)$$

Если Δ_i некоррелированы — $\mathbf{M}(\Delta_i\Delta_j) = 0$ $(i, j = \overline{1, m}, i \neq j)$, то

$$\check{\mathbf{D}}\eta = \sum_{j=1}^{m} \left(\frac{\partial f}{\partial \eta_j}\right)^2 D_{0j}. \qquad (8.14')$$

Задача о невязке. При одновременном измерении группы показателей $\xi_1, \xi_2, \ldots, \xi_m$, связанных некоторым соотношением, иногда возникает *задача о невязке* — исправлении результатов измерений в соответствии с этим соотношением. Например, при химическом анализе образцов пород сумма содержаний всех окислов каждой пробы должна составлять 100% — $\sum_{i=1}^{m} \xi_i = 100\%$, тогда как сумма результатов измерений $\sum_{i=1}^{m} \eta_i$ отклоняется в ту или иную сторону от 100%.

Рассмотрим этот случай. Обозначим, как и раньше, Δ_i — отклонение результата измерения η_i от истинного значения ξ_i: $\eta_i = \xi_i + \Delta_i$. Пусть для любых возможных значений $\xi_1, \xi_2, \ldots, \xi_m$, $\sum_{i=1}^{m} \xi_i = A$. Так как $\sum_{i=1}^{m} \eta_i = \sum_{i=1}^{m} \xi_i + \sum_{i=1}^{m} \Delta_i = A + \sum_{i=1}^{m} \Delta_i$, невязка суммы результатов измерений η_i составит $\Delta = \sum_{i=1}^{m} \Delta_i = \sum_{i=1}^{m} \eta_i - A$. Введем для каждого измерения η_i поправки вида $\alpha_i\Delta$ $(\alpha_i \geqslant 0, i = \overline{1, m})$ так, чтобы сумма исправленных величин $\tilde{\eta}_i = \eta_i - \alpha_i\Delta$ составила A, $\sum_{i=1}^{m} \tilde{\eta}_i = A$, а сумма дисперсий отклонений $\tilde{\eta}_i$ от истинных величин ξ_i была минимальна. Считая Δ_i некоррелированными — $\mathbf{M}(\Delta_i\Delta_j) = 0$ $(i, j = \overline{1, m}, \quad i \neq j)$, имеем:

$$\mathbf{M}\left[\sum_{i=1}^{m} (\tilde{\eta}_i - \xi_i)^2\right] = \mathbf{M}\left[\sum_{i=1}^{m} \left(\xi_i + \Delta_i - \alpha_i \sum_{j=1}^{m} \Delta_j - \xi_i\right)^2\right] =$$

$$= \mathbf{M}\left[\sum_{i=1}^{m} \left(\Delta_i^2 - 2\alpha_i\Delta_i \sum_{j=1}^{m} \Delta_j + \alpha_i^2 \sum_{j=1}^{m} \Delta_j^2\right)\right] =$$

$$= \sum_{i=1}^{m} \left(D_{0i} - 2\alpha_i D_{0i} + \alpha_i^2 \sum_{j=1}^{m} D_{0j}\right),$$

где D_{0i} — дисперсия Δ_i. Приравняв производные по α_i нулю, получим

$$\alpha_i = D_{0i}\left(\sum_{j=1}^{m} D_{0j}\right)^{-1} \quad (i = \overline{1, m}).$$

Эти коэффициенты удовлетворяют необходимому условию $\sum_{i=1}^{m} \alpha_i =$

$= 1$. Поправка измерения η_i составит часть невязки, прямо пропорциональную дисперсии ошибок измерений D_{0i}:

$$\delta_i = \Delta D_{0i} \left(\sum_{j=1}^{m} D_{0j}\right)^{-1} = \left(\sum_{i=1}^{m} \eta_i - A\right) D_{0i} \left(\sum_{j=1}^{m} D_{0j}\right)^{-1}. \qquad (8.16)$$

Для приведения невязки к нулю необходимо из каждого η_i вычесть δ_i.

Пример 8.2 Независимо друг от друга измерены теплофизические характеристики образца: удельная теплоемкость $C = 0{,}18 \frac{кал}{г \cdot град}$; температуропроводность $a = 0{,}014 \frac{см^2}{сек}$; плотность $\rho = 2{,}65 \frac{г}{см^3}$; теплопроводность $\lambda = 6 \cdot 10^{-3} \frac{кал}{см \cdot сек \cdot град}$. Средние квадратические отклонения ошибок измерений этих параметров составляют: $\sigma_C = 0{,}005 \frac{кал}{г \cdot град}$, $\sigma_a = 1{,}5 \cdot 10^{-3} \frac{см^2}{сек}$, $\sigma_\rho \approx 0$, $\sigma_\lambda = 0{,}4 \cdot 10^{-3} \frac{кал}{см \cdot сек \cdot град}$. Внести поправки в результаты измерений, используя соотношение $Ca\rho = \lambda$.

Решение. В соответствии с указанной зависимостью $\lg C + \lg a + \lg \rho - \lg \lambda = 0$. По формуле (8.14) дисперсия $\lg C$, $\mathbf{D}(\lg C) \approx \lg^2 e \cdot \frac{1}{C^2} \sigma_C^2 \approx (0{,}434)^2 \times \frac{25 \cdot 10^{-6}}{324 \quad 10^{-4}} \approx 1{,}45 \cdot 10^{-4}$. Аналогично, $\mathbf{D}(\lg a) \approx 21{,}62 \cdot 10^{-4}$; $\mathbf{D}(\lg \rho) = 0$; $\mathbf{D}(\lg \lambda) \approx 8{,}37 \cdot 10^{-4}$. Сумма этих дисперсий $S \approx 31{,}44 \cdot 10^{-4}$. Невязка составляет $\Delta = \lg 0{,}18 + \lg 0{,}014 + \lg 2{,}65 - \lg 0{,}006 \approx -0{,}745 - 1{,}854 + 0{,}423 + 2{,}222 = 0{,}046$. Поправки, по (8.16): для $\lg C$ — $\delta_1 = \Delta \mathbf{D}(\lg C) S^{-1} = 0{,}002$; для $\lg a$ — $\delta_2 = \Delta \mathbf{D}(\lg a) S^{-1} \approx 0{,}032$; для $\lg \rho$ — $\delta_3 = 0$; для $\lg \lambda$ — $\delta_4 \approx -0{,}012$. Исправленные значения: $C = 10^{\lg 0{,}18 - \delta_1} \approx 0{,}179$; $a = 10^{\lg 0{,}014 - \delta_2} \approx 0{,}013$; $\rho = 2{,}65$; $\lambda = 10^{\lg 0{,}006 - \delta_4} = 6{,}17 \cdot 10^{-3}$.

§ 2. Расчет числовых характеристик распределений геолого-геофизических показателей пород с учетом ошибок измерений

Если дисперсия ошибок измерений сопоставима с дисперсией изучаемого распределения, расчет числовых характеристик, интерпретируемых в дальнейшем в качестве параметров распределения действительных значений показателя, необходимо проводить с учетом этих ошибок.

Числовые характеристики одномерных распределений. В соответствии со схемой (8.1), оценкой дисперсии ξ истинных значений показателя будет

$$\check{D}_\xi = \check{D}_\eta - \check{D}_0, \qquad (8.17)$$

где $\check{D}_\eta$ — оценка, получаемая по исходной выборке $x_1, x_2, \ldots, x_n$: $\check{D}_\eta = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$; $\check{D}_0$ — оценка дисперсии ошибок ($\check{D}_\eta > D_0$). Если эта

оценка получена по N независимым контрольным измерениям одной и той же пробы (8.6), то точность оценки природной дисперсии (8.17) можно определить в виде приближенных доверительных пределов:

$$D_\xi^- = \check{D}_\xi - u_{\frac{1+q}{2}} \sqrt{\mathbf{D}(\check{D}_\xi)}, \quad D_\xi^+ = \check{D}_\xi + u_{\frac{1+q}{2}} \sqrt{\mathbf{D}(\check{D}_\xi)}, \qquad (8.18)$$

где, по (4.17'), $\mathbf{D}(\check{D}_\xi) = \mathbf{D}(\check{D}_\eta) + \mathbf{D}(\check{D}_0) \approx \frac{1}{n}(\mu_{4\eta} - D_\eta^2) + \frac{1}{N}(\mu_{40} - D_0^2)$; $\mu_{4\eta}$ и μ_{40} – четвертые центральные моменты η и Δ соответственно; $u_{\frac{1+q}{2}}$ — квантиль порядка $\frac{1+q}{2}$ (0; 1)-нормального распределения *.

При нормальном законе распределения η и Δ $\mathbf{D}(\check{D}_\xi) = \frac{2D_\eta^2}{n} + \frac{2D_0^2}{N}$ и относительная погрешность оценки $\check{D}_\xi$ при доверительной вероятности q —

$$a_q(\check{D}_\xi) = u_{\frac{1+q}{2}} \sqrt{\frac{2}{n}(1+k^2) + \frac{2}{N}k^2} \cdot 100\% \qquad \left(k = \frac{D_0}{D_\eta - D_0}\right). \qquad (8.19)$$

Оценка среднего квадратического отклонения ξ —

$$s_\xi = \sqrt{\check{D}_\eta - \check{D}_0}. \qquad (8.20)$$

Точность этой оценки можно определить также через доверительные пределы, вычислив ее дисперсию, по аналогии с (4.22), в виде

$$\mathbf{D}(s_\xi) = \frac{1}{D_\eta - D_0}\left(\frac{\mu_{4\eta} - D_\eta^2}{4n} + \frac{\mu_{40} - D_0^2}{4N}\right) = \frac{1}{D_\eta - D_0}\left(\frac{D_\eta^2}{2n} + \frac{D_0^2}{2N}\right). \qquad (8.21)$$

Оценка математического ожидания ξ ввиду $\mathbf{M}\eta = \mathbf{M}\xi$ остается без изменений: $\check{\mathbf{M}}\xi = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Оценка коэффициента вариации — $\check{V}_\xi = \frac{s_\xi}{\bar{x}}$; коэффициента асимметрии —

$$\check{A}_\xi = \check{A}\left(\frac{\check{D}_\eta}{\check{D}_\xi}\right)^{\frac{3}{2}} - A_0\left(\frac{\check{D}_0}{\check{D}_\xi}\right)^{\frac{3}{2}}, \qquad (8.22)$$

где $\check{A}$ — оценка коэффициента асимметрии вида (6.22), вычисляемая по наблюдениям x_i; A_0 — коэффициент асимметрии распределения ошибок измерений. Формула (8.22) следует из того, что третий центральный момент

$$\mu_{3\eta} = A\sigma_\eta^3 = \mathbf{M}(\eta - \mathbf{M}\eta)^3 = \mathbf{M}(\xi + \Delta - \mathbf{M}\xi)^3 = \mu_{3\xi} + \mu_{30} = A_\xi\sigma_\xi^3 + A_0\sigma_0^3,$$

где $\mu_{3\xi}$ — третий центральный момент распределения ξ; μ_{30} — третий центральный момент распределения ошибок; $\sigma_0 = \sqrt{D_0}$.

* Здесь и далее n и N предполагаются достаточно большими.

При схеме (8.5) с логарифмически нормальным законом распределения коэффициента ζ, интерпретирующего случайные погрешности измерений, будут испытывать смещение не только среднее квадратическое отклонение, коэффициенты вариации и асимметрии, но и математическое ожидание. В соответствии с (1.47) и (2.39)

$$\mathbf{M}\eta = \mathbf{M}\xi\mathbf{M}\zeta = \mathbf{M}\xi \exp\left(\mu_0 + \frac{1}{2}\delta_0^2\right) (\mu_0 = \mathbf{M}\ln\zeta,\ \delta_0^2 = \mathbf{D}\ln\zeta). \quad (8.23)$$

Будем считать, что измерительная шкала приведена в соответствие с условием $\mu_0 = 0$. В качестве оценки математического ожидания природного распределения следует взять

$$\check{\mathbf{M}}\xi = \bar{x} \exp\left(-\frac{1}{2}\check{\delta}_0^2\right); \quad (8.24)$$

$\check{\delta}_0^2$ вычисляется по контрольным измерениям —

$$\check{\delta}_0^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\ln y_i - \check{\mu}_0)^2, \quad \check{\mu}_0 = \frac{1}{N}\sum_{i=1}^{N}\ln y_i.$$

Так как $\mathbf{M}\ln\eta = \mathbf{M}\ln\xi + \mathbf{M}\ln\zeta = \mathbf{M}\ln\xi$, оценкой математического ожидания $\ln\xi$ будет $\check{\mu} = \frac{1}{n}\sum_{i=1}^{n}\ln x_i$, оценкой дисперсии $\ln\xi$ — $\check{\delta}^2 = \frac{1}{n}\sum_{i=1}^{n}(\ln x_i - \check{\mu})^2 - \check{\delta}_0^2$ (при $\check{\delta}^2 > \check{\delta}_0^2$). Эти величины можно использовать, например, для расчета квантилей природного распределения.

Дисперсия наблюденных значений η

$$D_\eta = \mathbf{M}(\xi\zeta - \mathbf{M}\xi\mathbf{M}\zeta)^2 = \mathbf{M}(\xi\zeta - \zeta\mathbf{M}\xi + \zeta\mathbf{M}\xi - \mathbf{M}\xi\mathbf{M}\zeta)^2 = \mathbf{M}\zeta^2\mathbf{D}\xi + \\ + (\mathbf{M}\xi)^2\mathbf{D}\zeta = \mathbf{D}\xi\mathbf{D}\zeta + (\mathbf{M}\xi)^2\mathbf{D}\zeta + \mathbf{D}\xi(\mathbf{M}\xi)^2, \quad (8.25)$$

поэтому оценка дисперсии ξ

$$\check{D}_\xi = \frac{\check{D}_\eta - \check{D}_\zeta(\check{\mathbf{M}}\xi)^2}{\check{D}_\zeta + (\check{\mathbf{M}}\zeta)^2} = \frac{\check{D}_\eta - \check{D}_\zeta(\check{\mathbf{M}}\xi)^2}{(\check{\mathbf{M}}\zeta)^4} \quad (8.26)$$

(используя (2.40), $\check{\mathbf{D}}_\zeta = \exp(2\check{\delta}_0^2) - \exp\check{\delta}_0^2 = (\check{\mathbf{M}}\zeta)^4 - (\check{\mathbf{M}}\zeta)^2$; $\check{\mathbf{M}}\zeta = \exp\frac{\check{\delta}_0^2}{2}$); предполагается, что $\check{D}_\eta \geqslant \check{D}_\xi(\mathbf{M}\xi)^2$.

Оценка коэффициента вариации ξ

$$\check{V}_\xi = \sqrt{\frac{\check{D}_\eta - \check{D}_\zeta(\check{\mathbf{M}}\xi)^2}{(\check{\mathbf{M}}\xi)^2[D_\zeta + (\check{\mathbf{M}}\zeta)^2]}} = \sqrt{\frac{\check{D}_\eta(\check{\mathbf{M}}\xi)^{-2}(\check{\mathbf{M}}\zeta)^{-2} - \check{V}_\zeta^2}{\check{V}_\zeta^2 + 1}} = \frac{\sqrt{\check{V}_\eta^2 - \check{V}_\zeta^2}}{\check{\mathbf{M}}\zeta}, \quad (8.27)$$

где $\check{V}_\zeta$ — оценка коэффициента вариации ошибок измерений ζ: $\check{V}_\zeta = \sqrt{\check{D}_\zeta}(\check{\mathbf{M}}\zeta)^{-1} = \sqrt{(\check{\mathbf{M}}\zeta)^2 - 1}$.

Оценка коэффициента асимметрии природного распределения —

$$\check{A}_\xi = \frac{\check{A}_\eta (\check{V}_\xi^2 \check{V}_\zeta + \check{V}_\zeta^2 + \check{V}_\xi^2)^{\frac{3}{2}} - \check{A}_\zeta \check{V}_\zeta^3 (1 + 3\check{V}_\xi^2) - 6\check{V}_\xi^2 \check{V}_\zeta^2}{\check{V}_\xi^3 (\check{A}_\zeta \check{V}_\zeta^3 + 3\check{V}_\zeta^2 + 1)}, \qquad (8.28)$$

где $\check{A}_\eta$ — оценка коэффициента асимметрии величины η, вычисляемая непосредственно по наблюдениям x_i; $\check{A}_\zeta$ — оценка коэффициента асимметрии ошибок измерений ζ, вычисляемая по (2.44).

Если метод измерений имеет определенный порог чувствительности a, при расчете оценок параметров распределения η могут возникнуть трудности из-за того, что отдельные данные, оказавшись ниже порога чувствительности, не имеют точного количественного выражения. Если распределение η предполагается нормальным, то используя формулы (7.100), $\mathbf{M}\eta$ и $\mathbf{D}\eta$ можно оценить по схеме усеченного нормального распределения. Использование оценки степени усечения $\alpha = \frac{n_0}{n}$ (n_0 — количество наблюдений, оказавшихся меньше a), дает возможность получить уточненные оценки с помощью таблиц для односторонне не полностью определенной выборки из нормальной совокупности [28]. Подобным образом можно действовать и при логнормальном распределении η, выполняя необходимые расчеты по логарифмам наблюдений.

Учет ошибок измерений при анализе парных связей. Пусть ξ_1 и ξ_2 — два показателя, η_1 и η_2 — результаты их измерений; σ_1 и σ_2 — средние квадратические отклонения η_1 и η_2; σ_{01} и σ_{02} — средние квадратические отклонения ошибок измерений при схеме (8.1) для обоих показателей — $\eta_1 = \xi_1 + \Delta_1$, $\eta_2 = \xi_2 + \Delta_2$, причем Δ_1 и Δ_2 не зависят от ξ_1 и ξ_2; $\sigma_{\xi 1} = \sqrt{\sigma_1^2 - \sigma_{01}^2}$ и $\sigma_{\xi 2} = \sqrt{\sigma_2^2 - \sigma_{02}^2}$ — средние квадратические отклонения ξ_1 и ξ_2 (природных распределений); R — коэффициент корреляции η_1 и η_2; r — коэффициент природной корреляции (ξ_1 и ξ_2); r_0 — коэффициент корреляции ошибок измерений Δ_1 и Δ_2. Так как $R\sigma_1\sigma_2 = \mathbf{M}[(\eta_1 - \mathbf{M}\eta_1)(\eta_2 - \mathbf{M}\eta_2)] = \mathbf{M}[(\xi_1 + \Delta_1 - \mathbf{M}\xi_1) \times (\xi_2 + \Delta_2 - \mathbf{M}\xi_2)] = r\sigma_{\xi 1}\sigma_{\xi 2} + r_0\sigma_{01}\sigma_{02}$, R имеет вид

$$R = \frac{r\sigma_{\xi 1}\sigma_{\xi 2} + r_0\sigma_{01}\sigma_{02}}{\sigma_1\sigma_2}. \qquad (8.29)$$

Таким образом, коэффициент корреляции двух показателей должен оцениваться с учетом не только дисперсий ошибок измерений каждого, но и корреляции этих ошибок. В качестве оценки коэффициента корреляции природных распределений можно взять

$$\check{r} = \frac{\check{R}s_1s_2 - \check{r}_0 s_{01}s_{02}}{\sqrt{(s_1^2 - s_{01}^2)(s_2^2 - s_{02}^2)}}, \qquad (8.30)$$

использовав оценки величин R, σ_{0j} и σ_j:

$$\check{R} = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)}{s_1 s_2}; \qquad s_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2},$$

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n}x_{ij}, \quad s_{0j} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_{ij}-\bar{y}_j)^2}, \quad \bar{y}_j = \frac{1}{N}\sum_{i=1}^{N}y_{ij}, \quad j = 1, 2; \tag{8.31}$$

$$\check{r}_0 = \frac{1}{N}\sum_{i=1}^{N}\frac{(y_{i1}-\bar{y}_1)(y_{i2}-\bar{y}_2)}{s_{01}s_{02}}.$$

Здесь x_{i1} и x_{i2} — наблюдения η_1 и η_2 $(i = \overline{1, n})$; y_{i1} и y_{i2} — результаты измерений показателей ξ_1 и ξ_2 по контрольной пробе, взятой для оценки ошибок измерений $(i = \overline{1, N})$; $s_j^2 > s_{0j}^2$.

Как видно из (8.29), при некоррелированных ошибках измерений Δ_1 и Δ_2 $(r_0 = 0)$ обычная оценка коэффициента корреляции будет заниженной по абсолютной величине. Если дисперсия ошибок измерений значительно превышает природную — $\sigma_{0j} > \sigma_{\xi j}$, оценка $\check{R}$ будет определяться главным образом корреляцией ошибок измерений, а не природной связью ξ_1 и ξ_2. В этом случае, впрочем, исправленную оценку (8.30) можно применять лишь при больших n и N.

