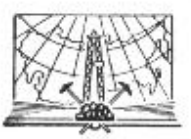
ПРИМЕНЕНИЕ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ В ГЕОЛОГИИ

И. П. Шарапов

И. П. ШАРАПОВ

ПРИМЕНЕНИЕ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ В ГЕОЛОГИИ

(СТАТИСТИЧЕСКИЙ АНАЛИЗ ГЕОЛОГИЧЕСКИХ ДАННЫХ)



ИЗДАТЕЛЬСТВО «НЕДРА»
Москва 1965

У.Д.К. 519.24/.27 : 55

ОТ РЕДАКТОРА

За последнее пятилетие среди геологов значительно возрос интерес к статистическим методам анализа. Это связано с тем, что статистические приемы обработки эмпирических данных позволяют получать обоснованные выводы в условиях неопределенности, что, в свою очередь, повышает эффективность геологических исследований. Поэтому появление настоящей книги весьма своевременно.

В задачу автора не входило систематическое изложение основ теории вероятностей, которые можно найти в обширной отечественной и зарубежной математической литературе. В книге описаны наиболее распространенные приемы статистической обработки фактических данных с наглядной иллюстрацией применения этих приемов к решению геологических задач. Автор уделил основное внимание наиболее простым статистическим критериям, которые не требуют при вычислении большой затраты труда и могут быть использованы геологами в повседневной работе.

Книга рассчитана на читателей, знакомых с математикой в объеме курса геологического вуза, что делает ее доступной для широкого круга геологов.

Следует отметить, что книга И. П. Шарапова является первым опытом создания подобного практического руководства в нашей стране. Имеющаяся обширная отечественная литература по применению математических методов в геологии носит характер тематических исследований и не ставит целью систематическое изложение статистических приемов и их применения при геологических исследованиях. В 1963 г. в США вышла книга Миллера и Кана «Статистический анализ в геологических науках», которая издается на русском языке издательством «Мир» в 1965 г. Однако несмотря на то, что эта книга является более полной, чем работа И. П. Шарапова, она значительно сложнее и ее использование на первых стадиях применения статистики будет затруднительным.

Д. А. РОДИОНОВ

ОТ АВТОРА

Эта книга подготовлена к печати при содействии Пермского научно-исследовательского угольного института, в связи с чем автор выражает глубокую благодарность директору института Л. В. Кучерскому, а также Э. Д. Литвиновой, Э. Ф. Бачурину, Л. А. Леденцову, Г. А. Леденцовой, В. М. Прощенковой, Н. И. Коионенко, М. И. Денисову, И. А. Косицыной, А. Г. Юрьеву, М. С. Трушниковой, Е. П. Емельяновой, М. В. Мартенбаум, В. Т. Шевчуку, В. А. Мантурову, Г. А. Тиховой и А. Н. Костылевой за помощь в оформлении рукописи.

Автор также благодарит коллег по работе Л. Н. Краснопевцева, Л. А. Ренделя и А. И. Тихого за полезные советы, рецензентов П. А. Шехтмана, Ю. В. Прохорова, Л. Ф. Дементьева, Ю. В. Воронина, В. А. Кутолина, Н. А. Добрецова, В. В. Шарапова за внимательный разбор работы, редактора Д. А. Родионова за исправления, сделанные в рукописи, и сотрудников Пермской Центральной Научно-технической библиотеки — за содействие в подборе литературы.

Без помощи названных лиц рукопись еще долго не могла бы быть опубликована.

Все замечания по книге автор просит направлять ему через издательство «Недра» или непосредственно в ПермНИУИ.

«Никакой достоверности нет в науках там, где нельзя приложить ни одной из математических наук, и в том, что не имеет связи с числом».

Леонардо да Винчи

ВВЕДЕНИЕ

Собирая коллекции окаменелостей, минералов, горных пород и делая большое количество однородных наблюдений, геологи интуитивно использовали статистику. Жорж Бюффон в трудах «Теория Земли» (1749 г.) и «Эпохи природы» (1778 г.), М. В. Ломоносов в работе «О слоях земных» (1763 г.), В. М. Севергин в «Опыте минералогического землеописания Российского государства» (1809 г.), Чарльз Лайель в «Принципах геологии» (1833 г.), Д. И. Соколов в «Курсе геогнозии» (1839 г.) и другие выдающиеся ученые сделали некоторые обобщения, имеющие по существу статистический характер.

Статистический анализ геологических данных был проведен впервые в 1899 г. сибирским золоторазведчиком Н. Псаревым, который, исходя из предположения о нормальном распределении содержаний золота, вычислил ошибку оценки среднего содержания его в россыпи и определил количество проб, необходимых для оценки с заданной точностью среднего содержания металла на месторождении.

Позднее статистическим анализом геологических данных занимались С. Ю. Доборжинский (1908, 1909, 1911, 1912), П. Н. Чирвинский (1911, 1918, 1946, 1952, 1953, 1955), Д. В. Наливкин (1925), В. В. Билибин (1930), В. Г. Соловьев (1934, 1937, 1938, 1939, 1946, 1952), Л. И. Шаманский (1935, 1936, 1937, 1938, 1948), И. П. Шарапов (1936—1964), Д. А. Казаковский (1937, 1941, 1948, 1951, 1952, 1954, 1957, 1957, 1959, 1960), П. А. Рыжов (1938, 1940, 1943, 1952, 1954, 1957, 1961), Н. К. Разумовский (1939, 1940, 1941, 1948, 1962), А. Б. Вистелиус (1944—1963), В. В. Богацкий (1948, 1954, 1960, 1962, 1963), В. М. Кузьмин (1952, 1955, 1956, 1957, 1958, 1960, 1961), Л. Ф. Дементьев (1958, 1959, 1960, 1961), Д. А. Родионов (1959, 1961, 1962, 1963), В. П. Бухарцев (1961—1962) в СССР; Крамбейн (Krumbein, 1934—1963), Чейз (Chayes, 1944—1962), Барма (Burma, 1948—1953), Миллер (Miller, 1949—1962), Гриффитс (Griffiths, 1953—1960), Уиттен (Whitten, 1961—1962) в США; Дюваль (Duval, 1949—1962), Матерон (Matheron, 1955—1963), Ломбар (Lombard, 1959—1960), Мюрар (Murard, 1956—1960) во Франции; Сичил (Sichel, 1947—1958), Криг (Krige, 1952—1963), Аренс (Ahrens, 1953—1963) в ЮАР; Янс (Jahns, 1950—1961) в ФГР; Бинтиг (Bintig, 1959—1961) в ГДР; Трембецкий (Trembecki, 1958—1961), Зубржицкий (Zubrzycki, 1959) в Польше; Де-Вийс (De Wijs, 1948—1953) в Нидерландах; Захариев (1952) в Болгарии; Ямагути (Yamaguchi, 1958) в Японии; Асватанаряна (Aswathanarayana, 1956) в Индии; Шайдигер (Scheidegger, 1959—1960) в Канаде и др.

Количество работ, в которых методы математической статистики использованы при решении той или иной геологической задачи, достигло в 1964 г. 2000. Опубликовывались они на русском, английском, польском, немецком, французском, украинском, голландском, болгар-

ском, итальянском, испанском, японском, китайском, чешском, венгерском, румынском, сербском, датском, норвежском, португальском, турецком, литовском и других языках.

Математическая статистика, методы которой нашли широкое применение в геологии, — это наука о количественных закономерностях в природе и человеческом обществе, целью которой является предсказание некоторых явлений.

Практической основой математической статистики являются совокупности однородных величин, получаемые в результате организованного наблюдения, а методологической основой служит теория вероятностей.

Под статистикой понимают: 1) количественные сведения о каком-либо явлении, полученные с помощью метода массового наблюдения; 2) практическую деятельность некоторых учреждений и организаций по сбору данных о массовых явлениях; 3) исследование массовых явлений с целью отыскания закономерных связей между ними и 4) теорию метода массовых наблюдений.

Термин «статистика» употребляют еще и во множественном числе. Под статистиками подразумевают некоторые средние показатели, характеризующие с количественной стороны совокупность однородных величин (замеров, предметов и пр.).

Главнейшие проблемы, изучаемые математической статистикой, следующие: 1) выявление законов распределения случайных величин; 2) разработка методов получения оценок параметров распределения случайных величин; 3) создание способов проверки статистических гипотез; 4) выявление связи между признаками, определение ее характера и измерение ее силы; 5) планирование экспериментов, относящихся к массовым явлениям и могущих дать наиболее точные результаты при наименьшей затрате сил.

Слово «статистик» встречается в трудах ученых IV в. Не менее древним является слово «статистический», но термин «статистика» (status — по-латыни — состояние) впервые был употреблен в труде Дж. Бильфельда «Основы универсальной эрудиции» (1770 г.). Одна из глав этой книги получила название «Статистика». Под статистикой Дж. Бильфельд понимал учение о политическом устройстве государства (Дж. Юл и М. Кендэл, 1960).

Более широкое понятие о статистике дано Готтфридом Ахенваллем (1794 г.), который этим термином назвал возникшую тогда описательную науку о государствоведении, рассматривавшую вопросы, впоследствии вошедшие в экономическую географию (данные о населении страны, о политических условиях, о продукции, производимой в стране, и т. д.).

В дальнейшем статистика стала не только описательной, но и аналитической наукой, что оказалось возможным благодаря применению методов теории вероятностей.

Теория вероятностей, без которой сейчас невозможна ни кинетическая теория газов, ни теория стрельбы, ни теория ошибок, ни ядерная физика, ни расчет космических полетов, ни многие другие теории, как ни странно, возникла при решении задач, связанных с азартными играми.

Арабское слово «азар» по-русски можно перевести как «трудный». Если какая-либо комбинация очков при бросании одновременно двух или трех игральных костей могла возникнуть одним единственным способом из числа многих возможных, то игроки говорили: «азар», т. е. трудный случай.

Многие выдающиеся математики в XV—XVII вв. и позднее трудились над теорией азартных игр (игры в кости, рулетку, карты, «орлянку» и т. д.). Предрасчет трудных комбинаций очков или карт, определение шансов игроков и другие задачи этой теории доставляли наслаждение математическому интеллекту. Позднее теория азартных игр оказалась могучим орудием при исследованиях в артиллерийском деле, астрономии,

страховом деле и т. д. В наше время теория азартных игр дополнилась более общей теорией игр, или теорией конфликтных ситуаций.

Одним из первых математиков, проявивших научный интерес к азартным играм, был итальянец Лука Пачиоло (1445—1514 гг.). Он искал решение задачи о разделении ставки между игроками. Так как с этой задачей связаны лучшие работы по теории вероятностей, то приведем ее краткое описание.

Два игрока уговорились сыграть ряд партий. Победителем должен считаться тот, кто выиграет определенное число (S) партий. Начинается игра, но вдруг она почему-либо прекращается, когда один из игроков выиграл a, а другой b партий, причем и a и b меньше S. Как в этом случае разделить ставку между игроками?

Пачиоло считал, что ставка должна быть разделена пропорционально числам a и b.

Позднее Джероламо Кардано (1501—1576 гг.) в своем сочинении «Об азартной игре» при решении той же задачи учитывал дополнительно и то число партий, которое не хватает каждому игроку до выигрыша. Он дал другое решение задачи, хотя и оно было неудовлетворительным. Кардано, кроме того, занимался подсчетом числа способов (из всех возможных) для получения суммы того или иного количества очков при одновременном бросании двух или трех игральных костей.

Математик Никколо Тарталья (1499—1557 гг.) подсчитывал различные случаи выпадения того или иного количества очков при игре в кости.

Задачу о разделении ставки между игроками решили французские математики Блез Паскаль (1623—1662 гг.) и Пьер Ферма (1601—1665 гг.), а также голландец Христиан Гюйгенс (1629—1695 гг.), причем все эти решения были достигнуты различными способами (Гнеденко, 1954, стр. 362—363). Гюйгенс, кроме задачи о разделении ставки между игроками, сформулировал много других задач, связанных с азартными играми. Некоторые из этих задач ему удалось решить.

Новый этап в развитии теории вероятностей связан с трудами швейцарского математика Якоба Бернулли (1654—1705 гг.). Последний решил некоторые задачи Гюйгенса, а одну из них (задачу о разорении игрока) ему решить не удалось, но он ее более точно сформулировал, что помогло новым поколениям математиков в ее решении. Важнейшей заслугой Я. Бернулли является доказательство теоремы, носящей теперь его имя, — одной из основных в теории вероятностей.

Английский ученый Абрахам де Муавр (1667—1754 гг.) в труде «Учение о случае» (1718 г.) развил методы решения задач из теории азартных игр. В другом его сочинении («Аналитический сборник», 1730 г.) доказана теорема, носящая теперь имя Муавра—Лапласа (Лаплас ее обобщил).

Английский математик Томас Симпсон (1710—1761 гг.) в книге «Природа и законы случая» (1740 г.) рассмотрел задачу, связанную с контролем качества продукции (вычисление вероятностей того, что в отобранной браковщиком пробе окажется определенное соотношение сортов).

Интересную задачу сформулировал Николай Бернулли (1687—1759 гг.). Математикам всего мира она известна под названием «петербургской игры». Сущность ее такова: «Петр подбрасывает монету раз за разом до тех пор, пока она не ляжет «орлом» кверху. С первым выпадением «орла» игра считается законченной. Если «орел» выпадет при первом бросании, то Павел, другой игрок, выплачивает Петру рубль. Если «орел» выпадет при втором бросании монеты, то Павел платит два рубля, при третьем — четыре рубля, при четвертом — восемь рублей и т. д. Вообще, если «орел» впервые выпадет только при n-ом бросании, то Петр получает 2^{n-1} рублей. Для того чтобы игра была безобидной (т. е., чтобы игроки имели равные шансы на выигрыш), Петр заранее дает Павлу какую-то

сумму денег». Какой должна быть эта сумма? В этом как раз и состоит задача.

Русский академик Даниил Бернулли (1700—1782 гг.), французские ученые Жан Лерон Даламбер (1717—1783 гг.), Жорж Луи Бюффон (1707—1788 гг.), Жан Антуан де Кондорсе (1743—1794 гг.), Симон Дени Пуассон (1781—1840 гг.) и многие другие долго искали решение этой задачи. Одно из таких решений приведено в книге Б. В. Гнеденко (1954).

Естествоиспытатель Жорж Луи Бюффон в молодости увлекался математикой. Его труды «Теория Земли» (1749 г.) и «Эпохи природы» (1778) оставили глубокий след в истории геологии. В 1777 г. Бюффон издал труд «Опыт моральной арифметики», в котором теория вероятностей была применена в геометрии, в частности для определения числа π. Для этой цели Бюффон много раз бросал иглу на пол и подсчитывал число случаев пересечения границы между двумя соседними половицами в общем числе бросаний. Игла имела длину, меньшую, чем ширина половицы, а все половицы были одной ширины. Задача состоит в том, чтобы найти вероятность того, что игла пересечет границу между половицами.

Число π Бюффон вычислял по формуле

$$\pi = \frac{2l}{ap},$$

где l — половина длины иглы;

a — половина ширины половицы;

p — отношение числа случаев пересечения иглой границы между половицами к общему числу бросаний.

Пересечение границы связано с кратчайшим расстоянием x от середины иглы до ближайшей границы и с углом φ между направлением иглы и направлением границы. При этом $x = l \sin \varphi$, а вероятность пересечения

$$p = \frac{1}{a\pi} \int_0^{\pi} l \sin \varphi \, d\varphi.$$

Из этой формулы видно, что для вычисления π не требуется измерять ни x, ни φ, а достаточно подсчитать p.

Б. В. Гнеденко приводит следующую таблицу результатов эксперимента с иглой (табл. 1):

Таблица 1

Экспериментатор	Год	Число бросаний иглы	Экспериментальное значение числа π
Вольф	1850	5000	3,1596
Смит	1855	3204	3,1553
Фокс	1894	1120	3,1419
Лаццарини	1901	3408	3,1415929

Задача с бросанием иглы имеет важное значение не только для определения значения π, но и для решения некоторых проблем в теории стрельбы. В настоящее время задача об игле используется в теории оценки зернистой структуры металла, т. е. в металлографии.

Петербургский академик Леонард Эйлер (1707—1783 гг.) решил ряд задач из теории азартных игр, страхового дела, демографии и других, связанных с теорией вероятностей.

Англичанин Томас Бейес в середине XVIII в. вывел формулу вероятности гипотез.

Жан Антуан де Кондорсе пытался применить теорию вероятностей к определению степени справедливости судебных приговоров, вынесенных по большинству голосов.

Очень важное значение для теории вероятностей имеют исследования французского математика Пьера Симона Лапласа (1749—1827 гг.). Первая его работа из серии исследований по теории вероятностей была им опубликована еще в молодости (1774 г.). Его большой труд «Аналитическая теория вероятностей», впервые изданная в 1812 г. и затем много раз переиздававшаяся, является классическим трудом в этой области. В разделе этой книги «Опыт философии теории вероятностей» Лаплас пишет: «Наступит день, когда благодаря длившемуся несколько столетий изучению вещи, ныне скрытые, явятся со всей своей очевидностью; и потомки наши удивятся, что столь очевидные истины ускользнули от нас» (Гнеденко, 1954, стр. 370).

В математической статистике большое значение имеет способ наименьших квадратов, разработанный Лапласом и Адриеном Мари Лежандром (1752—1833 гг.).

Этот же способ, независимо от Лапласа и Лежандра, предложил в 1821—1823 гг. немецкий математик Карл Фридрих Гаусс (1777—1855).

Выдающийся русский математик Н. И. Лобачевский (1792—1856 гг.) написал две работы, относящиеся к теории вероятностей. В «Пангеометрии» Н. И. Лобачевский высказался за вероятностный метод решения некоторых геометрических проблем.

Бельгийский астроном Адольф Кетле (1799—1874 гг.) издал в 1846 г. большой труд по теории вероятностей. По инициативе этого ученого в 1853 г. в Брюсселе был созван первый в истории статистический конгресс.

Академик М. В. Остроградский (1801—1861 гг.) написал три работы по теории вероятностей и три по математической статистике. В статье «Об одном вопросе о вероятностях» (1846 г.) Остроградский пишет, что с помощью некоторых приведенных им формул можно раз в двадцать уменьшить работу по проверке качества «. . .очень большого числа мешков с мукой или кусков сукна. . .».

В. Я. Буняковский (1804—1889 гг.), бывший главным экспертом русского правительства по вопросам статистики, издал в 1846 г. первый в России курс теории вероятностей («Основания математической теории вероятностей»).

Огромное значение для теории вероятностей имели труды русского ученого П. Л. Чебышева (1821—1894 гг.), который ввел в науку и широко использовал понятие случайной величины и создал метод моментов. П. Л. Чебышев создал русскую школу теории вероятностей.

Английский ученый Фрэнсис Гальтон (1826—1911 гг.) впервые применил корреляционный анализ статистических величин. Гальтон совместно с Уэлдоном и Карлом Пирсоном основал в 1891 г. журнал «Биометрика», в котором печатались статьи по применению методов математической статистики в биологии и по общим вопросам теории вероятностей. Этот журнал сыграл важную роль в развитии математической статистики. Впоследствии возникли подобные журналы: «Эконометрика», «Психометрика», «Технометрика» и др.

Немецкий статистик Вильгельм Лексис (1837—1914 гг.) изучал «устойчивость», т. е. постоянство или повторяемость статистических показателей. Критерий Лексиса, выведенный в 1877 г., используется для проверки испытаний на их независимость и постоянство вероятностей.

Русский экономист А. И. Чупров написал в 1886 г. учебник по статистике, по которому впоследствии учились Г. В. Плеханов и В. И. Ленин.

В конце XIX в. математическая статистика увлеклась психологическими и моральными проблемами. Тогда, например, в Западной Европе была опубликована статистика душевных свойств человека, статистика умственных способностей школьников, уголовная статистика и т. п. В одном из подобных сочинений исследовались характерные черты гекзаметра Виргилия с помощью статистики. В настоящее время академик

А. Н. Колмогоров исследует статистическими методами ямб Пушкина, Лермонтова и других русских поэтов.

Ученик П. Л. Чебышева А. А. Марков (1856—1922 гг.) разработал теорию «испытаний, связанных в цепь». В настоящее время «цепи» Маркова лежат в основе теории радиоактивного распада, а А. Б. Вистелиус использовал эту теорию для создания математической модели слоеобразования.

Другой ученик Чебышева А. М. Ляпунов (1857—1918 гг.) доказал одну предельную теорему, ныне носящую его имя, и создал метод характеристических функций, примененный для исследования законов распределения.

Англичанин Карл Пирсон (1857—1936 гг.) развил новое, биометрическое направление в математической статистике, основанное Гальтоном. Им впервые введен в математическую статистику коэффициент вариации.

В 1911 г. издан учебник статистики петербургского профессора А. А. Кауфмана «Теория и методы статистики».

Профессор берлинского университета В. И. Борткевич (1868—1931 гг.) успешно изучал вероятности редких событий.

Русский ученый А. А. Чупров (1874—1926 гг.) написал много трудов по теории математической статистики. Его труд по теории корреляции, изданный в 1926 г., имеет и ныне важное значение.

В. М. Обухов (1874—1945 гг.) успешно применял методы математической статистики в метеорологии и агрономии.

Советский математик Е. Е. Слуцкий (1880—1948 гг.) разработал теорию случайных функций и теорию корреляции.

Выдающимся статистиком был советский ученый В. И. Романовский (1879—1954 гг.), написавший монографию «Дискретные цепи Маркова» и много других исследований. Его фундаментальный курс математической статистики (Романовский, 1938) и сейчас имеет большую ценность.

Английский статистик Вильям Госсет, писавший свои работы под псевдонимом Стьюдент, разработал теорию малой выборки.

Стационарные случайные процессы исследовал советский математик А. Я. Хинчин (1894—1959 гг.).

В настоящее время в области теории вероятностей и математической статистики работают такие крупные ученые, как А. Н. Колмогоров, С. Н. Бернштейн, Ю. В. Линник, Б. В. Гнеденко, Н. В. Смирнов, Е. Б. Дынкин, Ю. В. Прохоров, О. В. Сарманов (СССР), Г. Крамер (Швеция), Р. Фишер, Е. Пирсон, Д. Дуб (США), М. Кендэл (Англия).

В последнее время успешно развиваются отпочковавшиеся от теории вероятностей самостоятельные дисциплины — теория информации, связанная с кибернетикой, и теория игр.

I. ОСНОВНЫЕ ПОЛОЖЕНИЯ ТЕОРИИ ВЕРОЯТНОСТЕЙ

Основой математической статистики является теория вероятностей, изучающая количественные закономерности массовых случайных явлений.

Понятиями, с которыми приходится часто встречаться в теории вероятностей, считаются испытание, событие, случайная величина.

И с п ы т а н и е м называется такое действие, при котором реализуется некоторый комплекс условий, необходимых для возникновения какого-либо явления. Более упрощенно под испытанием понимают реализацию комплекса условий. Эту реализацию называют также опытом. Не следует смешивать данное (вероятностное) понятие опыта с его обычным, например физическим или химическим, понятием.

Классическим примером испытания в теории вероятностей является извлечение наугад шара из урны, содержащей большое число шаров. В геологии примером испытания (с некоторой долей условности) будет взятие пробы аллювия и ее анализ.

Совсем не обязательно, чтобы испытания производились людьми или наблюдались ими. Есть многие явления, протекающие без какого-либо участия и даже без присутствия человека и которые можно рассматривать как последовательность испытаний. К ним относятся, например, вулканические извержения, работа морского прибоя, перемещение льдинами эрратических валунов и т. п.

Явление, возникшее в результате испытания, называется исходом испытания, или с о б ы т и е м.

Классическим примером события в теории вероятностей считается извлечение из урны с черными и белыми шарами шара того или иного цвета.

В геологии, например, можно считать событием присутствие (или отсутствие) алмаза в одной из многих проб речного песка, обрыв штанги в процессе бурения одной из многих скважин, появление недопустимо большой ошибки при анализе одной из многих геологических проб и т. д. Во всех этих и им подобных случаях событие будет заключаться в появлении (или непоявлении) признака в одном из многих испытаний.

События бывают трех типов. Одни из них неизбежно возникают при каждом испытании данного вида. Это д о с т о в е р н ы е с о б ы т и я. Другие, наоборот, никогда не появляются. Это н е в о з м о ж н ы е с о б ы т и я.

События третьего типа характеризуются тем, что они в данном испытании могут произойти, а могут и не произойти. Если испытание повторяется многократно, то эти события в одних случаях произойдут, а в других нет. В каких именно случаях они произойдут, а в каких нет — заранее сказать нельзя. Такие события называются с л у ч а й н ы м и.

На практике обычно отступают от приведенного здесь строгого определения достоверного и невозможного события и говорят о практически достоверном и практически невозможном событиях.

Так, например, встреча вечной мерзлоты одним, наудачу заданным шурфом, будет событием практически достоверным для района г. Алдана и практически невозможным для района Железноводска. Слово «практически» в обоих этих случаях употребляется потому, что в районе Алдана кое-где вечная мерзлота отсутствует (участки таликов), а в районе Железноводска местами даже в середине лета сохраняется лед, образующийся при переохлаждении паров воды в результате резкого падения давления в струе углекислого газа, поднимающегося с глубин земли.

Приведем пример случайного события в геологии. На золотой россыпи пробито 100 шурфов. Все они полностью пересекли золотоносный пласт. Осмотр вынутой из шурфов породы показал, что в некоторых шурфах встречаются валуны. Если из всех 100 шурфов наудачу выбрать один, то в нем могут оказаться валуны. Однако предсказать точно, заранее это событие (до того как шурф пройден) нельзя. Точно так же нельзя сказать уверенно, что в шурфе не будут встречены валуны. Оба эти события в результате единичного эксперимента (выбора наудачу одного шурфа) могут произойти, а могут и не произойти и их следует рассматривать как случайные.

Теория вероятностей и математическая статистика имеют дело с массовыми явлениями (событиями), т. е. такими, которые могут воспроизводиться очень большое число раз.

В геологии массовыми явлениями (событиями) можно считать многократное измерение мощности пласта, определение содержания металла во многих пробах руды и т. д.

События бывают простые и сложные.

Простое событие не разлагается на другие, более мелкие события. Так, появление валуна в одном наудачу взятом шурфе предыдущего примера будет простым событием.

Сложные события представляют собой комбинации простых событий. Так, например, при однократном бросании игральной кости может наступить одно из шести простых событий, заключающихся в выпадении числа очков 1, 2, 3, 4, 5, 6. Событие же, заключающееся в выпадении четного числа очков, будет сложным, состоящим из трех простых событий: или 2, или 4, или 6.

Если наступление события A обязательно влечет за собой наступление события B, то событие B является сложным. Это соотношение записывается как

$$A \subset B,$$

что значит: A влечет за собою B или A содержится в B.

Если при этом и B содержится в A, т. е. если и

$$A \supset B,$$

то оба эти события совпадают и называются эквивалентными. Это записывается так:

$$A = B.$$

События бывают совместимые и несовместимые. Два или более событий называются совместимыми, если они могут одновременно наступить при осуществлении одного испытания. Иными словами, это события, которые содержат одни и те же простые события. Если, например, событие A состоит из событий E_1, E_2, E_3, а событие B — из E_3 и E_4, то события A и B будут совместимыми, поскольку в каждое из них входит событие E_3.

Несовместимыми называют такие события, которые не могут наступить одновременно при одном опыте, т. е. они не содержат ни одного общего события. Если событие A состоит из событий E_1, E_2 и E_3, а событие B — из E_4 и E_5, таких, что ни одно из событий в A не совпадает с событиями из B, то события A и B несовместимы.

Поясним теперь, что означает выражение «событие A состоит из E_1, E_2 и E_3». Это означает, что в результате одного испытания событие A будет представлять собою или событие E_1, или событие E_2, или, наконец, событие E_3. Это можно записать так:

$$A = E_1 + E_2 + E_3,$$

или

$$A = E_1 \cup E_2 \cup E_3.$$

В левой части этого равенства стоит название сложного события A, а в правой — сумма возможных элементарных исходов (результатов) испытания. Знаки $+$ и $\cup$ означают здесь «или».

На основании сказанного приведенные выше примеры совместимых и несовместимых событий можно записать как новые события C_1 и C_2:

$$C_1 = A + B = (E_1 + E_2 + E_3) + (E_3 + E_4) = E_1 + E_2 + E_3 + E_4;$$

$$C_2 = A + B = (E_1 + E_2 + E_3) + (E_4 + E_5).$$

Совокупность всех возможных событий, которые могут произойти в результате одного испытания, называется множеством возможных событий. Математическое представление об этом множестве дано в классическом труде Э. Бореля «Случай» и называется борелевским полем событий или просто полем событий. Оно содержит не только случайные события, но также невозможное и достоверное, т. е. образует полную систему событий.

Если поле событий построено исходя из n элементарных событий (E_1, E_2, . . ., E_n), то их сумма будет являться достоверным событием. Сущность последнего будет заключаться в том, что одно испытание обязательно даст нам один какой-либо исход из числа всех возможных, т. е. из числа $E_1 + E_2 . . . E_n$. Следует отметить, что сумма элементов множества элементарных событий входит как составная часть в поле событий.

Если какое-либо событие одновременно принадлежит двум сложным событиям, то это символически обозначается как произведение двух событий. Так, если $A = E_1 + E_2 + E_3$, а $B = E_3 + E_4$, то произведение сложных событий $AB = E_3$. Это произведение E_3 иначе называют пересечением событий A и B.

Сказанное иллюстрируется на рис. 1. Поле событий, изображенное здесь, состоит из девяти событий, обозначенных номерами от *1* до *9*. Из основного множества событий, т. е. из этих 9 элементарных событий, выделено два сложных (A и B) и два простых (C и D) события. Сложное событие A состоит из трех простых событий (*1*, *4* и *7*), а сложное событие B представлено четырьмя простыми (*4*, *5*, *7* и *8*). Простое событие C имеет номер *3*, а простое событие D — номер *8*. События A, C и D несовместимы, а A и B совместимы. Событие D влечет событие B, т. е. $D \subset B$. Произведение событий: $AB = E_4 + E_7$ (знак E на рисунке не представлен, указаны только номера событий). Сумма событий: $A + B = E_1 + E_4 + E_5 + E_7 + E_8$; сумма: $B + D = B = E_4 + E_5 + E_7 + E_8$; произведение: $BD = D = E_8$.

Взаимоотношение событий можно иллюстрировать не только точечным, но и площадным способом. На рис. 2 показано 9 случаев взаимоотношения событий: а) сумма или объединение событий A и B, т. е. событие $A + B$ ($A \cup B$); *б*) произведение или пересечение событий A и B, т. е. событие AB ($A \cap B$); *в*) симметричная разность событий A и B (или объединение неповторяющихся элементов двух событий), т. е. собы-

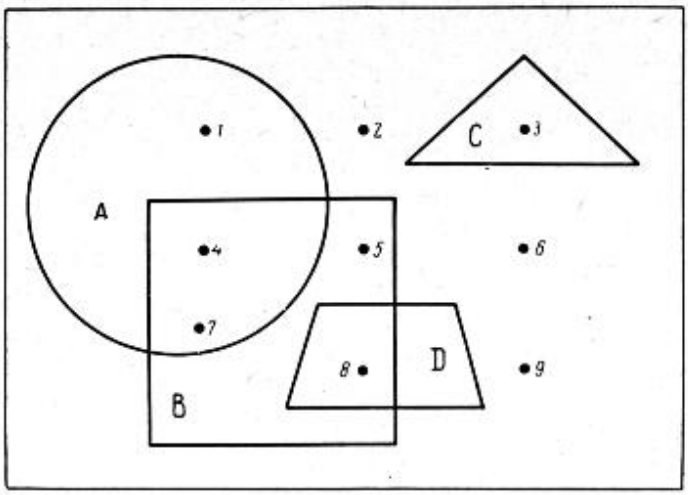


Рис. 1. Графическая интерпретация поля событий (по Н. В. Смирнову и И. В. Дунину-Барковскому)

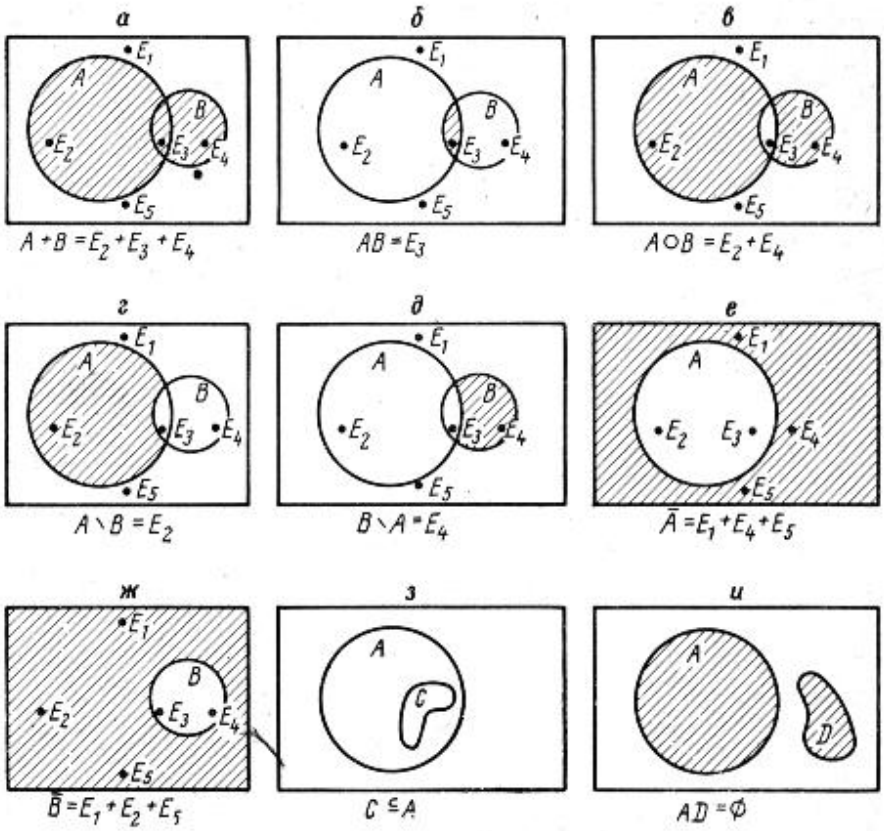


Рис. 2. Графическая интерпретация соотношений между событиями

тие $A \cup B$; *г*) разность событий A и B (или элементы события A, не принадлежащие одновременно событию B), т. е. событие $A \setminus B$; *д*) разность событий B и A (или элементы события B, не принадлежащие одновременно событию A), т. е. событие $B \setminus A$; *е*) отрицание события A, т. е. событие $\overline{A}$; *ж*) отрицание события B, т. е. событие $\overline{B}$; *з*) перекрытие события C событием A, т. е. $C \subseteq A$; *и*) несовместимость или непересечение событий A и D, т. е. $AD = \varnothing$ (знак $\varnothing$ показывает пустое множество, обозначаемое $\wedge$ и $\cup$).

Приведем примеры различных событий. Допустим, что испытанием является опробование добытого оптического гипса. Каждый кристалл, обломок кристалла или сросток кристаллов, словом, образец гипса, завернут в бумагу. Все образцы гипса уложены в ящики. При сдаче этого сырья оптическому заводу необходимо определить процент выхода бездефектных блоков. Для этого требуется взвесить образец и замерить (без резания) объем оптически годной части образца. Изучение каждого образца без исключения отняло бы слишком много времени. Для этой цели можно отобрать наудачу некоторое число образцов и по результатам их исследования судить о всей совокупности. Единичным испытанием здесь будет отбор одного образца и его исследование. Достоверное событие будет заключаться в том, что образец окажется гипсом и ничем иным. Невозможное событие — появление не гипса, а какого-либо другого минерала. Случайное событие — годность кристалла или, наоборот, непригодность его в качестве оптического сырья.

Если в приведенном примере объем бездефектной области кристалла разделить на объем всего образца, а частное от этого деления выразить в процентах, то получим содержание бездефектной области. Величина этого содержания будет меняться от образца к образцу, причем заранее нельзя предсказать, какое значение это содержание примет в результате одного испытания.

Переменная величина, принимающая в результате испытания то или иное заранее неизвестное значение, называется случайной величиной.

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

Дискретная величина может принимать в заданном интервале только конечное или счетное число значений. Примером такой величины служит число буровых станков. Число это всегда целое. Нет и не может быть, например, пяти с четвертью станков.

Непрерывная величина в заданном интервале может принимать бесконечное множество значений. Так, содержание глинозема в пробе, взятой на бокситовом (гидраргиллитовом) месторождении, может служить примером непрерывной величины (в интервале от 0 до 65%).

Выше говорилось, что случайным называется такое событие, которое при данном испытании имеет возможность произойти или не произойти. Одни из случайных событий в данном эксперименте имеют большую возможность произойти, а другие — малую. В качестве меры возможности появления случайного события используется величина, называемая вероятностью.

Вероятность события A — это число, которое характеризует возможность появления этого события. Оно обозначается $P(A)$, где в скобках стоит обозначение случайного события A.

Иногда вероятность обозначают просто строчной буквой p, т. е.

$$p = P(A).$$

Существует несколько определений вероятности, из которых для практики наибольший интерес представляют два — классическое и статистическое.

Классическое определение вероятности основано на понятии равновозможности исходов, а статистическое определение связано с частотой появления данного события при очень большом повторении испытаний.

Понятие о равновозможности лучше всего показать на примере.

Пример. Пусть в рюкзаке геолога лежит пять одинаковых мешочков с образцами — 2 гранита и 3 сиенита. Если наугад вынуть один мешочек, то каждый из пяти мешочков имеет одинаковые шансы быть вынутым — это и есть равновозможность. Общее число возможных исходов испытания, т. е. извлечения образцов, в данном случае равно пяти (n). Число возможных исходов, способствующих появлению события (A) (извлечение гранита) равно 2 (m). Отношение числа благоприятных исходов к общему числу исходов, как представляет собой вероятность события A, в ее классическом выражении, т. е.

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

В нашем случае имеем

$$P(A) = \frac{2}{5}.$$

Это и будет искомая вероятность извлечения гранита.

В нашем примере вынут один образец, но можно вынимать и два и большее число образцов. В общем виде это можно обозначить так: число всех образцов в рюкзаке N, среди них образцов гранита M, число вынутых образцов n. Требуется вычислить вероятность события (A), заключающегося в том, что среди вынутых образцов будет m образцов гранита. Общее число возможных исходов будет равно C_N^n, число благоприятных случаев — $C_M^n C_{N-M}^{n-m}$, а искомая вероятность составит

$$P(A) = \frac{C_M^n C_{N-M}^{n-m}}{C_N^n},$$

где C_M^n — число сочетаний из M по n.

Свяжем теперь классическое определение вероятности с тем, что выше говорилось о поле событий.

Б. В. Гнеденко определяет эту связь следующим образом:

«Рассмотрим какую-либо группу G, состоящую из n попарно несовместимых равновозможных событий (назовем их элементарными событиями) $E_1, E_2, \ldots, E_n$ и рассмотрим систему S, состоящую из невозможного события V, всех событий E_k группы G, и всех событий A, которые могут быть подразделены на частные случаи, входящие в состав группы G.

Например, если группа G состоит из трех событий E_1, E_2 и E_3, то в систему S входят события V, E_1, E_2, E_3, $E_1 + E_2$, $E_2 + E_3$, $E_1 + E_3$, $U = E_1 + E_2 + E_3$.

Легко установить, что система S и есть поле событий. В самом деле, очевидно, что сумма, разность и произведение событий из S входят в S; невозможное событие V входит в S по определению, а достоверное событие U входит в S, так как оно представляется в виде

$$U = E_1 + E_2 + \ldots\ldots + E_n.$$

В соответствии с приведенным определением каждому событию A, принадлежащему к построенному сейчас полю событий S, приписывается вполне определенная вероятность

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

где m есть число тех событий E_l исходной группы G, которые являются частными случаями события A. Таким образом, вероятность $P(A)$ можно

рассматривать как функцию от события A, определенную на множестве событий S.»

Эта функция имеет семь следующих свойств:

1. Для каждого события A поля S

$$P(A) \geqslant 0,$$

т. е. вероятность не может быть числом отрицательным, так как отношение $\frac{m}{n}$ не может быть отрицательным.

2. Для каждого события A поля S

$$P(A) \leqslant 1,$$

т. е. вероятность не может быть больше единицы, так как m не может быть больше n.

3. Для невозможного события V

$$P(V) = 0.$$

4. Для достоверного события U

$$P(U) = 1,$$

так как такому событию благоприятствуют все n исходов и в этом случае $m = n$.

5. Если событие A подразделяется на несовместимые события B и C и все три события A, B и C принадлежат полю S, то

$$P(A) = P(B) + P(C).$$

Эта формула выражает собою теорему сложения вероятностей. Доказывается эта теорема следующим образом.

Пусть событию B благоприятствует m_1, а событию C — m_2 событий E_i, т. е. элементарных событий из группы G. По условию теоремы события B и C несовместимы, поэтому среди элементарных событий, благоприятствующих событию B, нет ни одного элементарного события, благоприятствующего событию C, а среди элементарных событий, способствующих событию C, нет ни одного элементарного события, благоприятствующего событию B. Общее число элементарных событий, благоприятствующих событиям B и C (событию $A = B + C$), поэтому равно $m_1 + m_2$, т. е.

$$P(A) = \frac{m_1 + m_2}{n} = \frac{m_1}{n} + \frac{m_2}{n} = P(B) + P(C),$$

что и требовалось доказать.

Из сформулированной выше теоремы сложения вытекает следующее важное следствие: если несовместимые события A_1, A_2, . . ., A_n единственно возможны, то сумма их вероятностей равна единице, т. е.

$$P(A_1) + P(A_2) + \ldots + P(A_n) = 1.$$

6. Для события $\bar{A}$, противоположного событию A, вероятность

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

Доказывается это равенство так:

а) сумма событий $A + \bar{A} = U$, но так как выше было доказано, что $P(U) = 1$, то имеем

$$P(A + \bar{A}) = 1;$$

б) поскольку события A и $\bar{A}$ несовместимы, то по теореме сложения событий получим

$$P(A + \bar{A}) = P(A) + P(\bar{A}),$$

или

$$1 = P(A) + P(\bar{A}).$$

Перенеся в левую часть равенства $P(A)$, а в правую 1, получим доказываемое выражение.

7. Если событие A влечет за собой событие B, то

$$P(A) \leqslant P(B).$$

Доказывается это выражение следующим образом.

Событие B может быть представлено как сумма событий A и $\bar{A}B$. Используя ранее доказанные положения, получим

$$P(B) = P\left(A + \bar{A}B\right) = P(A) + P\left(\bar{A}B\right) \geqslant P(A).$$

Перейдем теперь к рассмотрению статистического понятия вероятности. Последнее, как уже говорилось выше, связано с частотой события.

В природе и обществе есть много явлений, повторяющихся с удивительным постоянством, т. е. с почти неизменной частотой. Но что такое частота события? Ответим на этот вопрос примером.

Пример. Пусть на какой-либо территории в течение последних 10 лет произошло 12 сильных и 38 слабых землетрясений. Число 12 здесь является частотой, а отношение $\frac{12}{12+38} = 0{,}24$ — частостью (относительной частотой) сильных землетрясений. Частость — это доля данных событий в общей массе событий. Если частости сильных землетрясений для определенной территории, вычисленные по десятилетиям, будут, например, равны 0,24; 0,23; 0,25; 0,24; 0,23 и т. д., то мы имеем дело с устойчивым явлением.

Если в результате какого-либо испытания, повторяемого много раз в одинаковых условиях, может произойти или не произойти событие A, то мы можем определить относительную частоту появления этого события в серии испытаний. Если далее серия таких же испытаний будет повторена достаточно большое число раз, а относительная частота появления события A в серии будет приблизительно одной и той же, то среднее значение этой величины можно считать статистической вероятностью события A. Таким образом, статистическую вероятность можно рассматривать как предел, к которому стремится частное от деления числа появлений события в серии n испытаний на n (при n, стремящемся к бесконечности).

В случае статистического определения также имеют место следующие свойства вероятности.

1. Статистическая вероятность не должна быть меньше нуля или больше единицы.

2. Если сложное событие C является суммой конечного числа несовместимых событий $A_1, A_2, \ldots, A_n$, причем вероятность каждого из них имеется, то вероятность события C существует и может быть определена по формуле

$$P(C) = P(A_1) + P(A_2) + \ldots + P(A_n).$$

Вероятности бывают безусловными и условными. Первые относятся к таким событиям, которые происходят при реализации только одного, строго определенного комплекса условий. Вторые же связаны с событиями, которые вызываются реализацией такого же комплекса условий и какого-либо дополнительного условия.

Пользуясь ранее введенной символикой, обозначим безусловную вероятность события A как $P(A)$. Если вероятность того же события A должна быть определена при дополнительном условии, заключающемся в том, что уже произошло событие B, то это будет условная вероятность и обозначается она $P_B(A)$ или $P(A/B)$. Читается такая запись так: «Вероятность события A при условии, что произошло событие B». Приведем примеры вычисления условной вероятности.

Пример 1. Имеется 10 шлифов гранита, сиенита, диорита, известняка и песчаника — каждой породы по 2 шлифа. Шлифы изверженных пород имеют красную метку. Наудачу взято 2 шлифа, после чего было замечено, что на каждом из них есть красная метка. Требуется определить вероятность того, что один из этих шлифов сделан из гранита, а другой из сиенита.

Сложное событие, состоящее в том, что один шлиф гранита, а второй сиенита, обозначим A. Наличие же красной метки, т. е. тот факт, что шлифы сделаны именно из изверженной породы, назовем событием B.

Возможные варианты взятия шлифов (всего 25) приведены в табл. 2.

Таблица 2

Гранит Гранит	Сиенит Гранит	Диорит Гранит	Известянк Гранит	Песчаник Гранит
Гранит Сиенит	Сиенит Сиенит	Диорит Сиенит	Известняк Сиенит	Песчаник Сиенит
Гранит Диорит	Сиенит Диорит	Диорит Диорит	Известняк Диорит	Песчаник Диорит
Гранит Известняк	Сиенит Известняк	Диорит Известняк	Известняк Известняк	Песчаник Известняк
Гранит Песчаник	Сиенит Песчаник	Диорит Песчаник	Известняк Песчаник	Песчаник Песчаник

Из данных таблицы видно, что число случаев, благоприятствующих событию A, равно 2, а число случаев, благоприятных для события B, равно 9. Безусловная вероятность события A

$$P(A) = \frac{2}{25}.$$

Поскольку произошло событие B, то осуществилась одна из 9 возможностей (оба шлифа по изверженной породе). Поэтому условная вероятность равна

$$P_B(A) = \frac{2}{9}.$$

Пример 2. Необходимо двумя станками пробурить гидрогеологические скважины. На геологической карте этого района имеется всего 22 приблизительно равных по площади участка, в том числе 4 несмежных участка, показанные как площади распространения юрских отложений. Точки для бурения скважин выбирают наугад поочередно, но так, чтобы на один и тот же участок не попали обе скважины. Нужно определить вероятность того, что точка для второй скважины попадет на участок распространения юрских отложений, если известно, что первая скважина уже задана на породы этого возраста.

Если первый участок — верхняя юра, то невыбранных осталось 3 участка распространения юрских отложений среди 21 участка. Отсюда условная вероятность равна

$$P_B(A) = \frac{3}{21} = \frac{1}{7}.$$

Если через $P(AB)$ обозначить вероятность совместного наступления двух событий A и B, то условные вероятности $P_A(A)$ и $P_A(B)$ можно записать так:

$$P_B(A) = \frac{P(AB)}{P(B)}$$

и

$$P_A(B) = \frac{P(AB)}{P(A)}.$$

Эти формулы теряют смысл в том случае, если в знаменателе каждой из них будет нуль (если B в первом случае и A во втором являются невозможными событиями).

Из этих формул (даже в случаях, когда A или B — невозможные события) выводится теорема умножения вероятностей, выражаемая следующей формулой:

$$P(AB) = P(A)\,P_A(B) = P(B)\,P_B(A),$$

т. е. вероятность произведения двух событий равна произведению вероятности одного из этих событий на условную вероятность другого при условии, что первое уже произошло.

Используя понятия об условных вероятностях, можно привести следующую формулу, выражающую условие независимости события A от события B:

$$P_B(A) = P(A).$$

Это значит, что вероятность события A не зависит от наступления события B.

Также справедливо выражение

$$P_B(B) = P(B).$$

В случае независимости событий A и B теорема умножения вероятностей выражается проще, а именно:

$$P(AB) = P(A) \cdot P(B).$$

Формулируется она так: «Вероятность совместного появления независимых событий A и B равна вероятности появления одного из них, умноженной на вероятность появления другого».

С помощью этой теоремы можно более точно определить вероятность наступления хотя бы одного из двух совместимых событий. Для этого можно привести такую формулу:

$$P(A + B) = P(A) + P(B) - P(AB),$$

знак минус здесь означает обычное вычитание.

Предположим, что $A_1, A_2, \ldots, A_n$ — последовательность событий, $P(A_i)$ — соответствующие им вероятности, B — событие, вероятность появления которого зависит от A_i. Обозначим условную вероятность появления события B при условии, что A_i наступило через $P_{A_i}(B)$. Тогда безусловную вероятность $P(B)$ можно вычислить по формуле

$$P(B) = \sum_{i=1}^{n} P(A_i)\,P_{A_i}(B).$$

Это выражение принято называть формулой полной вероятности.

Пример. Имеется 6 ящиков с рудным керном. Скв. № 1 представлена двумя ящиками, в каждом из которых пять отделений; скв. № 2 — одним ящиком с шестью отделениями и скв. № 3 — тремя ящиками с четырьмя отделениями каждый. В ящиках помещен керн со сплошным и вкрапленным оруденением. Причем в каждом ящике со скв. № 1 в двух отделениях лежат вкрапленные руды, а в трех — сплошные; в ящиках со скв. № 2 весь керн с вкрапленными рудами; в ящиках со скв. № 3 по три отделения заняты сплошными рудами и по одному — с вкрапленными. Для химического анализа на элементы-примеси отобран наугад керн из одного отделения одного ящика. Требуется определить вероятность того, что проба будет представлена сплошными рудами.

Событие, состоящее в том, что взятый керн окажется из скв. № 1, обозначим A_1, из скв. № 2 — A_2 и скв. № 3 — A_3. Событие же, состоящее в том, что взятый керн окажется по сплошной руде, выразим через B.

Описанные события можно представить в виде

$$B = A_1B + A_2B + A_3B;$$

$$P(B) = P(A_1)\,P_A(B) + P(A_2)\,P_{A_2}(B) + P(A_3)\,P_{A_3}(B).$$

Из условия задачи получаем

$$P(A_1) = \frac{2}{6}; \quad P(A_2) = \frac{1}{6}; \quad P(A_3) = \frac{3}{6}; \quad P_{A_1}(B) = \frac{2}{5};$$

$$P_{A_2}(B) = 0; \quad P_{A_3}(B) = \frac{3}{4}.$$

Подставляя эти данные в формулу полной вероятности, получим

$$P(B) = \frac{2}{6}\cdot\frac{2}{5} + \frac{1}{6}\cdot 0 + \frac{3}{6}\cdot\frac{3}{4} = \frac{61}{120}.$$

По теореме умножения имеем

$$P(A_iB) = P(B)\,P_B(A_i) = P(A_i)\,P_{A_i}(B).$$

Отсюда находим

$$P_B(A_i) = \frac{P(A_i)\,P_{A_i}(B)}{P(B)}.$$

Подставив значение $P(B)$ по формуле полной вероятности, получим

$$P_B(A_i) = \frac{P(A_i)\,P_{A_i}(B)}{\sum_{j=1}^{n} P(A_j)\,P_{A_j}(B)}.$$

Полученное выражение носит название формулы Бейеса. Она называется также формулой вероятности гипотез. Формула Бейеса часто используется артиллеристами для пристрелки орудий.

Допустим, что событие B произошло под влиянием некоторых условий, относительно действия которых выдвигается n гипотез (A_1, A_2, ..., A_n). Вероятности этих гипотез известны до испытания. Пусть гипотеза A_i связана с тем, что событие B имеет вероятность $P_{A_i}(B)$. Пусть даже событие B наступило. Вероятность гипотезы A_i теперь нужно переоценить. Для этого и предлагается формула Бейеса.

Сумма вероятностей гипотез как до, так и после испытания должна равняться единице.

Приведем пример использования формулы Бейеса.

Пример. На 6 полках находятся в одинаковых капсулях 85 измельченных проб вольфрамовой руды, в том числе на первой полке 12 проб, на второй 12, на третьей 12, на четвертой 14, на пятой 14 и на шестой 21. При этом на первых трех полках лежат пробы, проанализированные в лабораториях Москвы, на следующих двух — в лаборатории Одессы и на последней — в лаборатории Иркутска.