Ошибки измерений могут исказить и форму связи. При схеме (8.1) для η_1 и η_2, если природная связь между ξ_1 и ξ_2 линейна, т. е. $\xi_1 = \alpha\xi_2 + \beta + \varphi_1$ (φ_1 — не зависящая от ξ_2 случайная величина с нулевым средним), в качестве оценки α следует взять

$$\tilde{\alpha} = \frac{\check{r}s_{\xi 1}}{s_{\xi 2}} = \frac{\check{r}\sqrt{s_1^2 - s_{01}^2}}{\sqrt{s_2^2 - s_{02}^2}} = \frac{\check{R}s_1 s_2 - \check{r}_0 s_{01}s_{02}}{s_2^2 - s_{02}^2}. \tag{8.32}$$

Если $\check{\alpha} = \check{R}s_1 s_2^{-1}$ — обычная оценка, то $\tilde{\alpha} = (\check{\alpha}s_2^2 - \check{r}_0 s_{01}s_{02})(s_2^2 - s_{02}^2)^{-1}$, так что при некоррелированных ошибках измерений $(r_0 = 0)$ обычная оценка коэффициента α будет заниженной по абсолютной величине. Оценка β вычисляется в виде

$$\tilde{\beta} = \bar{x}_1 - \tilde{\alpha}\bar{x}_2. \tag{8.33}$$

Для косвенных измерений (показателя ξ_1 по ξ_2) необходимо использовать регрессию, связывающую истинные значения одного (ξ_1) с результатами измерений другого (η_2). Коэффициент корреляции ξ_1 и η_2

$$r' = \mathbf{M}[(\xi_1 - \mathbf{M}\xi_1)(\eta_2 - \mathbf{M}\eta_2)](\sigma_{\xi 1}\sigma_2)^{-1} = \mathbf{M}[(\xi_1 - \mathbf{M}\xi_1) \times$$

$$\times(\xi_2 + \Delta_2 - \mathbf{M}\xi_2)](\sigma_{\xi 1}\sigma_2)^{-1} = \frac{r\sigma_{\xi 2}}{\sigma_2}. \tag{8.34}$$

Коэффициенты линейной регрессии $\xi_1 = a\eta_2 + b$ оцениваются в виде

$$a = \check{r}'\frac{s_{\xi 1}}{s_2} = \frac{\check{r}s_{\xi 1}s_{\xi 2}}{s_2^2}, \qquad \check{b} = \bar{x}_1 - \check{a}\bar{x}_2. \tag{8.35}$$

Оценка среднего квадратического отклонения ошибки косвенного измерения

$$\check{\sigma}_p = s_{\xi 1}\sqrt{1-(\check{r}')^2} = \sqrt{(s_1^2 - s_{01}^2)(1-\check{r}'^2)}. \qquad (8.36)$$

При схеме (8.5) для обоих показателей, $\eta_1 = \xi_1\zeta_1$ и $\eta_2 = \xi_2\zeta_2$ с логнормальным распределением ζ_1 и ζ_2, коэффициент корреляции η_1 и η_2 имеет вид

$$R = (r\sigma_{\xi 1}\sigma_{\xi 2}r_0\sigma_{\zeta 1}\sigma_{\zeta 2} + r\sigma_{\xi 1}\sigma_{\xi 2}\mathbf{M}\zeta_1\mathbf{M}\zeta_2 + r_0\sigma_{\zeta 1}\sigma_{\zeta 2}\mathbf{M}\xi_1\mathbf{M}\xi_2)(\sigma_1\sigma_2)^{-1},$$

где $\sigma_{\zeta i}$ — среднее квадратическое отклонение ζ_i $(i = 1, 2)$; r_0 — коэффициент корреляции ζ_1 и ζ_2 (ζ_i и ξ_i предполагаются независимыми). Таким образом, в качестве оценки коэффициента природной корреляции можно взять

$$\check{r} = \frac{\check{R}\check{\sigma}_1\check{\sigma}_2 - \check{r}_0\check{\sigma}_{\zeta 1}\check{\sigma}_{\zeta 2}\check{\mathbf{M}}\xi_1\check{\mathbf{M}}\xi_2}{\check{\sigma}_{\xi 1}\check{\sigma}_{\xi 2}(\check{r}_0\check{\sigma}_{\zeta 1}\check{\sigma}_{\zeta 2} + \check{\mathbf{M}}\zeta_1\check{\mathbf{M}}\zeta_2)}$$

с использованием оценок вида (8.24), (8.26) и соблюдая условие $|\check{r}| \leqslant 1$. Оценка коэффициента корреляции ζ_1 и ζ_2 вычисляется по данным измерений контрольной пробы:

$$\tilde{r}_0 = \frac{\exp(\check{\rho}_0\check{\delta}_{01}\check{\delta}_{02}) - 1}{\sqrt{(\exp\check{\delta}_{01}^2 - 1)(\exp\check{\delta}_{02}^2 - 1)}} = \frac{\exp(\check{\rho}_0\check{\delta}_{01}\check{\delta}_{02}) - 1}{\sqrt{((\check{\mathbf{M}}\zeta_1)^2 - 1)((\check{\mathbf{M}}\zeta_2)^2 - 1)}}, \qquad (8.37)$$

где $\check{\delta}_{0j}^2 = \frac{1}{N}\sum_{i=1}^{N}(\ln y_{ij} - \check{\mu}_{0j})^2$, $\check{\mu}_{0j} = \frac{1}{N}\sum_{i=1}^{N}\ln y_{ij}$, $j = 1, 2$; y_{i1}, y_{i2} — результаты измерений показателей ξ_1 и ξ_2 по контрольной пробе $(i = \overline{1, N})$; $\check{\rho}_0$ — выборочный коэффициент корреляции, вычисленный по $\ln y_{i1}$, $\ln y_{i2}$ (предполагается, что $\mu_{0j} = \mathbf{M}\ln\zeta_j = 0$).

Пример 8.3. Для изучения связи между намагниченностью насыщения I_s и концентрацией магнетита C_{mt} проведены независимые измерения этих показателей по выборке образцов. По данным измерений вычислены средние арифметические: C_{mt} — $\bar{x}_1 = 0{,}075\%$, I_s — $\bar{x}_2 = 70 \cdot 10^{-3}\frac{эс}{г}$; выборочные средние квадратические отклонения: C_{mt} — $s_1 = 0{,}033\%$, I_s — $s_2 = 35 \cdot 10^{-3}\frac{эс}{г}$. Выборочный коэффициент корреляции составил $\check{R} = 0{,}55$. Оценки средних квадратических отклонений контрольного образца составили: C_{mt} — $s_{01} = 0{,}019\%$; I_s — $s_{02} = 25 \cdot 10^{-3}\frac{эс}{г}$. Вычислить: 1) оценки средних квадратических отклонений $s_{\xi 1}$, $s_{\xi 2}$ истинных значений $C_{mt}^{(u)}$ и $I_s^{(u)}$; 2) оценку $\check{r}$ коэффициента корреляции $C_{mt}^{(u)}$ и $I_s^{(u)}$ и уравнение регрессии $C_{mt}^{(u)}$ на $I_s^{(u)}$, полагая ошибки измерений C_{mt} и I_s некоррелированными; 3) построить уравнение регрессии для косвенных измерений $C_{mt}^{(u)}$ по I_s с учетом ошибок измерений.

Решение. 1) Оценки средних квадратических отклонений $s_{\xi 1}$, $s_{\xi 2}$, по (8.20):

$$s_{\xi 1} = \sqrt{0{,}033^2 - 0{,}019^2} \approx 0{,}027\%. \quad s_{\xi 2} = \sqrt{35^2 - 25^2}\ 10^{-3} \approx 24{,}5 \cdot 10^{-3}\ \frac{\text{г}\text{о}}{\text{г}}.$$

2) Оценка коэффициента корреляции $\check{r}$, по (8.30), с учетом $r_0 = 0$

$$\check{r} = \frac{0{,}55 \cdot 0{,}033 \cdot 35 \cdot 10^{-3}}{0{,}027 \cdot 24{,}5 \cdot 10^{-3}} \approx 0{,}96.$$

Коэффициенты уравнения $C_{mt}^{(u)} = \alpha I_s^{(u)} + \beta$ регрессии истинных значений $C_{mt}^{(u)}$ на $I_s^{(u)}$, по (8.32) и (8.33)

$$\tilde{\alpha} = \check{r}\frac{s_{\xi 1}}{s_{\xi 2}} = 0{,}96\ \frac{0{,}027}{24{,}5 \cdot 10^{-3}} \approx 1{,}06; \quad \tilde{\beta} = \bar{x}_1 - \tilde{\alpha}\bar{x}_2 =$$

$$= 0{,}075 - 1{,}06 \cdot 70 \cdot 10^{-3} \approx 0{,}001.$$

3) Коэффициенты уравнения $C_{mt}^{(u)} = aI_s + b$ регрессии для косвенных измерений $C_{mt}^{(u)}$ по I_s, согласно (8.35),

$$\check{a} = \check{r}\frac{s_{\xi 1} s_{\xi 2}}{s_2^2} = 0{,}96\ \frac{0{,}027 \cdot 24{,}5 \cdot 10^{-3}}{1{,}225 \cdot 10^{-3}} \approx 0{,}52; \quad \check{b} = \bar{x}_1 - \check{a}\bar{x}_2 \approx 0{,}039.$$

§ 3. Случайная функция как модель распределения показателя

Результаты измерений показателя в точках объекта исследования можно считать обусловленными действием факторов, не поддающихся индивидуальному учету и относящихся к источникам случайного, а также факторов, которые могут проявлять свое влияние в виде определенных тенденций в распределении показателя в пределах объекта. Благодаря последним, известное значение показателя в некоторой точке объекта позволяет с большей уверенностью судить о его возможных значениях в близких к данной точках, чем в более удаленных. Это свидетельствует о статистической зависимости между значениями одного и того же показателя в относительно близких точках. Такую зависимость можно объяснить и статистической неоднородностью поля значений показателя.

Чтобы учесть эти особенности, в качестве общей модели, включающей как частный случай случайную величину, можно использовать *случайный процесс* или *случайное поле*.

Случайные функции. В отличие от случайной величины, принимающей при испытании одно из возможных числовых значений, случайный процесс, или *случайная функция*, $\xi(t)$ реализуется в результате испытания в одну из возможных функций переменной t. Эти функции называют *реализациями* случайного процесса $\xi(t)$. При каждом фиксированном t случайная функция $\xi(t)$ представляет собой случайную величину. Распределения этих случайных величин зависят от t.

Основными характеристиками случайного процесса являются *математическое ожидание* и *корреляционная*, или *автокорреляционная, функция*. Математическое ожидание случайной функции $\xi(t)$

представляет собой неслучайную функцию, $\mathbf{M}\xi(t) = m(t)$, которая при каждом фиксированном t_0 равна математическому ожиданию случайной величины $\xi(t_0)$. Корреляционной функцией называют функцию двух переменных t_1 и t_2, представляющую собой зависимость ковариации величин $\xi(t_1)$ и $\xi(t_2)$ от аргументов t_1 и t_2:

$$r(t_1, t_2) = \mathbf{M}\{[\xi(t_1) - m(t_1)][\xi(t_2) - m(t_2)]\} = \mathbf{M}\xi(t_1)\xi(t_2) - m(t_1)m(t_2). \quad (8.38)$$

Из этого определения следует, что $r(t_1, t_2) = r(t_2, t_1)$. Если $\xi(t_1)$ и $\xi(t_2)$ независимы, то $r(t_1, t_2) = 0$. При $t_1 = t_2 = t$, $r(t, t) = D(t)$ — дисперсия $\xi(t)$. *Нормированная корреляционная функция* —

$$r_0(t_1, t_2) = \frac{r(t_1, t_2)}{\sqrt{D(t_1)D(t_2)}}. \quad (8.39)$$

Ее значения — коэффициенты корреляции величин $\xi(t_1)$ и $\xi(t_2)$.

Если математическое ожидание постоянно, $\mathbf{M}\xi(t) = m$, а корреляционная функция зависит лишь от разности аргументов, $r(t_1, t_2) = r(t_2 - t_1) = r(t_1 - t_2)$, то случайный процесс называют *стационарным* в *широком смысле*. У такого процесса дисперсия $D(t) = r(0)$ также постоянна. Случайный процесс *вполне стационарен*, если при любых $t_1, t_2, \ldots, t_n$ многомерное распределение $\xi(t_1), \xi(t_2), \ldots, \xi(t_n)$ не изменяется от добавления к каждому аргументу t_i произвольного τ. Такой процесс стационарен и в широком смысле. Стационарный случайный процесс обладает *свойством эргодичности*, если с вероятностью, равной единице, $\lim\limits_{T\to\infty} \frac{1}{T}\int\limits_0^T \xi(t)\,dt = m = \mathbf{M}\xi(t)$ ($(0, T)$ — интервал наблюдения $\xi(t)$). Достаточное условие эргодичности — $\lim\limits_{T\to\infty} \frac{1}{T^2}\int\limits_0^T\int\limits_0^T r(t-s)\,dt\,ds = 0$. Оно, в частности, выполняется, если интеграл $\int\limits_0^\infty |r(u)|\,du$ конечен. Свойство эргодичности обеспечивает возможность получения оценки математического ожидания по одной реализации с любой точностью за счет увеличения интервала ее наблюдения. Распространяемое на процесс $\eta(t) = \xi(t)\xi(t+\tau)$ при каждом фиксированном $\tau \geqslant 0$, оно обеспечивает такую возможность и для корреляционной функции.

В отличие от случайного процесса случайное поле $\xi(\mathbf{x})$ задается на множествах значений $\mathbf{x} = \{x_1, x_2, \ldots, x_k\}'$ размерности k, большей единицы. Если при каждом фиксированном $\mathbf{x}$ $\xi(\mathbf{x})$ — многомерные случайные величины одной и той же размерности, случайное поле $\xi(\mathbf{x})$ называют многомерным.

Случайное поле $\xi(\mathbf{x})$ *однородно*, если оно имеет постоянное математическое ожидание — $\mathbf{M}\xi(\mathbf{x}) = m$ и его корреляционная функция $r(\mathbf{x}, \mathbf{y}) = \mathbf{M}[(\xi(\mathbf{x}) - m)(\xi(\mathbf{y}) - m)]$ зависит лишь от разности аргументов: $r(\mathbf{x}, \mathbf{y}) = r(\mathbf{x} - \mathbf{y}) = r(\tau)$ ($\tau = \mathbf{x} - \mathbf{y}$). Однородное случайное поле

изотропно, если корреляционная функция не изменяется от поворота аргумента $\tau = x - y$ вокруг начала координат. Таким будет случайное поле, корреляционная функция которого зависит лишь от расстояния между x и y: $r(x, y) = r(d)$, где $d = \sqrt{(x-y)'(x-y)}$. Случайный процесс $\xi(t)$ или случайное поле $\xi(x)$, для которых при каждом фиксированном наборе значений аргумента $t_1, t_2, \ldots, t_n$ $(x_1, x_2, \ldots, x_n)$ многомерная случайная величина $\{\xi(t_1), \xi(t_2), \ldots, \xi(t_n)\}$ $(\{\xi(x_1), \xi(x_2), \ldots, \xi(x_n)\})$ распределена по нормальному закону, называются *гауссовскими*. Для гауссовского случайного процесса понятия стационарности в широком и точном смысле совпадают.

Оценка математического ожидания и корреляционной функции случайного процесса. Если имеется n реализаций случайного процесса $\xi(t) - x_1(t), x_2(t), \ldots, x_n(t)$, то его математическое ожидание и корреляционная функция оцениваются в виде

$$\check{m}(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t); \qquad \check{r}(t_1, t_2) = \frac{1}{n}\sum_{i=1}^{n}[x_i(t_1) - \check{m}(t_1)][x_i(t_2) - \check{m}(t_2)] =$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i(t_1)\, x_i(t_2) - \check{m}(t_1)\check{m}(t_2). \qquad (8.40)$$

Часто, однако, имеют единственную реализацию случайного процесса, которую в лучшем случае можно получить повторно с точностью до ошибок измерений. Так обстоит дело, например, с интерпретацией значений геофизических параметров, содержаний минералов, химических элементов по профилю или скважине как реализаций случайных функций. Обычно приходится наделять используемые в качестве моделей случайные функции правдоподобными свойствами, обеспечивающими возможность хотя бы упрощенного статистического анализа.

Подобной моделью служит представление результата измерения показателя как значения реализации случайного процесса* $\xi(t)$, имеющего вид

$$\xi(t) = m(t) + \delta(t), \qquad (8.41)$$

где $m(t)$ — обычная (неслучайная) функция; $\delta(t)$ — стационарный случайный процесс с нулевым математическим ожиданием, t — аргумент случайного процесса, представляющий собой признак линейного упорядочивания наблюдений. Математическое ожидание $\xi(t)$, $\mathbf{M}\xi(t) = m(t)$, а корреляционная функция зависит лишь от разности аргументов,

$$r_\xi(t_1, t_2) = r_\delta(t_1, t_2) = r(t_2 - t_1) = r(\tau) \quad (\tau = t_2 - t_1). \qquad (8.42)$$

---

* Здесь выражение «случайный процесс $\xi(t)$» обозначает математическую модель, в которой t является некоторым параметром (координатой по профилю, скважине и т. д.), хотя в некоторых задачах последний можно ввести и как величину, пропорциональную времени.

Зная математическое ожидание процесса $\xi(t)$ и имея значения его реализации $x(t_i)$ в n точках t_i ($i = \overline{1, n}$) с равными промежутками между ними h, можно оценить значение корреляционной функции для произвольного $\tau = kh$ (k — целое число, причем $n > k \geqslant 0$)

$$\check{r}_\xi(t,\ t+\tau) = \check{r}(\tau) = \frac{1}{n-k}\sum_{i=1}^{n-k}[x(t_i) - m(t_i)]\,[x(t_{i+k}) - m(t_{i+k})]. \tag{8.43}$$

По этой формуле оценка дисперсии процесса

$$\check{D}_\xi(t) = \check{r}(0) = \frac{1}{n}\sum_{i=1}^{n}[x(t_i) - m(t_i)]^2. \tag{8.43'}$$

Если математическое ожидание $m(t)$ неизвестно, но постоянно — $m(t) = m = \text{const}$, можно построить его оценку в виде

$$\check{m}(t) = \check{m} = \frac{1}{n}\sum_{i=1}^{n} x(t_i), \tag{8.44}$$

а подставив ее в (8.43), получим и оценку $r(\tau)$. Дисперсия оценки $\check{m}$, $\mathbf{D}\check{m} = \frac{1}{n^2}\sum_{i,\ j=1}^{n} r_\xi(t_i, t_j)$.

Если $m(t) \neq \text{const}$, при построении оценки используют параметрический вид $m(t)$ или, в крайнем случае, какие-либо аппроксимирующие функции (например, тригонометрические многочлены). Удовлетворительный результат можно получить уже методом наименьших квадратов.

Оценка корреляционной функции по непрерывной реализации $x(t)$ на интервале $(0, T)$

$$\tilde{r}(\tau) = \frac{1}{T-\tau}\int_0^{T-\tau}[x(t)-m(t)]\,[x(t+\tau)-m(t+\tau)]\,dt \quad (T > \tau \geqslant 0); \tag{8.45}$$

математического ожидания, в случае $m(t) = \text{const}$ —

$$\tilde{m}(t) = \tilde{m} = \frac{1}{T}\int_0^T x(t)\,dt. \tag{8.44'}$$

Дисперсия оценки $\tilde{m}$

$$\mathbf{D}\tilde{m} = \frac{1}{T^2}\int_0^T\int_0^T r(t, s)\,dtds.$$

Аналогично построение оценок и для случайных полей. Случайным полем, подобным (8.41), будет

$$\xi(\mathbf{x}) = m(\mathbf{x}) + \delta(\mathbf{x}), \tag{8.46}$$

где $m(\mathbf{x})$ — обычная (неслучайная) функция, $\delta(\mathbf{x})$ — однородное изотропное случайное поле с нулевым математическим ожиданием. Корреляционная функция $r(d)$ ($d^2 = (\mathbf{x} - \mathbf{y})'(\mathbf{x} - \mathbf{y})$) оценивается аналогично (8.43) по наблюдениям $\xi(\mathbf{x})$ в точках, отстоящих друг от друга на расстоянии d. На рис. 44 изображены виды корреляционных функций случайных полей концентраций химических элементов и характеристик физических свойств.

Для учета ошибок Δ измерений формула (8.46) несколько видоизменяется: $\eta(\mathbf{x}) = m(\mathbf{x}) + \delta(\mathbf{x}) + \Delta$ ($\eta(\mathbf{x})$ — результат измерения показателя ξ в точке $\mathbf{x}$).

Взаимная корреляционная функция. При изучении двух и более случайных функций используют, помимо уже упоминавшихся характеристик, *взаимную корреляционную функцию*, описывающую связь (точнее, корреляцию) двух случайных функций. Для случайных процессов $\xi(t)$ и $\eta(t)$ она определяется в виде

$$r_{\xi,\eta}(t_1, t_2) = \mathbf{M}\{[\xi(t_1) - m_\xi(t_1)] \times [\eta(t_2) - m_\eta(t_2)]\}, \quad (8.47)$$

где $m_\xi(t)$ и $m_\eta(t)$ — математические ожидания $\xi(t)$ и $\eta(t)$.