До того как попасть на полки, все пробы были взяты с разных участков и обработаны одним способом. В числе участков есть один (западный), особо интересующий нас. Пробы с этого участка брались постепенно и поэтому они попали на разные полки. Всего таких проб 50, в том числе на первой полке 10, на второй 10, на третьей 10, на четвертой 7, на пятой 7

и на шестой 6. С какой-то полки произвольно взята одна проба, оказавшаяся с западного участка. Спрашивается, какова вероятность, что:

1) на полке, с которой взята проба, было 10 проб с западного участка, т. е. что это была полка первая, вторая или третья?

2) на полке, с которой взята проба, было 7 проб с западного участка, т. е., что это была четвертая или пятая полка?

3) на полке, с которой взята проба, было 6 проб с западного участка, т. е. что это была шестая полка?

Полки нас интересуют по той причине, что разные лаборатории дали разные аналитические результаты и нам нужно сопоставить новое определение со старым.

Чтобы ответить на эти вопросы сначала составим следующую вспомогательную таблицу (табл. 3):

Таблица 3

i	Гипотеза A_i	Вероятность гипотезы до испытания $P(A_i)$	Вероятность того, что проба анализировалась в лаборатории Москвы, $P_{A_i}(B)$	$P(A_i)\,P_{A_i}(B)$	Вероятность гипотезы после испытаний, т. е. после взятия пробы, анализировавшейся в лаборатории Москвы, $P(A_i)$
1	Проба, взята с первой, второй или третьей полки	$\frac{3}{6}$	$\frac{10}{12}$	$\frac{3}{6}\cdot\frac{10}{12}=\frac{30}{72}$	$\frac{30}{72}:0{,}674=$ $=0{,}618$
2	Проба взята с четвертой или пятой полки	$\frac{2}{6}$	$\frac{7}{14}$	$\frac{2}{6}\cdot\frac{7}{14}=\frac{14}{84}$	$\frac{14}{84}:0{,}674=$ $=0{,}247$
3	Проба взята с шестой полки	$\frac{1}{6}$	$\frac{6}{21}$	$\frac{1}{6}\cdot\frac{6}{21}=\frac{6}{66}$	$\frac{6}{66}:0{,}674=$ $=0{,}135$
	Всего	$\sum P(A_i)=$ $=1$	—	$\sum P(A_i)\cdot P_{A_i}\times$ $\times(B)=0{,}674$	$\sum P(A_i)=1$

В этой таблице вероятность гипотезы $P(A_i)$ до испытания определена по числу полок. Вероятность того, что проба взята с одной из первых трех полок, т. е. из числа проб, проанализированных в Москве, равна 0,618, хотя с первого взгляда могло бы показаться, что эта вероятность равна 0,5 (отношение числа полок 3 : 6). Вероятность того, что проба взята из числа проб, проанализированных в Одессе (т. е. с четвертой или пятой полки) оказалась равной 0,247, хотя отношения полок $\frac{2}{6}=$ $=0{,}33$. Вероятность последней гипотезы, а именно того, что проба взята с шестой полки (из проб, проанализированных в Иркутске), 0,135, хотя отношение числа полок $\frac{1}{6}=0{,}167$.

Если испытание повторить g раз, то вызываемое им событие A может произойти μ раз. Очевидно, что $\mu \leqslant g$. Вероятность того, что μ будет

иметь максимальное значение ($\mu = g$), может быть определена по формуле

$$P[(\mu = g) = [P(A)]^g.$$

Вероятность появления события в одних случаях может оставаться неизменной, сколько бы испытаний ни производилось (она, следовательно, не зависит от порядкового номера испытания), а в других случаях меняется от испытания к испытанию (зависит от номера испытания). Про испытания первого рода говорят, что они проводятся по «схеме возвращенного шара» (вынутый из урны шар после записи результата снова возвращается в урну), а про испытания второго рода — по «схеме невозвращенного шара» (вынутый шар после записи результата не возвращается в урну).

Пример. Имеется 6298 кристаллов кварца. Необходимо узнать среднюю длину кристаллов, хотя мы не имеем времени измерить все кристаллы. Тогда мы решаем измерить один процент от их числа, т. е. 63 кристалла, взятых наугад. Каждый из 6298 кристаллов завернут в бумагу. На бумаге снаружи стоит номер кристалла и записано место его взятия. Заменим кристаллы одинаковыми бирками с номерами кристаллов, положим их в ящик, перемешаем, вынем наугад одну бирку и положим ее на стол. Затем вынем вторую, третью бирки и т. д., до тех пор пока у нас не наберется 63 бирки. После извлечения очередной бирки ящик необходимо встряхивать и бирки перемешивать. В конце испытания в ящике останется 6298 — 63 = 6235 бирок. Затем находятся кристаллы с номерами вынутых бирок, которые мы и будем измерять.

Этот метод отбора или испытания называется «схемой невозвращенного шара», или безповторным отбором.

Другой метод отбора («схема возвращенного шара», или повторный отбор) отличается от первого тем, что номер каждой бирки после извлечения ее из ящика записывается, а сама бирка сразу же кладется обратно в ящик. При этом методе в выборку 63 номеров могут попасть одни и те же номера по два, три и больше раз.

Если в процессе испытания вероятность p изменяется («схема невозвращенного шара», или бесповторный отбор), то вероятность $P_{m,n}$ того, что событие произойдет m раз при n испытаниях, можно вычислить по следующей приближенной формуле:

$$P_{m,n} = \frac{1}{\sqrt{2\pi npq\left(1-\frac{n}{N}\right)}} e^{-\frac{x^2}{2npq\left(1-\frac{n}{N}\right)}},$$

где N — объем совокупности, из которой отбирается n ее членов, а $x = m - np$. В том случае, когда $\frac{n}{N}$ близко к нулю, это выражение будет близко к

$$P_{m,n} = \frac{1}{\sqrt{2\pi npq}} e^{-\frac{(m-np)^2}{2npq}}.$$

Когда шар после извлечения возвращается в урну, вычисление вероятностей ведется по так называемой схеме Бернулли. Сущность схемы Бернулли можно показать на двух следующих примерах.

Пример 1. Анализами очень большого числа проб сильвинитовой руды Соликамска было определено, что в среднем 45% общего числа проб содержат нерастворимый остаток в количестве менее 1,5% и в 55% проб нерастворимого остатка 1,5% и более. Какова вероятность того, что из трех наудачу взятых проб по той же руде в двух пробах содержание остатка будет менее 1,5%?

Событие, заключающееся в том, что выбрана проба с содержанием нерастворимого остатка менее 1,5%, обозначим как A, а выбор пробы,

содержащей нерастворимого остатка 1,5% или выше, выразим как B. Событие, вероятность которого требуется вычислить, состоит в том, что из трех проб две окажутся с содержанием нерастворимого остатка менее 1,5%. Обозначим его C.

Сложное событие (две первые пробы будут содержать нерастворимого остатка менее 1,5%) обозначим как AAB. Возможны также такие исходы как BBA, ABA и т. п. (их обозначения понятны без пояснений).

Событие C может наступить по одной из трех схем — AAB, ABA и BAA.

Эти три события попарно несовместимы. Используя теорему сложения вероятностей, получим

$$P(C) = P(AAB) + P(ABA) + P(BAA).$$

Перейдем к конкретным числам. Из условий задачи имеем

$$P(A) = 0{,}45,$$

$$P(B) = 0{,}55.$$

Вероятность каждого из событий (AAB, ABA и BAA), согласно теоремы умножения вероятностей, будет: $P(AAB) = P(ABA) = P(BAA) = 0{,}45 \cdot 0{,}45 \cdot 0{,}55 = 0{,}11$, поэтому

$$P(C) = 0{,}11 + 0{,}11 + 0{,}11 = 0{,}33.$$

Пример 2. На основании исследования многих тысяч кристаллов кварца из 8 месторождений, расположенных в разных странах, оценена вероятность встречи правых кристаллов $p = 0{,}49$ и левых $1 - p = 0{,}51$. Пусть из этой совокупности наудачу взято 10 кристаллов кварца. Требуется найти вероятность P того, что среди этих 10 кристаллов окажется не более трех правых.

Что означает — не более трех? Это значит в одном случае не будет встречено ни одного правого кристалла, в другом будет встречен 1 правый кристалл, в третьем 2 и в четвертом 3. Всего, таким образом, имеется четыре возможных варианта, удовлетворяющих поставленному условию. Если вероятность первого варианта обозначить P_0, второго — P_1, третьего — P_2 и четвертого — P_3, то искомая вероятность P_0 может быть найдена суммированием четырех вероятностей:

$$P = P_0 + P_1 + P_2 + P_3.$$

Определим значение каждой из четырех вероятностей.

Вероятность того, что среди 10 кристаллов не будет ни одного правого, равна вероятности того, что из 10 кристаллов все 10 будут левыми.

По ранее приведенной формуле

$$P_0 = 0{,}51^{10} = 0{,}0012.$$

Найдем теперь величину P_1, т. е. вероятность того, что из 10 кристаллов один окажется правым. Обозначим через A событие, заключающееся в том, что вынутый наудачу кристалл окажется правым, а через B — противоположное событие, заключающееся в том, что этот кристалл будет левым. Если наудачу извлекаются десять кристаллов, то событию, заключающемуся в том, что из десяти кристаллов один окажется правым, благоприятствуют 10 возможных исходов:

$$ABBB\ldots B,$$

$$BABB\ldots B,$$

$$BBAB\ldots B,$$

$$\ldots\ldots\ldots$$

$$BBBB\ldots A.$$

Все эти последовательности равновероятны, и вероятность появления любой из них можно представить как

$$P(A) \cdot P(B \ldots B) = p \cdot (1-p)^9 = 0{,}49 \cdot 0{,}51^9 = 0{,}00114.$$

Вероятность появления хотя бы одной из этих последовательностей

$$P_1 = P(m = 1) = 10 \cdot 0{,}49 \cdot 0{,}51^9 = 0{,}0114.$$

Величину P_2, т. е. вероятность того, что из 10 кристаллов два окажутся правыми, определим по формуле

$$P_2 = P(m = 2) = C_{10}^2 \cdot p^2 (1-p)^8$$

или

$$P_2 = C_{10}^2 \cdot 0{,}49^2 \cdot 0{,}51^{10-2},$$

где C_{10}^2 — число сочетаний из 10 по 2 (число исходов, благоприятствующих данному событию).

$$P_2 = 45 \cdot 0{,}49^2 \cdot 0{,}51^{10-2} = 0{,}0495.$$

Подобным способом для P_3 получим

$$P_3 = C_{10}^3 \cdot 0{,}49^3 \cdot 0{,}51^{10-3} = 0{,}1266.$$

Искомая же вероятность, т. е. вероятность того, что среди 10 кристаллов окажется не более трех правых, составит

$$P = P_0 + P_1 + P_2 + P_3 = 0{,}0012 + 0{,}0114 + 0{,}0495 + 0{,}1266 = 0{,}1887.$$

Формулы, по которым вычислялись величины P_0, P_1, P_2, P_3, можно представить в общем виде так:

$$P_m = C_n^m \cdot p^m \cdot (1-p)^{n-m},$$

где P_m — вероятность того, что в n независимых испытаниях интересующее нас событие произойдет ровно m раз, а противоположное событие произойдет $n - m$ раз (при условии, если вероятность p появления события в отдельном испытании одна и та же).

Необходимо заметить, что в ряде курсов математической статистики вместо C_n^m пишут $\binom{n}{m}$.

Приведенная выше формула для P_m носит имя Бернулли.

Число сочетаний C_n^m можно определить по формуле

$$C_n^m = \frac{n!}{m!\,(n-m)!}.$$

Вычисление C_n^m по этой формуле представляет собой довольно трудоемкую операцию, хотя в ряде математических справочников и имеются таблицы факториалов. Иногда пользуются более легким методом вычисления, основанным на приближенном равенстве:

$$n! \simeq n^n e^{-n} \sqrt{2\pi n}.$$

Это равенство справедливо для достаточно больших n.

Для небольших значений n величину C_n^m можно брать по треугольнику Паскаля (прилож. 2).

По исходным данным примера 2 (стр. 24) можно вычислить также вероятность того, что из $n = 10$ кристаллов окажется m правых для всех значений m от 0 до 10. Результаты вычисления приведены в табл. 4 и на рис. 3.

Из данных табл. 4 видно, что при $m = 5$ получается максимальное значение P_m, равное 0,254. Таким образом, в данном примере (см. рис. 3) наиболее вероятным является случай, когда из 10 кристаллов 5 будет правых и 5 левых. Вправо и влево от этого числа значения вероятностей располагаются почти симметрично, так как вероятность встречи правого

кристалла ($p = 0{,}49$) очень близка к 0,5. Если же эта вероятность будет сравнительно далека от середины интервала возможных значений вероятности вообще (0—1), то вероятности, отвечающие значениям m при уме-

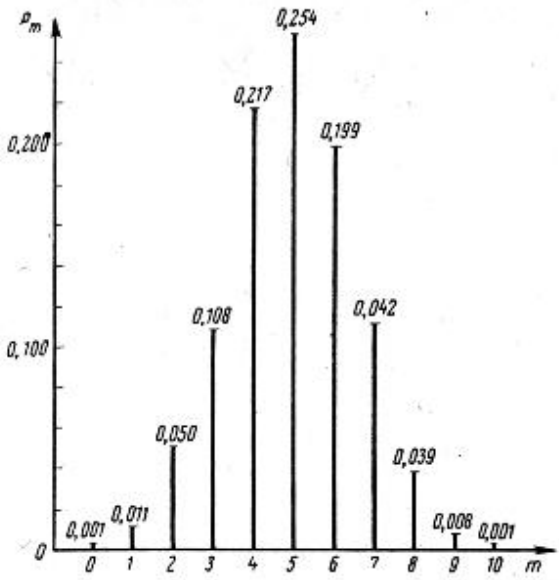


Рис. 3. Вероятности появления правых кристаллов кварца в выборе объема $n = 10$

Таблица 4

m	P_m
0	0,001
1	0,011
2	0,050
3	0,108
4	0,217
5	0,254
6	0,199
7	0,112
8	0,039
9	0,008
10	0,001
Сумма	1,000

ренных значениях n, будут располагаться асимметрично.

Пример. Три четверти территории района представляют собой солонец. Если взять 10 образцов почвы в 10 наудачу выбранных точках, то из них несколько образцов могут оказаться солонцовыми. Нужно найти наиболее вероятное число появления солонцовых образцов в выборке 10 образцов.

Таблица 5

m	P_m
0	0,000001
1	0,00003
2	0,00039
3	0,00309
4	0,01622
5	0,05866
6	0,14597
7	0,25019
8	0,28147
9	0,18767
10	0,05631
Всего	1,00000

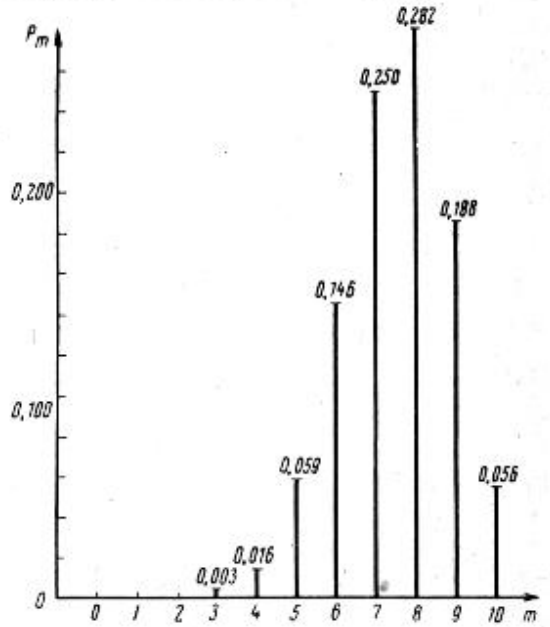


Рис. 4. Вероятности появления образцов солонца в выборе объема $n = 10$

Обозначив через $p = 0{,}75$ вероятность того, что в одной, наудачу выбранной точке, будет встречен солонец, через $n = 10$ — общее число образцов и через m — число солонцовых образцов среди них, найдем вероятности, соответствующие различным значениям m (рис. 4, табл. 5).

Наиболее вероятным здесь будет число m, равное 8. Для любого другого значения m вероятность P_m меньше, чем для $P_8 = 0{,}28147$.

II. ФУНКЦИИ РАСПРЕДЕЛЕНИЯ СЛУЧАЙНЫХ ВЕЛИЧИН

Выше (глава I) было сказано, что переменная величина, которая в результате испытания принимает то или иное заранее неизвестное значение, называется случайной величиной. Сделав n испытаний, мы получим n значений случайной величины. В результате новой серии n испытаний мы получим новый ряд значений случайной величины. Как правило, эти ряды будут отличаться друг от друга, причем различия определяются не только ошибками наблюдений, но и самой природой изучаемого явления.

Примером различия рядов значений случайной величины можно считать результаты опробования какого-либо массива гранита. Если процентное содержание кремнезема в образцах гранита считать случайной величиной, то ряд наблюдаемых значений этой величины для одной серии образцов будет одним, для другой — другим, для третьей — третьим и т. д., хотя бы все серии и брались на одном и том же массиве и в одних и тех же условиях.

В предыдущей главе были даны определения дискретной и непрерывной случайных величин.

Приведем примеры, связанные с этими понятиями.

Пример 1. В одной геологической экспедиции в течение четырех лет бурились скважины станками одного типа и приблизительно в одинаковых условиях, касающихся глубины и диаметра скважины, характера бурового инструмента, состава пород и т. д. Каждый буровой станок после окончания бурения одной скважины перебрасывался на другую, затем на третью и т. д.

В процессе бурения на многих скважинах случались аварии (обрыв штанг, прихват инструмента и др.)

В табл. 6 показано распределение числа аварий по скважинам.

Таблица 6

Число аварий, приходящихся на одну скважину	0	1	2	3	4	Всего
Количество скважин с данным числом аварий	47	39	18	5	1	110

Каждая авария характеризуется временем, затраченным на ее устранение.

Распределение времени, потерянного из-за аварий, по скважинам (табл. 7):

Таблица 7

Длительность аварий на одной скважине, ч	0—100	100—200	200—300	300—400	400—500	500—600	600—700	700—800	Всего
Количество скважин с данной продолжительностью потерянного времени	61	31	9	4	2	—	2	1	110

В первой из этих таблиц (табл. 6) случайная величина (число аварий на скважине) дискретная, во второй (длительность аварий на скважине) — непрерывная.

Пример 2. В табл. 8 приводятся содержания кремнезема в гранитах по результатам химических анализов.

Случайная величина, показанная в этой таблице (содержание кремнезема в граните), тоже непрерывная, как и в предыдущем примере продолжительность аварий на скважине.

Если в таблицах 6, 7, 8 вместо числа наблюдений случайного события вычислить значения вероятностей этого события, то получим таблицы распределения случайной величины.

Таблица 8

Содержание кремнезема, %	71,0—71,25	71,25—71,50	71,50—71,75	71,75—72,00	72,00—72,25	72,25—72,50	72,50—72,75	Всего
Число образцов	1	—	4	1	7	15	6	34

По Б. В. Гнеденко и А. Я. Хинчину, построить такую таблицу, т. е. определить все возможные значения случайной величины вместе с их вероятностями, означает задать закон распределения этой случайной величины. Для более полной характеристики случайной величины необходимо, во-первых, привести все фактически встреченные, объединенные в группы (или одиночные) ее значения и, во-вторых, указать вероятности появления этих значений.

Обозначим случайную величину через ξ, а любое заданное значение этой величины через x. Рассмотрим событие $\xi < x$, заключающееся в том, что случайная величина ξ примет в результате единичного эксперимента значение, меньшее или равное x. Вероятность P события $\xi < x$ называется функцией распределения случайной величины ξ и обычно обозначается $F(x)$, т. е. $P(\xi < x) = F(x)$. Естественно, что эта функция может быть определена для всех значений x.

Рассмотрим событие $x_1 < \xi < x_2$, которое заключается в том, что случайная величина ξ примет значение в интервале от x_1 до x_2. Вероятность события $P(x_1 < \xi < x_2)$ можно представить как разность $P(\xi < x_2) - P(\xi < x_1)$. Естественно, эта вероятность стремится к нулю при уменьшении интервала x_1, x_2 и к единице при беспредельном увеличении этого интервала.

Функция распределения случайной величины всегда однозначна. Это значит, что для данного (фиксированного) значения аргумента функция имеет одно и только одно значение. Эта функция всегда положительна (или равна нулю) и является, кроме того, неубывающей функцией, принимающей значения в интервале от нуля до единицы.

Функция распределения непрерывной случайной величины также непрерывна, и для нее может быть определена производная

$$\frac{dF(x)}{dx} = p(x).$$

Функцию $p(x)$ принято называть плотностью вероятности распределения случайной величины. Величина $p(x)\,dx = dF(x)$ — это вероятность того, что случайная величина ξ примет значение в интервале x, $x + dx$. Вероятность, что непрерывная случайная величина примет фиксированное значение x_a, равна нулю, так как

$$\int_{x_a}^{x_a} f(x)\,dx = 0.$$

В качестве примеров непрерывных распределений можно привести нормальное распределение, логарифмически нормальное распределение Коши, Стьюдента и другие.

Если производную $\frac{dF(x)}{dx} = p(x)$ интегрировать в интервале $-\infty$, x, то мы получим

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

Верхний предел этого интервала равен x. Это означает, что мы ищем вероятность того, что случайная величина будет меньше некоторой фиксированной величины x. Если верхний предел отодвинуть в бесконечность, т. е. взять интервал от $-\infty$ до $+\infty$, то величина вероятности будет равна единице, т. е.

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

Функции $p(x)$ и $F(x)$ можно изобразить графически. При этом получаются дифференциальная для $p(x)$ и интегральная для $F(x)$ кривые

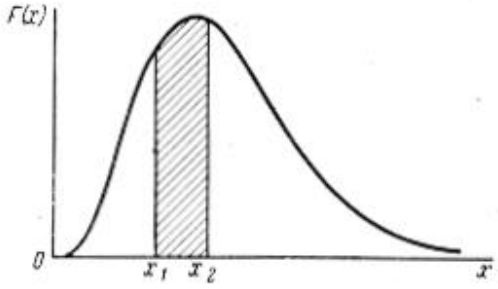


Рис. 5. Кривая плотности вероятности непрерывной случайной величины

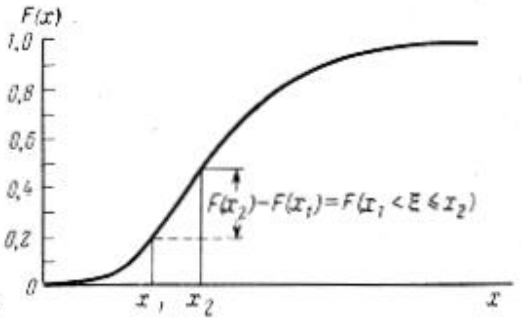


Рис. 6. Интегральная кривая распределения непрерывной случайной величины

распределения. Первая носит название кривой плотности вероятности, а вторая — кривой накопления, или кумулятивной кривой, или кривой функции распределения.

Вероятность попадания значения случайной величины в интервал x_1, x_2 на графике плотности вероятности представляет собой площадь, ограниченную справа ординатой x_2, сверху кривой $p(x)$, слева ординатой x_1, а снизу осью абсцисс (рис. 5). Та же вероятность на интегральной кривой представляет собой разность ординат, отвечающих точкам x_1 и x_2 (рис. 6).

Рассмотрим дискретные распределения. Дискретной называется такая функция распределения, которая имеет ступенчатый характер, т. е. делает положительные скачки в определенных значениях x, оставаясь неизменной в промежутке между двумя соседними значениями x. В этом случае вероятность сосредоточена в точках $\ldots, x_{-2}, x_{-1}, x_0, x_1, x_2, \ldots$.

Если для непрерывной случайной величины вероятность $P(\xi = x) = 0$, то для дискретной существуют такие значения x, в которых

$$P(\xi = x) > 0.$$

эта функция представляет собой вероятности появления отдельных значений x_i случайной величины ξ.

Накопленная же вероятность в этом случае может быть выражена

$$P(x_k) = \sum_{i=-\infty}^{k} p(x_i),$$

а сумма всех вероятностей

$$\sum_{i=-\infty}^{\infty} p(x_i) = 1.$$

Вероятности появления отдельных значений дискретной случайной величины представляют собой разности между соответствующими значениями функции распределения:

$$p(x_k) = P(x_k) - P(x_{k-1}).$$

Вероятность попадания случайной величины в определенный промежуток равна

$$P(x_a < x \leqslant x_b) = P(x_b) - P(x_a) = \sum_{i=a+1}^{i=b} p(x_i).$$

Графическое изображение вероятностей $p(x)$ появления отдельных значений дискретной случайной величины представляет собой ряд изолированных столбиков (рис. 7).

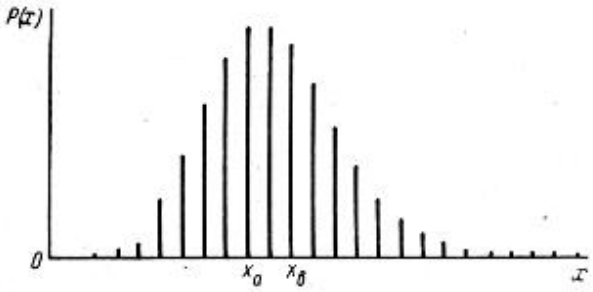


Рис. 7. Вероятности дискретного распределения

Дискретная функция распределения, если ее изобразить на графике, похожа на лесенку, ступеньки которой могут быть различными по высоте (рис. 8).

Для решения геологических задач наибольший интерес представляют дискретные распределения: биномиальное, полиномиальное, гипергеометрическое и Паскаля. В первых трех распределениях случайная величина ξ может принимать конечное число значений ($x = 0, 1, 2, \ldots, n$), а в распределении Паскаля — бесконечное число значений ($x = a, a+1, a+2, \ldots$).

Рассмотрим более подробно некоторые функции распределения, или, как их еще называют, законы распределения.

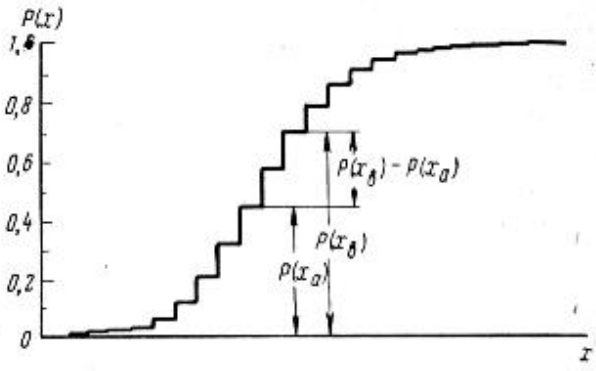


Рис. 8. Функция распределения дискретной случайной величины

В геологии может представлять интерес прежде всего биномиальный закон. Он появляется тогда, когда испытания производятся по схеме Бернулли (схема возвращенного шара). Напомним, что если производится n независимых испытаний, в процессе которых интересующее нас событие A появляется с неизменной вероятностью p и не появляется с вероятностью q (при этом $p + q = 1$), то такой порядок испытаний будет представлять схему Бернулли. Число появлений события A в n испытаниях представляет собой случайную величину ξ. Произведя большое количество серий испытаний по n, получим для каждой серии свое значение ξ.

Закон биномиального распределения был найден Я. Бернулли и описан им в «Ars Conjectandi» (1713 г.). По этому закону вероятность $P_{m,n}$ того, что событие A, появляющееся с вероятностью p при одном испытании, произойдет в серии n испытаний точно m раз, равна

$$P_{m,n} = \frac{n!}{m!\,(n-m)!}\, p^m q^{n-m}.$$

Эта формула в несколько измененном виде уже приводилась в главе I.

Коэффициенты $\frac{n!}{m!\,(n-m)!}$ при значениях $m = 0, 1, \ldots, n$ образуют ряд коэффициентов разложения бинома Ньютона, в связи с чем распределение такого вида называется биномиальным.

Эти коэффициенты можно найти по таблице для различных значений m и n, приведенной Купарадзе (1960) или по треугольнику Паскаля (приложение 2).

Для биномиального закона при любом значении n сумма всех вероятностей $P_{m,\,n}$ всегда равна единице, т. е.

$$\sum_{m=0}^{m=n} P_{m,\,n} = 1.$$

Так как числа n и m могут быть очень большими, вычисление $P_{m,\,n}$ по формуле представляет значительные трудности. Чтобы преодолеть их, можно, кроме применения способов, описанных в главе I, рекомендовать следующую приближенную формулу, предложенную Муавром:

$$P_{m,\,n} \simeq \frac{1}{\sqrt{2\pi npq}}\, e^{-\frac{1}{2}\cdot\frac{(m-np)^2}{npq}}.$$

Лапласом эта формула обобщена для любых значений p, отличных от 0 и 1. В результате получена теорема Лапласа, или, как ее называет Б. В. Гнеденко, теорема Муавра-Лапласа, формулируемая (по Б. В. Гнеденко) следующим образом.

Если вероятность наступления некоторого события A в n независимых испытаниях постоянна и равна p $(0 < p < 1)$, то вероятность $P_{m,\,n}$ того, что в этих испытаниях событие A наступит ровно m раз, удовлетворяет при $n \to \infty$ соотношению

$$\frac{\sqrt{npq}\,P_{m,\,n}}{\frac{1}{\sqrt{2\pi}}\,e^{-\frac{1}{2}t^2}} \to 1$$

равномерно для всех m, для которых t находится в любом конечном интервале.

В этой формуле

$$q = 1 - p, \text{ а}$$

$$t = \frac{m - np}{\sqrt{npq}}.$$

Исходя из приведенной выше теоремы Муавра-Лапласа, можно определить величину $P_{m,\,n}$:

$$P_{m,\,n} \simeq \frac{1}{\sqrt{2\pi npq}}\, e^{-\frac{1}{2}\cdot\frac{(m-np)^2}{npq}}.$$

Для облегчения вычисления по этой формуле составлена таблица (приложение 3) для функции

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

С помощью этой функции величина $P_{m,\,n}$ определяется так:

$$P_{m,\,n} \simeq \frac{1}{\sqrt{npq}}\, Z(t).$$

Приведем примеры вычисления вероятности $P_{m,\,n}$ по этой формуле.

Пример 1. Пусть $p = q = \frac{1}{2}$, $n = 100$, $m = 55$.

$$\sqrt{npq} = \sqrt{100 \cdot \frac{1}{2} \cdot \frac{1}{2}} = 5,$$

$$t = \frac{m - np}{\sqrt{npq}} = \frac{55 - 100 \cdot \frac{1}{2}}{5} = 1.$$

Затем по таблице (приложение 3) для $t = 1$ находим

$$Z(t) = 0{,}2420.$$

В заключение вычисляем

$$P_{55,100} = \frac{1}{\sqrt{npq}} Z(t) = \frac{1}{5} \cdot 0{,}2420 = 0{,}0484.$$

Пример 2. Пусть $n = 7500$, $m = 36$, $p = 0{,}006$, $q = 0{,}994$.

$$\sqrt{npq} = \sqrt{7500 \cdot 0{,}006 \cdot 0{,}994} = 6{,}68,$$

$$t = \frac{m - np}{\sqrt{npq}} = \frac{36 - 7500 \cdot 0{,}006}{6{,}68} = -1{,}35.$$

Функция $Z(t)$ одинакова как для положительных, так и для отрицательных значений t. В нашем примере

$$Z(-1{,}35) = 0{,}1606.$$

Искомая вероятность составляет

$$P_{36,7500} \simeq \frac{0{,}1604}{6{,}68} = 0{,}0240.$$

Не следует забывать, что величина $P_{m,n}$, получаемая таким путем, несколько отличается от истинного значения вероятности, но эта ошибка при достаточно больших значениях n невелика. Ее зависимость от n видна из данных, приведенных ниже:

n	Ошибка, %
25	6,5
100	3,0
400	0,4
1156	0,1

Все ошибки положительные (вычисленная величина вероятности всегда больше истинной).

Вероятность того, что в серии из n испытаний число наступлений события A примет значение в интервале от a до b, можно определить следующим образам. Введем обозначения:

$$\alpha = \frac{a - np}{\sqrt{npq}} \text{ или } a = np + \alpha\sqrt{npq};$$

$$\beta = \frac{b - np}{\sqrt{npq}} \text{ или } b = np + \beta\sqrt{npq}.$$

При достаточно больших значениях n будет иметь место приближенное равенство

$$P(a \leqslant \xi \leqslant b) \simeq \frac{1}{\sqrt{2\pi}} \int_{\alpha}^{\beta} e^{-\frac{1}{2}t^2} dt,$$

где $t = \frac{m - np}{\sqrt{npq}}$.

Это выражение называется формулой или интегралом Лапласа.

Пример. На месторождении меди было взято и проанализировано на медь, кобальт и другие элементы 7012 проб. Результаты анализов записывались на стандартной статистической карточке.

Из общего количества проб число проб с интересующим нас (кондиционным) содержанием кобальта равно 3514, т. е. 50,1%. Таким образом, частость интересующих нас проб равна $\sim 0{,}5$.

Необходимо определить вероятность того, что из 10 наудачу вынутых карточек 0, 1, 2, . . ., 10 окажутся с содержанием кобальта бо́льшим или равном кондиционному.

Вероятность того, что в 10 вынутых наудачу пробах будет 0, 1, 2, ..., 10 интересующих нас проб, определялась по формуле биномиального закона. При $n = 10$ и $p = 0{,}5$ искомая вероятность (умноженная на 1000) приведена в табл. 9.

Таблица 9

m	0	1	2	3	4	5	6	7	8	9	10
$P(\xi = m) \times 1000$	1	10	44	117	205	246	205	117	44	10	1

Выше было показано, что в случае биномиального распределения вероятности были получены путем разложения бинома $(p + q)^n$. Но бывают и такие испытания, когда возможны не два исхода (один с вероятностью p, другой — q), а три, четыре и большее число исходов с вероятностями p_1, p_2, p_3 и т. д. Сумма всех вероятностей, конечно, равна единице.

Вероятности исходов в таких испытаниях образуют особый вид распределения, называемый полиномиальным.

Поясним смысл полиномиального распределения на следующем примере (схема примера заимствована у Хальда, 1956).

Пример. В угленосном бассейне нарезано очень много (например, не менее 1000) лав. Горный комбайн какого-либо одного типового размера может успешно применяться лишь в таких лавах, где максимальная мощность угольного пласта превышает минимальную не более чем в 1,95 раза и где абсолютные значения минимума и максимума мощности тоже ограничены (например, абсолютный минимум 0,5 *м*, а абсолютный максимум 7 *м*).

Допустим, нас интересует только какой-то один типовой размер комбайна, например для мощности пласта от 1 до 1,95 *м*. Вероятность того, что в случайно выбранной лаве максимальная мощность пласта окажется более 1,95 *м*, пусть будет $p_1 = 0{,}06$ (при этом минимальная мощность может быть какой угодно). Вероятность, что в случайно выбранной лаве минимальная мощность пласта угля окажется менее 1 *м*, примем равной $p_2 = 0{,}02$ (при этом максимальная мощность может быть любой). Как получены вероятности p_1 и p_2, в данном случае не имеет значения, но, приняв их, мы тем самым определяем вероятность того, что лава по мощности пласта окажется пригодной для комбайна. Эта вероятность равна

$$p_3 = 1{,}00 - (0{,}06 + 0{,}02) = 0{,}92.$$

Выберем далее наудачу 100 лав. Вероятность того, что из них x_1 лав будут иметь мощность пласта более 1,95 *м*, x_2 — менее 1 *м* и x_3 — от 1 до 1,9 *м* ($x_1 + x_2 + x_3 = 100$), можно определить по формуле

$$P_{100}(x_1 + x_2 + x_3) = \frac{100!}{x_1! x_2! x_3!} 0{,}06^{x_1} 0{,}02^{x_2} 0{,}92^{x_3}.$$

По вероятностям, отвечающим полиномиальному закону при $p_1 = 0{,}06$; $p_2 = 0{,}02$; $p_3 = 0{,}92$, составлена табл. 10. Для упрощения запи-

Таблица 10

x_1	x_2								Частное распределение
	0	1	2	3	4	5	6	7	
0	0	1	1	0	0	0	0	0	2
1	2	3	4	2	1	1	0	0	13
2	5	11	11	8	4	2	1	0	42
3	11	23	24	16	8	3	1	0	86
4	17	35	37	25	13	5	2	1	135
5	21	44	45	30	15	6	2	1	164
6	22	45	45	30	15	6	2	1	166
7	19	39	39	26	13	4	2	0	142
8	15	29	29	19	9	3	1	0	105
9	10	19	19	12	6	2	1	0	69
10	6	11	11	7	3	1	0	0	39
11	3	6	6	3	2	1	0	0	21
12	1	3	3	2	1	0	0	0	10
13	1	1	1	1	0	0	0	0	4
14	0	1	1	0	0	0	0	0	2
Частное распределение	133	271	276	181	90	34	12	3	1000

сей в этой таблице все вероятности умножены на 1000, и число 45, например, означает вероятность, равную 0,045.

Совершенно очевидно, что если сделать много выборок по 100 лав, то среднее число лав с мощностью пласта более 1,95 *м* будет близко к $100 \cdot 0{,}06 = 6$, а среднее число лав с минимальной мощностью меньше 1 *м* близко к $100 \cdot 0{,}02 = 2$.

Вероятность того, что случайно выбранная лава окажется непригодной для работы принятого горного комбайна, по таблице (x_1, x_2) равна 0,045. Это максимальное значение $P_{100}(x_1, x_2, x_3)$. В таблице приведены частные распределения, отвечающие биномиальному закону (для x_1 и x_2).

По теореме сложения частная вероятность, что 6 лав будет с максимальной мощностью пласта более 1,95 *м*, составит

$$\sum_{x_2=0}^{94} P_{100}\,(6,\ x_2,\ 94 - x_2) = \frac{22 + 45 + \cdots + 1}{1000} = 0{,}166.$$

По биномиальному закону частная вероятность результата x_1 равна

$$C_{100}^{x_1} \cdot 0{,}06^{x_1} \cdot 0{,}94^{100 - x_1}.$$

По тому же закону для результата x_2 частная вероятность составит

$$C_{100}^{x_2} \cdot 0{,}02^{x_2} \cdot 0{,}98^{100 - x_2}.$$

Вероятность того, что из 100 лав не более 10 лав будут иметь максимальную мощность пласта свыше 1,95 *м* и не более 5 лав — минимальную мощность менее 1 *м*, равна сумме вероятностей (см. табл. 9) для тех значений x_1 и x_2, которые удовлетворяют неравенствам

$$0 \leqslant x_1 \leqslant 10,$$
$$0 \leqslant x_2 \leqslant 5.$$

Эта вероятность получается путем сложения всех значений вероятностей в той части табл. 9, где записаны вероятности для x_1 от 0 до 10 и для значений x_2 от 0 до 5.

Вероятность того, что не более 5 лав из 100 окажутся непригодными для работы комбайна по обоим (нижним и верхним) пределам мощности пласта, можно получить путем сложения вероятностей в той части таблицы, где параметры x_1 и x_2 удовлетворяют неравенству $0 \leqslant x_1 + x_2 \leqslant 5$.

Практически это будет сумма вероятностей, взятых по треугольнику в левой верхней части таблицы:

0	1	1	0	0	0	=2
2	3	4	2	1		=12
5	11	11	8			=35
11	23	24				=58
17	35					=52
21						=21
						180

Для нашего примера эта сумма равна 0,180.

В общем виде вероятность того, что из 100 лав x лав окажутся непригодными для комбайна данного типа, равна

$$C_{100}^{x} \cdot 0{,}08^{x} \cdot 0{,}92^{100-x},$$

так как $p_1 + p_2 = 0{,}06 + 0{,}02 = 0{,}08$.

Вероятность того, что число непригодных лав будет не более 5, составит

$$\sum_{x=0}^{5} C_{100}^{x} \cdot 0{,}08^{x} \cdot 0{,}92^{100-x}.$$

Полиномиальное распределение используется геологами при исследовании тренда реальной или воображаемой поверхности, выраженной системой изолиний (изогипс, изопахит и др.). По этому вопросу рекомендуются работы Уиттена (Whitten, 1959) и Крамбейна (Krumbein, 1959).

Разберем такую задачу. Пусть событие происходит с неизменной вероятностью v, а испытания производятся последовательно (независимо одно от другого) до тех пор, пока не наступит это событие, после чего испытания прекращаются. Необходимо узнать вероятность того, что число испытаний, произведенных до первого наступления события, будет равно x.

Эта задача решается по следующей формуле, иногда называемой распределением Паскаля:

$$P(x) = \left(C_{\vartheta}^{1-\vartheta}\right)^{a} C_{x-1}^{a-1} (1-\vartheta)^{x},$$

где a — постоянная величина, причем $x \geqslant a$ ($x = a$, $x = a + 1$, $x = a + 2$, . . .).

Распределение Паскаля, очевидно, можно использовать для предсказания числа испытаний, необходимых для первого наблюдения нужного события.

Пример. Если при изучении алмазных россыпей вероятность того, что взятая проба окажется хотя бы с одним алмазом, равна 0,02 (пробы берутся случайно и независимо одна от другой), то вероятность противоположного события (отсутствие алмазов) равна 0,98.

Вероятность того, что только $x + 1$ проба встретит хотя бы один алмаз (число x здесь означает число пустых проб) может быть определена, исходя из следующих данных:

x	$1-0{,}98^{x}$
10	0,184
20	0,332
50	0,636
100	0,867
200	0,982

Подобную таблицу можно составить и для других значений исходной вероятности p (x). По ней можно решить две задачи: 1) определить минимальное число проб, при котором встреча алмаза в $x + 1$-й пробе отвечала бы приемлемой вероятности; 2) определить вероятность, при которой

встреча алмаза в $x + 1$-й пробе отвечала бы исходной вероятности события (в приведенном примере исходная вероятность равна 0,02).

Пусть A и $\overline{A}$ — два противоположных события, появляющиеся с вероятностью p и $q = 1 - p$. Если вероятность одного из них (событий) близка к нулю, то асимптотическое представление вероятности числа m наступления этого события в серии n испытаний, даваемое теоремой Муавра—Лапласа, становится малопригодным и теряет смысл при $p = 0$. В связи с этим возникает необходимость в нахождении асимптотического выражения для вероятностей появления маловероятных событий в серии из n испытаний. Такое выражение было найдено Пуассоном.

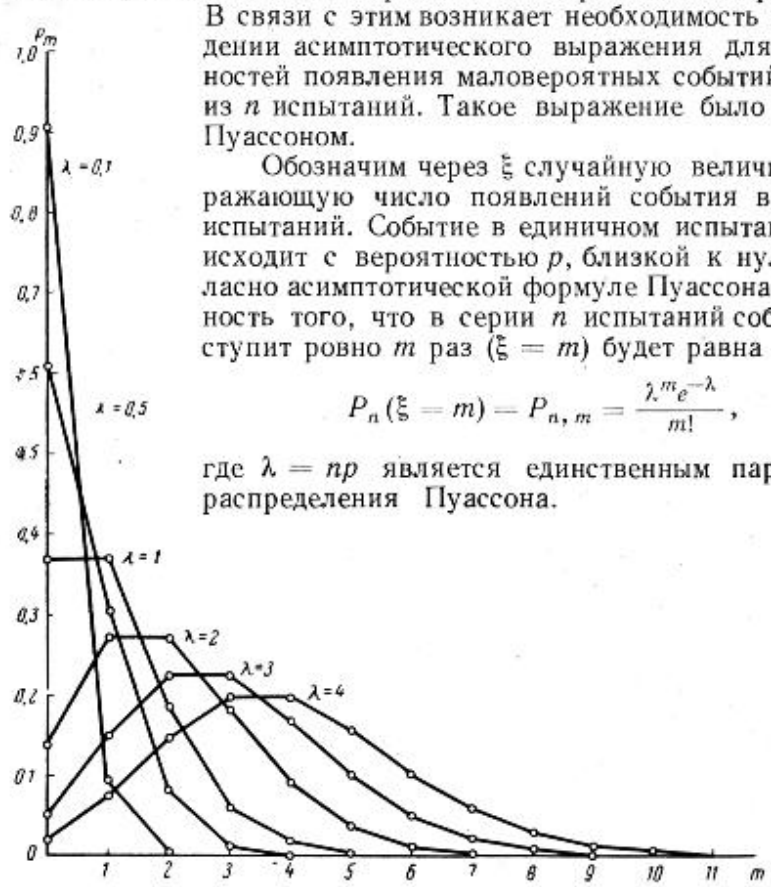


Рис. 9. Распределения Пуассона при различных значениях λ

Обозначим через ξ случайную величину, выражающую число появлений события в серии n испытаний. Событие в единичном испытании происходит с вероятностью p, близкой к нулю. Согласно асимптотической формуле Пуассона, вероятность того, что в серии n испытаний событие наступит ровно m раз ($\xi = m$) будет равна

$$P_n(\xi = m) = P_{n,m} = \frac{\lambda^m e^{-\lambda}}{m!},$$

где $\lambda = np$ является единственным параметром распределения Пуассона.

Функция распределения такой случайной величины представляет собой сумму

$$P_n(\xi \leqslant m) = \sum_{k=0}^{m} \frac{\lambda^k e^{-\lambda}}{k!},$$

где $k = 0, 1, \ldots, m$.

Распределение Пуассона приведено в таблице (приложение 11).

Графическое изображение распределения Пуассона дается на рис. 9. Рассматривая этот график, необходимо иметь в виду, что величина m, откладываемая по оси абсцисс, дискретна, что линии, соединяющие точки, служат только для удобства обозрения точек и что вероятность P_m соответствует только проставленным на рисунке точкам.

Сравнивая между собой шесть изображенных на графике кривых, легко можно заметить, что распределение Пуассона не имеет максимума при $\lambda \leqslant 1$. При $\lambda > 1$ кривая двусторонняя с вершиной. По мере роста λ асимметрия распределения уменьшается, а при $\lambda = 4$ кривая становится почти симметричной.

Иногда вместо вероятности $P_n(\xi \leqslant m)$ необходимо определить вероятность $P_n(\xi \geqslant m)$, т. е. вероятность того, что интересующее нас событие произойдет не менее m раз в серии n испытаний. Эта вероятность равна

$$P_n(\xi \geqslant m) = \sum_{k=m}^{n} \frac{\lambda^k e^{-\lambda}}{k!}.$$

Вычисленные значения вероятностей $P_n(\xi \leqslant m)$ и $P_n(\xi \geqslant m)$ сведены в таблицы (приложения 12 и 13). Более подробные значения этих функций приводятся в работах И. В. Дунина-Барковского и Н. В. Смирнова (1955) и А. К. Митропольского (1952).

Если количество последовательных независимых испытаний недостаточно велико, а единичная вероятность p недостаточно мала (например, больше 0,1), то суммарная вероятность P_m, вычисляемая по формуле Пуассона, будет содержать заметную погрешность. Чтобы уменьшить последнюю, А. Н. Колмогоров предложил следующую исправленную им формулу Пуассона:

$$P'_m \approx \frac{\lambda^m e^{-\lambda}}{m!} - \frac{b}{2}\frac{\lambda^{m-2}e^{-\lambda}}{2(m-2)!}\left(\frac{\lambda^2}{(m-1)m} - \frac{2\lambda}{m-1} + 1\right),$$

где b — величина, зависящая от изменяющихся единичных вероятностей.

Традиционная формула Пуассона рассчитана на неизменную единичную вероятность p. Формула же, предложенная А. Н. Колмогоровым, пригодна и для изменяющегося значения единичной вероятности.

Обозначим эту единичную вероятность появления события в i-том испытании через p_i. Тогда величина λ для s испытаний будет равна

$$\lambda = p_1 + p_2 + \cdots + p_i + \cdots + p_s,$$
$$b = p_1^2 + p_2^2 + \cdots + p_i^2 + \cdots + p_s^2.$$

Для случая неизменной единичной вероятности, т. е. для $p_1 = p_2 = \cdots = p_i = \cdots = p_s = p$, получим

$$\lambda = sp, \quad b = sp^2.$$

Сопоставление вероятностей, вычисленных по биномиальному закону $P_{m,n}$, по закону Пуассона P_m и по приведенной выше формуле Колмогорова P'_m дается в табл. 11. (Дунин-Барковский и Смирнов, 1955).

Таблица 11

m	$P_{m,n}$	P_m	P'_m	$100\,\frac{P_{m,n}-P_m}{P_{m,n}}$	$100\,\frac{P_{m,n}-P'_m}{P_{m,n}}$
0	0,16807	0,22313	0,17293	—32,8	—2,9
1	0,36015	0,33470	0,35980	+7,1	+0,1
2	0,38070	0,25102	0,29495	+18,7	+4,5
3	0,13230	0,12551	0,13493	—5,1	—2,0
4	0,02835	0,04707	0,03648	—60,0	—28,7
5	0,00243	0,01412	0,00388	—481,1	—59,7

Из данных таблицы видно, что погрешность определения вероятности P'_m во много раз меньше погрешности определения вероятности P_m.

Распределение Пуассона, как и биномиальное распределение, дискретное, но им можно пользоваться и при изучении распределения любой непрерывной величины, если ее трансформировать в дискретную величину. Содержание металла в руде, например, можно показывать не процентами, а баллами или номерами классов содержания.

В. М. Гудков (1959) разбил все реально встреченные значения содержания свинца в полиметаллической руде на шесть классов, а номера классов показал по нижней границе содержания в каждом классе. В ре-

зультате этой операции распределение содержаний свинца оказалось очень близким к закону распределения Пуассона. Этот исследователь также поступил с содержаниями серебра и цинка в той же руде.

В табл. 12 показана частость проб полиметаллической руды, обработанных В. М. Гудковым этим способом.

Таблица 12

x	Pb		Ag		Zn	
	$P_ф$	$P_т$	$P_ф$	$P_т$	$P_ф$	$P_т$
0	61,3	60,6	63,7	62,7	61,6	60,7
1	30,5	30,3	29,2	29,3	30,5	30,4
2	5,9	7,6	5,1	6,8	5,9	7,5
3	1,9	1,4	1,3	1,1	2,2	1,3
4	0,4	0,1	0,6	0,1	0,2	0,1
5	0,0	0,0	0,1	0,0	0,1	0,0

Примечание. x — номер класса по содержанию металла, $P_ф$ — фактическая частость, $P_т$ — теоретическая частость по закону Пуассона.

Пример. Содержание свинца (m) в руде одного из уральских медных месторождений, выраженное в виде номера класса, и фактическое число проб (n_m) с таким содержанием показано в табл. 13.

Таблица 13

m	0	1	2	3	4	5	6	7	8	9	Всего
n_m	24	75	110	113	84	50	23	10	4	1	494

Для того чтобы можно было судить о близости фактического распределения к распределению Пуассона, необходимо заметить, что в формуле Пуассона величина λ представляет собой среднее арифметическое из всех значений случайной переменной и что формальным признаком близости фактического распределения к распределению Пуассона является близость оценок для λ, σ^2 и μ_3 (среднее арифметическое из наблюдаемых значений случайной переменной x является приближенной характеристикой или оценкой для λ; σ^2 — дисперсия той же случайной переменной; μ_3 — третий центральный момент).

Характеристика величин σ^2 и μ_3 будет дана ниже, а здесь только заметим, что величина σ^2 указывает на рассеяние значений m, а величина μ_3 характеризует так называемую асимметрию распределения величины m, т. е. степень совпадения наиболее часто встречающегося значения m со средним арифметическим из всех этих значений.

Для приведенного примера с 494 пробами медной руды имеем: $\overline{m} = 2{,}97$; $\hat{\sigma}^2 = 2{,}86$; $\hat{\mu}_3 = 2{,}50$, т. е. эти три величины близки друг к другу, поэтому данное распределение близко к распределению Пуассона.

Вычисление теоретического числа проб (по закону Пуассона) производится следующим образом. Так как $\hat{\lambda} = \overline{m} = 2{,}97 \simeq 3$, находим по таблице (приложение 11) графу, где $\lambda = 3$; выписываем величину P_m для разных значений m и умножаем ее на общее число проб (494). В результате получаем теоретическое число проб n'. Данные табл. 14 иллюстрируют близость фактического числа проб (n) с теоретическим (n'_m).

Совпадение фактического числа проб с теоретическим (n_m и n'_m) довольно близкое (рис. 10). Ниже (глава V) описываются критерии согла-

Таблица 14

m	P_m	n'_m	n_m
0	04979	25	24
1	14936	74	75
2	22404	111	110
3	22404	111	113
4	16803	83	84
5	10082	50	50
6	05041	24	23
7	02161	11	10
8	00810	4	4
9	00270	1	1
Всего		494	494

сия и показано, как можно более точно измерить степень близости друг к другу двух рядов (фактического и теоретического).

Если случайная величина принимает значения только в некотором конечном интервале (a, b), а плотность вероятности для этих значений постоянна, то распределение такой случайной величины называется равномерным, или прямоугольным.