С использованием (8.47) корреляционная функция суммы двух случайных процессов, $\xi(t) + \eta(t)$, записывается в виде

$$r_{\xi+\eta}(t_1, t_2) = r_\xi(t_1, t_2) + r_\eta(t_1, t_2) + r_{\xi,\eta}(t_1, t_2) + r_{\xi,\eta}(t_2, t_1). \quad (8.48)$$

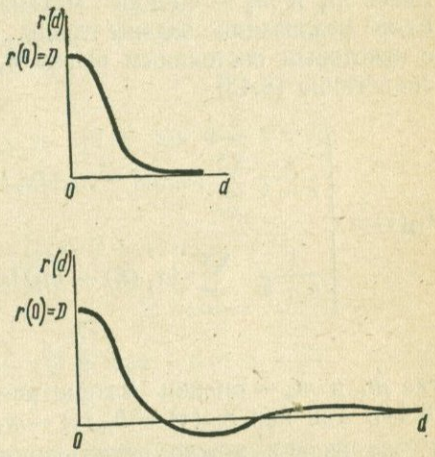


Рис. 44.

Если взаимная корреляционная функция $r_{\xi,\eta}(t_1, t_2)$ равна нулю для всех t_1 и t_2, то случайные процессы $\xi(t)$ и $\eta(t)$ называют *некоррелированными*. По (8.48) корреляционная функция суммы таких случайных процессов равна сумме их корреляционных функций.

Взаимная корреляционная функция используется при решении таких задач геофизики, как корреляция трасс сейсмограмм, корреляция данных каротажа по скважинам, корреляция геолого-гесфизических данных по профилям или разрезам. Обычно используется следующая схема. Имеются две стационарные случайные функции $\xi_1(t)$ и $\xi_2(t)$, взаимная корреляционная функция которых $r_{12}(t_1, t_2) = \mathbf{M}\{[\xi_1(t_1) - m_1][\xi_2(t_2) - m_2]\} = r_{12}(\tau)$ ($m_1 = \mathbf{M}\xi_1(t)$, $m_2 = \mathbf{M}\xi_2(t)$) зивисит лишь от разности $t_2 - t_1 = \tau$. Необходимо скоррелировать их реализации, т. е. оценить величину сдвига τ_s такую, при которой взаимная корреляционная функция получает максимальное значение: $r_{12}(\tau_s) = \max_\tau r(\tau)$.

Оценка взаимной корреляционной функции по непрерывным реализациям $x_1(t)$ и $x_2(t)$ на интервале $(0, T)$ случайных функций $\xi_1(t)$ и $\xi_2(t)$ имеет вид

$$\tilde{r}_{12}(\tau) = \begin{cases} \dfrac{1}{T-\tau} \displaystyle\int_0^{T-\tau} [x_1(t) - \tilde{m}_1][x_2(t+\tau) - \tilde{m}_2]\,dt \text{ при } 0 \leqslant \tau < T, \\ \dfrac{1}{T+\tau} \displaystyle\int_{-\tau}^{T} [x_1(t) - \tilde{m}_1][x_2(t+\tau) - \tilde{m}_2]\,dt \text{ при } 0 > \tau > -T. \end{cases} \tag{8.49}$$

Здесь $\tilde{m}_1$ и $\tilde{m}_2$ — оценки математических ожиданий вида (8.44′). Если реализации заданы лишь в дискретных точках $t_1, t_2, \ldots, t_n$ с некоторым постоянным шагом h, то для $\tau = kh$ (k — целое число), аналогично (8.43)

$$\check{r}_{12}(\tau) = \begin{cases} \dfrac{1}{n-k} \displaystyle\sum_{i=1}^{n-k} [x_1(t_i) - \check{m}_1][x_2(t_i + kh) - \check{m}_2] \text{ при } 0 \leqslant k < n, \\ \dfrac{1}{n+k} \displaystyle\sum_{n=1-k}^{n} [x_1(t_i) - \check{m}_1][x_2(t_i + kh) - \check{m}_2] \text{ при } 0 > k > -n, \end{cases} \tag{8.50}$$

где $\check{m}_1$ и $\check{m}_2$ — оценки математических ожиданий, вычисляемые по (8.44). Так как $r_{12}(\tau) = b_{12}(\tau) - m_1 m_2$, где $b_{12}(\tau) = \mathbf{M}[\xi_1(t)\,\xi_2(t+\tau)]$, ту же задачу можно решать, приводя к максимуму оценку $\check{b}_{12}(\tau)$ функции $b_{12}(\tau)$:

$$\check{b}_{12}(\tau) = \frac{1}{n-k} \sum_{i=1}^{n-k} x_1(t_i)\, x_2(t_i + kh) \text{ (при } 0 \leqslant k < n),$$

$$\check{b}_{12}(\tau) = \frac{1}{n+k} \sum_{i=1-k}^{n} x_1(t_i)\, x_2(t_i + kh) \text{ (при } 0 > k > -n).$$

Если $|k|$ мало́ по сравнению с n, то $\check{b}_{12} \approx \check{r}_{12}(\tau) + \check{m}_1 \check{m}_2$.

С другой стороны, так как

$$\rho(\tau) = \mathbf{M}[\xi_1(t) - \xi_2(t+\tau)]^2 = r_1(0) + r_2(0) - 2r_{12}(\tau) + (m_1 - m_2)^2,$$

величину τ_s, приводящую к максимуму $r_{12}(\tau)$, можно определять по минимуму «расстояния» $\rho(\tau)$ между $\xi_1(t)$ и $\xi_2(t+\tau)$ при сдвиге $\xi_1(t)$ относительно $\xi_2(t)$ на τ. Оценка $\rho(\tau)$,

$$\check{\rho}(\tau) = \check{\rho}(kh) = \begin{cases} \frac{1}{n-k} \sum_{i=1}^{n-k} [x_1(t_i) - x_2(t_i + kh)]^2 \text{ при } 0 \leqslant k < n, \\ \frac{1}{n+k} \sum_{i=1-k}^{n} [x_1(t_i) - x_2(t_i + kh)]^2 \text{ при } 0 > k > -n. \end{cases} \tag{8.51}$$

Оценка $\tau_s = k_s h$ будет определяться числом k_s, при котором значение функции $\rho(\tau)$ минимально. Оценки, аналогичные (8.50), (8.51), используются и в случае непрерывных реализаций —

$$\tilde{b}_{12}(\tau) = \frac{1}{T-\tau} \int_0^{T-\tau} x_1(t)\, x_2(t+\tau)\, dt,$$

$$\tilde{\rho}(\tau) = \frac{1}{T-\tau} \int_0^{T-\tau} [x_1(t) - x_2(t+\tau)]^2\, dt \text{ при } 0 \leqslant \tau < T;$$

$$\tilde{b}_{12}(\tau) = \frac{1}{T+\tau} \int_{-\tau}^{T} x_1(t)\, x_2(t+\tau)\, dt,$$

$$\tilde{\rho}(\tau) = \frac{1}{T+\tau} \int_{-\tau}^{T} [x_1(t) - x_2(t+\tau)]^2\, dt \text{ при } 0 > \tau > -T. \tag{8.52}$$

Оценки (8.50) и (8.51) находят применение, в частности, при обработке геофизических данных с целью обнаружения на исследуемой площади выдержанных по простиранию аномалий с применением метода скользящего среднего, для определения оптимальной формы окна осреднения (наклона сторон окна либо величины смещения каждого последующего профиля относительно предыдущего) и числа профилей, включаемых в окно.

Случайное поле при оценке среднего по объекту. Будем считать величину показателя $\xi_i = \xi(x_i, y_i)$ в пункте наблюдения с координатами x_i, y_i значением реализации случайного поля $\xi(x, y)$ с математическим ожиданием $\mathbf{M}\,\xi(x, y) = m(x, y)$. Рассматриваются для простоты наблюдения на плоскости, хотя приведенные ниже выводы можно легко распространить на трехмерный случай.

Функцию $m(x, y)$ можно считать постоянной лишь при отсутствии закономерной тенденции в распределении показателя, проявляющейся на достаточно больших (по сравнению с ячейкой сети наблюдений) участках исследуемого объекта G. Будем считать корреляционную функцию равной нулю для минимально возможных расстояний между пунктами наблюдений (x_i, y_i): $r(x_i, y_i; x_j, y_j) = \mathbf{M}\{[\xi(x_i, y_i) - m(x_i, y_i)][\xi(x_j, y_j) - m(x_j, y_j)]\} = 0$ $(i, j = \overline{1, n}, i \neq j)$; $n -$

количество пунктов наблюдений. Задача состоит в оценке M — среднего по G значения показателя:

$$M = \frac{1}{S}\iint\limits_G m(x, y)\,dxdy \tag{8.53}$$

(S — площадь G).

Обозначим s_i — площади ячеек сети наблюдений с центрами (x_i, y_i) $(i = \overline{1, n})$. Считая $m(x, y)$ гладкой функцией, используем приближенное равенство

$$M \approx M_n = \frac{1}{S}\sum_{i=1}^{n} m(x_i, y_i)\, s_i \left(S = \sum_{i=1}^{n} s_i\right).$$

Оценкой M_n будет

$$\check{M} = \sum_{i=1}^{n} \xi_i s_i \frac{1}{S}, \tag{8.54}$$

причем ее дисперсия составит

$$\mathbf{D}\check{M} = \sum_{i=1}^{n} \frac{s_i^2}{S^2} D(x_i, y_i), \tag{8.55}$$

где $D(x_i, y_i)$ — дисперсия показателя ξ в ячейке* с центром (x_i, y_i). При равномерной сети наблюдений $\mathbf{D}\check{M} = \frac{1}{n^2}\sum_{i=1}^{n} D(x_i, y_i)$. Приближенные доверительные пределы для M уровня q, с использованием асимптотической нормальности оценки $\check{M}$: $\check{M} - u_{\frac{1+q}{2}}\sqrt{\mathbf{D}\check{M}}$, $\check{M} + u_{\frac{1+q}{2}}\sqrt{\mathbf{D}\check{M}}$.

Итак, при наличии статистической неоднородности поля значений показателя точность оценки среднего определяется не дисперсией показателя на всем объекте, а дисперсиями его $D(x_i, y_i)$ в ячейках сети наблюдений. Ясно, что в этом случае истинная точность оценки среднего будет выше, чем определяемая формулой (4.15), а необходимое для достижения заданной точности количество наблюдений, рассчитанное исходя из этой же формулы (табл. 4.1), будет завышенным.

Если дисперсии $D(x_i, y_i)$ неизвестны, оценку $\mathbf{D}\check{M}$ можно вычислить непосредственно по наблюдениям ξ_i, упорядочив их в ряд по

---

* Выражение (8.55) будет более точным, если в качестве $D(x_i, y_i)$ взять дисперсию отклонения поля от линейной функции координат (x, y), аппроксимирующей функцию $m(x, y)$ в пределах ячейки сети с центром (x_i, y_i).

степени близости пунктов наблюдений друг к другу и считая пренебрежимым изменение $m(x, y)$ в ячейке:

$$\mathbf{D}\check{M} = \frac{1}{2n(n-1)} \sum_{i=1}^{n-1} (\xi_{i+1} - \xi_i)^2. \qquad (8.56)$$

Условие оптимальности оценки среднего. Выясним условия, при которых распределение n пунктов наблюдений на исследуемом объекте обеспечивает наилучшую точность оценки (8.54). Предположим для простоты, что объект состоит из областей $G_1, G_2, \ldots, G_k$, в каждой из которых математическое ожидание и дисперсия поля $\xi(x, y)$ постоянны: $m(x, y) = m_j$, $D(x, y) = D_j$ для $(x, y) \in G_j (j = \overline{1, k})$. Обозначив n_j — количество наблюдений в G_j, L_j — площадь $G_j (j = \overline{1, k})$, $S = \sum_{j=1}^{k} L_j$, получим дисперсию (8.55) оценки среднего в таком виде:

$$\mathbf{D}\check{M} = \sum_{i=1}^{n} \frac{s_i^2}{S^2} D(x_i, y_i) = \sum_{j=1}^{k} D_j \left(\frac{L_j}{n_j S}\right)^2 n_j = \sum_{j=1}^{k} \frac{L_j^2 D_j}{S^2 n_j}.$$

Приравнивая производные по n_j $(j = \overline{1, k-1})$ нулю с учетом условия $n_k = n - \sum_{j=1}^{k-1} n_j$, получим

$$n_j = \frac{n L_j \sqrt{D_j}}{\sum_{j=1}^{k} L_j \sqrt{D_j}}, \quad \frac{n_j}{n L_j} = \frac{\sqrt{D_j}}{\sum_{j=1}^{k} L_j \sqrt{D_j}} \quad (j = \overline{1, k}). \qquad (8.57)$$

Плотность сети наблюдений должна быть пропорциональной среднему квадратическому отклонению поля.

Подобно этому можно рассчитать и оптимальное соотношение сторон ячеек прямоугольной сети наблюдений. Пусть $D(\delta_i)$ — дисперсия показателя в ячейке δ_i сети наблюдений. При фиксированном направлении профилей эта дисперсия, включающая и ошибку аппроксимации $m(x, y) \approx m_i = \text{const}$ в δ_i, будет функцией ее (δ_i) ширины u длины v: $D(\delta_i) = D(u, v)$. Ввиду относительной малости u и v можно приближенно считать

$$D(u, v) = \alpha u + \beta v + \gamma, \quad \alpha = \left.\frac{\partial D}{\partial u}\right|_{0,\,0}, \quad \beta = \left.\frac{\partial D}{\partial v}\right|_{0,\,0}, \quad \gamma = D(0, 0), \qquad (8.58)$$

причем $\alpha \geqslant 0$, $\beta \geqslant 0$, так как $D(u, v)$ — неубывающая функция по аргументам u и v. При фиксированной площади ячейки s_i оптимальное соотношение между u и v определяется из условия $\frac{d}{du}\left(\alpha u + \beta \frac{s_i}{u} + \gamma\right) = 0$, откуда

$$\frac{v}{u} = \frac{\alpha}{\beta} = \frac{\partial D}{\partial u} \Big/ \frac{\partial D}{\partial v}. \qquad (8.59)$$

Таким образом, отношение сторон оптимальной ячейки с заданными направлениями сторон обратно отношению проекций градиента функции $D(\delta_l)$ на эти направления. Все формы ячеек, при которых $D(\delta_l)$ не зависит от u и v ($\alpha = \beta = 0$), эквивалентны.

При выводе формулы (8.57) рассматривался случай, когда дисперсия показателя в ячейке не зависела от ее размеров в пределах возможных величин. Если такая зависимость имеет место*, то считая ее, как в (8.58), линейной и действуя так же, как при выводе (8.57) и (8.59), нетрудно получить необходимые соотношения и в этом более общем случае.

Влияние способов отбора проб на точность оценки среднего. В соответствии с формулой (8.55) отбор проб должен осуществляться так, чтобы обеспечить по возможности минимальные дисперсии отклонений результатов измерений от средних по ячейкам значений показателя. Рассмотрим *способ сборных проб*, при котором в каждом пункте опробования отбирается некоторое число k частных проб примерно одного веса p на небольших расстояниях друг от друга, объединяемых затем в сборную пробу. Этот способ нередко используется при геохимическом опробовании горных пород.

Пусть $\bar{\xi}_k(x, y) = \frac{1}{k}\sum_{i=1}^{k} \xi_i(x, y)$ — значение показателя в сборной пробе, отобранной в пункте (x, y); $\eta_k(x, y) = \bar{\xi}_k(x, y) + \Delta$ — результат измерения этого значения; $m(x, y) = \mathbf{M}\xi(x, y)$; $\xi(x, y)$ — случайное поле значений показателя в пробах весом p ($\xi_i(x, y)$ считаем его реализациями). Будем считать условие однородности и изотропности случайного поля $\delta(u, v) = \xi(u, v) - m(u, v)$ выполненным, по крайней мере приближенно, для точек (u, v) в пределах площади, на которую будет распространяться результат опробования в точке (x, y). Обозначим $D_\xi = \mathbf{D}\xi(x, y) = \mathbf{D}\delta(x, y)$ — дисперсию поля в пределах этой площади, $r_\xi(d)$ — корреляционную функцию (d — расстояние между точками — аргументами корреляционной функции в этой области). Дисперсия $\eta_k(x, y)$

$$\mathbf{D}[\eta_k(x, y)] = \mathbf{M}[\bar{\xi}_k(x, y) + \Delta - m(x, y)]^2 =$$

$$= \frac{D_\xi}{k} + \frac{2}{k^2}\sum_{i, j, i<j}^{k} r_\xi(d_{ij}) + D_0, \qquad (8.60)$$

где D_0 — дисперсия ошибок измерений; d_{ij} — расстояния между частными пробами.

Из монотонности $r_\xi(d)$ ($r_\xi(d_2) < r_\xi(d_1)$, если $d_2 > d_1 \geqslant 0$) следует, в соответствии с этой формулой, что сборные пробы предпочтительнее монолитных того же веса $P = kp$. Последние можно считать состоящими из частных проб, взятых вплотную друг к другу, так что

* Эта зависимость может быть обусловленной ошибками аппроксимации функции $m(x, y)$ в ячейке линейной функцией или средним значением по ячейке, в зависимости от способа аппроксимации.

дисперсия измеренных по монолитным пробам значений $\eta_P(x, y)$ будет иметь вид

$$\mathbf{D}[\eta_P(x, y)] = \frac{D_\xi}{k} + \frac{k-1}{k} r_\xi(+0) + D_0, \qquad (8.60')$$

где $r_\xi(+0)$ — значение корреляционной функции для двух проб, взятых вплотную друг к другу. Полученная формула показывает, что выигрыш в точности за счет увеличения веса монолитной пробы невелик, если значения показателя в двух пробах, отбираемых из одного и того же места, сильно связаны: $r_\xi(+0) \approx D_\xi$. Из формулы (8.60) следует, что при монотонной корреляционной функции наилучшей будет сборная проба с расстояниями $d_{ij} > d_0$, где d_0 — расстояние, при котором $r(d_0) = 0$. Оценку корреляционной функции $r_\xi(d)$ можно получить по данным опробования отдельных участков, полагая в них приближенно $m(x, y) = \text{const}$. При вычислении оценки дисперсии D_ξ по наблюденным значениям поля нужно иметь в виду, что дисперсия D_η поля $\eta_1(x, y)$ имеет вид $D_\eta = D_\xi + D_0$. Оценка D_ξ будет вычисляться в виде $\check{D}_\xi = \check{D}_\eta - D_0$, где $\check{D}_\eta$ — оценка D_η.

Формула (8.60′) позволяет рассчитывать дисперсию показателя в пробах различного веса без прямых измерений таких проб, используя оценки D_ξ, D_0, $r_\xi(+0)$. Подобный расчет для сборных проб различной структуры можно осуществить по (8.60).

Из (8.60′) следует, что при значительной дисперсии ошибок эффективность опробования следует повышать прежде всего за счет совершенствования методики измерений. Если метод измерений чувствителен, можно проводить независимые кратные измерения каждой сборной пробы с последующим осреднением результатов. Если l — кратность, то дисперсия таких средних $\bar{\eta}(x, y)$, при $r(d_{ij}) = 0$,

$$\mathbf{D}[\bar{\eta}(x, y)] = \frac{D_\xi}{k} + \frac{D_0}{l}. \qquad (8.61)$$

Если c_1 — стоимость отбора и доставки одной частной пробы при отборе их группой в одном пункте, c_2 — стоимость измерения, то фиксируя затраты на один пункт опробования $c = c_1 k + c_2 l$, получим условие минимума (8.61):

$$\frac{k}{l} = \sqrt{\frac{c_2 D_\xi}{c_1 D_0}}. \qquad (8.62)$$

Это условие обеспечивает наиболее рациональный режим кратных измерений при способе сборных проб*.

Спектральная плотность. Пусть стационарная случайная функция $\xi(t)$ задана на интервале $(0, T)$, ее корреляционная функция $r(t_1, t_2) = r(t_2 - t_1) = r(\tau)$ и математическое ожидание $\mathbf{M}\xi(t) = m$. Пользуясь четностью функции $r(\tau)$ и считая необходимые условия выпол-

* Если $\frac{l}{k} = t \geqslant 1$, t — целое число, то этот режим эквивалентен t кратным измерениям каждой частной пробы.

ненными, разложим $r(\tau)$ в ряд Фурье [5] на интервале $(-T, T)$:

$$r(\tau) = \sum_{k=0}^{\infty} D_k \cos \omega_k \tau = \sum_{k=0}^{\infty} (D_k \cos \omega_k t_1 \cos \omega_k t_2 + D_k \sin \omega_k t_1 \sin \omega_k t_2), \tag{8.63}$$

где

$$\omega_k = \frac{k\pi}{T}; \quad D_0 = \frac{1}{T}\int_0^T r(\tau)\, d\tau, \quad D_k = \frac{2}{T}\int_0^T r(\tau) \cos \omega_k \tau d\tau \quad (k \geqslant 1). \tag{8.64}$$

Из теоремы Карунена — одной из основных теорем о случайных функциях [6] следует: если $r(t_1, t_2)$ представляется в виде $r(t_1, t_2) = \sum_{i=1}^{\infty} D_i \varphi_i(t_1) \varphi_i(t_2)$, то случайную функцию можно представить в виде $\xi(t) = m + \sum_{i=1}^{\infty} V_i \varphi_i(t)$, где V_i — некоррелированные между собой случайные величины с нулевыми математическими ожиданиями и дисперсиями $\mathbf{D}V_i = D_i \quad (i = 1, 2, \ldots)$.