С законом равномерного распределения геологу приходится встречаться сравнительно редко; например, при изучении ошибок, возникающих при округлении чисел во время записи результатов каких-либо измерений. Допустим, мощность какого-то слоя по стенке штрека или по керну скважин измерялась с точностью до миллиметра, а потом было решено округлить эти результаты до сантиметра. При этом цифры 0, 1, 2, 3 и 4 в n-м порядке не изменяют цифры в $n-1$-м порядке, а цифры 5, 6, 7, 8 и 9 увеличивают цифру $n-1$-го порядка на единицу. Ошибки округления в этом случае будут одинаково частыми для всех цифр n-го порядка.

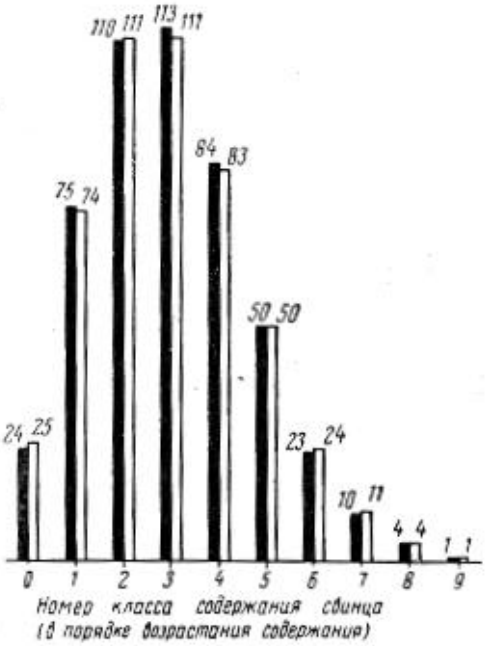


Рис. 10. Распределение содержаний свинца в медной руде одного уральского месторождения (фактическое — черные столбцы и теоретическое по закону Пуассона — белые столбцы)

Плотность равномерного распределения равна

$$P(x) = \frac{1}{a},$$

где $a = b - a$.

Случайная переменная принимает значения x, удовлетворяющие неравенству $a \leqslant x \leqslant b$.

В приведенном примере x — величина ошибки от округления, а величина $a = 1$.

Функция равномерного распределения выражается формулой

$$P(x) = \frac{x-a}{a} = \frac{x-a}{b-a},$$

причем величина x удовлетворяет приведенному выше неравенству.

Плотность вероятности равномерного распределения графически (рис. 11) представляет собой отрезок прямой линии, идущей параллельно оси x на расстоянии $\frac{1}{a}$ от нее. Функция равномерного распределения также имеет вид прямой, наклоненной к оси x с угловым коэффициентом $\frac{1}{a}$ (рис. 12).

Пусть ξ — случайная величина, а $P(x)$ — соответствующая ей плотность вероятности. Если

$$P(x) = \frac{1}{\pi} \frac{\lambda}{\lambda^2 + (x-\mu)^2},$$

где λ и μ — константы, причем $\lambda > 0$, то случайная величина ξ распределена по з а к о н у К о ш и. Кривая плотности вероятности этого распределения имеет один максимум и симметрична относительно точки $x = \mu$.

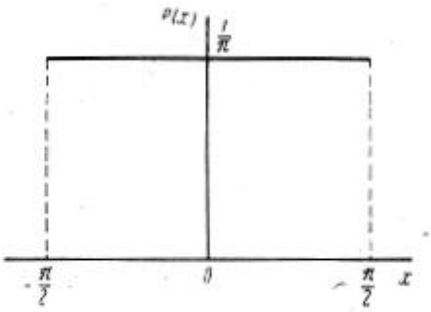


Рис. 11. Графическое выражение плотности вероятности равномерного распределения случайной величины в интервале $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$.

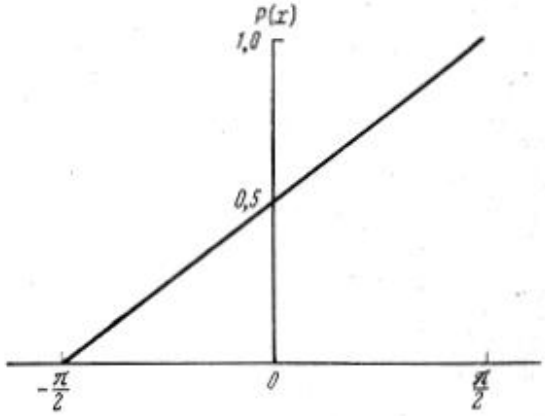


Рис. 12. Графическое изображение функции равномерного распределения случайной величины в интервале $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$

Следует особо отметить, что ни один из моментов положительного порядка (понятие о моменте см. в главе III) не существует. В частности, понятие среднего значения для этого распределения не имеет смысла.

Рассмотренные выше распределения (биномиальное, полиномиальное, распределение Пуассона) соответствуют схеме «возвращенного шара». Рассмотрим теперь распределение, соответствующее схеме «невозвращенного шара» и называемое г и п е р г е о м е т р и ч е с к и м. Сущность гипергеометрического распределения покажем на примере.

На одной золотой россыпи пробито N шурфов, M из которых показали промышленное содержание золота. Для контроля опробования необходимо вторично взять и промыть пробы песков по некоторому числу шурфов. Для этого наудачу выбирается (по схеме «невозвращенного шара») n шурфов. Требуется определить вероятность того, что из n шурфов ровно x окажется с промышленным содержанием золота при первом опробовании.

Число способов, по которым может быть отобрано n шурфов, равно числу сочетаний C_N^n, причем все эти способы равновероятны. Число же способов отбора x шурфов с промышленным содержанием, очевидно, равно C_N^x, а число способов отбора $n - x$ шурфов с непромышленным содержанием — C_{N-M}^{n-x}. При этом каждый из C_M^x способов может соче-

таться с каждым из C_{N-M}^{n-x} способов, так что общее число благоприятных случаев равно произведению $C_M^x \cdot C_{N-M}^{n-x}$. Искомая вероятность:

$$p(x) = \frac{C_M^x C_{N-M}^{n-x}}{C_M^n}.$$

Значение величины x удовлетворяет неравенствам

$$0 \leqslant x \leqslant M,$$
$$0 \leqslant n - x \leqslant N - M.$$

Вероятности $p(x)$, вычисленные описанным способом, образуют гипергеометрическое распределение, названное так ввиду связи его с гипергеометрической функцией.

Если $N \to \infty$, а отношение $M : N = \vartheta = \text{const}$, то

$$p(x) \to C_n^x \vartheta (1 - \vartheta)^{n-x}.$$

Таким образом, гипергеометрическое распределение стремится к биномиальному при беспредельном росте N.

Пример. Общее число шурфов $N = 1000$. Число шурфов с промышленным содержанием металла $M = 20$. Индивидуальная вероятность равна

$$\vartheta = \frac{20}{1000} = 0{,}02.$$

Вероятность того, что среди 100 шурфов, отобранных по схеме невозвращенного шара, x шурфов окажется с промышленным содержанием, равна

$$p(x) = \frac{C_{20}^x C_{1000-20}^{100-x}}{C_{1000}^{100}} \text{ при } 0 \leqslant x \leqslant 20.$$

В табл. 15 приведены вероятности $p(x)$, рассчитанные для этого примера.

Таблица 15

x	$p(x)$	$\sum p(x)$
0	0,1190	0,1190
1	0,2701	0,3891
2	0,2881	0,6772
3	0,1918	0,8690
4	0,0895	0,9585
5	0,0311	0,9896
6	0,0083	0,9979
7	0,0018	0,9997
8	0,0003	1,0000

Из этой таблицы видно, что наиболее вероятным будет $x = 2$ [$p(x) = 0{,}2881$].

Весьма распространенной моделью распределения непрерывных случайных величин является так называемый нормальный закон. Случайная величина ξ распределена нормально, если соответствующая ей плотность вероятности дается выражением

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

где a и σ^2 — параметры распределения. Величина a представляет собой точное среднее значение случайной величины ξ (математическое ожидание), а σ^2 является мерой рассеяния значений ξ вокруг a. Следует отметить, что параметры a и σ^2 независимы.

Графически кривая плотности вероятности нормального распределения показана на рис. 13. Она достигает максимума в точке $x = a$. По мере удаления от этой точки вправо и влево, т. е. в сторону уменьшения и в сторону увеличения значений x, $f(x)$ асимптотически приближается к оси x.

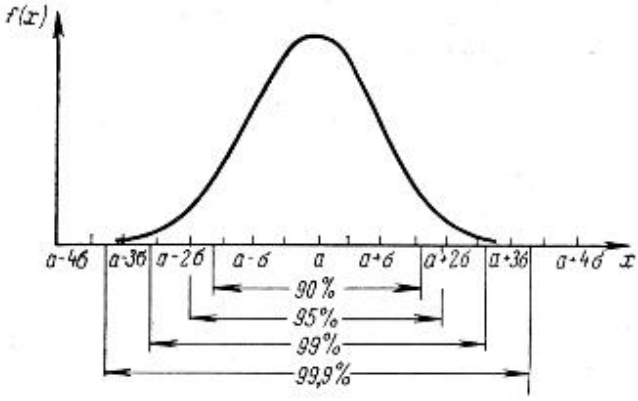


Рис. 13. Графическое выражение плотности вероятностей нормального распределения

Интегральная кривая нормального распределения вероятностей показана на рис. 14.

Если вместо случайной величины ξ рассмотреть $\tau = \frac{\xi - a}{\sigma^2}$, то новая случайная величина будет также распределена нормально со средним значением, равным нулю, и дисперсией, равной 1. Плотность вероятности $f(t)$ величины τ будет иметь вид

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

Это уравнение иногда называют именем Гаусса.

Функция $f(t)$ — это не что иное, как $Z(t)$, значения которых даны в приложении 3.

Функция нормального распределения обладает следующими свойствами:

1. Она всегда симметрична относительно ординаты в точке $x = a$, а в случае $f(t)$ — при $t = 0$.

2. При $t \to \infty$ или $-\infty$ функция $f(t)$ стремится к нулю (асимптотически приближается к оси ox). При увеличении $|t|$ хотя бы до 6 величина функции становится очень малой (при $|t| = 6$ $f(x) = 0{,}000000015$).

3. При $t = 0$ функция $f(t)$ максимальна, т. е.

$$f(0) = \frac{1}{\sqrt{2\pi}} = 0{,}3989.$$

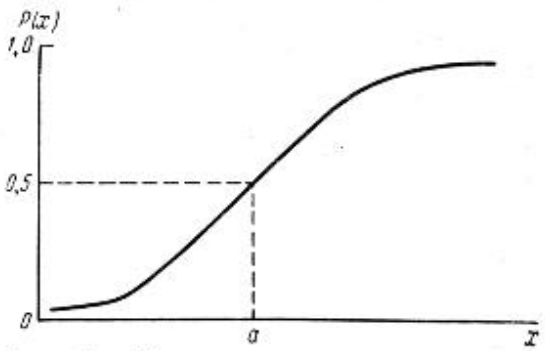


Рис. 14. Интегральная кривая нормального распределения

4. При $t = \pm 1$ кривая $f(t)$ имеет точки перегиба, или, что то же самое, нормальная функция $f(x)$ плотности вероятности с параметрами a, σ^2 имеет точки перегиба при $x = a \pm \sigma$.

5. Площадь, ограниченная кривой $f(x)$ и осью ot, на всей прямой $-\infty$, $+\infty$ равна единице.

Нормальная функция распределения $F(x)$, выражающая вероятность того, что случайная величина ξ не превзойдет фиксированного значения x, представлена выражением

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(x-a)^2}{2\sigma^2}} dx.$$

При $\tau = \frac{\xi - a}{\sigma}$

$$P(\tau \leqslant t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{z^2}{2}} dz.$$

Эта функция графически выражена на рис. 14.

Для определения площади, заключенной между нормальной кривой (на дифференциальном графике) и осью абсцисс, составлены таблицы. В приложении 4 приведены значения функции $\Phi\left(\frac{t}{\sqrt{2}}\right) = \Phi_м(z)$, формула которой имеет вид

$$\Phi_м(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt,$$

где z — нормированное отклонение x от средней a, деленное на $\sqrt{2}$. Иначе говоря, $\Phi_м(z)$ — это величина площади, ограниченной нормальной кривой, осью абсцисс и двумя ординатами, находящимися на расстоянии z вправо и влево от его среднего значения, принятого за нуль. Часто пользуются таблицами, составленными для функций:

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

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

т. е. для отрицательных (приложение 5) и положительных (приложение 6) значений z.

В приложении 5 даны значения функции от 0 до 0,5, а в приложении 6 — от 0,5 до 1.

Если необходимо иметь значения интегральной вероятности, подсчитываемые начиная не от $-\infty$, а от нуля, т. е.

$$\Phi^*(z) = \frac{1}{\sqrt{2\pi}} \int_0^z e^{-\frac{t^2}{2}} dt,$$

то можно воспользоваться приложением 7. В этой таблице интеграл вероятностей дает ряд значений от 0 до 0,5, причем взяты только положительные значения z. Для отрицательных значений величина вероятности будет такой же, так как функция симметрична.

Для определения вероятности того, что значение случайной величины ξ будет лежать внутри определенных границ от $-x$ до $+x$, дается в приложении 8, составленном для функции

$$\Phi(z) = \frac{2}{\sqrt{2\pi}} \int_0^z e^{-\frac{t^2}{2}} dt,$$

где

$$z = \frac{|x - a|}{\sigma}.$$

Нетрудно увидеть, что $\Phi(z) = 2\Phi^*(z)$.

В приложении 8 показана вероятность $\Phi(z)$, представляющая собой площадь, ограниченную нормальной кривой, осью абсцисс и ординатами в точках $-z$ и z.

В приложении 9 даны значения интеграла

$$1 - \Phi(z) = \frac{2}{\sqrt{2\pi}} \int_z^\infty e^{-\frac{t^2}{2}} dt,$$

т. е. величина площади под нормальной кривой за границами $\pm z$.

В связи с проблемой ураганных содержаний в пробах, интересующей разведчиков месторождений драгоценных минералов, редких и цветных металлов, редких элементов и некоторых других полезных ископаемых, интересно знать границы, вычисляемые от среднего значения признака,

вне которых находится только один член нормальной совокупности (одна проба, например). Эти границы даны в приложении 10. Кроме описанных существуют и другие таблицы для определения различных вероятностей в условиях нормального закона (Митропольский, 1952, т. 2).

Нормальный закон нашел очень широкое применение в естественных науках вообще и в геологии в особенности. Распределение ошибок замеров, анализов, наблюдений очень часто согласуется с нормальным законом.

Примеры согласованности эмпирических данных с нормальным законом в геологии будут приведены ниже (при описании критериев согласия).

Во всех рассмотренных выше случаях речь шла об одномерном распределении, т. е. о распределении одной какой-либо случайной величины. Однако в статистике часто используются двухмерные, трехмерные и вообще многомерные распределения, т. е. совместные распределения двух, трех и более случайных величин. Некоторые из таких распределений будут рассмотрены при описании методов корреляции.

В практике встречаются случаи, когда статистическому исследованию подвергается не вся генеральная совокупность, а лишь какая-то часть ее. При исследовании отрабатываемого угольного пласта, например, мы имеем дело только с рабочей мощностью. Нерабочая же мощность (например, меньше 0,5 или 0,6 *м*) остается за пределами шахтного поля. Исследуемая в таком случае мощность $x_i \leqslant 0{,}5$ *м* образует не полное, а усеченное распределение.

Подобный случай имеет место и тогда, когда нас интересует зольность угля. Участки угольных пластов с зольностью выше критической (более 40%, например) не включаются в шахтное поле, и зольность $A_i \leqslant 40\%$ образует тоже усеченное распределение.

Не включенные в изучаемое распределение значения случайной величины иногда остаются неизмеренными или вообще неопределенными.

Усечения могут быть односторонними, как в рассмотренных случаях, и двусторонними. Примером последних является совокупность измерений размера песчинок (при разведке месторождения стекольного песка). Участки месторождения, где преобладают слишком мелкие или слишком крупные зерна, остаются при подсчете запасов за границами промышленного контура.

Усеченное распределение применял Колмогоров (1949) при исследовании математической модели слоеобразования.

С методами статистического анализа усеченной совокупности можно познакомиться по книге А. Хальда (1956).

Эмпирические распределения, с которыми приходится сталкиваться геологу, нередко бывают асимметричны. Желание использовать удобные и полезные приемы анализа, созданные для нормального распределения, заставляет исследователей искать возможности трансформации асимметричных распределений в нормальное. В теории математической статистики (Хальд, 1956) такие возможности имеются.

Примером асимметричного распределения, которое нередко применяется в качестве модели распределения эмпирических данных, является так называемая логарифмически нормальная (логнормальная) функция. Мы говорим, что случайная величина ξ распределена логнормально, если случайная величина $\eta = \ln \xi$ распределена нормально. Обозначим плотность вероятности распределения случайной величины ξ через $f_\xi(x)$, а плотность распределения случайной величины η через $f_\eta(y)$. Пусть также μ и σ^2 — параметры распределения случайной величины η, т. е. μ — среднее значение логарифмов ξ, а σ^2 — их дисперсия. Тогда

$$f_\eta(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, \qquad f_\xi(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}.$$

Кривая $f_\eta(y)$, будучи нормальной, достигает максимума в точке $y = \mu$ и симметрична относительно ординаты в этой точке. Функция же $f_\xi(x)$ достигает максимума в точке $x = e^{\mu-\sigma^2}$ и является положительно асимметричной. Таким образом, максимум кривой $f_\xi(x)$ смещен влево относительно среднего значения случайной величины ξ, которое равно $e^{\mu+\frac{1}{2}\sigma^2}$.

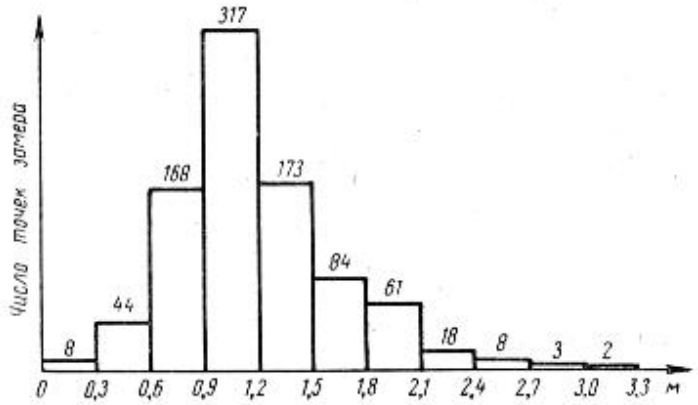


Рис. 15. Гистограмма распределения мощности угольного пласта № 11 по замерам в лавах Коспашского района Кизеловского бассейна

Обозначим функцию распределения логнормальной случайной величины ξ через $F_\xi(x)$. Ее можно представить как

$$P(\xi \leqslant x) = F_\xi(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_0^x \frac{1}{x} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}} dx.$$

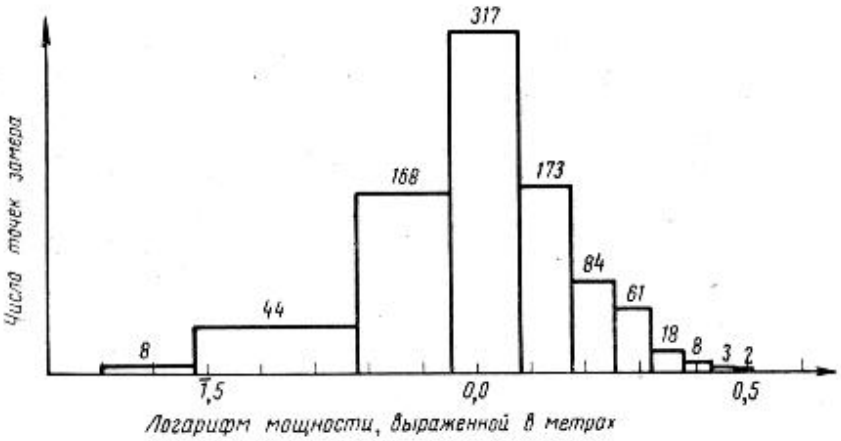


Рис. 16. График, показывающий распределение мощности угольного пласта № 11 по замерам в лавах Коспашского района Кизеловского бассейна

Нетрудно видеть, что

$$P(\xi \leqslant x) = P(e^\eta \leqslant e^y) = P(\eta \leqslant y).$$

Так как случайная величина $\eta = \ln \xi$ распределена нормально, то

$$P(\eta \leqslant \mu) = P(\eta \geqslant \mu) = 0{,}5.$$

Следовательно,

$$P(e^\eta \leqslant e^\mu) = P(e^\eta \geqslant e^\mu) = P(\eta \leqslant \mu) = 0{,}5.$$

Таким образом, величина e^{μ} в случае логнормального распределения случайной величины ξ является медианой этого распределения, т. е. делит его на две равные части. Следует отметить, что в условиях логнормального распределения величина e^{μ} больше наиболее часто наблюдаемого значения $e^{\mu-\sigma^2}$ случайной величины ξ и меньше среднего значения этой величины $\left(e^{\mu+\frac{1}{2}\sigma^2}\right)$.

На важную роль логнормального распределения в геологии впервые указал Н. К. Разумовский (1939), установивший, что распределение содержаний золота в россыпи близко к логнормальному.

Л. Аренс (Ahrens, 1953) также указывал на логнормальное распределение содержаний химических элементов в гранитах. Он назвал это явление основным законом геохимии.

Эффект логарифмического преобразования асимметричных распределений, приводящий к распределениям, близким к нормальным, можно продемонстрировать на примере распределения мощности угольного пласта № 11 (рис. 15) Кизеловского бассейна (Шарапов, 1964). Если мощность пласта выразить логарифмически (рис. 16), то вид диаграммы изменится — распределение будет выглядеть более симметричным.

III. ХАРАКТЕРИСТИКИ РАСПРЕДЕЛЕНИЯ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

Средняя

Выше говорилось, что распределение случайной величины дается в виде множества ее значений, каждому из которых соответствует вероятность. Если эти значения расположить в порядке возрастания, то получим вариационный ряд. Члены этого ряда иногда называются вариантами.

В практике очень часто возникает необходимость сравнения этих рядов друг с другом. Для этого необходимо найти надежные критерии для такого сравнения, а также иметь характеристики, дающие достаточно полное представление о главнейших свойствах каждого ряда.

Операция, приводящая к краткой записи статистических данных, характеризующих всю совокупность результатов наблюдения, называется свертыванием информации.

Прежде чем перейти к рассмотрению методов свертывания информации и сравнения совокупностей результатов наблюдений, кратко дадим описание некоторых наиболее важных характеристик распределения случайных величин.

Как и раньше, будем обозначать случайную величину через ξ, а через $f(x)$ и $F(x)$ — плотность вероятности и функцию ее распределения.

Очень важной характеристикой распределения случайной величины является ее среднее значение или, как его еще называют, математическое ожидание, которое обозначим через $M\xi$.

Если случайная величина ξ дискретна и $x_1, x_2, \ldots, x_i, \ldots, x_n$ — множество принимаемых ей значений, которым соответствуют вероятности $p_1, p_2, \ldots, p_i, \ldots, p_n$, то

$$M\xi = \sum_{i=1}^{n} p_i x_i.$$

Если же случайная величина ξ непрерывна, а $f(x)$ — соответствующая ей плотность вероятностей, то

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

Следует отметить, что для некоторых распределений математическое ожидание не имеет смысла. Пример такой величины приводит Е. С. Вент-

цель (1962). Если случайная величина ξ принимает значения: 2, 2^2, ..., 2^i, ..., 2^n с вероятностями P_i (соответственно) $\frac{1}{2}$, $\frac{1}{2^2}$, ..., $2\frac{1}{2^i}$, ..., $\frac{1}{2^n}$, то $\sum_{i=1}^{n} p_i = 1$, т. е. распределение имеет смысл, а $\sum_{i=1}^{\infty} x_i p_i$ представляет сумму членов расходящегося ряда. Кроме того, случайная величина, распределенная по закону Коши, не имеет математического ожидания.

В теории вероятностей имеется несколько теорем, касающихся математического ожидания. Простейшие из них приводятся ниже.

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

$$MA = A.$$

2. Математическое ожидание суммы конечного числа n любых как независимых, так и зависимых случайных величин равно сумме их математических ожиданий, т. е.

$$M(\xi_1 + \xi_2 + \cdots + \xi_n) = M\xi_1 + M\xi_2 + \cdots + M\xi_n.$$

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

$$M(\xi_1 \xi_2 \ldots \xi_n) = M\xi_1 M\xi_2 \ldots M\xi_n.$$

4. Постоянный множитель c можно выносить за знак математического ожидания, т. е.

$$M(c\xi) = cM\xi.$$

Математическое ожидание представляет собой теоретическое значение и не является реальной величиной. Так, например, если мы имеем два алмаза весом в 2 и 4 карата, то средний их вес будет равен 3 каратам. Однако алмаза с таким весом нет в натуре, он — воображаемый. К понятию среднего веса алмаза мы приходим путем такого рассуждения: «Если бы оба алмаза, весящие вместе 6 каратов, были одинаковыми по весу, то каждый из них весил бы 3 карата. Назовем эту величину средним весом».

Суть дела не изменится, если у нас будет не 2, а 102 или 1002 алмаза. В большой совокупности алмазов только случайно может оказаться один или несколько алмазов, вес которых равен среднему для всей совокупности. В общем же случае средним весом может не обладать ни один реально имеющийся алмаз.

Возьмем другой пример, иллюстрирующий природу средней величины. Имеется один алмаз. Десять лаборантов определяли его вес и получили разные значение последнего, но алмаз, повторяем, один и вес у него один. Таким образом, в результате наблюдений мы получим значения случайной величины η, представляющей сумму истинного веса алмаза a и случайной величины ξ — погрешности измерения, т. е. $\eta = a + \xi$. В случае, если $M\xi = 0$, то $M\eta = a$.

На практике в большинстве случаев точное значение математического ожидания случайной величины бывает неизвестно, и исследователь вынужден судить о нем по приближенной характеристике, полученной по наблюдаемым значениям этой величины. Такая приближенная характеристика называется оценкой или выборочной средней.

Пусть ξ — случайная величина и $M\xi = \theta$. Обозначим n выборочных значений случайной величины ξ через $x_1, x_2, \ldots, x_i, \ldots, x_n$. Оценкой $\bar{\theta}$ для неизвестной средней θ будем называть некоторую функцию $g(x_1, x_2, \ldots, x_n)$ от выборочных значений $x_1, x_2, \ldots, x_n$, т. е.

$$\bar{\theta} = g(x_1, x_2, \ldots, x_n).$$

Естественно, свойства оценки $\bar{\theta}$ во многом зависят от характера оценивающий функции g. Ясно, что извсего множества возможных оценок следует выбирать те, которые обеспечивают минимальную погрешность в характеристике неизвестной средней.

Оценка называется несмещенной, если ее математическое ожидание равно значению оцениваемого параметра, т. е.

$$M\bar{\theta} = \theta.$$

Геологу очень часто приходится встречаться с оценками средних. Так он оценивает среднюю мощность пласта, среднее содержание полезного компонента в руде, среднюю глубину залегания рудного тела, среднюю ошибку подсчета запасов и т. п.

В зависимости от способа вычисления, выборочные средние бывают простыми, или невзвешенными, и сложными, или взвешенными.

Простая средняя выводится тогда, когда каждому наблюдаемому значению приписывается одна и та же вероятность. Взвешенная же средняя выводится с учетом неодинаковой частоты появления отдельных значений.

Это условие вывода средних не всегда соблюдается. Простую среднюю иногда вычисляют и тогда, когда частота значений различная. При этом получается ошибка, величина которой зависит от степени различия вероятностей.

В связи с вопросом о взвешивании необходимо указать еще на одну ошибку, часто допускаемую геологами.

В геологии распространено понятие среднего содержания, взвешенного по мощности. В практике разведки часто бывает так, что пробы захватывают тело полезного ископаемого не на всю мощность. Часть тела по мощности остается неопробованной. Произведение весового содержания полезного компонента в кубометре пласта и мощности последнего обычно называют вертикальным запасом. Это понятие иногда путают с линейным запасом, т. е. с запасом в полосе шириной в 1 *м* и длиной по всей разведочной линии (на полную мощность пласта). Произведение процентного содержания полезного компонента и мощности называется метропроцентом.

Сущность упомянутой ошибки заключается в том, что, заменив в одном случае исходное содержание вертикальным запасом или в другом случае метро-процентом, геолог получает новую совокупность с другим законом распределения. Если при этом гипотезы и критерии, принятые для исходной совокупности, исследователь переносит на новую совокупность, это может привести к ошибочным выводам. Получив новую совокупность, геолог должен подходить к ней с новыми оценками.

Выборочные средние бывают степенными, показательными, логарифмическими и др., в зависимости от характера определяющей формулы средней (Боярский и др., 1930).

В практике нередко используется степенная средняя, которая в случае невзвешенной средней имеет следующий вид

$$\bar{x} = \sqrt[k]{\frac{\sum_{i=1}^{n} x_i^k}{n}},$$

а в случае взвешенной:

$$\bar{x} = \sqrt[k]{\frac{\sum_{i=1}^{l} m_i x^k}{\sum_{i=1}^{l} m_i}}.$$

В этих формулах $\bar{x}$ — выборочная средняя, k — постоянное целое число, определяющее вид средней, n — объем выборочной совокупности (число членов ряда), $n = \sum_{i=1}^{l} m_i$, m_i — частота или статистический вес значения x_i, l — число значений x_i.

Для k, последовательно принимающего значения $-1, 0, 1$, получаются следующие виды средней: средняя гармоническая, средняя геометрическая, средняя арифметическая, средняя квадратическая и средняя кубическая.

Формулы средней гармонической (соответственно невзвешенной и взвешенной) таковы:

$$\bar{x}_{\text{гарм}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x}}, \qquad \bar{x}_{\text{гарм}} = \frac{\sum_{i=1}^{l} m_i}{\sum_{i=1}^{l} \frac{m_i}{x_i}}.$$

Средняя гармоническая используется тогда, когда суммируемый признак представлен обратной величиной. Такие обратные величины в геологии встречаются часто. Так, например, степень разведанности определяется у одних авторов числом выработок (скважин), приходящихся на разведанную площадь, у других, наоборот, — площадью приходящейся на одну выработку (скважину). В этом случае можно было бы пользоваться средней гармонической.

Средняя гармоническая применяется также и тогда, когда изучается эффективность работы в единицу времени, или производительность труда, брак в работе, зарплата и другие экономические показатели.

Иногда при статистической обработке данных используется средняя геометрическая, которая вычисляется по формулам:

а) невзвешенная

$$\bar{x}_{\text{геом}} = \sqrt[n]{x_1 x_2 \cdots x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}};$$

б) взвешенная

$$\bar{x}_{\text{геом}} = \left(\prod_{i=1}^{l} x_i^{m_i}\right)^{\frac{1}{\sum_{i=1}^{l} m_i}},$$

где $\prod_{i=1}^{n}$ — знак произведения чисел x_i, от x_1 до x_n;

n — число наблюдений в выборке;

m_i — частота значений x_i.

Для упрощения вычислительной работы можно логарифмировать последние две формулы, после чего получим

$$\lg \bar{x}_{\text{геом}} = \frac{\sum_{i=1}^{n} \lg x_i}{n},$$

$$\lg \bar{x}_{\text{геом}} = \frac{\sum_{i=1}^{l} m_i \lg x_i}{\sum_{i=1}^{l} m_i}.$$

Средняя геометрическая часто используется для вычисления индексов, например для вывода среднего коэффициента роста, т. е. применяется как характеристика средних темпов изменения явления. Иногда для упрощения расчетов среднюю геометрическую, выражающую собой коэффициент роста, представляют так (Купарадзе, 1960):

$$h = \sqrt[n-1]{\frac{x_n}{x_a}},$$

где h — средний коэффициент роста;
x_n — конечный уровень;
x_a — начальный уровень;
n — число равных периодов.

Эта формула приводит к такому же результату, что и обычная средняя геометрическая, если изменение уровней от начального до конечного происходит более или менее плавно. Но в геологии уровни могут меняться и плавно, и скачкообразно, поэтому данная формула может привести к недопустимо большим ошибкам, почему ее и нельзя рекомендовать для использования. Г. К. Купарадзе предложил эту формулу для экономических исследований.

Для иллюстрации подобной ошибки используем пример первоисточника (Купарадзе, 1960).

Номер периода времени	Объем производства
1	$30{,}00 = x_a$
2	37,50
3	56,25
4	$92{,}80 = x_n$
Σ	216,65

Для этих данных $h = \sqrt[3]{\frac{92{,}8}{30{,}0}} = 1{,}46$.

Влияние промежуточных периодов этой формулой совершенно исключается, но если во втором периоде $x_2 = 90{,}00$, а в третьем $x_3 = 3{,}85$, то при неизменности общей суммы (216,65) и величины h, равной 1,46, будем иметь: в первых двух периодах $x_a + x_2 = 120{,}00$, а во вторых двух периодах $x_3 + x_n = 96{,}65$. Сопоставив эти удвоенные периоды, видим значительное падение величины x, тогда как величина h возрастает.

На основании сказанного можно сделать вывод, что величину h нельзя рекомендовать для измерения роста сильно изменчивого процесса.

Среднюю геометрическую иногда используют для характеристики среднего диаметра частицы в дробленой пробе (Андреев, Зверевич, Перов, 1961), среднего размера золотинки, а также для вычисления критического веса ураганной (выдающейся) пробы.

Следует отметить, что среднее геометрическое не является оценкой неизвестного математического ожидания, так как в пределе (при увеличении числа наблюдений) оно стремится к величине, меньшей, чем математическое ожидание.

Очень широко распространенной оценкой математического ожидания является среднее арифметическое. Весьма важным свойством этой величины является то, что она при увеличении числа наблюдений в пределе приближается к неизвестному значению математического ожидания, независимо от вида функции распределения случайной величины.

Средняя арифметическая вычисляется по формулам:

а) невзвешенная

$$\bar{x}_{ap} = \frac{\sum_{i=1}^{n} x_i}{n};$$

б) взвешенная

$$\bar{x}_{ар} = \frac{\sum_{i=1}^{l} m_i x_i}{\sum_{i=1}^{l} m_i}.$$

Среднее арифметическое обладает следующими свойствами:

1. Средняя из равных величин всегда равна их значению. Если, например,

$$x_1 = x_2 = x_3 = x_4 = x_5 = 4,$$

то и

$$\bar{x}_{ар} = \frac{\sum x_i}{n} = \frac{20}{5} = 4.$$

2. Сумма соответствующим образом взвешенных отклонений наблюдаемых значений от средней арифметической равна нулю:

$$\sum_{i=1}^{l} m_i \left(x_i - \bar{x}_i\right) = 0,$$

так как

$$\sum_{i=1}^{l} m_i \left(x_i - \bar{x}_i\right) = \sum_{i=1}^{l} m_i x_i - \bar{x} \sum_{i=1}^{l} m_i = n\bar{x} - \bar{x}n = 0,$$

где

$$n = \sum_{i=1}^{l} m_i.$$

3. Сумма $S = \sum_{i=1}^{l} m_i (x_i - a)^2$ достигает минимума для $a = \bar{x}$ (здесь a — любое число). Это свойство важно для способа наименьших квадратов. Сумма квадратов отклонений наблюдаемых значений x_i от их среднего арифметического $\bar{x}$ меньше, чем сумма квадратов отклонений тех же значений от любого другого числа a.

4. Если все наблюдаемые значения умножить на одно и то же постоянное число (целое или дробное), то их среднее арифметическое также окажется умноженным на это число. Это значительно облегчает вычисления.

5. Если из всех значений x_i вычесть одно и то же число x_0, то среднее арифметическое $\bar{x}'$, вычисленное по уменьшенным данным ($x - x_0 = x'$), окажется меньше $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ на ту же величину x_0.

В случае небольшого объема совокупности вычисление простой (невзвешенной) средней арифметической легко производится на конторских счетах или с помощью специальных таблиц А. Н. О'Рурка (1953).

Если же объем совокупности большой, а наблюдаемые значения выражены многозначными числами, то вычисление взвешенной средней арифметической отнимает много времени и часто сопровождается ошибками (просчетами). Чтобы сократить ошибки и облегчить труд вычислителя, рекомендуют пользоваться следующей формулой:

$$\bar{x} = a + \frac{\sum m_i (x_i - a)}{n},$$

где $\bar{x}$, m_i, n и x_i — имеют прежние значения, а число a может быть каким угодно, лишь бы разности $x_i - a$ — были возможно проще и меньше.

Выбор числа a производится на глаз. Наиболее удачным число a будет тогда, когда оно близко к $\bar{x}$.

Пример. Применение этой формулы показано на примере лоткового опробования золотой россыпи (табл. 16).

Таблица 16

x_i	m_i	$x_i - a$	$m_i (x_i - a)$
0	64	—2	—128
1	80	—1	—80
2	46	0	0
3	13	1	13
4	5	2	10
5	2	3	6
7	1	5	5
9	2	7	14
11	3	9	27
Σ	216		—133

В таблице x_i — число золотинок в пробе, m_i — количество проб. Число a принято равным 2.

Вычисление среднего числа золотинок в пробе по последней формуле дает

$$\bar{x} = 2 + \frac{-133}{216} = 2 - 0{,}62 = 1{,}38.$$

Таким образом, в среднем на каждую пробу приходится 1,38 золотинки.

В последнем примере рассматривалась дискретная случайная величина. Гораздо чаще в геологической практике встречается непрерывная случайная величина. Поэтому приведем пример вычисления средней арифметической и для этого случая.

Пример. При составлении геолого-углехимической карты Донбасса в 1936—1950 гг. по 2450 пробам был определен удельный вес угля. Результаты этого определения и вычисление среднего арифметического удельного веса показаны в табл. 17.

Таблица 17

i	x_i	m_i	x_i'	$x_i' - a$	$m_i (x_i' - a)$	e_i	$m_i e_i$
1	1,20—1,25	14	1,225	—0,200	—2,800	—4	—56
2	1,25—1,30	83	1,275	—0,150	—12,450	—3	—249
3	1,30—1,35	475	1,325	—0,100	—47,500	—2	—950
4	1,35—1,40	680	1,375	—0,050	—34,000	—1	—680
5	1,40—1,45	448	1,425	0	0	0	0
6	1,45—1,50	208	1,475	0,050	10,400	1	280
7	1,50—1,55	100	1,525	0,100	10,000	2	200
8	1,55—1,60	82	1,575	0,150	12,300	3	246
9	1,60—1,65	126	1,625	0,200	25,200	4	504
10	1,65—1,70	110	1,675	0,250	27,500	5	550
11	1,70—1,75	57	1,725	0,300	17,100	6	342
12	1,75—1,80	37	1,775	0,350	12,950	7	259
13	1,80—1,85	19	1,825	0,400	7,600	8	152
14	1,85—1,90	6	1,875	0,450	2,700	9	54
15	1,90—1,95	3	1,925	0,500	1,500	10	30
16	1,95—2,00	0	1,975	0,550	0	11	0
17	2,00—2,05	2	2,025	0,600	1,200	12	24
		2450	$a = 1{,}425$		31,700		706

В этой таблице i — номер интервала; x_i — границы интервалов удельного веса угля; m_i — число проб, попавших в i-тый интервал; x_i' — середина интервала. Число a, равное 1,425, взято на глаз, как наиболее близкое к искомому среднему удельному весу. Выбор этот, как увидим дальше, оказался удачным.

Искомое среднее арифметическое равно

$$\bar{x} = 1{,}425 + \frac{31{,}700}{2450} = 1{,}425 + 0{,}013 = 1{,}438.$$

Вычисление средней арифметической в этом примере можно еще более упростить и облегчить, если вместо x_i' ввести новую величину ε_i, связанную с x_i' следующим соотношением:

$$\varepsilon_i = \frac{x_i' - a}{0{,}050} = 20\,(x_i' - a).$$

Среднее арифметическое для ε_i равно

$$\bar{\varepsilon} = \frac{706}{2450} = 0{,}288.$$

С помощью этой величины средний удельный вес охарактеризуем следующим образом:

$$\bar{x} = a + \frac{1}{n}\sum_{i=1}^{t} m_i\,(x_i' - a) = a + \frac{1}{n \cdot 20}\sum_{i=1}^{t} m_i \varepsilon_i =$$

$$= 1{,}425 + \frac{1}{2450 \cdot 20} \cdot 706 = 1{,}425 + 0{,}013 = 1{,}438,$$

т. е. тот же результат.

Для еще большего облегчения вычисления среднего арифметического можно укрупнить интервалы, но это связано с некоторой ошибкой, величина которой зависит от характера распределения. Для ориентировочного определения среднего арифметического можно объединить интервалы так, чтобы их число уменьшилось в три раза. Тогда каждому из трех объединенных интервалов просто приписывается значение среднего из этих трех интервалов.

Пример. Сделаем этот подсчет с числом золотинок (табл. 18).

$$\bar{x} = 2 + \frac{-105}{216} = 2 - 0{,}49 = 1{,}51.$$

Таблица 18

x_i	m_i	$x_i - a$	$m_i\,(x_i - a)$
1	190	—1	—190
4	20	2	40
7	1	5	5
10	5	8	40
	216		—105

Полученная величина 1,51 отличается от более точной 1,38 на 11%. Это расхождение связано со значительной неравномерностью распределения, не учтенной последним, самым упрощенным способом.

В случае очень большого объема совокупности ее можно разбить на части, затем вычислить среднюю для каждой из них, а по этим средним определить общую среднюю, взвешенную по объемам частей.

Описанные способы вычисления выборочных средних практически очень важны в условиях полевой работы геолога, когда он не имеет никакой вычислительной техники.

Свойство выборочной средней при увеличении числа наблюдений приближаться к математическому ожиданию случайной величины, или, как его еще называют, генеральному среднему, имеет огромное значение для геологической разведки, в основе которой лежит выборочный метод.

Методы выборки будут рассмотрены ниже, а здесь покажем на примере близость выборочной средней к генеральной средней.

Пример. На одном месторождении меди проанализировано 1190 проб, которые рассматриваются как модель генеральной совокупности. Разрабатывая новую методику опробования комплексных руд на элементы-примеси, автор сделал из нее различные выборки. Результаты выборки и сравнение их с генеральной совокупностью показаны в табл. 19.

Таблица 19

Определение	Генеральная совокупность	Выборочные совокупности				
Число проб	1190	12	24	36	60	120
Среднее содержание цинка, %	3,79	4,17	3,70	2,93	4,17	3,59
Среднее содержание кадмия, *г/т*	117	174	105	105	125	108

Эффективность выборки и точность выборочных средних будут рассмотрены ниже, а здесь лишь отметим только, что расхождение между средними из 1190 проб и из 120 проб не очень велико (по цинку на 7 относительных процентов, а по кадмию — на 8 относительных процентов), а разница в затратах труда на взятие, обработку и анализ проб огромна (экономия труда десятикратная).

Перейдем к рассмотрению средней квадратической.

Невзвешенная квадратическая средняя вычисляется по формуле:

$$\bar{x}_{\text{кв}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}\,.$$

Взвешенная средняя квадратическая представлена выражением

$$\bar{x}_{\text{кв}} = \sqrt{\frac{1}{\sum_{i=1}^{l} m_i}\sum_{i=1}^{l} m_i x_i^2}\,.$$

Средняя квадратическая имеет важное значение в теории контрольных анализов геологических проб, в определении точности подсчитанных запасов и во многих других случаях разведочной практики.

Средняя кубическая невзвешенная вычисляется по формуле

$$\bar{x}_{\text{куб}} = \sqrt[3]{\frac{1}{n}\sum_{i=1}^{n} x_i^3}\,.$$

Для взвешенной же средней кубической существует формула

$$\sqrt[3]{\frac{1}{\sum_{i=1}^{l} m_i}\sum_{i=1}^{l} m_i x_i^3}\,.$$

Средняя кубическая иногда применяется при изучении крупности россыпного золота и частиц дробленой породы.

Выбор вида средней зависит от цели исследования, а иногда и от степени осторожности выбирающего, так как между этими видами средней есть заметная разница. Так, средняя гармоническая меньше любой другой средней (из числа вышеописанных). При подсчете запасов полезного ископаемого иногда проявляют осторожность в определении среднего содержания металла в руде и стараются поэтому найти «законное», математическое, обоснование для такой осторожности. При вычислении средней ошибки химического определения полезного компонента в руде, наоборот, допускают иногда некоторое заведомое преувеличение этой ошибки, чтобы застраховать себя от переоценки качества разведки.

Между разными видами средних существует такое соотношение:

$$\bar{x}_{\text{гарм}} < \bar{x}_{\text{геом}} < \bar{x}_{\text{ар}} < \bar{x}_{\text{кв}} < \bar{x}_{\text{куб}}$$

при $x_i \geqslant 0$.

Примером применения различных видов средней является вычисление среднего диаметра частицы при дроблении проб (Андреев, Зверевич, Перов, 1961).

В тех случаях, когда нас интересует время проявления какого-либо процесса, мы можем вычислить еще один вид средней — среднюю хронологическую (Купарадзе, 1960), которая определяется по формуле

$$\bar{x}_{\text{хр}} = \frac{\frac{1}{2}x_1 + x_2 + x_3 + \cdots + x_{n-1} + \frac{1}{2}x_n}{n},$$

где $\bar{x}_{\text{хр}}$ — средняя хронологическая,

x_1, x_2 и т. д. — члены ряда (от первого до n-ого).

n — число членов ряда (или интервалов).

В геологии время играет значительную роль, например, при изучении дебита источников, газовых струй, нефти и т. д.

Средняя хронологическая иногда бывает полезна и тогда, когда время в нашей совокупности не фигурирует, но когда веса крайних членов ряда необходимо брать половинными. Такой случай может быть, например, в разведочной линии, ограниченной крайними выработками (Шарапов, 1952).

В некоторых случаях, когда распределение резко асимметрично, как, например, распределение размера золотинок, некоторые исследователи применяют показательную среднюю.

Ее величина определяется из формулы

$$e^{n\xi} = e^{\Sigma x_i},$$

где e — основание натуральных логарифмов;

n — число членов ряда;

ξ — средняя показательная;

x_i — изменяющееся значение признака.

Показательную среднюю использовал Н. К. Разумовский (1939) для создания нового метода подсчета запасов золота.

Медиана

Особой мерой положения распределения случайной величины является медиана, под которой понимают то значение признака, которое разбивает весь упорядоченный ряд на две равные (по числу значений) части. Если объем совокупности выражается нечетным числом, то выборочной медианой будет значение, находящееся в середине упорядоченного ряда. Если же объем совокупности выражается четным числом, то медианой будет среднее арифметическое между двумя соседними значениями, лежащими на равном удалении (по числу членов ряда) от начала и от конца

ряда. Иначе говоря, под выборочной медианой понимается то значение случайной величины, которое разбивает всю область наблюдаемых значений, расположенных в возрастающем порядке, на две равные по частоте части.

Для вычисления медианы в интервальном вариационном ряду пользуются следующей формулой:

$$Me = N + \frac{c\,\frac{\sum m_i}{2} - S_n}{m_{Me}},$$

где Me — медиана,

N — нижняя граница интервала, в котором лежит медиана,

c — длина интервала,

S_n — сумма частот всех интервалов, от самого нижнего до медианного (но без него),

m_{Me} — частота медианного интервала.

Медиана удобна тогда, когда крайние интервалы в совокупности неопределенны.

Пример. Вычислить медиану по данным табл. 20.

Таблица 20

x	m_i	$\sum m_i$
1,20—1,25	14	14
1,25—1,30	83	97
1,30—1,35	475	572
1,35—1,40	680	1252
1,40—1,45	448	1700
1,45—1,50	208	1908
1,50—1,55	100	2008
1,55—1,60	82	2090
1,60—1,65	126	2216
1,65—1,70	110	2326
1,70—1,75	57	2383
1,75—1,80	37	2420
1,80—1,85	19	2439
1,85—1,90	6	2445
1,90—1,95	3	2448
1,95—2,00	0	2448
2,00—2,05	2	2450
	2450	

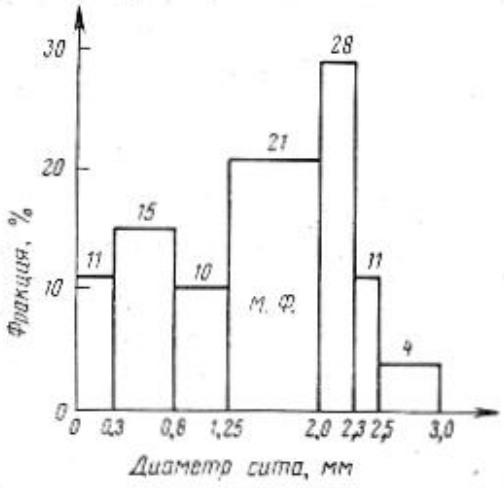


Рис. 17. Гистограмма результатов ситового анализа. *Мф* — медианная фракция

Медиана, т. е. величина x, делящая все число (2450) пополам, находится в интервале 1,35—1,40. По последней формуле имеем

$$Me = 1{,}35 + \frac{0{,}05\left(\frac{2450}{2} - 572\right)}{680} = 1{,}397 \simeq 1{,}40.$$

Медиану можно найти также графическим путем. На графике интегрального распределения медианой будет то значение признака, которому соответствует вероятность 0,5.

В случае гистограммы сначала находим тот интервал, в котором лежит медиана. Для этого складываем числа наблюдений в интервалах одно за другим, начиная с самого нижнего, и тот интервал, нижней границе которого соответствует сумма, меньшая 50% всего объема совокупности, а верхней — большая 50%, будет медианным интервалом. На рис. 17 он показан. До него имеется $11 + 15 + 10 = 36\%$ накопления. После него $28 + 11 + 4 = 43\%$.

Медиана иногда бывает величиной неопределенной, точнее не строго фиксированной, могущей принимать разные значения для одной и той же

совокупности. Так, например, если число золотинок в отдельных пробах будет равно 1, 2, 3, 4, 5, 8, 9, 10, 11, 12, то медиана числа золотинок будет неопределенной — 6 или 7. Любое из этих двух чисел можно принимать за медиану.

В непрерывном распределении медиана тоже может быть неопределенной, что видно из табл. 21.

Медиана здесь может быть принята какой угодно в интервале 4—7%.

Таблица 21

Содержание свинца, %	0—1	1—2	2—3	3—4	4—7	7—8	8—9	9—10
Число проб . . .	3	4	1	2	0	6	3	1

Мода

Особым видом меры положения распределения случайной величины является мода. Под модой понимается то значение признака, для которого кривая плотности вероятности достигает максимума. Значения случайной величины, близкие к моде, встречаются наиболее часто. По числу вершин кривых распределения различают безвершинные, одновершинные, двухвершинные и вообще многовершинные. Такие кривые распределения называются бимодальными, тримодальными и вообще полимодальными. Если вершина одна, то кривая будет мономодальной. Кривая, имеющая хотя бы одну моду, называется модальной. В противном случае она будет амодальной. Частным случаем амодальной кривой является кривая, имеющая одну «впадину» и ни одной вершины. Если две, три и больше самых высоких вершин имеют одинаковую высоту, то все они будут модами.

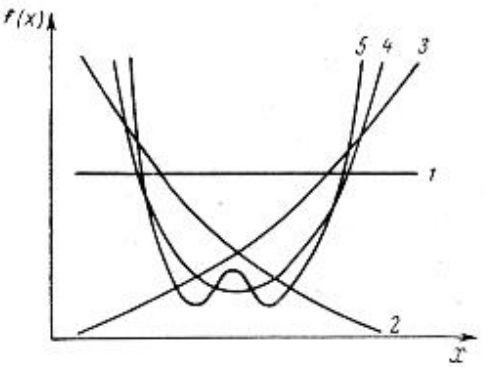


Рис. 18. Схематический вид амодальных кривых плотности вероятности

На рис. 18 схематически показано пять разновидностей амодальных кривых распределения: равномерное распределение (*1*), однобокие распределения (*2*, *3*), U-образное распределение (*4*) и биантимодальное распределение (*5*) (вершина является модой только в том случае, если правая и левая ветви кривой не поднимаются выше вершины).

В. В. Померанцев (1957) приводит из геологической практики кривые распределения, которые можно считать антимодальными.

Возможны сочетания разного числа вершин с разным числом мод на кривых распределения: одновершинное амодальное, двухвершинное мономодальное, трехвершинное бимодальное и т. д.

Л. И. Шаманский (1936) считает, что наличие двух или большего числа вершин на кривой распределения металла в руде свидетельствует о двух или большем числе генетических фаз (т. е. фаз рудообразования).

Оценить моду для мономодальных не очень асимметричных распределений можно по следующей формуле К. Пирсона:

$$Mo = \bar{x} + 3\left(Me - \bar{x}\right) = 3Me - 2\bar{x},$$

где Mo — мода,
x — среднее арифметическое,
Me — медиана.

При непрерывном распределении, а также в сгруппированном дискретном распределении найти моду не всегда легко. Она может быть внутри какого-то интервала.

Моду можно также оценить графическим путем. Если на гистограмме интервалы равные, то моду надо искать в том интервале, которому отвечает самая большая ордината.