В соответствии с этим, с учетом (8.63), стационарную случайную функцию можно представить [5] в виде:

$$\xi(t) = m + \sum_{k=0}^{\infty} (U_k \cos \omega_k t + V_k \sin \omega_k t) = m + \sum_{k=0}^{\infty} \sqrt{U_k^2 + V_k^2} \cos (\omega_k t + \varphi_k); \tag{8.65}$$

$$\cos \varphi_k = U_k / \sqrt{U_k^2 + V_k^2}, \quad \sin \varphi_k = V_k / \sqrt{U_k^2 + V_k^2}, \quad \omega_k = \frac{k\pi}{T}$$

$$(k = 0, 1, 2, \ldots),$$

где U_k, V_k — случайные величины с нулевыми средними, причем U_0, V_0, U_1, V_1, ... некоррелированы и $\mathbf{D}U_k = \mathbf{D}V_k = D_k$. Таким образом, случайная функция $\xi(t)$ представляется в виде суммы гармоник различных частот ω_k со случайными амплитудами и фазами.

Из (8.63) дисперсия $\xi(t)$, $\mathbf{D}\xi(t) = r(0) = \sum_{k=0}^{\infty} D_k$, причем, как следует из сходимости ряда $\sum_{k=0}^{\infty} D_k$, $D_k \to 0$ при $k \to \infty$ — с ростом частоты дисперсии коэффициентов U_k и V_k убывают. Для различных случайных функций D_k могут быть различными; распределение дисперсий D_k по частотам ω_k называют *спектром* стационарной случайной функции. Спектр дает возможность выяснить, какие гармонические колебания оказывают преобладающее влияние в разложении (8.65). Представление случайной функции $\xi(t)$ в виде (8.65) называется ее *спектральным разложением* по координатным функциям $\cos \omega_k t$, $\sin \omega_k t$. Средние квадраты амплитуд гармоник $U_k \cos \omega_k t$, $V_k \sin \omega_k t$ спектрального

разложения равны дисперсиям D_k. Если $\xi(t)$ — гауссовский случайный процесс, математические ожидания амплитуд $|U_k|$, $|V_k|$ составят $\sqrt{\frac{2}{\pi} D_k}$.

При больших T различие между соседними значениями частот спектра мало́: $\omega_{k+1} - \omega_k = \frac{\pi}{T}$. В этом случае оперируют не дискретным спектром, а средней плотностью дисперсии, определяемой как дисперсия, приходящаяся на единицу длины интервала частот

$$S(\omega_k) = \frac{D_k}{\Delta\omega}, \tag{8.66}$$

где $\Delta\omega = \frac{\pi}{T}$, D_k — дисперсия, соответствующая частоте ω_k. Если рассматривать $\xi(t)$ в бесконечном интервале $(0, \infty)$, то перейдя к пределу при $\Delta\omega \to 0$, получим функцию $S(\omega)$, которая имеет смысл плотности распределения дисперсии по частотам ω непрерывного спектра [5]. Эту функцию и называют *спектральной плотностью* стационарной случайной функции.

Подобно свойству спектра $\sum_{k=0}^{\infty} D_k = \mathbf{D}\xi(t) = r(0)$, спектральная плотность обладает следующим свойством: $\int_0^{\infty} S(\omega)\, d\omega = \mathbf{D}\xi(t)$. Подставив в формулу (8.63) выражение D_k из (8.66) и перейдя к пределу при $T \to \infty$ $(\Delta\omega \to 0)$, найдем выражение корреляционной функции через спектральную плотность с помощью косинус-преобразования Фурье

$$r(\tau) = \int_0^{\infty} S(\omega) \cos \omega\tau d\omega, \tag{8.67}$$

а умножив D_k в (8.64) на $\frac{1}{\Delta\omega} = \frac{T}{\pi}$ и также перейдя к пределу при $T \to \infty$, получим выражение спектральной плотности через корреляционную функцию:

$$S(\omega) = \frac{2}{\pi} \int_0^{\infty} r(\tau) \cos \omega\tau d\tau. \tag{8.68}$$

Как и спектр, спектральная плотность дает возможность оценить преобладающие частоты случайной функции и по (8.66) — математические ожидания квадратов амплитуд, соответствующих частотам из отдельных интервалов.

Статистические задачи теории случайных функций. Современная теория обобщает методы получения оценок и проверки гипотез на случайные функции. Разработка методов статистического анализа случайных функций в применении к геологическим и, особенно, геофизическим исследованиям представляет собой актуальное и перспективное направление. В форме определения оценки математического

ожидания случайной функции можно ставить задачу нахождения по реализации «сигнала» определенной формы, вид которого можно задать параметрическим семейством функций. В форме проверки гипотез о математическом ожидании можно сопоставлять вероятные в данной реализации формы сигналов и определять наиболее правдоподобную, проверять наличие сигналов.

Задача прогноза состоит в определении оценки значения случайной функции или случайного поля в некоторой точке по известным их реализациям на интервале или в области, к которым точка не принадлежит. Пример подобной задачи — прогнозирование значений количественных характеристик пород на глубинах, превышающих глубину скважины.

К задаче прогноза близка по содержанию задача *фильтрации*. Предполагается, что наблюдаемая случайная функция представляется в виде $\eta(t) = \xi(t) + \delta(t)$, где $\xi(t)$ — случайная функция, которую необходимо «отфильтровать» от «шума» — случайной функции $\delta(t)$, т. е. оценить с наибольшей точностью значение $\xi(t_0)$ в произвольной точке t_0 интервала $[0, T]$ по реализации $\eta(t)$ на этом интервале. Методы решения такой задачи применяются в исследованиях, связанных с поиском слабых аномалий, выделением сигналов на фоне помех — при обработке сейсмограмм, данных грави- и магниторазведки.

Подробное изучение этой группы вопросов не входит в нашу задачу. Мы ограничимся лишь общей постановкой двух важных родственных задач — прогноза и фильтрации и наметим принципы их решения.

Рассмотрим способы определения наилучшего линейного прогноза, когда искомые оценки определяются в виде линейных комбинаций значений $\xi(t)$, взятых в различных точках области наблюдений, либо пределов таких линейных комбинаций*.

Пусть случайный процесс $\xi(t) = m(t) + \delta(t)$ задан на интервале $[0, T]$, $m(t)$ — математическое ожидание $\xi(t)$, $\delta(t)$ — случайный процесс с нулевым средним — $\mathbf{M}\delta(t) = 0$. Необходимо найти наилучшую оценку $\check{\xi}(s)$ значения $\xi(s)$ в точке $s > T$.

Простейшим прогнозом будет, очевидно, $\check{\xi}(s) = m(s)$. Ясно, что такой прогноз можно уточнить, предсказывая отклонение $\xi(s)$ от математического ожидания, $\delta(s) = \xi(s) - m(s)$, на основании той связи, которую выражает корреляционная функция. Фактически задача сводится к прогнозу именно этого отклонения.

Рассмотрим сначала случай, когда заданы N значений реализации $\xi(t)$ в дискретных точках $t_1, t_2, \ldots, t_N$ интервала $[0, T]$. Обозначим эти значения $\xi_i = \xi(t_i)$ $(i = \overline{1, N})$. Линейный прогноз имеет вид

$$\check{\xi}(s) = \sum_{i=1}^{N} a_i [\xi_i - m(t_i)] + m(s). \qquad (8.69)$$

*В теории случайных процессов доказано, что для гауссовского случайного процесса наилучший прогноз будет линейным.

Задача заключается в определении коэффициентов c_i, при которых прогноз будет наилучшим, т. е. дисперсия отклонения $\hat{\xi}(s)$ от $\xi(s)$ минимальна. Продифференцировав выражение этой дисперсии

$$\mathbf{M}[\hat{\xi}(s) - \xi(s)]^2 = \sum_{i,\,j=1}^{N} c_i c_j r(t_i,\ t_j) - 2\sum_{i=1}^{N} c_i r(t_i,\ s) + r(s,\ s) \qquad (8.70)$$

по c_j $(j = \overline{1,\ N})$, получим систему уравнений для определения c_j:

$$\sum_{i=1}^{N} c_i r(t_i,\ t_j) = r(t_j,\ s), \qquad j = \overline{1,\ N}. \qquad (8.70')$$

Систему уравнений (8.70′) можно записать в форме *условия ортогональности* $\hat{\xi}(s) - \xi(s)$ ко всем величинам $\delta_j = \xi_j - m(t_j)$ $(j = \overline{1, N})$, по которым строится прогноз:

$$\mathbf{M}\{[\hat{\xi}(s) - \xi(s)][\xi_j - m(t_j)]\} = \sum_{i=1}^{N} \mathbf{M}\{[c_i(\xi_i - m(t_i)) + m(s) -$$

$$- \xi(s)][\xi_j - m(t_j)]\} = \sum_{i=1}^{N} c_i r(t_i,\ t_j) - r(t_j,\ s) = 0,\ j = \overline{1,\ N}.$$

Как доказано в теории случайных процессов, аналогичному условию (оно приведено ниже) должен удовлетворять оптимальный линейный прогноз, построенный и по непрерывной реализации.

Дисперсию отклонения прогноза (8.69) от $\xi(s)$, характеризующую его ошибку, получим по (8.70) с использованием (8.70′):

$$\mathbf{M}[\hat{\xi}(s) - \xi(s)]^2 = r(s,\ s) - \sum_{i=1}^{N} c_i r(t_i,\ s).$$

Если $\delta(t)$ — стационарная случайная функция, т. е. $r(t,\ s) = r(s - t)$ и шаг между точками t_i постоянен, $t_{i+1} - t_i = h$ $(i = \overline{1,\ N-1})$, система уравнений (8.70′) примет вид

$$\sum_{i=1}^{N} c_i r(h(j - i)) = r(s - t_j), \qquad j = \overline{1,\ N}.$$

Рассмотрим теперь случай, когда $\xi(t)$ наблюдается непрерывно на $[0,\ T]$. Прогноз будем искать в виде

$$\hat{\xi}(s) = \int_0^T [\xi(u) - m(u)]\, c_s(u)\, du + m(s), \qquad (8.71)$$

где $c_s(u)$ — некоторая функция на $[0,\ T]$. Условие ортогональности—

$$\mathbf{M}\{[\hat{\xi}(s) - \xi(s)][\xi(t) - m(t)]\} = 0, \quad 0 \leqslant t \leqslant T, \qquad (8.72)$$

из которого получим интегральное уравнение относительно $c_s(t)$:

$$\int_0^T r(t,\ u)\, c_s(u)\, du = r(s,\ t), \quad 0 \leqslant t \leqslant T. \qquad (8.73)$$

Уравнение такого типа уже встречалось в задаче оценки плотности природного распределения (8.3).

Иногда наилучший линейный прогноз удается построить непосредственно по условию ортогональности. Так, для стационарного случайного процесса с корреляционной функцией $r(t, s) = De^{-\beta|s-t|}$ ($D > 0$, $\beta > 0$) наилучший линейный прогноз вычисляется в виде $\check{\xi}(s) = m + [\xi(T) - m]\, e^{-\beta(s-T)}$ ($m = \mathbf{M}\xi(t)$), который, как легко проверить, удовлетворяет условию (8.72).

Условие, аналогичное (8.72), используется в качестве исходного соотношения и при решении задачи фильтрации. Пусть на интервале $[0, T]$ задана случайная функция $\eta(t) = \xi(t) + \delta(t)$; $\eta(t)$, $\xi(t)$ и $\delta(t)$ — случайные функции с математическими ожиданиями $\mathbf{M}\xi(t) = \mathbf{M}\eta(t) = m(t)$, $\mathbf{M}\delta(t) = 0$ и корреляционными функциями $r_\eta(t_1, t_2)$, $r_\xi(t_1, t_2)$ и $r_\delta(t_1, t_2)$. Наилучшая линейная оценка $\check{\xi}(t_0)$ значения $\xi(t)$ в точке t_0 ($0 \leqslant t_0 \leqslant T$) определяется условием

$$\mathbf{M}\{[\check{\xi}(t_0) - \xi(t_0)]\,[\eta(t) - m(t)]\} = 0, \quad 0 \leqslant t \leqslant T. \tag{8.74}$$

Будем считать $\xi(t)$ и $\delta(t)$ некоррелированными, так что $r_\eta(t_1, t_2) = r_\xi(t_1, t_2) + r_\delta(t_1, t_2)$. Если значения $\xi(t)$ известны в N дискретных точках интервала $[0, T]$ $t_1, t_2, \ldots, t_N$, то линейный *фильтр* — оценка $\xi(t_0)$ имеет вид

$$\check{\xi}(t_0) = m(t_0) + \sum_{i=1}^{N} c_i\,[\eta(t_i) - m(t_i)], \tag{8.75}$$

причем по условию ортогональности, аналогично (8.74),

$$\mathbf{M}\{[\sum_{i=1}^{N} c_i\,(\eta(t_i) - m(t_i)) - \xi(t_0) + m(t_0)]\,[\eta(t_j) - m(t_j)]\} = 0,$$

$$j = \overline{1, N}.$$

Коэффициенты c_i определяются из системы N уравнений

$$\sum_{i=1}^{N} c_i r_\eta(t_i, t_j) = r_\xi(t_0, t_j), \qquad j = \overline{1, N}. \tag{8.76}$$

Если $\eta(t)$ наблюдается непрерывно на $[0, T]$, причем $\xi(t)$ и $\delta(t)$ стационарны, для решения задачи фильтрации можно воспользоваться спектральным разложением. Рассмотрим оценку $\check{\xi}(t_0)$ вида

$$\check{\xi}(t_0) = m + \int_0^T c_0(u)\,[\eta(u) - m]\,du \quad (m = \mathbf{M}\eta(t)). \tag{8.77}$$

По условию ортогональности (8.74) получим интегральное уравнение для определения $c_0(u)$:

$$\int_0^T c_0(u) r_\eta(u, t)\,du = r_\xi(t_0, t), \quad 0 \leqslant t \leqslant T. \tag{8.78}$$

Решение уравнения (8.78) даже приближенными методами, как правило, требует кропотливой работы. Один из способов использования спектрального разложения состоит в следующем. Будем искать оценку $\overset{\vee}{\xi}(t_0)$ в виде

$$\overset{\vee}{\xi}(t_0) = \sum_{k=0}^{\infty} f_k (U_k \cos \omega_k t_0 + V_k \sin \omega_k t_0) + m, \qquad (8.79)$$

где U_k и V_k — коэффициенты спектрального разложения $\eta(t)$; $\omega_k = \frac{k\pi}{T}$. Очевидно,

$$\mathbf{M}\{[\overset{\vee}{\xi}(t_0) - m][\eta(t) - m]\} = \sum_{k=0}^{\infty} f_k D_{k\eta} (\cos \omega_k t_0 \cos \omega_k t + \sin \omega_k t_0 \sin \omega_k t) = \sum_{k=0}^{\infty} f_k D_{k\eta} \cos \omega_k (t - t_0),$$

где $D_{k\eta}$ — коэффициенты разложения (8.63) функции $r_\eta(t_1, t_2)$. Ввиду того, что $r_\eta(t_1, t_2) = r_\xi(t_1, t_2) + r_\delta(t_1, t_2)$, $D_{k\eta} = D_{k\xi} + D_{k\delta}$, где $D_{k\xi}$ и $D_{k\delta}$ — коэффициенты разложения (8.63) функций $r_\xi(t_1, t_2)$ и $r_\delta(t_1, t_2)$. Так как $\mathbf{M}\{[\xi(t_0) - m][\eta(t) - m]\} = r_\xi(t_0, t) = \sum_{k=0}^{\infty} D_{k\xi} \cos \omega_k (t - t_0)$, условие (8.74) соблюдается при $f_k = D_{k\xi}(D_{k\xi} + D_{k\delta})^{-1}$.

Пользуясь спектральными разложениями $\xi(t)$, $\delta(t)$ и учитывая, что разложение $\eta(t)$ представляет собой сумму этих разложений, нетрудно показать, что дисперсия разности $\overset{\vee}{\xi}(t_0) - \xi(t_0)$, характеризующая качество фильтра, составит

$$\mathbf{M}[\overset{\vee}{\xi}(t_0) - \xi(t_0)]^2 = \sum_{k=0}^{\infty} \frac{D_{k\xi} D_{k\delta}}{D_{k\xi} + D_{k\delta}} \leqslant \min[\mathbf{D}\xi(t), \mathbf{D}\delta(t)]. \qquad (8.80)$$

Для получения оценки $\overset{\vee}{\xi}(t_0)$, выраженной через реализацию $\eta(t)$, можно воспользоваться разложением в ряд Фурье на конечном интервале четной функции $f(\omega) = D_\xi(\omega)[D_\xi(\omega) + D_\delta(\omega)]^{-1}$, положив

$$D_\xi(\omega) = D_{k\xi},\ D_\delta(\omega) = D_{k\delta} \text{ при } \omega_k - \frac{\pi}{2T} \leqslant \omega < \omega_k + \frac{\pi}{2T} \quad (k = 0, \pm 1, \ldots).$$

Так как $\mathbf{D}U_k = \mathbf{D}V_k = D_{k\eta} \to 0$ при $k \to \infty$ (влияние U_k и V_k с ростом частоты ω_k становится исчезающе малым) и $f_k = D_{k\xi}(D_{k\xi} + D_{k\delta})^{-1} \leqslant 1$, в (8.79) можно ограничиться суммированием не до ∞, а до такого номера N, при котором $\sum_{k=N+1}^{\infty} D_{k\eta} = \mathbf{D}\eta(t) - \sum_{k=0}^{N} D_{k\eta}$ пренебрежимо мала. Тогда тем более мала дисперсия остатка ряда

(8.79) — $\sum_{k=N+1}^{\infty} f_k D_{k\eta} \leqslant \sum_{k=N+1}^{\infty} D_{k\eta}$. Приняв $G = \frac{\pi N}{T}$, разложим функцию $f(\omega)$ в ряд Фурье на интервале $(-G, G)$:

$$f(\omega) = \sum_{\nu=0}^{\infty} b_\nu \cos g_\nu \omega, \text{ где } g_\nu = \frac{\nu\pi}{G} \ (\nu = 0, 1, 2, \ldots); \ b_0 = \frac{1}{G}\int_0^G f(\omega)\, d\omega;$$

$$b_\nu = \frac{2}{G}\int_0^G f(\omega) \cos g_\nu \omega d\omega \ (\nu = 1, 2, \ldots).$$

Используем для аппроксимации $f(\omega)$ часть ряда: $f(\omega) \approx$ $\approx \sum_{\nu=0}^{n} b_\nu \cos g_\nu \omega$. Нетрудно убедиться, что при $g_\nu < t_0 < T - g_\nu$ $(\nu = \overline{1, n})$ в качестве приближенного фильтра можно взять

$$\xi(t_0) = m + \sum_{k=0}^{\infty}\sum_{\nu=0}^{n} b_\nu \cos g_\nu \omega_k (U_k \cos \omega_k t_0 + V_k \sin \omega_k t_0) =$$

$$= m + \sum_{\nu=0}^{n} \frac{1}{2} b_\nu [\eta(t_0 + g_\nu) + \eta(t_0 - g_\nu)]. \qquad (8.81)$$

§ 4. Разделение неоднородной совокупности

С задачей разделения неоднородной совокупности по комплексу показателей приходится иметь дело при геологическом картировании, построении геологических разрезов немых толщ, выделении геологических границ по геофизическим и геохимическим данным и решении других подобных вопросов. От рассматривавшейся в гл. 7 задачи классификации она отличается отсутствием эталонных данных — групп результатов измерений показателей из тех классов, которые должны получиться после разделения. Сложность и многогранность этой проблемы обусловливают ее актуальность как самостоятельного направления [21]. Рассмотрим некоторые методы решения.

Исходные данные представляют собой группу N k-мерных наблюдений $y_1, y_2, \ldots, y_N$, каждое из которых составлено из результатов измерений k показателей $\{\xi_1, \xi_2, \ldots, \xi_k\}' = \xi : y_i = \{y_{i1}, y_{i2}, \ldots, y_{ik}\}'$ $(i = \overline{1, N})$. Можно выделить два типа задач разделения неоднородной совокупности: а) по линейно упорядоченным данным (y_i пронумерованы в соответствии со следованием их друг за другом по профилю, скважине, возрастанию или убыванию какого-либо показателя и т. п.); б) по неупорядоченным данным. Решение задачи состоит в построении последовательности (возможно, иерархической) классификаций векторов y_i на однородные классы при фиксируемых последовательно возрастающих количествах этих классов J и дальнейшей интерпретации зависимости от J меры однородности получаемых

классов. На каждом этапе построения, т. е. при каждом фиксированном J необходимо производить оптимальное разделение.

В задаче с упорядоченными наблюдениями используем такую модель. Будем считать $\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_N$ наблюдениями в точках $t_1, t_2, \ldots, t_N$ k-мерной случайной функции $\xi(t) = \{\xi_1(t), \xi_2(t), \ldots, \xi_k(t)\}'$, математическое ожидание которой имеет вид $\mathbf{m}(t) = \mathbf{M}\xi(t) = \mathbf{m}_j$ при $t \in Q_j$, $j = \overline{1, J}$. Классам Q_j соответствуют участки шкалы упорядоченности (профиля, скважины и т. п.), составляющие в сумме интервал, на котором наблюдается $\xi(t)$. Для простоты примем корреляционные и взаимные корреляционные функции случайных функций $\xi_i(t)$ $(i = \overline{1, k})$ равными нулю при возможных для данных пунктов наблюдений пар значений t_i, t_j $(i, j = \overline{1, N}$; для автокорреляционных функций $i \neq j)$.

В одномерном случае $(k = 1)$ можно применить метод наименьших квадратов, отождествив с границами между классами максимальные номера r_j наблюдений каждого класса Q_j. Наиболее вероятными будут те положения границ $r_1, r_2, \ldots, r_{J-1}$, при которых обращается в минимум $\rho_1 = \sum_{j=1}^{J} \sum_{i=r_{j-1}+1}^{r_j} (y_i - \bar{y}_j)^2$. Здесь $r_0 = 0$, $r_J = N$; $\bar{y}_j = \frac{1}{r_j - r_{j-1}} \sum_{i=r_{j-1}+1}^{r_j} y_i$ — среднее по наблюдениям j-й группы (с номерами $r_{j-1} + 1, r_{j-1} + 2, \ldots, r_j$).

Величина ρ_1 характеризует компактность классов при различных положениях границ. Если y_i — k-мерные наблюдения $(k \geqslant 2)$, т. е. решается задача выделения границ по комплексу признаков, меру компактности определяют так, чтобы учесть поведение всех измеряемых показателей $\xi_1, \xi_2, \ldots, \xi_k$. Классы разграничивают по общему принципу минимума или максимума такой меры, в зависимости от ее структуры. Способы задания меры могут быть различными. По аналогии с (7.91), например, ее можно определить в виде

$$\rho_k = \sum_{j=1}^{J} \sum_{i=r_{j-1}+1}^{r_j} (\mathrm{y}_{i\alpha} - \bar{\mathrm{y}}_{j\alpha})' (\mathrm{y}_{i\alpha} - \bar{\mathrm{y}}_{j\alpha}), \qquad (8.82)$$

где $\mathrm{y}_{i\alpha} = \{\alpha_1 y_{i1}, \alpha_2 y_{i2}, \ldots, \alpha_k y_{ik}\}'$, $\bar{\mathrm{y}}_{j\alpha} = \frac{1}{r_j - r_{j-1}} \sum_{i=r_{j-1}+1}^{r_j} \mathrm{y}_{i\alpha}$; α_s — коэффициенты масштаба, обратно пропорциональные усредненным по классам средним квадратическим отклонениям ξ_s $(s = \overline{1, k})$. Коэффициенты α_s можно приближенно оценить, используя метод последовательных квадратов. Если отличия математических ожиданий $\mathbf{m}_j$ и $\mathbf{m}_{j+1}$ при переходе от Q_j к Q_{j+1} и число границ невелики, можно взять α_s^2 обратно пропорциональными величинам g_s^2, вычисляемым как g^2 в (6.1) по рядам наблюдений $y_{1s}, y_{2s}, \ldots, y_{Ns}$ каждого ξ_s, исключая,

возможно, резко выделяющиеся наблюдения, а также пары наблюдений, интервалы между которыми явно содержат границы. Минимум ρ_k по $r_1, r_2, \ldots, r_{J-1}$ определяет наиболее вероятное положение границ при каждом J. По зависимости $\rho_k(J)$ минимальных ρ_k от J производят окончательную интерпретацию: определяют вероятные количества классов по скачкам $\rho_k(J)$ (для этого удобно использовать функцию $\rho_k^0(J) = \frac{\rho_k(J+1) - \rho_k(J)}{\rho_k(J)}$) и строят иерархическую последовательность разделений на все более мелкие классы. Для сравнения оптимальных разделений при значительно отличающихся друг от друга значениях J целесообразно взять в качестве меры компактности $\frac{1}{N-J}\rho_k(J)$.

Основная трудность в решении рассмотренной задачи — большое число возможных комбинаций границ. При количестве внутренних границ $J-1$ оно составит C_{N-1}^{J-1}, поэтому полный перебор всех комбинаций возможен только на быстродействующих ЭВМ и при небольших J или N. Для сокращения вычислительной процедуры векторы $\mathbf{y}_i$ группируются в подклассы по метрике (8.82); в дальнейших операциях подклассы неразрывны. Если предполагаемые классы протяженны по шкале упорядоченности, перебор границ можно проводить с некоторым шагом, с последующим разделением интервалов, содержащих границы. Еще один способ, не гарантирующий, впрочем, получение оптимального решения, основан на включении границ между классами, полученных на предыдущем этапе классификации, в число последующих, определяемых при последовательно возрастающих значениях J.

В задаче разделения неоднородной совокупности по неупорядоченным данным (известной еще под названиями *задачи группирования*, *кластер-анализа*, *анализа групп*) используется такая модель. Векторы $\mathbf{y}_i$ рассматриваются как наблюдения k-мерной случайной величины $\boldsymbol{\xi}$, плотность распределения которой при фиксированном числе J классов Q_j имеет вид: $p(\mathbf{x}) = \sum_{j=1}^{J} a_j p_j(\mathbf{x})$, где a_j — доли однородных совокупностей, образующих классы $(a_j \geqslant 0,\ j = \overline{1, J};\ \sum_{j=1}^{J} a_j = 1)$, $p_j(\mathbf{x})$ — плотности распределения в них. В соответствии с этой схемой, в частности, $\mathbf{M}\boldsymbol{\xi} = \sum_{j=1}^{J} a_j \mathbf{M}(\boldsymbol{\xi}/j)$, где $\mathbf{M}(\boldsymbol{\xi}/j)$ — математическое ожидание $\boldsymbol{\xi}$ в Q_j. Неоднородность может выражаться в многомодальности плотностей распределения отдельных показателей, если распределения их в различных классах значительно отличаются друг от друга.

Параметрический способ решения задачи, состоящий в оценке параметров a_j и плотностей $p_j(\mathbf{x})$ с последующей классификацией

векторов y_i, на практике реализовать трудно. Сравнение всех возможных вариантов разделения по простейшим мерам однородности классов (полный перебор) также обычно невозможен даже на быстродействующих ЭВМ. Уже при небольших количествах объектов число этих вариантов чрезвычайно велико. Количество способов, которыми можно разделить N объектов на два класса, составляет $2^{N-1}-1$, а на t классов — $\frac{1}{t!}\sum_{j=0}^{t}(-1)^j C_t^j (t-j)^N$. Поэтому обычно ограничиваются вычислительными процедурами, дающими приближение к оптимальному разделению, хотя достаточная степень такого приближения не всегда может быть гарантирована.

Если в задаче используется только один показатель, наблюдения можно упорядочить по его значениям. Общее решение — по комплексу k показателей — можно реализовать на ЭВМ по такому алгоритму. 1) Для совокупности, подлежащей разделению, строится покрытие, представляющее собой последовательность N окрестностей фиксированного размера вокруг точек y_i $(i=\overline{1, N})$ k-мерного пространства. Размер окрестностей определяется величиной меры сходства включаемых в них точек с центрами окрестностей y_i. 2) Окрестности объединяются в компактные подпокрытия в соответствии с определенной заранее мерой компактности объединений пересекающихся окрестностей. В качестве такой меры можно взять, например, плотность точек y_j в таких объединениях, либо, в простейшем варианте, число общих точек y_j этих окрестностей. Объединяемые вместе с окрестностями их центры — точки y_i — образуют классы. 3) Последовательно увеличивая размер окрестностей, начиная с такого минимального, при котором число классов равно N, определяют последовательность разбиений, каждому из которых ставится в соответствие средняя мера компактности классов. 4) По скачкам зависимости $\mu(J)$ этой меры от числа классов, которые определяются по виду функции $\mu_0(J)=\frac{\mu(J+1)-\mu(J)}{\mu(J)}$, находят наиболее вероятные количества классов.

В простой процедуре подобного рода, при которой группируются точки с пересекающимися окрестностями, предварительно вычисляются расстояния вида (7.91) между всеми парами точек, $d(i, j)=\rho(y_i, y_j)$ $(i<j)$. Полученные значения, число которых $L=C_N^2=\frac{N(N-1)}{2}$, упорядочиваются по возрастанию: $d(i_1, j_1)\leqslant d(i_2, j_2)\leqslant \ldots \leqslant d(i_L, j_L)$. Группирование производится в L шагов. На первом объединяются y_{i_1}, y_{j_1}; на m-м шаге анализируются номера i_m и j_m: если один из них есть в списке номеров одной из групп, имевшихся после $(m-1)$-го шага, то точка с оставшимся номером включается в эту группу; если оба номера оказались среди номеров одной и той же группы, то списки номеров остаются без изменений; если i_m и j_m после $(m-1)$-го шага оказались в списках разных групп, последние объединяются; в остальных случаях y_{i_m} и y_{j_m} образуют отдельную

группу. На каждом шаге вычисляется мера компактности, по значениям которой строится зависимость меры от числа классов.

Методы, разработанные подобно описанной процедуре, по принципу построения иерархической последовательности объединений, представляют самостоятельный интерес. На первом этапе решения такими методами объединяются два наиболее близких между собой по используемой мере сходства вектора y_i, y_j, на втором — либо образуется новая пара наиболее сходных векторов, либо к уже имеющейся группе присоединяется еще один вектор и т. д. На каждом этапе, в соответствии с принципом наибольшей компактности групп, либо объединяются два вектора, либо объединяются две группы предыдущего этапа, либо вектор присоединяется к одной из таких групп. Для проведения всех $N-1$ этапов требуется C^3_{N+1} попарных сравнений. Результаты группирования представляют в виде *дендрограмм* — графиков, изображающих дерево классификации в соответствии с ее иерархической структурой. Уровни объединения на дендрограмме откладываются по высоте сообразно с мерами компактности групп, что дает возможность наглядно оценивать однородность объединений.

Описанные методы используют не только для классификации многомерных наблюдений (Q-тип анализа), но и для группирования показателей, которые представлены этими наблюдениями (R-тип анализа). В последнем в качестве меры сходства можно взять, например, выборочный коэффициент корреляции. Дендрограммы, получаемые в результате подобного анализа, можно использовать при выделении ассоциаций взаимосвязанных химических элементов, минералов и других показателей.

§ 5. Поисковые сети и вероятности пересечения

Исходя из понятия геометрической вероятности, можно предложить простую модель, позволяющую рассчитывать вероятности пересечения профилями односвязных (ограниченных замкнутой кривой) областей заданной формы и размера.

Случай непрерывного прослеживания по профилям. Обозначим L — расстояние между соседними профилями, f_L — проекцию области F на прямую, перпендикулярную к направлению профилей (рис. 45). При $f_L < L$ вероятность пересечения области F, $p = f_L L^{-1}$. Если $f_L \geqslant L$, то $p = 1$. Если целая часть $f_L L^{-1}$, $[f_L L^{-1}] = k \geqslant 1$, то вероятность пересечения F $(k+1)$ непрерывными профилями $p_{k+1} = (f_L - Lk) L^{-1}$, а вероятность пересечения k профилями $p_k = 1 - p_{k+1}$.

Для линейно вытянутых областей F можно приближенно считать $f_L = f|\cos\alpha|$, где α — угол между прямой, перпендикулярной к профилям, и направлением вытянутости; f — длина области в этом направлении. Если α и f неизвестны, то для расчета необходима информация, по крайней мере, о распределении α и f в пределах исследуемого района. Если $t(\alpha, f)$ — плотность совместного распределения α и f $(0 \leqslant \alpha < \pi)$, то полагая $\bar{p}(\alpha, f) = f|\cos\alpha| L^{-1}$ при $f|\cos\alpha| <$

$< L$ и $\bar{p}(\alpha, f) = 1$ при $f|\cos\alpha| \geqslant L$, получим $p = \int_0^{\pi} \int_0^{\infty} \bar{p}(\alpha, f) t(\alpha, f) \times$ $\times\, d\alpha\, df$. Распределение числа областей, которые окажутся пересеченными, при общем количестве областей n и их независимом расположении, описывается схемой Бернулли с параметрами p и n.

Точечные наблюдения по профилям. Обозначим h шаг по профилю прямоугольной сети наблюдений, L — расстояние между соседними профилями. Рассмотрим сначала случай $f_L < L$, где f_L — проекция F на прямую ab, перпендикулярную к профилям. Построим систему координат xOy так, чтобы ось Ox была параллельной ab и обе оси координат касались границы области F (рис. 46). В этой системе точка u пересечения ближайшего к Oy справа профиля с осью

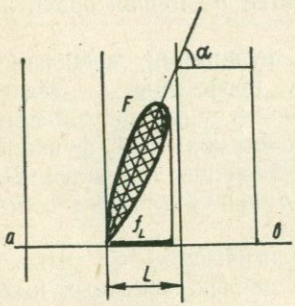


Рис. 45.

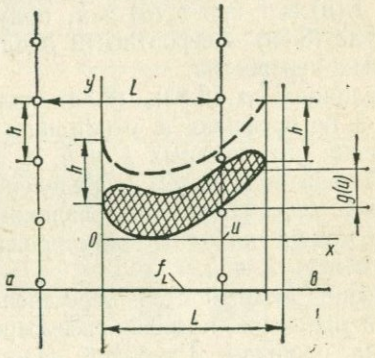


Рис. 46.

Ox распределена равномерно на $[0, L)$ с плотностью распределения $\frac{1}{L}$. Обозначим $g(u)$ длину пересечения F при положении профиля u (рис. 46). Вероятность попадания хотя бы одной точки профиля в F при этом его положении будет $\bar{p}(u) = g(u) h^{-1}$, если $g(u) < h$ и $\bar{p}(u) = 1$, если $g(u) \geqslant h$. Искомая вероятность пересечения — попадания хотя бы одной точки сети наблюдений в F — определится как

$$p = \int_0^{f_L} \bar{p}(u) \frac{du}{L}. \tag{8.83}$$

На рис. 46 показано, как эту величину найти графически: построив на F полосу шириной h, как на рис. 46, вычисляют S_h — площадь части F, заключенной в этой полосе. Вероятность $p = \frac{S_h}{Lh}$.

С помощью формулы (8.83) можно рассчитать вероятность p пересечения и в том случае, когда ориентировка F неизвестна, а имеется для данного района лишь оценка плотности $t(\alpha)$ распределения α — угла наклона оси, фиксирующей эту ориентировку. Положив функ-

цию $p(\alpha)$ равной величине p, вычисляемой по (8.83) при фиксированном α $(0 \leqslant \alpha < \pi)$, получим:

$$p = \int_0^{\pi} p(\alpha) t(\alpha) d\alpha. \tag{8.84}$$

Если $f_L \geqslant L$, для расчета p используем ту же формулу (8.83), приняв в ней $f_L = L$ и определив функцию $\bar{p}(u)$ следующим образом. На отрезок профиля, проходящего через точку u оси Ox (рис. 46), проектируются пересечения остальных профилей ($2u$, $3u$, ..., ku) с областью F, после чего наложением всех таких проекций определяется общая проекция $f_k(u)$. Для простоты рассмотрим случай, когда в результате наложения образуется один отрезок длиной $f_k(u)$. Подставив в (8.83) $f_L = L$ и $\bar{p}(u) = f_k(u) h^{-1}$ при $f_k(u) < h$, $\bar{p}(u) = 1$ при $f_k(u) \geqslant h$, получим искомую вероятность p, а по формуле (8.84) — вероятности p для областей F, точная ориентировка которых неизвестна.

Величины p (8.83), (8.84) являются функциями, зависящими от шага h по профилю и расстояния между профилями L. Задача нахождения оптимальных L и h, а также наилучшего направления профилей решается путём определения максимумов этих функций при условии $Lh = \text{const}$ для различных направлений профилей. Максимумы можно находить табулированием p при различных соотношениях между L и h.

Расчет вероятностей пересечения геологических тел сетью скважин в принципе подобен рассмотренному случаю точечных наблюдений на площади. Глубиной залегания тел можно оперировать как дополнительным параметром. Распределение глубин учитывается также, как распределение параметра α в формуле (8.84).

Объекты поиска на плоскости нередко с достаточной степенью приближения описываются эллиптической формой. В таких случаях при планировании поисковой сети можно воспользоваться таблицами вероятностей пересечения, приведенными в [23].

При сложной конфигурации и многосвязности тел можно использовать метод моделирования на ЭВМ *случайных чисел*. Моделирование случайных чисел состоит в искусственной выработке наблюдений случайной величины, следующей определенному закону распределения. По наблюдениям моделируемых на ЭВМ двух случайных величин, распределенных равномерно на отрезках $[0, L]$ и $[0, h]$, определяются случайные координаты начального пункта ориентированной сети наблюдений, а по ним все остальные. Модель F задается совокупностью пар чисел, определяющих положение элементарных ячеек области F. Повторяя опыт с наложением сети на эту модель и определяя каждый раз результат, можно за счет большой скорости счета на ЭВМ с любой точностью оценить вероятность пересечения в виде обычной оценки вероятности события по результатам независимых испытаний.

Таблица 1. Распределение Пуассона. Вероятности $P\{\xi = i\} = e^{-\lambda}\frac{\lambda^i}{i!}$

i	λ									
	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
0	0,905	0,819	0,741	0,670	0,607	0,549	0,497	0,449	0,407	0,368
1	0,090	0,164	0,222	0,268	0,303	0,329	0,348	0,359	0,366	0,368
2	0,005	0,016	0,033	0,054	0,076	0,099	0,122	0,144	0,165	0,184
3	0,000	0,001	0,003	0,007	0,013	0,020	0,028	0,038	0,049	0,061
4	—	0,000	0,000	0,001	0,002	0,003	0,005	0,008	0,011	0,015
5	—	—	—	0,000	0,000	0,000	0,001	0,001	0,002	0,003
6	—	—	—	—	—	—	0,000	0,000	0,000	0,001

i	λ									
	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2,0
0	0,333	0,301	0,273	0,247	0,223	0,202	0,183	0,165	0,150	0,135
1	0,366	0,361	0,354	0,345	0,335	0,323	0,311	0,298	0,284	0,271
2	0,201	0,217	0,230	0,242	0,251	0,258	0,264	0,268	0,270	0,271
3	0,074	0,087	0,100	0,113	0,126	0,138	0,150	0,161	0,171	0,180
4	0,020	0,026	0,032	0,039	0,047	0,055	0,064	0,073	0,081	0,090
5	0,004	0,006	0,008	0,011	0,014	0,018	0,022	0,026	0,031	0,036
6	0,001	0,001	0,002	0,003	0,004	0,005	0,006	0,008	0,010	0,012
7	0,000	0,000	0,000	0,001	0,001	0,001	0,001	0,002	0,003	0,003
8	—	—	—	0,000	0,000	0,000	0,000	0,000	0,001	0,001

i	λ									
	2,1	2,2	2,3	2,4	2,5	2,6	2,7	2,8	2,9	3,0
0	0,122	0,111	0,100	0,091	0,082	0,074	0,067	0,061	0,055	0,050
1	0,257	0,244	0,231	0,218	0,205	0,193	0,181	0,170	0,160	0,149
2	0,270	0,268	0,265	0,261	0,257	0,251	0,245	0,238	0,231	0,224

* Табл. 1, 3, 4, 6—14 в сокращенном виде взяты из [2], табл. 2 и 15 — из [28].