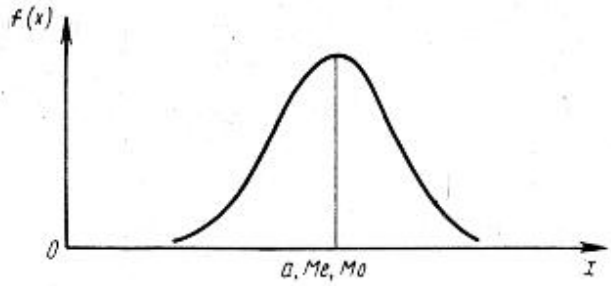


Рис. 19. Схематический вид совпадения среднего (*a*), медианы (*Me*) и моды (*Mo*) на кривой плотности вероятности симметричного распределения

Положение среднего (*a*), медианы *Me* и моды *Mo* на дифференциальной кривой распределения зависит от характера распределения (рис. 19, 20, 21).

В геологии мода интересует нас тогда, когда мы хотим знать наиболее часто встречающийся размер зерен песка, наивероятнейший размер ошибки анализа и т. д.

В случае симметричных распределений, в том числе и в случае нормального, мода, медиана и математическое ожидание совпадают.

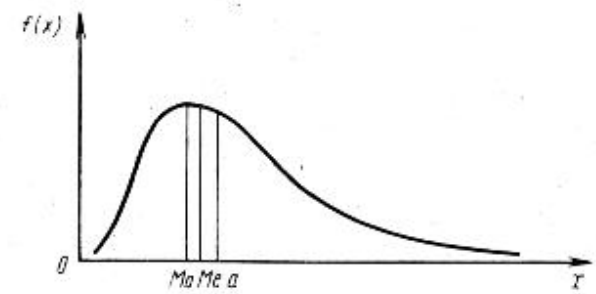


Рис. 20. Схематический вид расположения среднего (*a*), медианы (*Me*) и моды (*Mo*) на кривой плотности вероятности положительно асимметричного распределения

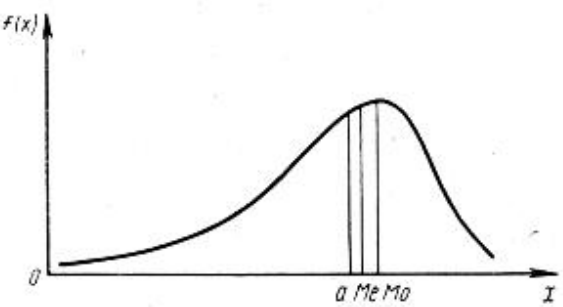


Рис. 21. Схематическое расположение среднего (*a*), медианы (*Me*) и моды (*Mo*) на кривой плотности вероятности отрицательно асимметричного распределения

Дисперсия

Одних лишь мер положения, т. е. средних, медианы и моды, недостаточно для характеристики распределений случайной величины и для решения ряда практических задач, связанных с этими распределениями. В таких случаях требуется знать, насколько далеки значения случайной величины друг от друга, т. е. каково рассеяние этих значений.

Меру такого рассеяния или отклонения значений случайной величины от ее среднего называют дисперсией. Более точно дисперсией случайной величины ξ называется математическое ожидание квадрата отклонения ξ от $M\xi$. Дисперсию величины ξ обычно обозначают $D\xi$ или σ_ξ^2.

Для дискретного распределения дисперсию определяют по формуле

$$D\xi = \sum_{i=1}^{n} p_i (x_i - M\xi)^2, \text{ где}$$

$$p_i = P(\xi = x_i).$$

Если случайная величина ξ непрерывна, а $f_\xi(x)$ — ее плотность вероятности, то

$$M(\xi - M\xi)^2 = D\xi = \int_{-\infty}^{\infty} (x - M\xi)^2 f_\xi(x)\, dx.$$

Формулу дисперсии можно дать в следующем виде:

$$D\xi = M\,[\xi^2 - 2\xi M\xi + (M\xi)^2] = M\xi^2 - 2M\xi M\xi + (M\xi)^2,$$

$$D\xi = M\xi^2 - (M\xi)^2.$$

Приведем пример, иллюстрирующий значение изучения дисперсии в разведке полезных ископаемых.

Качество песка для стекольного сырья зависит не только от среднего размера (диаметра) песчинок. Если песок хорошо отсортирован, с большим преобладанием зерен среднего размера и с малой примесью как очень мелких, так и очень крупных зерен, то он может применяться для изготовления стеклоизделий. Если же песок плохо отсортирован, если в нем много очень мелких и очень крупных зерен, то, хотя средний диаметр песчинок и будет таким же, как в первом случае, такой песок для стеклоизделий непригоден.

Для абразивных целей нужен песок другой степени отсортированности, для инертных добавок в бетон требуется песок со своей дисперсией диаметра зерен.

Значит дисперсия, наряду со средним диаметром песчинок, характеризует качество этого вида сырья. Их значения могут в одних случаях привести к изменению направления разведки, в других — к тому, что месторождение будет забраковано, а в третьих — к его положительной оценке.

Дисперсия содержания металла в руде учитывается при определении категории запасов. Если эта дисперсия будет очень высокой, категории запасов будут понижены, а в обратном случае — повышены.

Дисперсия ошибок анализов также может повлиять на точность разведочных работ.

Чтобы оценить дисперсию, обычно вычисляют квадраты разностей $(x_i - \bar{x})^2$ для всех x_i (здесь x_i — наблюдаемые значения случайной величины, $\bar{x}$ — среднее арифметическое из всех x_i), которые затем суммируют.

Вместо термина «дисперсия» иногда употребляют термины «рассеяние», «варианс», «флуктуация» и «девиата», но все это одно и то же. В. Романовский предложил термин «рассеяние» употреблять для самого явления расхождения значений признаков, а термин «дисперсия» — для измерения этого явления.

Оценкой дисперсии принято называть средний квадрат отклонения, вычисляемый по формуле

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

Корень квадратный из дисперсии называют средним квадратическим отклонением, или стандартом случайной величины. Условно принимается, что этот корень надо брать только со знаком плюс. Оценка среднего квадратического отклонения производится по формуле

$$s = \sqrt{\frac{\sum_{i=1}^{l} n_i (x_i - \bar{x})^2}{n-1}}, \quad \text{где} \quad n = \sum_{i=1}^{l} n_i.$$

Оценки дисперсии и среднего квадратического отклонения по приведенным выше формулам проще вычисляются при малом числе наблюде-

ний, т. е. при малом n, но с ростом последнего трудности вычислительной работы резко возрастают. Поэтому формулу преобразуют:

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

В случае, если n достаточно велико, то

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

Здесь a может быть любым числом, но лучше всего его выбрать так, чтобы разности $x_i - a$ были возможно меньшими.

Пример. На золотоносной россыпи взято 216 лотковых проб, в которых определялось число золотинок x_i. Необходимо вычислить оценку дисперсии. Данные для расчета приведены в табл. 22.

Таблица 22

x_i	n_i	$x_i - a$	$n_i (x_i - a)$	$n_i (x_i - a)^2$
0	64	—2	—128	256
1	80	—1	—80	80
2	46	0	0	0
3	13	1	13	13
4	5	2	10	20
5	2	3	6	18
7	1	5	5	25
9	2	7	14	98
11	3	9	27	243
	216		—133	753

Здесь x_i — число золотинок, n_i — число проб, при этом принято $a = 2$. Строки с нулевой частотой в этой таблице пропущены, но при нумерации оставшихся строк $(x_i - a)$ они учтены.

Среднее арифметическое равно

$$\bar{x} = \frac{\sum_{i=1}^{l} n_i (x_i - a)}{n} + a = \frac{-133}{216} + 2 = 1{,}38,$$

$$\bar{x} - a = 1{,}38 - 2 = -0{,}62.$$

Из последней формулы имеем

$$s^2 = \frac{753}{216} - (0{,}62)^2 = 3{,}49 - 0{,}38 = 3{,}11,$$

отсюда $s = \sqrt{3{,}11} = 1{,}76$.

Дисперсия обладает следующими общими свойствами:

1. Оценка дисперсии s_a^2 — минимальная, когда $a = M\xi$ (такая дисперсия была рассмотрена выше; она обозначена через s_x).

2. Если некоторая величина z связана с x отношением

$$z = \alpha x + \beta,$$

где α и β — некоторые постоянные, то

$$s_z^2 = \alpha^2 s_x^2,$$

где s_x^2 и s_z^2 — оценки дисперсии x и z соответственно.

Пусть A — совокупность объектов и ξ — случайная величина, характеризующая изучаемый признак. Допустим, что A разбита на непересекающиеся подмножества $A_1, A_2, \ldots, A_i, \ldots, A_m$. Пусть также вес каждого подмножества A_i в общей совокупности A равен $\frac{1}{m}$. Обозначим через ξ_i величину признака при условии, что наблюдения ведутся над объектами A_i. Пусть также $M\xi_i = \mu_i$, а дисперсия $D\xi_i = \sigma_i^2$. Тогда генеральная дисперсия $D\xi = \sigma^2$ может быть выражена следующим образом:

$$\sigma^2 = \frac{1}{m}\sum_{i=1}^{m}\sigma_i^2 + \frac{1}{m}\sum_{i=1}^{m}(\mu_i - \mu)^2,$$

где $\mu = M\xi$.

Если $\mu_1 = \mu_2 = \cdots = \mu_m = \mu$, то

$$\sigma^2 = \frac{1}{m}\sum_{i=1}^{m}\sigma_i^2,$$

так как второе слагаемое приведенного выражения обратится в нуль.

В геологии часто приходится сравнивать изменчивость двух и большего числа рядов, хотя это не всегда оказывается возможным, так как результаты измерения признаков в разных рядах могут выражаться в разных единицах измерения. Так, например, мы хотим знать, какое свойство месторождения песков наиболее изменчиво (мощность пласта, содержание кремнезема или средний размер песчинок). Но мощность выражается в метрах, саженях, аршинах, футах, ярдах и др., содержание — в процентах, а размер зерен — в миллиметрах.

Чтобы получить сравнимые меры рассеяния, нужно вместо дисперсий воспользоваться безразмерными величинами.

Такой величиной является коэффициент вариации, или коэффициент изменчивости, представляющий собой стандартное отклонение, выраженное в единицах среднего. Его оценка (V) вычисляется по формуле

$$V = 100\frac{s_x}{\bar{x}}.$$

Здесь число 100 введено для того, чтобы величина V была выражена в процентах.

Коэффициент вариации был предложен К. Пирсоном, а в геологии он был впервые использован Д. В. Наливкиным (1925). В дальнейшем коэффициент вариации нашел широкое применение при разведке полезных ископаемых для характеристики изменчивости месторождений и обоснования густоты разведочной сетки.

В качестве меры рассеяния иногда пользуются средним абсолютным отклонением, оценка которого вычисляется по формуле

$$\theta = \frac{1}{n}\sum_{i=1}^{l} n_i \left| x_i - \bar{x} \right|.$$

Вычисляются средние абсолютные отклонения довольно просто. Это показано на следующем примере (табл. 23).

В этой таблице i — порядковые номера наблюдений; x_i — результаты наблюдений (вес кристаллов пьезокварца, добытых на одном месторождении, в $г$); n_i — число кристаллов (частота); x_0 — условная величина, равная 100; $n = \sum_{i=1}^{l} n_i$.

Таблица 23

i	x_i	n_i	$x_i - x_0$	$n_i(x_i - x_0)$	$\lvert x_i - \bar{x}_0 \rvert$	$n_i \lvert x_i - \bar{x} \rvert$
1	76	1	—24	—24	21	21
2	79	2	—21	—42	18	36
3	80	1	—20	—20	17	17
4	82	1	—18	—18	15	15
5	83	4	—17	—68	14	56
6	85	7	—15	—105	12	84
7	86	5	—14	—70	11	55
8	87	3	—13	—33	0	0
9	88	10	—12	—120	1	10
10	89	12	—11	—132	8	96
11	90	8	—10	—80	7	56
12	93	5	—7	—35	4	20
13	94	18	—6	—108	3	54
14	95	22	—5	—110	2	44
15	96	19	—4	—76	1	19
16	99	26	—1	—26	2	52
17	101	26	1	26	4	104
18	102	16	2	32	5	80
19	104	10	4	40	7	70
20	107	4	7	28	10	40
21	113	1	13	13	16	16
22	116	2	16	32	19	38
23	122	1	22	22	25	25
24	150	1	50	50	53	53
25	158	1	58	58	61	61
26	181	1	81	81	84	84
27	214	1	114	114	117	117
		208		—581		1323

Средний вес кристаллов

$$\bar{x} = x_0 + \frac{1}{n}\sum_{i=1}^{l} n_i (x_i - x_0) = 100 +$$

$$+ \frac{1}{208}(-581) = 100 - \frac{581}{208} = 100 - 2{,}8 = 97{,}2 \simeq 97 \text{ г}.$$

Оценка среднего абсолютного отклонения

$$\vartheta = \frac{1323}{208} = 6{,}351 \simeq 6{,}4 \text{ г}.$$

Если распределение слабо уклоняется от нормального, то

$$\vartheta \simeq s\sqrt{\frac{2}{\pi}} = 0{,}798s, \quad \text{или} \quad s \simeq \frac{\vartheta}{0{,}798}.$$

Величина ϑ, как видно из этой формулы, меньше, чем s.

Применение ϑ для вычисления оценки дисперсии s^2 дает весьма неточные результаты. Так, по Р. Фишеру, нужно 800 испытаний (проб) для того чтобы, оценивая ϑ — неизвестную дисперсию нормальной совокупности, получить столь же точное приближение, какого можно добиться, пользуясь обычной оценкой из 700 испытаний.

Таким образом, среднее абсолютное отклонение ϑ имеет как преимущества, так и недостатки, по сравнению с выборочным средним квадратическим отклонением s. В некоторых случаях его можно использовать для решения методических задач разведки. И. П. Шарапов (1952) предложил коэффициент выдержанности месторождений, выведенный на базе среднего абсолютного отклонения.

Квантили

Величину рассеяния можно измерить также с помощью квартилей, т. е. четвертей. Различают первую, вторую и третью квартили. Их обозначают через Q_1, Q_2 и Q_3.

Первая квартиль — это то значение случайной величины, которое отсекает от самого нижнего значения этой величины в упорядоченном ряду ровно четвертую часть всего объема совокупности. Вторая квартиль — есть не что иное, как медиана. Третья квартиль отсекает три четверти всего объема совокупности в упорядоченном ряду, начиная с самого нижнего значения случайной величины.

Если разность между третьей и первой квартилями поделить пополам, то получим так называемое вероятное отклонение аргумента Q, т. е.

$$Q = \frac{Q_3 - Q_1}{2}.$$

Для нормального распределения

$$Q = 0{,}67449\sigma,$$

откуда

$$\sigma = 1{,}4826Q.$$

Квартиль является частным случаем квантиля. Понятие о квантиле можно выразить следующим образом. Возьмем какой-либо вариационный ряд, состоящий, например, из 1000 членов. Расположим эти члены в порядке возрастания (точнее, неубывания). Тогда 250-й член ряда будет иметь величину признака, равную первой квартили, т. е. первой четверти, 500-й второй квартили, 750-й — третьей квартили. Таким образом, вся совокупность членов ряда в этом случае разбита на 4 части.

Если же ее разбить не на 4, а на 10 частей, то получим уже не квартили, а децили. Четвертый дециль, например, будет представлен значением признака у 400-го члена ряда.

Рассеяние можно охарактеризовать также размахом, или широтой, распределения, т. е. разностью между максимальным и минимальным выборочными значениями ($x_{max} - x_{min}$). Между средним размахом распределения и объемом совокупности в распределениях, не очень отличающихся от нормального, существует зависимость, показанная ниже:

n	Средн. $\frac{(x_{max} - x_{min})}{s}$
2	1,128
5	2,326
10	3,078
50	4,498
100	5,015
500	6,073

Эту зависимость можно использовать для нахождения оценки дисперсии.

Моменты

Математическое ожидание и дисперсия случайной величины представляют собой лишь частные случаи других, более общих характеристик, называемых моментами распределения или просто моментами.

Моментом μ_n порядка n случайной величины ξ относительно константы A называется математическое ожидание случайной величины $(\xi - A)^n$, т. е.

$$\mu_n = M(\xi - A)^n = \int_{-\infty}^{\infty} (x - A)^n \, dF_\xi(x),$$

где $F_\xi(x)$ — функция распределения случайной величины ξ.

Если величина ξ дискретна, то интеграл заменяется знаком суммы. При $A = 0$ момент называется начальным, а при $A = M\xi$ центральным.

Таким образом, $M\xi$ — первый начальный момент случайной величины, а $D\xi$ — второй центральный момент.

Оценку момента порядка n, или эмпирический момент, можно определить по формуле

$$\hat{\mu}_n = \frac{\sum_{i=1}^{l} m_i (x_i - A)^n}{\sum_{i=1}^{l} m_i},$$

где $\hat{\mu}_n$ — эмпирический момент n-го порядка,

m_i — частота или частость,

x_i — наблюдаемые значения случайной величины,

A — постоянная величина.

Оценки моментов вычисляются как для дискретных, так и для непрерывных рядов. В последнем случае x_i означает середину i-того интервала при сгруппированных данных.

Если A равно среднему арифметическому, то оценка момента, получаемая по формуле, представляет собой характеристику центрального момента.

Эмпирические начальные моменты вычисляются по следующей общей формуле:

$$\hat{\mu}_n = \frac{1}{\sum_{i=1}^{l} m_i} \sum_{i=1}^{l} m_i x_i^n.$$

При $n = 0$ имеем начальный момент нулевого порядка, который всегда равен единице.

При $n = 1$ получим начальный момент первого порядка

$$\hat{\mu}_1 = \frac{1}{\sum_{i=1}^{l} m_i} \sum_{i=1}^{l} m_i x_i = \bar{x}.$$

Этот выборочный момент представляет собой среднее арифметическое. При $n = 2$ имеем начальный момент второго порядка

$$\hat{\mu}_2 = \frac{1}{\sum_{i=1}^{l} m_i} \sum_{i=1}^{l} m_i x_i^2 = \bar{x}^2.$$

Попутно заметим, что не следует путать средний квадрат $\bar{x}^2$ с квадратом среднего $(\bar{x})^2$. Это же замечание относится и к другим степеням x.

Начальные моменты третьего и четвертого порядков даются выражениями:

$$\hat{\mu}_3 = \frac{\sum_{i=1}^{l} m_i x^3}{\sum_{i=1}^{l} m_i} = \bar{x}^3,$$

$$\hat{\mu}_4 = \frac{\sum_{i=1}^{l} m_i x^4}{\sum_{i=1}^{l} m_i} = \bar{x}^4.$$

Оценки центральных моментов (обозначим их $\hat{\nu}_n$) вычисляются по следующей общей формуле:

$$\hat{\nu}_n = \frac{\sum_{i=1}^{l} m_i (x_i - \bar{x})^n}{\sum_{i=1}^{l} m_i}.$$

При $n = 0$ получим оценку центрального момента нулевого порядка

$$\hat{\nu}_0 = \frac{\sum_{i=1}^{l} m_i (x - \bar{x})^0}{\sum_{i=1}^{l} m_i} = 1.$$

Если $n = 1$, получим эмпирический центральный момент первого порядка

$$\hat{\nu}_1 = \frac{\sum_{i=1}^{l} m_i (x - \bar{x})}{\sum_{i=1}^{l} m_i} = 0.$$

Он всегда равен нулю, так как отрицательные отклонения от среднего арифметического погашаются положительными отклонениями.

При $n = 2$ получим оценку центрального момента второго порядка

$$\hat{\nu}_2 = \frac{\sum_{i=1}^{l} m_i (x - \bar{x})^2}{\sum_{i=1}^{l} m_i}.$$

При достаточно большом значении $\sum_{i=1}^{l} m_i$ величина $\hat{\nu}_2$ является практически несмещенной оценкой дисперсии, т. е. $\hat{\nu} \simeq s^2$, откуда

$$s \simeq \sqrt{\hat{\nu}_2}.$$

Если $n = 3$, получим эмпирический центральный момент третьего порядка

$$\hat{\nu}_3 = \frac{\sum_{i=1}^{l} m_i (x - \bar{x})^3}{\sum_{i=1}^{l} m_i}.$$

Этот момент служит мерой асимметрии статистического распределения. Если распределение симметрично, то математическое ожидание случайной величины $\hat{\nu}_3$ равно 0.

При $n = 4$ имеем оценку центрального момента четвертого порядка

$$\hat{\nu}_4 = \frac{\sum_{i=1}^{l} m_i (x - \bar{x})^4}{\sum_{i=1}^{l} m_i}.$$

Подобным образом определяются эмпирические центральные моменты любых порядков.

В практике геологических работ вычисление оценок моментов пятого и еще более высокого порядка, по-видимому, не является необходимостью.

Оценки моментов вычисляются двумя способами — по способу произведений и по способу сумм. При этом сначала вычисляют оценки начальных моментов и оценки моментов относительно заданной величины x_0, а затем по формулам перехода, которые приведены ниже, вычисляют оценки центральных моментов.

Пример. Вычислим по способу произведений оценки начальных моментов распределения числа аварий на буровых скважинах, на основании данных табл. 24.

Таблица 24

x_i	m_i	$m_i x_i$	$m_i x_i^2$	$m_i x_i^3$	$m_i x_i^4$
1	1	1	1	1	1
2	3	6	12	24	48
3	4	12	36	108	324
4	5	20	80	320	1280
5	2	10	50	250	1250
6	1	6	36	216	1296
	16	55	215	919	4195

Здесь x_i — значение числа аварий на скважине;

m_i — количество скважин с числом аварий x_i.

Оценки начальных моментов для этого распределения следующие:

$$\hat{\mu}_0 = 1,$$

$$\hat{\mu}_1 = \frac{55}{16} = 3{,}44,$$

$$\hat{\mu}_2 = \frac{215}{16} = 13{,}44,$$

$$\hat{\mu}_3 = \frac{919}{16} = 57{,}40,$$

$$\hat{\mu}_4 = \frac{4195}{16} = 262{,}2.$$

Величина x_i нередко выражается многозначными числами. Возведение их в квадрат, в третью и в четвертую степень — работа трудоемкая. Если к тому же и частоты m_i довольно значительные, то вычисление оценок начальных моментов без вычислительной техники по приведенному в последнем примере способу затруднительно.

Для упрощения вычислений можно вместо x_i брать разность $x_i - x_0$, а затем ввести соответствующие поправки в результат.

Пример. Вычислить начальные моменты относительно x_0 (по способу произведений), исходя из данных табл. 25.

В этой таблице x_i — содержание хлористого натрия в каменной соли изучаемого месторождения, а m_i — число проб с содержанием x_i. В качестве x_0 можно было бы принять любую величину, например, 91%. Тогда вместо 92% пришлось бы писать 1%, вместо 93% — значение 2% и т. д. При этом все числа были бы положительными. Но еще лучше в качестве x_0 принять одно из значений x_i, близких к моде, как это и сделано в третьем столбце данной таблицы. При таком выборе величины x_0 получились как положительные, так и отрицательные числа. При подсчете суммы такие числа взаимно погашаются.

Таблица 25

x_i	m_i	$x_i - x_0$	$m_i(x_i - x_0)$	$m_i(x_i - x_0)^2$	$m_i(x_i - x_0)^3$	$m_i(x_i - x_0)^4$
92	9	—4	—36	144	—576	2304
93	12	—3	—36	108	—324	972
94	17	—2	—34	68	—136	272
95	16	—1	—16	16	—16	16
96	27	0	0	0	0	0
97	41	1	41	41	41	41
98	35	2	70	140	280	560
99	10	3	30	90	270	810
169			19	607	461	4975

Расчет оценок моментов относительно x_0 приводится ниже:

$$\hat{\mu}_0 = 1,$$

$$\hat{\mu}_1 = \frac{19}{169} = 0{,}11,$$

$$\hat{\mu}_2 = \frac{607}{169} = 3{,}59,$$

$$\hat{\mu}_3 = \frac{-461}{169} = -2{,}73,$$

$$\hat{\mu}_4 = \frac{4975}{169} = 29{,}41.$$

При большом размахе (напомним, размах — это разность между максимумом и минимумом значений случайной величины) распределения разность $x_i - x_0$ иногда бывает довольно большой. Поэтому вычисление моментов в таких случаях все же затруднительно. Между тем, есть возможность упростить вычисления. Для этого надо выбрать число c и на него разделить все разности $(x_i - x_0)$. Вычисление оценок моментов в таком случае нужно вести по формуле

$$\hat{\mu}_n = \frac{\sum_{i=1}^{l} m_i \left(\frac{x_i - x_0}{c}\right)^n}{\sum_{i=1}^{l} m_i} c^n.$$

Пример. Вычислить оценки моментов по следующим данным (табл. 26).

Таблица 26

x_i	m_i	$x_i - x_0$	$\frac{x_i - x_0}{C} = x_{i*}$	$m_i x_{i*}$	$m_i x_{i*}^2$	$m_i x_{i*}^3$	$m_i x_{i*}^4$
10	1	—40	—4	—4	16	—64	256
20	3	—30	—3	—9	27	—81	243
30	6	—20	—2	—12	24	—48	96
40	8	—10	—1	—8	8	—8	8
50	8	0	0	0	0	0	0
60	2	10	1	2	2	2	2
70	1	20	2	2	4	8	16
	29			—29	80	—191	621

В этой таблице x_i — мощность жилы в *см*, а m_i — число таких жил. Величина c принята равной 10. Поэтому вводится новая величина

$$x_{i*} = \frac{x_i - x_0}{c} = \frac{x_i - x_0}{10}.$$

Оценки начальных моментов по новым величинам x_{i*} вычисляются так:

$$\hat{\mu}_0^* = 1,$$

$$\hat{\mu}_1^* = \frac{-29}{29} = -1,$$

$$\hat{\mu}_2^* = \frac{80}{29} = 2{,}78,$$

$$\hat{\mu}_3^* = \frac{-191}{29} = -6{,}59,$$

$$\hat{\mu}_4^* = \frac{621}{29} = 21{,}41.$$

Приняв во внимание, что $x_{i*} = \frac{x_i - 50}{10} = 0{,}1x_i - 5$, найдем оценки моментов относительно x_0 (в нашем примере $x_0 = 50$, а $c = 10$).

$$\hat{\mu}_0 = 1 \cdot 10^0 = 1,$$

$$\hat{\mu}_1 = -1 \cdot 10^1 = -10,$$

$$\hat{\mu}_2 = 2{,}78 \cdot 10^2 = 278,$$

$$\hat{\mu}_3 = -6{,}59 \cdot 10^3 = -6590,$$

$$\hat{\mu}_4 = 21{,}41 \cdot 10^4 = 214100.$$

Далее по формулам перехода вычислим оценки центральных моментов. Общая формула оценок центральных моментов следующая:

$$\hat{\nu}_n = \hat{\mu}_n - C_n^1 \mu_{n-1} \hat{\mu}_1^1 + C_n^2 \hat{\mu}_{n-2} \hat{\mu}_1^2 - C_n^3 \hat{\mu}_{n-3} \hat{\mu}_1^3 + \cdots + (-\hat{\mu}_1)^n.$$

Здесь и ниже $\hat{\mu}_n$ — оценки начальных моментов.