Продолжение таблицы 1

t	λ									
	2,1	2,2	2,3	2,4	2,5	2,6	2,7	2,8	2,9	3,0
3	0,189	0,197	0,203	0,209	0,214	0,218	0,220	0,222	0,224	0,224
4	0,099	0,108	0,117	0,125	0,134	0,141	0,149	0,156	0,162	0,168
5	0,042	0,048	0,054	0,060	0,067	0,074	0,080	0,087	0,094	0,101
6	0,015	0,017	0,021	0,024	0,028	0,032	0,036	0,041	0,045	0,050
7	0,004	0,005	0,007	0,008	0,010	0,012	0,014	0,016	0,019	0,022
8	0,001	0,002	0,002	0,002	0,003	0,004	0,005	0,006	0,007	0,008
9	0,000	0,000	0,000	0,001	0,001	0,001	0,001	0,002	0,002	0,003
10	—	—	—	—	—	0,000	0,000	0,000	0,001	0,001

t	λ									
	3,1	3,2	3,3	3,4	3,5	3,6	3,7	3,8	3,9	4,0
0	0,045	0,041	0,037	0,033	0,030	0,027	0,025	0,022	0,020	0,018
1	0,140	0,130	0,122	0,113	0,106	0,098	0,091	0,085	0,079	0,073
2	0,216	0,209	0,201	0,193	0,185	0,177	0,169	0,162	0,154	0,147
3	0,224	0,223	0,221	0,219	0,216	0,212	0,209	0,205	0,200	0,195
4	0,173	0,178	0,182	0,186	0,189	0,191	0,193	0,194	0,195	0,195
5	0,107	0,114	0,120	0,126	0,132	0,138	0,143	0,148	0,152	0,156
6	0,056	0,061	0,066	0,072	0,077	0,083	0,088	0,094	0,099	0,104
7	0,025	0,028	0,031	0,035	0,039	0,042	0,047	0,051	0,055	0,060
8	0,010	0,011	0,013	0,015	0,017	0,019	0,022	0,024	0,027	0,030
9	0,003	0,004	0,005	0,006	0,007	0,008	0,009	0,010	0,012	0,013
10	0,001	0,001	0,002	0,002	0,002	0,003	0,003	0,004	0,005	0,005
11	0,000	0,000	0,000	0,001	0,001	0,001	0,001	0,001	0,002	0,002
12	—	—	—	0,000	0,000	0,000	0,000	0,000	0,001	0,001

t	λ									
	4,1	4,2	4,3	4,4	4,5	4,6	4,7	4,8	4,9	5,0
0	0,017	0,015	0,014	0,012	0,011	0,010	0,009	0,008	0,007	0,007
1	0,068	0,063	0,058	0,054	0,050	0,046	0,043	0,040	0,036	0,034
2	0,139	0,132	0,125	0,119	0,112	0,106	0,100	0,095	0,089	0,084
3	0,190	0,185	0,180	0,174	0,169	0,163	0,157	0,152	0,146	0,140
4	0,195	0,194	0,193	0,192	0,190	0,188	0,185	0,182	0,179	0,175
5	0,160	0,163	0,166	0,169	0,171	0,173	0,174	0,175	0,175	0,175
6	0,109	0,114	0,119	0,124	0,128	0,132	0,136	0,140	0,143	0,146
7	0,064	0,069	0,073	0,078	0,082	0,087	0,091	0,096	0,100	0,104
8	0,033	0,036	0,039	0,043	0,046	0,050	0,054	0,058	0,061	0,065
9	0,015	0,017	0,019	0,021	0,023	0,026	0,028	0,031	0,033	0,036
10	0,006	0,007	0,008	0,009	0,010	0,012	0,013	0,015	0,016	0,018
11	0,002	0,003	0,003	0,004	0,004	0,005	0,006	0,006	0,007	0,008
12	0,001	0,001	0,001	0,001	0,002	0,002	0,002	0,003	0,003	0,003
13	0,000	0,000	0,000	0,000	0,001	0,001	0,001	0,001	0,001	0,001
14	—	—	—	—	0,000	0,000	0,000	0,000	0,000	0,000

Продолжение таблицы 1

i	λ									
	5,5	6,0	6,5	7,0	7,5	8,0	8,5	9,0	9,5	10,0
0	0,004	0,002	0,002	0,001	0,001	0,000	0,000	0,000	0,000	—
1	0,022	0,015	0,010	0,006	0,004	0,003	0,002	0,001	0,001	0,000
2	0,062	0,043	0,032	0,022	0,016	0,011	0,007	0,005	0,003	0,002
3	0,113	0,089	0,069	0,052	0,039	0,029	0,021	0,015	0,011	0,008
4	0,156	0,134	0,112	0,091	0,073	0,057	0,044	0,034	0,025	0,019
5	0,171	0,161	0,145	0,128	0,109	0,092	0,075	0,061	0,048	0,038
6	0,157	0,161	0,157	0,149	0,137	0,122	0,107	0,091	0,076	0,063
7	0,123	0,138	0,146	0,149	0,146	0,140	0,129	0,117	0,104	0,090
8	0,085	0,103	0,119	0,130	0,137	0,140	0,138	0,132	0,123	0,113
9	0,052	0,069	0,086	0,101	0,114	0,124	0,130	0,132	0,130	0,125
10	0,029	0,041	0,056	0,071	0,086	0,099	0,110	0,119	0,124	0,125
11	0,014	0,023	0,033	0,045	0,059	0,072	0,085	0,097	0,107	0,114
12	0,007	0,011	0,018	0,026	0,037	0,048	0,060	0,073	0,084	0,095
13	0,003	0,005	0,009	0,014	0,021	0,030	0,040	0,050	0,062	0,073
14	0,001	0,002	0,004	0,007	0,011	0,017	0,024	0,032	0,042	0,052
15	0,000	0,001	0,002	0,003	0,006	0,009	0,014	0,019	0,027	0,035
16	—	0,000	0,001	0,001	0,003	0,005	0,007	0,011	0,016	0,022
17	—	—	0,000	0,001	0,001	0,002	0,004	0,006	0,009	0,013
18	—	—	—	0,000	0,000	0,001	0,002	0,003	0,005	0,007
19	—	—	—	—	—	0,000	0,001	0,001	0,002	0,004
20	—	—	—	—	—	—	0,000	0,001	0,001	0,002
21	—	—	—	—	—	—	—	0,000	0,000	0,001

Таблица 2. Функция (0; 1)-нормального распределения $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt$

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
—0,0	0,500	0,496	0,492	0,488	0,484	0,480	0,476	0,472	0,468	0,464
—0,1	460	456	452	448	444	440	436	433	429	425
—0,2	421	417	413	409	405	401	397	394	390	386
—0,3	382	378	374	371	367	363	359	356	352	348
—0,4	345	341	337	334	330	326	323	319	316	312
—0,5	309	305	302	298	295	291	288	284	281	278
—0,6	274	271	268	264	261	258	255	251	248	245
—0,7	242	239	236	233	230	227	224	221	218	215
—0,8	212	209	206	203	200	198	195	192	189	187
—0,9	184	181	179	176	174	171	169	166	164	161
—1,0	159	156	154	152	149	147	145	142	140	138
—1,1	136	134	131	129	127	125	123	121	119	117
—1,2	115	113	111	109	107	106	104	102	100	099
—1,3	097	095	093	092	090	089	087	085	084	082
—1,4	081	079	078	076	075	074	072	071	069	068
—1,5	067	066	064	063	062	061	059	058	057	056
—1,6	055	054	053	052	051	050	048	047	046	046
—1,7	045	044	043	042	041	040	039	038	038	037

Продолжение таблицы 2

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
—1,8	036	035	034	034	033	032	031	031	030	029
—1,9	029	028	027	027	026	026	025	024	024	023
—2,0	023	022	022	021	021	020	020	019	019	018
—2,1	018	017	017	017	016	016	015	015	015	014
—2,2	014	014	013	013	013	012	012	012	011	011
—2,3	011	010	010	010	010	009	009	009	009	008
—2,4	008	008	008	008	007	007	007	007	007	006
—2,5	006	006	006	006	006	005	005	005	005	005
—2,6	005	005	004	004	004	004	004	004	004	004
—2,7	003	003	003	003	003	003	003	003	003	003
—2,8	003	002	002	002	002	002	002	002	002	002
—2,9	002	002	002	002	002	002	002	001	001	001
—3,0	001	001	001	001	001	001	001	001	001	001
—3,1	001	001	001	001	001	001	001	001	001	001
—3,2	001	001	001	001	001	001	001	001	001	001
—3,3	000	001	000	000	000	000	000	000	000	000
0,0	0,500	0,504	0,508	0,512	0,516	0,520	0,524	0,528	0,532	0,536
0,1	540	544	548	552	556	560	564	567	571	575
0,2	579	583	587	591	595	599	603	606	610	614
0,3	618	622	626	629	633	637	641	644	648	652
0,4	655	659	663	666	670	674	677	681	684	688
0,5	691	695	698	702	705	709	712	716	719	722
0,6	726	729	732	736	739	742	745	749	752	755
0,7	758	761	764	767	770	773	776	779	782	785
0,8	788	791	794	797	800	802	805	808	811	813
0,9	816	819	821	824	826	829	831	834	836	839
1,0	841	844	846	849	851	853	855	858	860	862
1,1	864	867	869	871	873	875	877	879	881	883
1,2	885	887	889	891	893	894	896	898	900	901
1,3	903	905	907	908	910	912	913	915	916	918
1,4	919	921	922	924	925	926	928	929	931	932
1,5	933	934	936	937	938	939	941	942	943	944
1,6	945	946	947	948	949	951	952	953	954	954
1,7	955	956	957	958	959	960	961	962	962	963
1,8	964	965	966	966	967	968	969	969	970	971
1,9	971	972	973	973	974	974	975	976	976	977
2,0	977	978	978	979	979	980	980	981	981	982
2,1	982	983	983	983	884	984	985	985	985	986
2,2	986	986	987	987	987	988	988	988	989	989
2,3	989	990	990	990	990	991	991	991	991	992
2,4	992	992	992	992	993	993	993	993	993	994
2,5	994	994	994	994	994	995	995	995	995	995
2,6	995	995	996	996	996	996	996	996	996	996
2,7	997	997	997	997	997	997	997	997	997	997
2,8	997	998	998	998	998	998	998	998	998	998
2,9	998	998	998	998	998	998	998	999	999	999
3,0	999	999	999	999	999	999	999	999	999	999
3,1	999	999	999	999	999	999	999	999	999	999
3,2	999	999	999	999	999	999	999	999	999	999
3,3	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000

Примечание. Квантиль u_q порядка q определяется как значение z, для которого $\Phi(z) = q$.

Таблица 3. Квантили $\chi^2_q(n)$ распределения χ^2

n	q							
	0,01	0,025	0,05	0,10	0,90	0,95	0,975	0,99
1	0,00	0,00	0,00	0,02	2,71	3,84	5,02	6,64
2	0,02	0,05	0,10	0,21	4,61	5,99	7,37	9,21
3	0,12	0,22	0,35	0,58	6,25	7,82	9,35	11,35
4	0,30	0,48	0,71	1,06	7,78	9,49	11,14	13,28
5	0,55	0,83	1,15	1,61	9,24	11,07	12,83	15,09
6	0,87	1,24	1,64	2,20	10,65	12,59	14,45	16,81
7	1,24	1,69	2,17	2,83	12,02	14,07	16,01	18,48
8	1,65	2,18	2,73	3,49	13,36	15,51	17,54	20,09
9	2,09	2,70	3,33	4,17	14,68	16,92	19,02	21,67
10	2,56	3,25	3,94	4,87	15,99	18,31	20,48	23,21
11	3,05	3,82	4,58	5,58	17,28	19,68	21,92	24,73
12	3,57	4,40	5,23	6,30	18,55	21,03	23,34	26,22
13	4,11	5,01	5,89	7,04	19,81	12,36	24,74	27,69
14	4,66	5,63	6,57	7,79	21,06	23,69	26,12	29,14
15	5,23	6,26	7,26	8,55	22,31	25,00	27,49	30,58
16	5,81	6,91	7,96	9,31	23,54	26,30	28,85	32,00
17	6,41	7,56	8,67	10,09	24,77	27,59	30,19	33,41
18	7,02	8,23	9,39	10,87	25,99	28,87	31,53	34,81
19	7,63	8,91	10,12	11,65	27,20	30,14	32,85	36,19
20	8,26	9,59	10,85	12,44	28,41	31,41	34,17	37,57
21	8,90	10,28	11,59	13,24	29,62	32,67	35,48	38,93
22	9,54	10,98	12,34	14,04	30,81	33,92	36,78	40,29
23	10,20	11,69	13,09	14,85	32,01	35,17	38,08	41,64
24	10,86	12,40	13,85	15,66	33,20	36,42	39,36	42,98
25	11,52	13,12	14,61	16,47	34,38	37,65	40,65	44,31
26	12,20	13,84	15,38	17,29	35,56	38,89	41,92	45,64
27	12,88	14,57	16,15	18,11	36,74	40,11	43,19	46,96
28	13,57	15,31	16,93	18,94	37,92	41,34	44,46	48,28
29	14,26	16,05	17,71	19,77	39,09	42,56	45,72	49,59
30	14,95	16,79	18,49	20,60	40,26	43,77	46,98	50,89
31	15,66	17,54	19,28	21,43	41,42	44,99	48,23	52,19
32	16,36	18,29	20,07	22,27	42,59	46,19	49,48	53,49
33	17,07	19,05	20,87	23,11	43,75	47,40	50,73	54,78
34	17,79	19,81	21,67	23,95	44,90	48,60	51,97	56,06
35	18,51	20,57	22,47	24,80	46,06	49,80	53,20	57,34
36	19,23	21,34	23,27	25,64	47,21	51,00	54,44	58,62
37	19,96	22,11	24,08	26,49	48,36	52,19	55,67	59,89
38	20,69	22,88	24,88	27,34	49,51	53,38	56,90	61,16
39	21,43	23,65	25,70	28,20	50,66	54,57	58,12	62,43
40	22,16	24,43	26,51	29,05	51,81	55,76	59,34	63,69
41	22,91	25,22	27,33	29,91	52,95	56,94	60,56	64,95
42	23,65	26,00	28,14	30,77	54,09	58,12	61,78	66,21
43	24,40	26,79	28,97	31,63	55,23	59,30	62,99	67,41
44	25,15	27,58	29,79	32,49	56,37	60,48	64,20	68,71
45	25,90	28,37	30,61	33,35	57,51	61,66	65,41	69,96
46	26,66	29,16	31,44	34,22	58,64	62,83	66,62	71,20
47	27,42	29,96	32,27	35,08	59,77	64,00	67,82	72,44
48	28,18	30,76	33,10	35,95	60,91	65,17	69,02	73,68
49	28,94	31,56	33,93	36,82	62,04	66,34	70,22	74,92
50	29,71	32,36	34,76	37,69	63,17	67,51	71,42	76,15
51	30,48	33,16	35,60	38,56	64,30	68,67	72,62	77,39
52	31,25	33,97	36,44	39,43	65,42	69,83	73,81	78,62

Продолжение таблицы 3

n	q							
	0,01	0,025	0,05	0,10	0,90	0,95	0,975	0,99
53	32,02	34,78	37,28	40,31	66,55	70,99	75,00	79,84
54	32,79	35,59	38,12	41,18	67,67	72,15	76,19	81,07
55	33,57	36,40	38,96	42,06	68,80	73,31	77,38	82,29
56	34,35	37,21	39,80	42,94	69,92	74,47	78,57	83,51
57	35,13	38,03	40,65	43,82	71,04	75,62	79,75	84,73
58	35,91	38,84	41,49	44,70	72,16	76,78	80,94	85,95
59	36,70	39,66	42,34	45,58	73.28	77,93	82,12	87,17
60	37,49	40,48	43,19	46,46	74,40	79,08	83,30	88,38
61	38,27	41,30	44,04	47,34	75,51	80,23	84,48	89,59
62	39,06	42,13	44,89	48,23	76,63	81,38	85,65	90,80
63	39,86	42,95	45,74	49,11	77,75	82,53	86,83	92,01
64	40,65	43,78	46,60	50,00	78,86	83,68	88,00	93,22
65	41,44	44,60	47,45	50,88	79,97	84,82	89,18	94,42
66	42,24	45,43	48,31	51,77	81,09	85,97	90,35	95,63
67	43,04	46,26	49,16	52,66	82,20	87,11	91,52	96,83
68	43,84	47,09	50,02	53,55	83,31	88,25	92,69	98,03
69	44,64	47,92	50,88	54,44	84,42	89,39	93,86	99,23
70	45,44	48,76	51,74	55,33	85,53	90,53	95,02	100,43
71	46,25	49,59	52,60	56,22	86,64	91,67	96,19	101,62
72	47,05	50,43	53,46	57,11	87,74	92,81	97,35	102,82
73	47,86	51,27	54,33	58,01	88,85	93,95	98,52	104,01
74	48,67	52,10	55,19	58,90	89,96	95,08	99,68	105,20
75	49,48	52,94	56,05	59,80	91,06	96,22	100,84	106,39
76	50,29	53,78	56,92	60,69	92,17	97,35	102,00	107,58
77	51,10	54,62	57,79	61,59	93,27	98,48	103,16	108,77
78	51,91	55,47	58,65	62,48	94,37	99,62	104,32	109,96
79	52,73	56,31	59,52	63,38	95,48	100,75	105,75	111,14
80	53,54	57,15	60,39	64,28	96,58	101,88	106,63	112,33
81	54,36	58,00	61,26	65,18	97,68	103,01	107,78	113,51
82	55,17	58,85	62,13	66,08	98,78	104,14	108,94	114,70
83	55,99	59,69	63,00	66,98	99,88	105,27	110,09	115,88
84	56,81	60,54	63,88	67,88	100,98	106,40	111,24	117,06
85	57,63	61,39	64,75	68,78	102,08	107,52	112,39	118,24
86	58,46	62,24	65,62	69,68	103,18	108,65	113,54	119,41
87	59,28	63,09	66,50	70,58	104,28	109,77	114,69	120,59
88	60,10	63,94	67,37	71,48	105,37	110,90	115,84	121,77
89	60,93	64,79	68,25	72,39	106,47	112,02	116,99	122,94
90	61,75	65,65	69,13	73,29	107,57	113,15	118,14	124,12
91	62,58	66,50	70,00	74,20	108,66	114,27	119,28	125,29
92	63,41	67,36	70,88	75,10	109,76	115,39	120,43	126,46
93	64,24	68,21	71,76	76,01	110,85	116,51	121,57	127,63
94	65,07	69,07	72,64	76,91	111,94	117,63	122,72	128,80
95	65,90	69,93	73,52	77,82	113,04	118,75	123,86	129,97
96	66,73	70,78	74,40	78,73	114,13	119,87	125,00	131,14
97	67,56	71,64	75,28	79,63	115,22	120,99	126,14	132,31
98	68,40	72,50	76,16	80,54	116,32	122,11	127,28	133,48
99	69,23	73,36	77,05	81,45	117,41	123,23	128,42	134,64
100	70,07	74,22	77,93	82,36	118,50	124,34	129,56	135,81

Примечание. q — порядок квантиля, n — число степеней свободы

Таблица 4. Двухсторонние критические границы для количества серий

m	n	Уровни значимости α							
		0,10		0,05		0,02		0,01	
2	2	1	5	1	5	1	5	1	5
	3	1	6	1	6	1	6	1	6
	4	1	6	1	6	1	6	1	6
	5	1	6	1	6	1	6	1	6
	6	1	6	1	6	1	6	1	6
	7	1	6	1	6	1	6	1	6
	8	2	6	1	6	1	6	1	6
	9	2	6	1	6	1	6	1	6
	10	2	6	1	6	1	6	1	6
	11	2	6	1	6	1	6	1	6
	12	2	6	2	6	1	6	1	6
	13	2	6	2	6	1	6	1	6
	14	2	6	2	6	1	6	1	6
	15	2	6	2	6	1	6	1	6
	16	2	6	2	6	1	6	1	6
	17	2	6	2	6	1	6	1	6
	18	2	6	2	6	1	6	1	6
	19	2	6	2	6	2	6	1	6
	20	2	6	2	6	2	6	1	6
3	3	1	7	1	7	1	7	1	7
	4	1	7	1	8	1	8	1	8
	5	2	8	1	8	1	8	1	8
	6	2	8	2	8	1	8	1	8
	7	2	8	2	8	1	8	1	8
	8	2	8	2	8	1	8	1	8
	9	2	8	2	8	2	8	1	8
	10	3	8	2	8	2	8	1	8
	11	3	8	2	8	2	8	1	8
	12	3	8	2	8	2	8	2	8
	13	3	8	2	8	2	8	2	8
	14	3	8	2	8	2	8	2	8
	15	3	8	3	8	2	8	2	8
	16	3	8	3	8	2	8	2	8
	17	3	8	3	8	2	8	2	8
	18	3	8	3	8	2	8	2	8
	19	3	8	3	8	2	8	2	8
	20	3	8	3	8	2	8	2	8
4	4	2	8	1	9	1	9	1	9
	5	2	9	2	9	1	9	1	10
	6	3	9	2	9	2	10	1	10
	7	3	9	2	10	2	10	1	10
	8	3	10	3	10	2	10	2	10
	9	3	10	3	10	2	10	2	10
	10	3	10	3	10	2	10	2	10
	11	3	10	3	10	2	10	2	10
	12	4	10	3	10	3	10	2	10
	13	4	10	3	10	3	10	2	10
	14	4	10	3	10	3	10	2	10
	15	4	10	3	10	3	10	3	10
	16	4	10	4	10	3	10	3	10
	17	4	10	4	10	3	10	3	10
	18	4	10	4	10	3	10	3	10
	19	4	10	4	10	3	10	3	10
	20	4	10	4	10	3	10	3	10
5	5	3	9	2	10	2	10	2	11
	6	3	10	3	10	2	11	2	11
	7	3	10	3	11	2	11	2	12
	8	3	11	3	11	2	12	2	12
	9	4	11	3	12	3	12	2	12
	10	4	11	3	12	3	12	3	12
	11	4	12	4	12	3	12	3	12
	12	4	12	4	12	3	12	3	12
	13	4	12	4	12	3	12	3	12
	14	5	12	4	12	3	12	3	12
	15	5	12	4	12	4	12	3	12
	16	5	12	4	12	4	12	3	12
	17	5	12	4	12	4	12	3	12
	18	5	12	5	12	4	12	4	12
	19	5	12	5	12	4	12	4	12
	20	5	12	5	12	4	12	4	12
6	6	3	11	3	11	2	12	2	12
	7	4	11	3	12	3	12	2	13
	8	4	12	3	12	3	13	3	13
	9	4	12	4	13	3	13	3	14
	10	5	12	4	13	3	14	3	14
	11	5	13	4	13	4	14	3	14
	12	5	13	4	13	4	14	3	14
	13	5	13	5	14	4	14	3	14
	14	5	13	5	14	4	14	4	14
	15	6	14	5	14	4	14	4	14
	16	6	14	5	14	4	14	4	14
	17	6	14	5	14	5	14	4	14
	18	6	14	5	14	5	14	4	14
	19	6	14	6	14	5	14	4	14
	20	6	14	6	14	5	14	4	14
7	7	4	12	3	13	3	13	3	13
	8	4	13	4	13	3	14	3	14
	9	5	13	4	14	4	14	3	15
	10	5	13	5	14	4	15	3	15
	11	5	14	5	14	4	15	4	15
	12	6	14	5	14	4	15	4	16
	13	6	14	5	15	5	16	4	16
	14	6	14	5	15	5	16	4	16
	15	6	15	6	15	5	16	4	16
	16	6	15	6	16	5	16	5	16
	17	7	15	6	16	5	16	5	16
	18	7	15	6	16	5	16	5	16
	19	7	15	6	16	6	16	5	16
	20	7	15	6	16	6	16	5	16
8	8	5	13	4	14	4	14	3	15
	9	5	14	5	14	4	15	3	15

Продолжение таблицы 4

m	n	Уровни значимости α							
		0,10		0,05		0,02		0,01	
	10	6	14	5	15	4	15	4	16
	11	6	15	5	15	5	16	4	16
	12	6	15	6	16	5	16	4	17
	13	6	15	6	16	5	17	5	17
	14	7	16	6	16	5	17	5	17
	15	7	16	6	16	5	17	5	18
	16	7	16	6	17	6	17	5	18
	17	7	16	7	17	6	18	5	18
	18	8	16	7	17	6	18	6	18
	19	8	16	7	17	6	18	6	18
	20	8	17	7	17	6	18	6	18
9	9	6	14	5	15	4	16	4	16
	10	6	15	5	16	5	16	4	17
	11	6	15	6	16	5	17	5	17
	12	7	16	6	16	5	17	5	18
	13	7	16	6	17	6	18	5	18
	14	7	17	7	17	6	18	5	18
	15	8	17	7	18	6	18	6	19
	16	8	17	7	18	6	18	6	19
	17	8	17	7	18	7	19	6	19
	18	8	18	8	18	7	19	6	20
	19	8	18	8	18	7	19	6	20
	20	9	18	8	18	7	19	7	20
10	10	6	16	6	16	5	17	5	17
	11	7	16	6	17	5	18	5	18
	12	7	17	7	17	6	18	5	19
	13	8	17	7	18	6	19	5	19
	14	8	17	7	18	6	19	6	19
	15	8	18	7	18	7	19	6	20
	16	8	18	8	19	7	20	6	20
	17	9	18	8	19	7	20	7	20
	18	9	19	8	19	7	20	7	21
	19	9	19	8	20	8	20	7	21
	20	9	19	9	20	8	20	7	21
11	11	7	17	7	17	6	18	5	19
	12	8	17	7	18	6	19	6	19
	13	8	18	7	19	6	19	6	20
	14	8	18	8	19	7	20	6	20
	15	9	19	8	19	7	20	7	21
	16	9	19	8	20	7	21	7	21
	17	9	19	9	20	8	21	7	22
	18	10	20	9	20	8	21	7	22
	19	10	20	9	21	8	22	8	22
	20	10	20	9	21	8	22	8	22
12	12	8	18	7	19	7	19	6	20
	13	9	18	8	19	7	20	6	21
	14	9	19	8	20	7	21	7	21
	15	9	19	8	20	8	21	7	22
	16	10	20	9	21	8	22	7	22
	17	10	21	9	21	8	22	8	22
	18	10	21	9	21	8	22	8	23

m	n	Уровни значимости α							
		0,10		0,05		0,02		0,01	
	19	10	21	10	22	9	23	8	23
	20	11	21	10	22	9	23	8	23
13	13	9	19	8	20	7	21	7	21
	14	9	20	9	20	8	21	7	22
	15	10	20	9	21	8	22	7	22
	16	10	21	9	21	8	22	8	23
	17	10	21	10	22	9	23	8	23
	18	11	21	10	22	9	23	8	24
	19	11	22	10	23	9	24	9	24
	20	11	22	10	23	10	24	9	24
14	14	10	20	9	21	8	22	7	23
	15	10	21	9	22	8	23	8	23
	16	11	21	10	22	9	23	8	24
	17	11	22	10	23	9	24	8	24
	18	11	22	10	23	9	24	9	25
	19	12	23	11	23	10	24	9	25
	20	12	23	11	24	10	25	9	25
15	15	11	21	10	22	9	23	8	24
	16	11	22	10	23	9	24	9	24
	17	11	22	11	23	10	24	9	25
	18	12	23	11	24	10	25	9	25
	19	12	23	11	24	10	25	10	26
	20	12	24	12	25	11	26	10	26
16	16	11	23	11	23	10	24	9	25
	17	12	23	11	24	10	25	9	26
	18	12	24	11	25	10	26	10	26
	19	13	24	12	25	11	26	10	27
	20	13	25	12	25	11	26	10	27
17	17	12	24	11	25	10	26	10	26
	18	13	24	12	25	11	26	10	27
	19	13	25	12	26	11	27	10	27
	20	13	25	13	26	11	27	11	28
18	18	13	25	12	26	11	27	11	27
	19	14	25	13	26	12	27	11	28
	20	14	26	13	27	12	28	11	29
19	19	14	26	13	27	12	28	12	29
	20	14	77	13	27	12	29	12	29
20	20	15	27	14	28	13	29	12	30

m = n	Уровни значимости α							
	0,10		0,05		0,02		0,01	
21	16	27	15	28	14	29	13	30
22	17	28	16	29	14	31	14	31
23	17	30	16	31	15	32	14	33
24	18	31	17	32	16	33	15	34
25	19	32	18	33	17	34	16	35
26	20	33	19	34	18	35	17	36

Продолжение таблицы 4

$m=n$	Уровни значимости α							
	0,10		0,05		0,02		0,01	
27	21	34	20	35	19	36	18	37
28	22	35	21	36	19	38	18	39
29	23	36	22	37	20	39	19	40
30	24	37	22	39	21	40	20	41
31	25	38	23	40	22	41	21	42
32	25	40	24	41	23	42	22	43
33	26	41	25	42	24	43	23	44
34	27	42	26	43	24	45	23	46
35	28	43	27	44	25	46	24	47
36	29	44	28	45	26	47	25	48
37	30	45	29	46	27	48	26	49
38	31	46	30	47	28	49	27	50
39	32	47	30	49	29	50	28	51
40	33	48	31	50	30	51	29	52
41	34	49	32	51	31	52	29	54
42	35	50	33	52	31	54	30	55
43	35	52	34	53	32	55	31	56
44	36	53	35	54	33	56	32	57
45	37	54	36	55	34	57	33	58
46	38	55	37	56	35	58	34	59
47	39	56	38	77	36	59	35	60
48	40	57	38	59	37	60	35	62
49	41	58	39	60	38	61	36	63
50	42	59	40	61	38	63	37	64
51	43	60	41	62	39	64	38	65
52	44	61	42	63	40	65	39	66
53	45	62	43	64	41	66	40	67
54	45	64	44	65	42	67	41	68
55	46	65	45	66	43	68	42	69
56	47	66	46	67	44	69	42	71
57	48	67	47	68	45	70	43	72
58	49	68	47	70	46	71	44	73
59	50	69	48	71	46	73	45	74
60	51	70	49	72	47	74	46	75
61	52	71	50	73	48	75	47	76
62	53	72	51	74	49	76	48	77
63	54	73	52	75	50	77	49	78

$m=n$	Уровни значимости α							
	0,10		0,05		0,02		0,01	
64	55	74	53	76	51	78	49	80
65	56	75	54	77	52	79	50	81
66	57	76	55	78	53	80	51	82
67	58	77	56	79	54	81	52	83
68	58	79	57	80	54	83	53	84
69	59	80	58	81	55	84	54	85
70	60	81	58	83	56	85	55	86
71	61	82	59	84	57	86	56	87
72	62	83	60	85	58	87	57	88
73	63	84	61	86	59	88	57	90
74	64	85	62	87	60	89	58	91
75	65	86	63	88	61	90	59	92
76	66	87	64	89	62	91	60	93
77	67	88	65	90	63	92	61	94
78	68	89	66	91	64	93	62	95
79	69	90	67	92	64	95	63	96
80	70	91	68	93	65	96	64	97
81	71	92	69	94	66	97	65	98
82	71	94	69	96	67	98	66	99
83	72	95	70	97	68	99	66	101
84	73	96	71	98	69	100	67	102
85	74	97	72	99	70	101	68	103
86	75	98	73	100	71	102	69	104
87	76	99	74	101	72	103	70	105
89	77	100	75	102	73	104	71	106
89	78	101	56	103	74	105	72	107
90	79	102	77	104	74	107	73	108
91	80	103	78	105	75	108	74	109
92	81	104	79	106	76	109	75	110
93	82	105	80	107	77	110	75	112
94	83	106	81	108	78	111	76	113
95	84	107	82	109	79	112	77	114
96	85	108	82	111	80	113	78	115
97	86	109	83	112	81	114	79	116
98	87	110	84	113	82	115	80	117
99	87	112	85	114	83	116	81	118
100	88	113	86	115	84	117	82	119

Примечание. В табл. 4 нижние границы $g\left(\frac{\alpha}{2}, m, n\right)$ количества серий — левые числа столбцов таблицы, верхние границы $G\left(\frac{\alpha}{2}, m, n\right)$ — правые числа столбцов; m и n — количества элементов каждого типа в последовательности из $m+n$ элементов; α — уровни значимости двухсторонних пределов.

Таблица 5. Критические значения $k_{0,05}$ и $k_{0,10}$ для серий максимальной длины (при $m = n$)

$N = m + n$	$k_{0,05}$	$k_{0,10}$	N	$k_{0,05}$	$k_{0,10}$	N	$k_{0,05}$	$k_{0,10}$
2	4,28	3,24	52	8,98	7,94	110	10,06	9,02
4	5,28	4,24	54	9,03	7,99	120	10,18	9,14
6	5,87	4,83	56	9,09	8,02	130	10,30	9,26
8	6,28	5,24	58	9,13	8,09	140	10,41	9,37
10	6,60	5,56	60	9,18	8,14	150	10,50	9,46
12	6,86	5,82	62	9,23	8,19	160	10,60	9,56
14	7,09	6,05	64	9,28	8,24	170	10,68	9,64
16	7,28	6,24	66	9,32	8,28	180	10,76	9,72
17	7,45	6,41	68	9,37	8,33	190	10,85	9,81
20	7,60	6,56	70	9,41	8,37	200	10,95	9,91
22	7,74	6,70	72	9,45	8,41	220	11,06	10,02
24	7,86	6,82	74	9,49	8,45	240	11,18	10,16
26	7,98	6,94	76	9,53	8,49	260	11,30	10,26
28	8,09	7,05	78	9,56	8,52	280	11,40	10,36
30	8,18	7,14	80	9,60	8,56	300	11,50	10,46
32	8,28	7,24	82	9,63	8,59	320	11,60	10,56
34	8,36	7,32	84	9,67	8,63	340	11,69	10,65
36	8,45	7,41	86	9,70	8,66	360	11,77	10,73
38	8,53	7,49	88	9,74	8,70	380	11,84	10,80
40	8,60	7,56	90	9,77	8,73	400	11,92	10,88
42	8,67	7,63	92	9,80	8,76	420	11,99	10,95
44	8,74	7,70	94	9,83	8,79	440	12,06	11,02
46	8,80	7,76	96	9,86	8,82	460	12,12	11,08
48	8,86	7,82	98	9,89	8,85	480	12,18	11,14
50	8,92	7,88	100	9,92	8,88	500	12,24	11,20

Таблица 6. Значения математического ожидания d_n размаха (0; 1)-нормальной совокупности и коэффициенты d_n^{-1} для расчета оценки среднего квадратического отклонения по размаху

n	d_n	d_n^{-1}	n	d_n	d_n^{-1}
2	1,13	0,886	12	3,26	0,307
3	1,69	0,591	13	3,34	0,300
4	2,06	0,486	14	3,41	0,294
5	2,33	0,430	15	3,47	0,288
6	2,53	0,395	16	3,53	0,283
7	2,70	0,370	17	3,59	0,279
8	2,85	0,351	18	3,64	0,275
9	2,97	0,337	19	3,69	0,271
10	3,08	0,325	20	3,73	0,268
11	3,37	0,315			

Примечание. В табл. 6 и 7 n — число наблюдений выборки.

Таблица 7. Квантили $d_n(q)$ размаха (0; 1)-нормального распределения

n	q						n	q					
	0,005	0,025	0,05	0,95	0,975	0,995		0,005	0,025	0,05	0,95	0,975	0,995
2	0,01	0,04	0,09	2,77	3,17	3,97	12	1,55	1,88	2,07	4,62	4,92	5,54
3	0,13	0,30	0,43	3,31	3,68	4,42	13	1,64	1,97	2,16	4,68	4,99	5,60
4	0,34	0,59	0,76	3,63	3,98	4,69	14	1,72	2,06	2,24	4,74	5,04	5,65
5	0,55	0,85	1,03	3,86	4,20	4,89	15	1,80	2,14	2,32	4,80	5,09	5,70
6	0,75	1,06	1,25	4,03	4,36	5,03	16	1,88	2,21	2,39	4,85	5,14	5,74
7	0,92	1,25	1,44	4,17	4,49	5,15	17	1,94	2,27	2,45	4,89	5,18	5,78
8	1,08	1,41	1,60	4,29	4,61	5,26	18	2,01	2,34	2,51	4,93	5,22	5,82
9	1,21	1,55	1,74	4,39	4,70	5,34	19	2,07	2,39	2,57	4,97	5,26	5,95
10	1,33	1,67	1,86	4,47	4,79	5,42	20	2,12	2,45	2,62	5,01	5,30	5,89
11	1,45	1,78	1,97	4,55	4,86	5,49							

Таблица 8. Пределы r_q допустимых значений для $|\check{r}|$ — модуля выборочного коэффициента корреляции независимых нормально распределённых случайных величин ($\mathrm{P}\{|\check{r}| < r_q\} = q$)

ν	q			ν	q		
	0,90	0,95	0,99		0,90	0,95	0,99
1	0,988	0,997	1,000	16	0,400	0,468	0,590
2	900	950	0,990	17	389	456	575
3	805	878	959	18	378	444	561
4	729	811	917	19	369	433	549
5	669	754	875	20	360	423	537
6	0,621	0,707	0,834	25	0,323	0,381	0,487
7	582	666	798	30	296	349	449
8	549	632	765	35	275	325	418
9	521	602	735	40	257	304	393
10	497	576	708	45	243	288	372
11	0,476	0,553	0,684	50	0,231	0,273	0,354
12	457	532	661	60	211	250	325
13	441	514	641	70	195	232	302
14	426	497	623	80	183	217	283
15	412	482	606	90	173	205	267
				100	0,164	0,195	0,254

Примечание. $\nu = n - 2$, где n — число пар наблюдений, по которым вычисляется $\check{r}$. Пределы для выборочного частного коэффициента корреляции определяются по значениям $\nu = n - k - 2$, где k — число фиксируемых компонент.

Таблица 9. Квантили t_q распределения Стьюдента

n	q					n	q				
	0,90	0,95	0,975	0,99	0,995		0,90	0,95	0,975	0,99	0,995
1	3,078	6,314	12,706	31,821	63,657	3	1,638	2,353	3,182	4,541	5,841
2	1,886	2,920	4,303	6,965	9,925	4	1,533	2,132	2,776	3,747	4,604

Продолжение таблицы 9

n	q					n	q				
	0,90	0,95	0,975	0,99	0,995		0,90	0,95	0,975	0,99	0,995
5	1,476	2,015	2,571	3,365	4,032	30	1,310	1,697	2,042	2,457	2,750
6	1,440	1,943	2,447	3,143	3,707	32	1,309	1,694	2,037	2,449	2,739
7	1,415	1,895	2,365	2,998	3,500	34	1,307	1,691	2,032	2,441	2,728
8	1,397	1,860	2,306	2,897	3,355	36	1,306	1,688	2,028	2,435	2,720
9	1,383	1,833	2,262	2,821	3,250	38	1,304	1,686	2,024	2,429	2,712
10	1,372	1,813	2,228	2,764	3,169	40	1,303	1,684	2,021	2,423	2,705
11	1,363	1,796	2,201	2,718	3,106	42	1,302	1,682	2,018	2,419	2,698
12	1,356	1,782	2,179	2,681	3,055	44	1,301	1,680	2,015	2,414	2,692
13	1,350	1,771	2,160	2,650	3,012	46	1,300	1,679	2,013	2,410	2,687
14	1,345	1,761	2,145	2,625	2,977	48	1,299	1,677	2,011	2,407	2,682
15	1,341	1,753	2,131	2,603	2,947	50	1,299	1,673	2,009	2,403	2,678
16	1,337	1,746	2,120	2,584	2,921	55	1,297	1,667	2,004	2,396	2,668
17	1,333	1,740	2,110	2,567	2,898	60	1,296	1,671	2,000	2,390	2,660
18	1,330	1,734	2,101	2,552	2,878	65	1,295	1,669	1,997	2,385	2,654
19	1,328	1,729	2,093	2,540	2,861	70	1,294	1,667	1,994	2,381	2,648
20	1,325	1,725	2,086	2,528	2,845	80	1,292	1,664	1,990	2,374	2,639
21	1,323	1,721	2,080	2,518	2,831	90	1,291	1,662	1,987	2,369	2,632
22	1,321	1,717	2,074	2,508	2,819	100	1,290	1,660	1,984	2,364	2,626
23	1,320	1,714	2,069	2,500	2,807	120	1,289	1,658	1,980	2,358	2,617
24	1,318	1,711	2,064	2,492	2,797	150	1,287	1,655	1,976	2,352	2,609
25	1,316	1,708	2,060	2,485	2,787	200	1,286	1,653	1,972	2,345	2,601
26	1,315	1,706	2,056	2,479	2,779	250	1,285	1,651	1,970	2,341	2,596
27	1,314	1,703	2,052	2,473	2,771	300	1,284	1,650	1,968	2,339	2,592
28	1,313	1,701	2,048	2,467	2,763	400	1,284	1,649	1,966	2,336	2,588
29	1,311	1,699	2,045	2,462	2,756	500	1,283	1,648	1,965	2,334	2,586

Примечание. n — число степеней свободы, q — порядок квантиля.

Таблица 10. 95%-ные квантили $F_{0,95}(\nu_1, \nu_2)$ распределения Фишера с ν_1 и ν_2 степенями свободы

ν_2	ν_1									
	1	2	3	4	5	6	7	8	9	10
1	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,39	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60

Продолжение таблицы 10

ν_2	ν_1									
	1	2	3	4	5	6	7	8	9	10
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,30
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,37
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,22
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83

Продолжение таблицы 10

ν_2	ν_1								
	12	15	20	24	30	40	60	120	∞
1	243,91	245,95	248,01	249,05	250,09	251,14	252,20	253,25	254,32
2	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,37
6	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,29
8	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	2,31	2,23	2,16	2,11	2,07	2,03	2,98	1,93	1,88
20	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73

Продолжение таблицы 10

ν_2	ν_1								
	12	15	20	24	30	40	60	120	∞
25	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

Таблица 11. 97,5%-ные квантили $F_{0,975}(\nu_1, \nu_2)$ распределения Фишера с ν_1 и ν_2 степенями свободы

ν_2	ν_1								
	1	2	3	4	5	6	7	8	9
1	647,79	799,50	864,16	899,58	921,85	937,11	948,22	956,66	963,28
2	38,51	39,00	39,17	39,25	39,30	39,33	39,36	39,37	39,39
3	17,44	16,04	15,44	15,10	14,89	14,74	14,62	14,54	14,47
4	12,22	10,65	9,98	9,60	9,36	9,20	9,07	8,98	8,90
5	10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44
13	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31
14	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21
15	6,20	4,77	4,15	3,80	3,58	3,42	3,29	3,20	3,12
16	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05
17	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98
18	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93
19	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,88
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84
21	5,83	4,42	3,82	3,48	3,25	3,09	2,97	2,87	2,80
22	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76
23	5,75	4,35	3,75	3,41	3,18	3,02	2,90	2,81	2,73
24	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70
25	5,69	4,29	3,69	3,35	3,13	2,97	2,85	2,75	2,68
26	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,65
27	5,63	4,24	3,65	3,31	3,08	2,92	2,80	2,71	2,63
28	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,61
29	5,59	4,20	3,61	3,27	3,04	2,88	2,76	2,67	2,59
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33
120	5,15	3,80	3,23	2,89	2,67	2,52	2,39	2,30	2,22
∞	5,02	3,69	3,12	2,79	2,57	2,41	2,29	2,19	2,11

Продолжение таблицы 11

ν_2 \ ν_1	10	12	15	20	24	30	40	60	120	∞
1	968,63	976,71	984,87	993,10	997,25	1001,4	1005,6	1009,8	1014,0	1018,3
2	39,40	39,42	39,43	39,45	39,46	39,47	39,47	39,48	39,49	39,50
3	14,42	14,34	14,25	14,17	14,12	14,08	14,04	13,99	13,95	13,90
5	8,84	8,75	8,66	8,56	8,51	8,46	8,41	8,36	8,31	8,26
6	6,62	6,52	6,42	6,33	6,28	6,23	6,18	6,12	6,07	6,02
6	5,46	5,37	5,27	5,17	5,12	5,07	5,01	4,96	4,90	4,85
7	4,76	4,67	4,57	4,47	4,42	4,36	4,31	4,25	4,20	4,14
8	4,30	4,20	4,10	4,00	3,95	3,88	3,84	3,78	3,73	3,67
9	3,96	3,87	3,77	3,67	3,61	3,56	3,51	3,45	3,39	3,33
10	3,72	3,62	3,52	3,42	3,37	3,31	3,26	3,20	3,14	3,08
11	3,53	3,43	3,33	3,23	3,17	3,12	3,06	3,00	2,94	2,88
12	3,37	3,28	3,18	3,07	3,02	2,96	2,91	2,85	2,79	2,72
13	3,25	3,15	3,05	2,95	2,89	2,84	2,78	2,72	2,66	2,60
14	3,15	3,05	2,95	2,84	2,79	2,73	2,67	2,61	2,55	2,49
15	3,06	2,96	2,86	2,76	2,70	2,64	2,59	2,52	2,46	2,40
16	2,99	2,89	2,79	2,68	2,63	2,57	2,51	2,45	2,38	2,32
17	2,92	2,82	2,72	2,62	2,56	2,50	2,44	2,38	2,32	2,25
18	2,87	2,77	2,67	2,55	2,50	2,44	2,38	2,32	2,26	2,19
19	2,82	2,72	2,62	2,51	2,45	2,39	2,33	2,27	2,20	2,13
20	2,77	2,68	2,57	2,46	2,41	2,35	2,29	2,22	2,16	2,09
21	2,73	2,64	2,53	2,42	2,37	2,31	2,25	2,18	2,11	2,04
22	2,70	2,60	2,50	2,39	2,33	2,27	2,21	2,14	2,08	2,00
23	2,67	2,57	2,47	2,36	2,30	2,24	2,18	2,11	2,04	1,97
24	2,64	2,54	2,44	2,33	2,27	2,21	2,15	2,08	2,01	1,94
25	2,61	2,51	2,41	2,30	2,24	2,18	2,12	2,05	1,98	1,91
26	2,59	2,49	2,39	2,28	2,22	2,16	2,09	2,03	1,95	1,88
27	2,57	2,47	2,36	2,25	2,20	2,13	2,07	2,00	1,93	1,85
28	2,55	2,45	2,34	2,23	2,17	2,11	2,05	1,98	1,91	1,83
29	2,53	2,43	2,32	2,21	2,15	2,09	2,03	1,96	1,89	1,81
30	2,51	2,41	2,31	2,20	2,14	2,07	2,01	1,94	1,87	1,79
40	2,39	2,29	2,18	2,07	2,01	1,94	1,88	1,80	1,72	1,64
60	2,27	2,17	2,06	1,94	1,88	1,82	1,74	1,67	1,58	1,48
120	2,16	2,05	1,95	1,82	1,76	1,69	1,61	1,53	1,43	1,31
∞	2,05	1,94	1,83	1,71	1,64	1,57	1,48	1,39	1,27	1,00

Таблица 12. Квантили $\zeta_n(q)$ распределения наибольшего по абсолютной величине нормированного отклонения $\zeta_n = \max_i |y_i - \bar{y}|/s$, вычисленного по выборке из нормальной совокупности

n	q = 0,995	0,99	0,95	0,90	n	q = 0,995	0,99	0,95	0,90
3	1,41	1,41	1,41	1,41	10	2,68	2,62	2,41	2,29
4	1,73	1,73	1,71	1,69	11	2,76	2,69	2,47	2,34
5	1,98	1,97	1,92	1,87	12	2,83	2,75	2,52	2,39
6	2,18	2,16	2,07	2,00	13	2,89	2,81	2,56	2,43
7	2,34	2,31	2,18	2,09	14	2,95	2,86	2,60	2,46
8	2,48	2,43	2,27	2,17	15	3,00	2,91	2,64	2,49
9	2,59	2,53	2,35	2,24	16	3,04	2,95	2,67	2,52

Продолжение таблицы 12

n	q				n	q			
	0,995	0,99	0,95	0,90		0,995	0,99	0,95	0,90
17	3,08	2,98	2,70	2,55	35	3,49	3,36	3,02	2,85
18	3,12	3,02	2,73	2,58	36	3,51	3,38	3,03	2,86
19	3,16	3,05	2,75	2,60	37	3,52	3,39	3,04	2,87
20	3,19	3,08	2,78	2,62	38	3,53	3,40	3,06	2,89
21	3,22	3,11	2,80	2,64	39	3,55	3,41	3,07	2,89
22	3,25	3,13	2,82	2,66	40	3,56	3,42	3,08	2,90
23	3,27	3,16	2,84	2,68	41	3,57	3,44	3,08	2,91
24	3,30	3,18	2,86	2,70	42	3,58	3,45	3,09	2,92
25	3,32	3,20	2,88	2,72	43	3,59	3,46	3,10	2,93
26	3,34	3,22	2,90	2,73	44	3,60	3,47	3,11	2,94
27	3,36	3,24	2,91	2,75	45	3,61	3,47	3,12	2,95
28	3,38	3,26	2,93	2,76	46	3,62	3,48	3,13	2 96
29	3,40	3,28	2,94	2,78	47	3,63	3,49	3,14	2,96
30	3,42	3,29	2,96	2,79	48	3,64	3,50	3,15	2,97
31	3,43	3,31	2,97	2,81	49	3,65	3,51	3,15	2,98
32	3,45	3,32	2,99	2,82	50	3,66	3,52	3,16	2,99
33	3,47	3,34	3,00	2,83	51	3,67	3,53	3,17	2,99
34	3,48	3,35	3,01	2,84	52	3,67	3,53	3,18	3,00

Примечание. q — порядок квантиля $\zeta_n(q)$ ($\mathbf{P}\{\zeta_n < \zeta_n(q)\} = q$); n — количество наблюдений в выборке.

Таблица 13. **Верхние критические границы для отношений** $\frac{y_n - y_{n-1}}{y_n - y_1}$, $\frac{y_n - y_{n-1}}{y_n - y_2}$, $\frac{y_n - y_{n-2}}{y_n - y_1}$

n	α				n	α			
	0,005	0,01	0,05	0,10		0,005	0,01	0,05	0,10
3	0,99	0,99	0,94	0,89	8	0,63	0,59	0,47	0,40
	1,00	1,00	1,00	1,00		0,73	0,68	0,55	0,48
	1,00	1,00	1,00	1,00		0,75	0,71	0,61	0,55
4	0,93	0,89	0,77	0,68	9	0,60	0,56	0,44	0,37
	1,00	0,99	0,96	0,91		0,68	0,64	0,51	0,44
	1,00	0,99	0,97	0,94		0,70	0,67	0,57	0,51
5	0,82	0,78	0,64	0,56	10	0,57	0,53	0,41	0,35
	0,94	0,92	0,81	0,73		0,64	0,60	0,48	0,41
	0,95	0,93	0,85	0,78		0,66	0,63	0,53	0,47
6	0,74	0,70	0,56	0,48	11	0,54	0,50	0,39	0,33
	0,84	0,81	0,69	0,61		0,61	0,57	0,45	0,39
	0,87	0,84	0,74	0,67		0,63	0,60	0,50	0,45
7	0,68	0,64	0,51	0,43	12	0,52	0,48	0,38	0,32
	0,78	0,74	0,61	0,53		0,58	0,54	0,43	0,37
	0,81	0,78	0,66	0,60		0,61	0,48	0,48	0,43

Продолжение таблицы 13

n	α				n	α			
	0,005	0,01	0,05	0,10		0,005	0,01	0,05	0,10
15	0,48	0,44	0,34	0,29	24	0,40	0,37	0,28	0,23
	0,52	0,49	0,38	0,32		0,43	0,40	0,31	0,26
	0,55	0,52	0,43	0,38		0,46	0,43	0,35	0,31
20	0,43	0,39	0,30	0,25	30	0,37	0,34	0,26	0,22
	0,46	0,43	0,33	0,28		0,40	0,37	0,28	0,24
	0,49	0,46	0,37	0,33		0,43	0,40	0,32	0,29

Примечание. α — уровень значимости, n — количество наблюдений; y_1 — минимальное, y_n — максимальные наблюдения выборки $y_1, y_2, \ldots, y_n$; y_2, y_{n-1} — минимальное и максимальное наблюдения этой выборки без y_1 и y_n; y_{n-2} — максимальное наблюдение выборки без y_n и y_{n-1}.

Таблица 14. **Критерий Кочрена. Верхние пятипроцентные критические значения** $g(\alpha, l, J-1)$ **для статистики** $G = \dfrac{s^2_{\max}}{s_1^2 + s_2^2 + \ldots + s_l^2}$, **построенной по** l **независимым оценкам дисперсии, каждая из которых вычислена по** J **наблюдениям** ($\alpha = 0{,}05$)

l	$J-1$						
	1	2	3	4	5	6	7
2	0,999	0,975	0,939	0,906	0,877	0,853	0,833
3	967	871	798	746	707	677	653
4	907	768	684	629	590	560	537
5	0,841	0,684	0,598	0.544	0,506	0,478	0,456
6	781	616	532	480	445	418	398
7	727	561	480	431	397	373	354
8	0,680	0,516	0,438	0,391	0,360	0,336	0,319
9	639	478	403	358	329	307	290
10	602	445	373	331	303	282	267
12	0,541	0,392	0,326	0,288	0,262	0,244	0,230
15	471	335	276	242	220	203	191
20	389	271	221	192	174	160	150
24	0,343	0,235	0,191	0,166	0,149	0,137	0,129
30	293	198	159	138	124	114	106
40	237	158	126	108	097	089	083
60	0,174	0,113	0,090	0,077	0,068	0,062	0,058
120	0,100	063	050	042	037	034	031

Продолжение таблицы 14

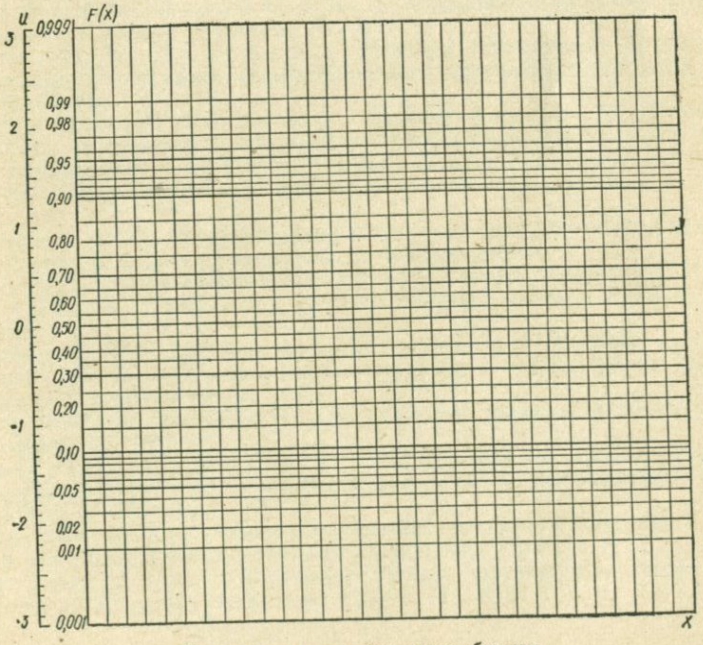
l	$J-1$						
	8	9	10	16	36	144	∞
2	0,816	0,801	0,788	0,734	0,660	0,581	0,500
3	633	617	603	547	475	403	333
4	518	502	488	437	372	309	250
5	0,439	0,424	0,412	0,365	0,307	0,251	0,200
6	382	368	357	314	261	212	167
7	338	326	315	276	228	183	143
8	0,304	0,293	0,283	0,246	0,202	0,162	0,125
9	277	266	257	223	182	145	111
10	254	244	235	203	166	131	100
12	0,219	0,210	0,202	0,174	0,140	0,110	0,083
15	182	174	167	143	114	089	067
20	142	136	130	111	088	068	050
24	0,122	0,116	0,111	0,094	0,074	0,057	0,042
30	100	096	092	077	060	046	033
40	078	075	071	060	046	035	025
60	0,055	0,052	0,050	0,041	0,032	0,023	0,017
120	029	028	027	022	017	012	008

Таблица 15. Односторонне усеченное нормальное распределение. Значения функций $f(y)$, $g(z)$, $\mu_1(z)$ и $\mu_2(z)$

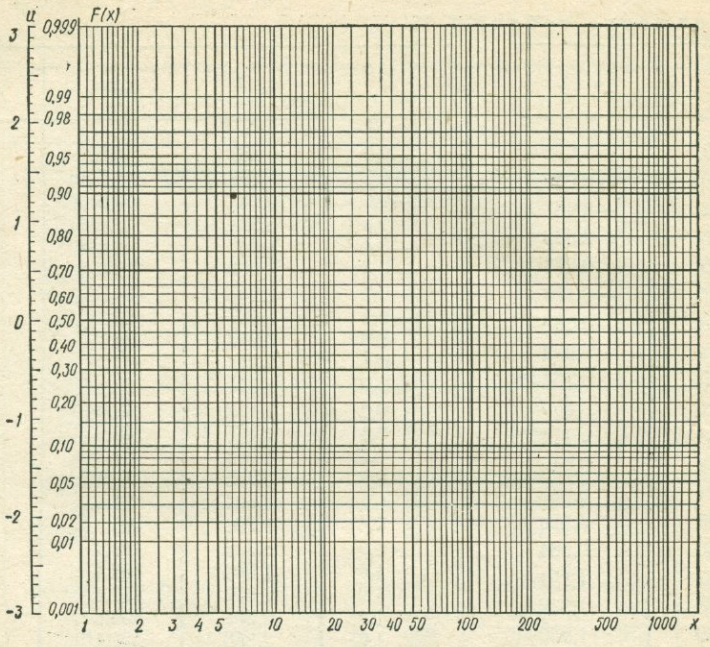
y	$z = f(y)$	z	$g(z)$	$\mu_1(z)$	$\mu_2(z)$
0,55	—3,145	—3,0	0,333	1,015	0,5363
0,56	—2,851	—2,8	0,356	1,026	0,5562
0,57	—2,613	—2,6	0,383	1,044	0,5842
0,58	—2,410	—2,4	0,413	1,074	0,6228
0,59	—2,232	—2,2	0,447	1,124	0,6747
0,60	—2,073	—2,0	0,487	1,204	0,7433
0,61	—1,928	—1,9	0,508	1,260	0,7852
0,62	—1,792	—1,8	0,531	1,332	0,8329
0,63	—1,665	—1,7	0,556	1,424	0,8871
0,64	—1,545	—1,6	0,582	1,540	0,9487
0,65	—1,429	—1,5	0,610	1,688	1,018
0,66	—1,318	—1,4	0,640	1,874	1,097
0,67	—1,209	—1,3	0,671	2,110	1,186
0,68	—1,103	—1,2	0,705	2,408	1,287
0,69	—0,999	—1,1	0,740	2,783	1,400
0,70	—0,896	—1,0	0,777	3,256	1,527
0,71	—0,794	—0,9	0,816	3,849	1,671

Продолжение таблицы 15

y	$z = f(y)$	z	$g(z)$	$\mu_1(z)$	$\mu_2(z)$
0,72	—0,692	—0,8	0,857	4,592	1,831
0,73	—0,589	—0,7	0,899	5,520	2,012
0,74	—0,487	—0,6	0,944	6,677	2,214
0,75	—0,383	—0,5	0,991	8,115	2,440
0,76	—0,277	—0,4	1,040	8,896	2,692
0,77	—0,170	—0,3	1,090	12,09	2,975
0,78	—0,060	—0,2	1,143	14,80	3,290
0,79	0,052	—0,1	1,197	18,12	3,641
0,80	0,168	0,0	1,253	22,19	4,031
0,81	0,289	0,1	1,311	27,14	4,465
0,82	0,414	0,2	1,371	33,16	4,947
0,83	0,545	0,3	1,432	40,44	5,481
0,84	0,683	0,4	1,495	49,23	6,072
0,85	0,829	0,5	1,560	59,81	6,725
0,86	0,984	0,6	1,626	72,49	7,447
0,87	1,151	0,7	1,694	87,63	8,242
0,88	1,332	0,8	1,762	105,7	9,118
0,89	1,530	0,9	1,833	127,1	10,08
0,90	1,749	1,0	1,904	152,4	11,14
0,91	1,994	1,2	2,051	217,4	13,57
		1,4	2,202	306,8	16,47
		1,6	2,358	428,1	19,92
		1,8	2,517	591,1	24,01
		2,0	2,679	807,6	28,81



Нормальная вероятностная бумага



Логнормальная вероятностная бумага

1. Андерсон Т. Введение в многомерный статистический анализ. «Физматгиз», М., 1963.
2. Большев Л. Н., Смирнов Н. В. Таблицы математической статистики. «Наука», М., 1965.
3. Боровко Н. Н. Статистический анализ пространственных геологических закономерностей. «Недра», Л., 1972.
4. Ван дер Варден. Математическая статистика. ИЛ, М., 1960.
5. Вентцель Е. С. Теория вероятностей. «Наука», М., 1964.
6. Гихман И. И., Скороход А. В. Введение в теорию случайных процессов. «Наука», М., 1965.
7. Гнеденко Г. В. Курс теории вероятностей. «Физматгиз», М., 1962.
8. Гренандер У. Случайные процессы и статистические выводы. ИЛ, М., 1961.
9. Дунин-Барковский И. В., Смирнов Н. В. Теория вероятностей и математическая статистика в технике (общая часть). Гос. изд. техн. лит., М., 1955.
10. Жданов М. С., Шрайбман В. И. Корреляционный метод разделения геофизических аномалий. «Недра», М., 1973.
11. Жуков М. Н., Дядюра В. О. Методика статистичного аналізу даних геохімічного випробування. Вісник Київського держуніверситету, серія геологія, № 10, 1968.
12. Жуков Н. Н., Молявко В. Г., Остафийчук И. М., Степченко С. Б. Определение фациальной принадлежности горных пород по данным геохимического опробования методом многомерной классификации на ЭВМ (на примере Волынского массива). Геологический журнал, т. 31, вып. I, 1971.
13. Каратаев Г. И. Корреляционная схема геологической интерпретации гравитационных и магнитных аномалий. Новосибирск, 1966.
14. Коуден Д. Статистические методы контроля качества. «Физматгиз», М., 1961,
15. Крамбейн У., Грейбилл Ф. Статистические модели в геологии. «Мир», М., 1969.
16. Крамер Г. Математические методы статистики. ИЛ, М., 1948.
17. Линник Ю. В. Метод наименьших квадратов и основы теории обработки наблюдений. «Физматгиз», М., 1962.
18. Миллер Р. Л., Кан Дж. С. Статистический анализ в геологических науках. «Мир», М., 1965.
19. Никитин А. А., Тархов А. Г. Статистические приемы извлечения информации при обработке геофизических данных. Геофизический сборник АН УССР, № 35, 1970.
20. Родионов Д. А. Функции распределений содержаний элементов и минералов в изверженных горных породах. «Наука», М., 1964.
21. Родионов Д. А. Статистические методы разграничения геологических объектов по комплексу признаков. «Недра», М., 1968.

22. Рыжов П. А., Гудков В. М. Применение математической статистики для разведки недр земли. «Недра», М., 1969.
23. Савинский И. Д. Таблицы вероятностей пересечения эллиптических объектов прямоугольной сетью наблюдений. «Недра», М., 1964.
24. Толстой М. И., Остафийчук И. М., Гудименко Л. М. К вопросу о типах кривых статистического распределения химических элементов в горных породах и способах расчета их параметров. «Геохимия», № 11, 1965.
25. Хальд А. Математическая статистика с техническими приложениями. ИЛ, М. 1956.
26. Шарапов И. П. Применение математической статистики в геологии. «Недра», М., 1965.
27. Шеффе Г. Дисперсионный анализ. «Физматгиз», М., 1963.
28. Янко Я. Математико-статистические таблицы. Госстатиздат, М., 1961.

СОДЕРЖАНИЕ

Николай Никанорович Жуков

Вероятностно-статистические методы анализа геолого-геофизической информации

Допущено Министерством высшего и среднего специального образования УССР в качестве учебного пособия для студентов геологических специальностей вузов

Издательское объединение «Вища школа»
Головное издательство

Редактор Л. Н. Троян
Обложка художника Д. Д. Грибова
Художественный редактор Н. Н. Панасюк
Технический редактор Л. Ф. Волкова
Корректор Г. М. Довгаль

Сдано в набор 19.04.1974 г. Подписано к печати 23.05.1975 г. Формат бумаги 60×90¹/₁₆. Бумага тип. № 2. Печ. л. 19. Уч.-изд. л. 20,58. Изд. № 1922. Тираж 1000. БФ 08557. Цена 86 коп. Зак. № 4-292

Головное издательство издательского объединения «Вища школа» 252054, Киев, 54, Гоголевская, 7

Книжная фабрика им. М. В. Фрунзе Республиканского производственного объединения «Полиграфкнига», Харьков, Донец-Захаржевская, 6/8.

86 коп.

ВШ