Знаки плюс и минус в этой формуле чередуются. Символы C_n^1, C_n^2 и т. д. обозначают число сочетаний из n по 1, из n по 2 и т. д. Для n, последовательно равного 0, 1, 2, 3 и 4, по этой формуле получаем такие выражения для оценок центральных моментов:

$$\hat{\nu}_0 = \hat{\mu}_0 = 1,$$

$$\hat{\nu}_1 = \hat{\mu}_1 - \hat{\mu}_1^1 = 0,$$

$$\hat{\nu}_2 = \hat{\mu}_2 - \hat{\mu}_1^2,$$

$$\hat{\nu}_3 = \hat{\mu}_3 - 3\hat{\mu}_2\hat{\mu}_1 + 2\hat{\mu}_1^3,$$

$$\hat{\nu}_4 = \hat{\mu}_4 - 4\hat{\mu}_3\hat{\mu}_1 + 6\hat{\mu}_2\hat{\mu}_2^2 - 3\hat{\mu}_1^4.$$

По данным приведенного выше примера вычислим оценки центральных моментов:

$$\hat{\nu}_0 = 1,$$

$$\hat{\nu}_1 = 0,$$

$$\hat{\nu}_2 = 278 - (-10)^2 = 178,$$

$$\hat{\nu}_3 = -6590 - 3 \cdot 278 \cdot (-10) + 2\,(-10)^3 = -6590 + 8340 - 2000 = -250,$$

$$\hat{\nu}_4 = 214\,100 - 4 \cdot (-6590) \cdot (-10) + 6 \cdot 278 \cdot (-10)^2 - 3 \cdot (10)^4 = 214\,100 - 263\,600 + 166\,800 - 30\,000 = 87\,300.$$

Тот же результат для $\hat{\nu}_3$ и $\hat{\nu}_4$ получим по двум последним формулам:

$$\hat{\nu}_3 = -6590 - 3 \cdot 178 \cdot (-10) - (-10)^3 = -6590 + 5340 + 1000 = -250;$$

$$\hat{\nu}_4 = 214\,100 - 4 \cdot (-250) \cdot (-10) - 6 \cdot 178 \cdot (-10)^2 - (-10)^4 = 214\,100 - 10\,000 - 106\,800 - 10\,000 = 87\,300.$$

Применение двух последних формул $\hat{\nu}_3$ и $\hat{\nu}_4$ несколько облегчает вычисления.

По сделанным расчетам имеем

$$x = 50 + (-10) = 40,$$

а оценка дисперсии будет равна

$$s^2 = 178; \quad s = 13{,}3.$$

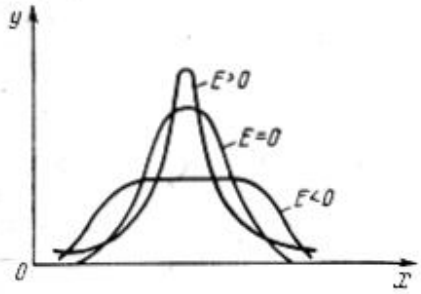


Рис. 22. Кривые плотности вероятности с различными значениями эксцесса

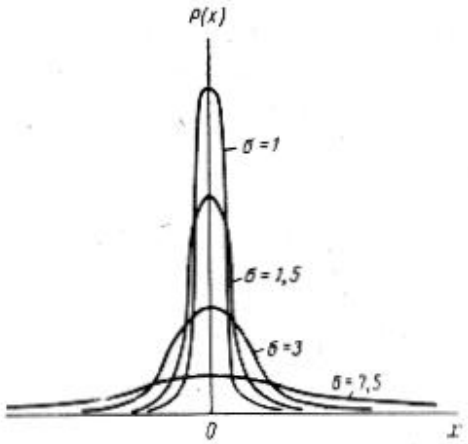


Рис. 23. Кривые плотности вероятности нормально распределенных случайных величин при одном значении среднего и различных стандартных

Если оценки центрального момента порядка k разделить на оценку дисперсии в степени $\frac{k}{2}$, то получим так называемую оценку нормированного момента. Обозначив ее через r, получим следующую формулу для k-того порядка:

$$r_k = \frac{\hat{\nu}_k}{\hat{\nu}_2^{\frac{k}{2}}} = \frac{\hat{\nu}_k}{s^k}.$$

Соответственно будем иметь:

$$r_1 = 0,$$

$$r_2 = 1,$$

$$r_3 = \frac{\nu_3}{s_3},$$

$$r_4 = \frac{\nu_4}{s_4}.$$

По данным предыдущего примера имеем:

$$r_1 = 0,$$
$$r_2 = 1,$$
$$r_3 = \frac{-250}{15{,}81^3} = -\frac{250}{3960} = -0{,}0632,$$
$$r_4 = \frac{87300}{62500} = 1{,}397.$$

Третий нормированный момент является мерой асимметрии кривой плотности вероятностей. Его называют также асимметрией.

Различают правую и левую асимметрию распределения. Первую называют также положительной, а вторую — отрицательной, так как в первом случае $v_3 > 0$, а во втором $v_3 < 0$. Для симметричных распределений асимметрия равна нулю, а среднее арифметическое находится вправо от моды, а в случае левой асимметрии — наоборот.

Четвертый нормированный момент является мерой крутизны кривой плотности вероятности, а разность $r_4 - 3 = E$ называется эксцессом.

Если E больше нуля, то вершина кривой выпуклая, острая. Если $E < 0$, то, наоборот, вершина плоская. Случай, когда $E = 0$, говорит о нормальности кривой (рис. 22, 23).

Не для всякого распределения можно вычислить моменты. Некоторые теоретические распределения не имеют конечных моментов. К таким распределениям можно отнести, например, распределение Коши, частный случай которого может быть представлен формулой

$$y = \frac{1}{\pi} \cdot \frac{1}{1 + x^2},$$

где y = плотность вероятности,
$\pi = 3{,}14159$,
x — заданное значение случайной величины.

IV. ПРАКТИЧЕСКИЕ ПРИЕМЫ ОБРАБОТКИ СТАТИСТИЧЕСКИХ ДАННЫХ

Тела полезных ископаемых, минералы, геологические явления, технические средства разведки или трудовые процессы разведчика, словом различные объекты исследования при изучении могут объединяться в статистические совокупности по одному или по нескольким изучаемым признакам. В зависимости от значений признаков изучаемые объекты образуют группы и классы. При этом определяется число объектов в каждой группе или в каждом классе, которые называются частотой, а сумма частот по всем классам — объемом совокупности.

Различают общую (генеральную) и выборочную совокупности. Первая представляет собой все множество изучаемых объектов. Частной, или выборочной совокупностью считается любая часть генеральной совокупности. Методы исследования выборочных совокупностей описываются ниже, а здесь мы дадим лишь общее представление о выборках. В качестве примера генеральной и выборочной совокупности можно привести следующие данные шурфовой разведки одного месторождения песка.

Пример. Месторождение песка имеет размеры 800 × 1000 *м*. Шурфы расположены по 50-метровой квадратной сетке. Общее число их 320. Все они пробиты до подстилающей пески глины. Шурфы имеют одинаковую площадь поперечного сечения, равную 2 *м*^{2}. Разведанный пласт песка имеет следующие интересующие нас признаки: мощность, размер зерен, наличие примесей, водоносность и др. Наименее выдержанный приз-

нак — это мощность. Поэтому возникает вопрос, по какой сетке следует разведывать другие, подобные этому, месторождения — по 100-метровой или 50-метровой сетке?

Имеющиеся 320 замеров мощности (по шурфам) мы условно будем считать генеральной совокупностью. Изучая «надежность» данных разведки возможной 100-метровой сеткой, мы можем (по методу разрежения сетки) составить четыре выборочные совокупности (по 80 шурфов). Характеристику мощности, установленную по каждой из этих выборочных совокупностей, мы можем сравнить как с эталоном, с характеристикой, полученной по генеральной совокупности, а по величине отклонения от эталона — сделать вывод о допустимости или недопустимости применения 100-метровой разведочной сетки.

Совокупность из 320 шурфов, рассматриваемая как генеральная, вообще также является выборочной из бесчисленного множества возможных шурфов на исследуемой площади.

Всякая разведка с точки зрения статистики есть не что иное, как взятие выборки, дающее возможность по части делать выводы о целом (целым в данном примере мы можем считать все месторождение).

Если в качестве изучаемого признака мы возьмем, например, содержание рутила в песке и будем подсчитывать запасы этого минерала, то мысленно можем все месторождение (пласт песка) разбить на равные прямоугольные участки площадью 2 $м^2$ каждый. Таких участков на месторождении будет 400 000. Из них разведчики вынули (шурфами) только 320, и на основании изучения этих 320 участков сделали заключение (подсчитали запасы) по всем 400 000 участкам. В этом случае выборочная совокупность будет состоять из 320 реальных объектов, а генеральная совокупность — из 400 000 возможных.

Этот пример показывает, что любую совокупность можно в одном случае (по отношению к другой, более значительной, совокупности, в которую входит данная совокупность) рассматривать как выборочную совокупность, а в другом (по отношению к другой, менее значительной, в нее входящей совокупности) — как генеральную. Необходимо заметить, что многие из генеральных совокупностей имеют бесконечно большой объем.

Выборочная статистическая совокупность бывает неупорядоченной и упорядоченной. В первом случае значения признака располагаются (при последовательной записи этих значений) беспорядочно. Во втором они расположены в порядке возрастания (точнее неубывания) или убывания (точнее невозрастания). Так, значения, приведенные в предыдущем примере, в порядке возрастания дадут такой ряд (последовательность): 0,13; 0,17; 0,64; 0,64; 1,12 *м*. Во втором случае те же значения расположатся в обратном порядке (1,12; 0,64; 0,64; 0,17; 0,13 *м*). Термины «неубывание» и «невозрастание» нужны в связи с тем, что в некоторых совокупностях встречается два или большее число одинаковых значений (например, в данном примере дважды встречается 0,64 *м*).

Когда в выборке наблюдается два или большее число одинаковых значений, то говорят о их частоте. Иначе говоря, частота — это число случаев, замеров, проб и других объектов, относящихся к тому или иному значению. Так, в приведенном примере значение 0,64 имеет частоту два, а все остальные — единицу.

Частота, выраженная в долях целого от объема совокупности (в этом случае объем принимается за единицу), называется частостью. Вместо долей целого мы можем брать также проценты — это также будет частость.

Для изучения характера распределения статистической совокупности пользуются разными приемами, в частности: приемом кумулирования (накапливания, суммирования) частот, причем начинают эту операцию в одних случаях с наименьшего значения, а в других — с наибольшего (восходящая и нисходящая, или прямая и обратная, кумуляция).

Накапливаться и убывать могут не только частоты, но и частости. Приведем пример кумуляции, взятый из практики разведки золотых россыпей (табл. 27).

Таблица 27

Определение	Число золотинок в пробе												Всего
	0	1	2	3	4	5	6	7	8	9	10	более 10	
Количество проб (частота)	64	80	46	13	5	2	—	1	—	2	—	3	216
Частость	0,296	0,371	0,213	0,060	0,023	0,009	—	0,005	—	0,009	—	0,014	1,000
То же, в %	29,6	37,1	21,3	6,0	2,3	0,9	—	0,5	—	0,9	—	1,4	100,0
Накопленная частота	64	144	190	203	208	210	210	211	211	213	213	216	—
Убывающая частота	216	152	72	26	13	8	6	6	5	5	3	3	—

При значительном объеме совокупности и сравнительно большом числе значений признака последние рекомендуется группировать. Данные, приведенные в табл. 27, можно сгруппировать следующим образом (табл. 28).

Таблица 28

Определение	Число золотинок							Всего
	0	1	2	3, 4	5, 6	7, 8, 9, 10	более 10	
Частота	64	80	46	18	2	3	3	216
Частость	0,296	0,371	0,213	0,083	0,009	0,014	0,014	1,000

В этом примере отдельные столбцы первоначального распределения (на которые падает мало проб) объединены. Такие объединения делаются по значениям признака, лежащим в каком-то интервале. Поэтому такое распределение называется интервальным.

В этом примере взяты дискретные значения признака, но группировку выборочных данных можно показать и для случая непрерывных распределений.

Пример. Анализ разведочных проб по одному месторождению показал следующие содержания (x_i) олова в процентах и число проб (n_i) для каждого из этих содержаний:

x_i	n_i
0	16
0,01	7
0,03	2
0,08	1
0,10	1
0,11	3
0,71	1
0,72	1
1,20	1
1,93	1
4,05	1
Всего	35

Интервальное распределение, построенное на основе этой таблицы, может быть следующим (табл. 29).

Таблица 29

Содержание олова, %	0	0,01	0,03—0,11	0,71—1,93	4,05	Всего
Число проб	16	7	7	4	1	35

Здесь, как и в предыдущем примере, объединение коснулось не всех, а лишь нескольких значений, так как распределение частот весьма неравномерное. Такие частично интервальные ряды на практике встречаются нередко, но они неудобны для статистических расчетов. В последнем примере между значениями 0,01 и 0,03—0,11 имеется пропуск (пропущено значение 0,02). Такие же пропуски есть и между другими значениями. Подобных пропусков следует по возможности избегать, но в данном примере этого сделать нельзя, так как число проб для содержания 0,01% большое.

В последних двух примерах интервалы взяты неравные. Это вызывается значительной неравномерностью распределения. На практике неравноинтервальные ряды встречаются довольно часто. Такой ряд, например, приведен В. И. Владимирским (1958), который собрал данные о 913 откачках воды из опытных скважин и составил следующую таблицу распределения дебита в зависимости от глубины динамического уровня:

Дебит, м³/ч	Частость
0,5—7	54,3
7—10	19,5
10—20	14,6
20—30	4,3
30—50	3,7
50—70	1,6
70—120	0,9
120—200	0,44
200—300	0,44
Всего	99,78
Невязка . . .	0,22

Более удобна для расчета равноинтервальная группировка. Она бывает с пропусками, как, например, у А. Б. Каждан и Н. И. Соловьева (1958) *, и без пропусков, например, в следующем примере (табл. 30):

Таблица 30

Мощность жилы, м	0—0,2	0,2—0,4	0,4—0,6	0,6—0,8	0,8—1,0	Всего
Число замеров . . .	3	7	6	4	1	21

Здесь все интервалы равны друг другу; между ними нет перерывов; нулевое значение не выделено, а верхний предел конечен.

Каждый интервал имеет нижнюю и верхнюю границу. В интервале 0,4—0,6 *м*, например, нижняя граница 0,4, верхняя 0,6.

При разбивке частот по интервалам надо иметь в виду, что если встретится значение, точно равное пограничному двух соседних интервалов,

* В таблице распределения содержания олова (387 проб) пропущены интервалы 66—585, 596—605 (условных единиц) и другие.

то его следует относить к нижнему интервалу. Так, например, если в приведенном выше примере имеется замер, показавший мощность, равную 0,4 *м*, то этот результат надо отнести к интервалу 0,2—0,4 *м*, а не к интервалу 0,4—0,6 *м*. Только в отчетах о выполнении плана работ и в других экономических работах (Купарадзе, 1960) иногда делают наоборот: вариант с пограничным значением относят к верхнему интервалу, например вариант 100% выполнения плана относят к интервалу 100—110%, а не к 90—100%.

Иногда интервал обозначается не его границами, а центральным значением. Так, интервал 0,4—0,6 *м* обозначают через 0,5 *м*. Соответственным образом и другие интервалы в последнем примере получают обозначение 0,1; 0,3; 0,7; 0,9.

Разность между наибольшим и наименьшим значениями в изучаемой генеральной совокупности называется размахом или амплитудой ряда. Размах бывает конечным и бесконечным. Последний имеет место тогда, когда наибольшее значение обозначено, например: «10 и более» или «более 10», и когда наименьшее значение выражено, например, так: «2 и менее» или «менее 2».

В рядах распределения бывают не только положительные, но и отрицательные значения, а также одни только положительные и одни только отрицательные. Наряду с этим встречаются целочисленные и дробные значения.

Отрицательные значения бывают, например, тогда, когда изучаемым признаком является температура горных пород, а дробные имеются всегда, когда изучается содержание металла в руде.

Длина интервала группировки и общее их число взаимосвязаны (чем больше длина, тем меньше число интервалов).

Число интервалов выбирается так, чтобы видны были особенности характера распределения. Число интервалов зависит от объема выборки, размаха и от некоторых других характеристик ряда. Влияние объема выборочной совокупности чаще всего самое определяющее. Практически берут примерно следующее число интервалов (k) в зависимости от объема выборки (N):

N	k
До 10	3
10—30	3—4
30—100	4—6
100—500	6—9
500—3000	9—13
Более 3000	13—18

Число интервалов практически очень редко превышает 15, так как с их увеличением резко возрастают трудности статистических расчетов.

Выборочный размах (он обозначается как разность $x_{max} - x_{min}$ между наибольшими и наименьшими значениями в выборке) оказывает не обратное, как объем N, а прямое влияние на длину интервала d. Это выражается формулой Стерджесса:

$$d = \frac{x_{max} - x_{min}}{1 + 3{,}322 \lg N}.$$

В правой части этой формулы числитель — размах выборки, а знаменатель — число интервалов, являющееся функцией объема N совокупности.

По этой формуле составлена номограмма в двух вариантах (рис. 24 и 25).

Формула Стерджесса используется также для определения величины

интервала группировки при вычислении критерия согласия χ^2 (описание этого и других критериев дается в главе V).

Применим эту формулу к данным табл. 30.

$$d = \frac{1{,}0 - 0}{1 + 3{,}22 \lg 21} = 0{,}19 \simeq 0{,}2.$$

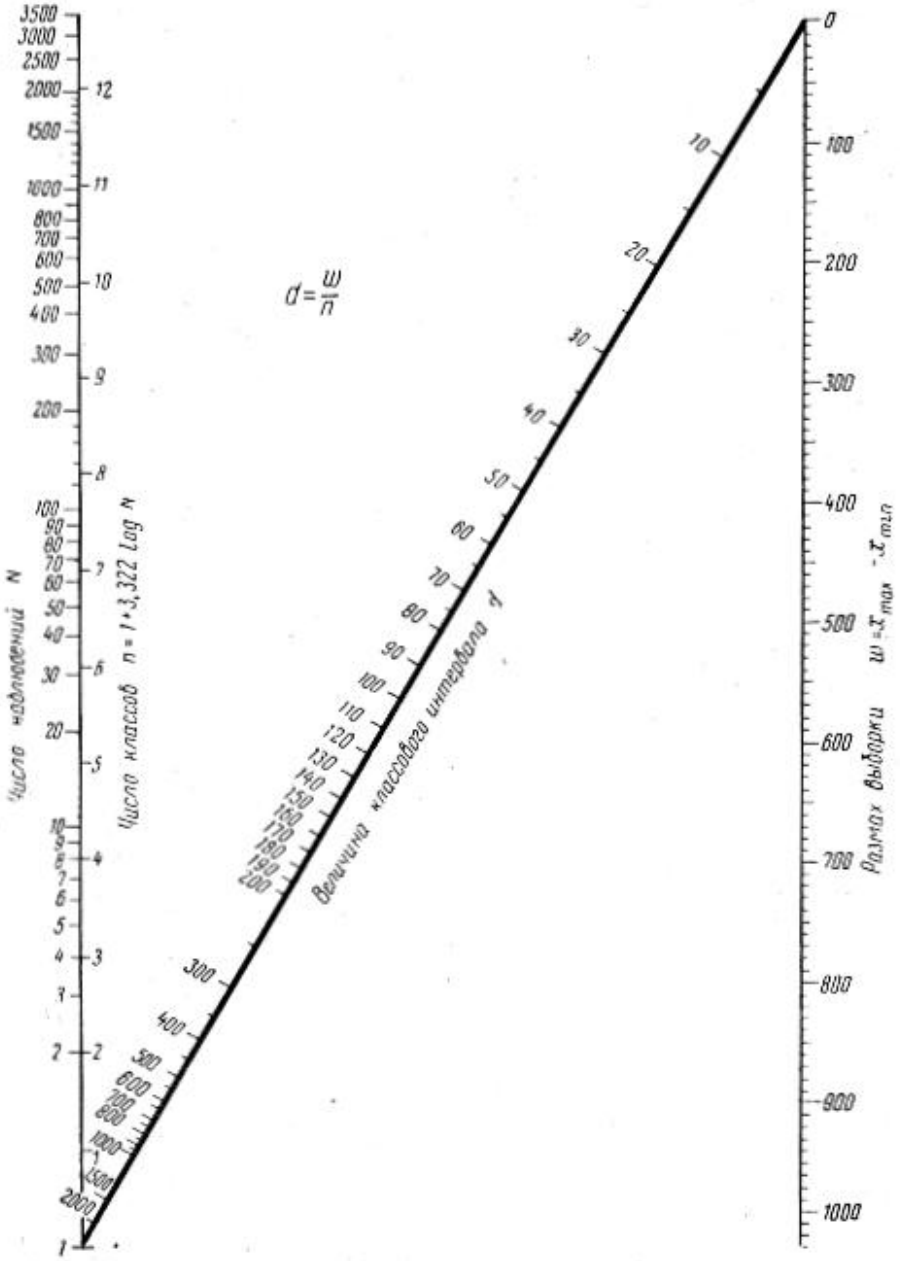


Рис. 24. Номограмма для определения величины интервала группировки (первый вариант)

Поскольку размах выборки в данном примере равен 1,0, а длина интервала равна 0,2, то число интервалов равно 5.

Приведенные расчеты применяются, конечно, для равных интервалов.

Выборочные данные можно представить не только в виде таблицы, но и графически. Графики очень полезны в статистике, так как позволяют сравнительно легко увидеть ту или иную особенность распределения, незаметную в таблице.

Существует четыре вида графического изображения статистических данных: гистограмма, полигон, кумулята и огива. Все чертежи составляются в прямоугольной системе координат.

Гистограмма распределения представляет собой ступенчатую (столбиковую) диаграмму, на оси абсцисс которой отложены нижние и верхние границы всех интервалов, а на оси ординат — отношения $\frac{n_i}{n\Delta_i}$, где

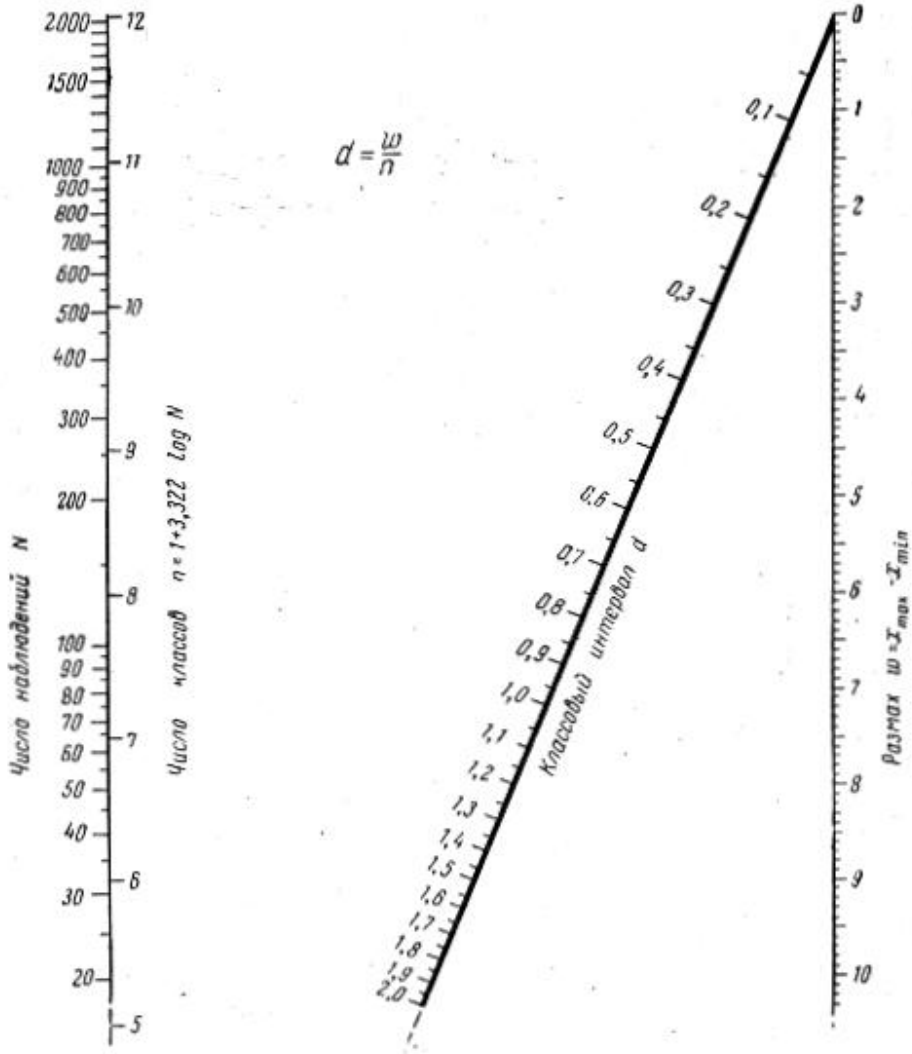


Рис. 25. Номограмма для определения величины интервала группировки (второй вариант)

n_i — число наблюдений, попавших в i-тый интервал, $n = \sum_{i=1}^{l} n_i$, Δ_i — длина i-того интервала. Если интервалы равны, то при построении гистограмм можно пользоваться величинами $\frac{n_i}{n}$ или n_i.

Пример гистограммы, построенной по данным распределения зольности угля, показан на рис. 26.

Полигон распределения представляет собой диаграмму, на оси абсцисс которой отложены центральные значения всех интервалов, а на оси ординат — частоты или частости, если интервалы равны, или $\frac{n_i}{n\Delta_i}$, если интервалы разные. По внешнему виду полигон представляет собой ломаную линию.

Пример полигона распределения показан на рис. 27.

От гистограммы можно перейти к полигону, если середину верхних сторон всех соседних прямоугольников, означающих частоту или частость, соединить между собой прямой линией, как это сделано на рис. 27, исходя из рис. 26.

Значение длины интервалов для наглядности можно показать на примере из практики. При разведке было сделано 38 пересечений одной жилы, по которым замерена ее мощность. В табл. 31 показаны результаты этих измерений.

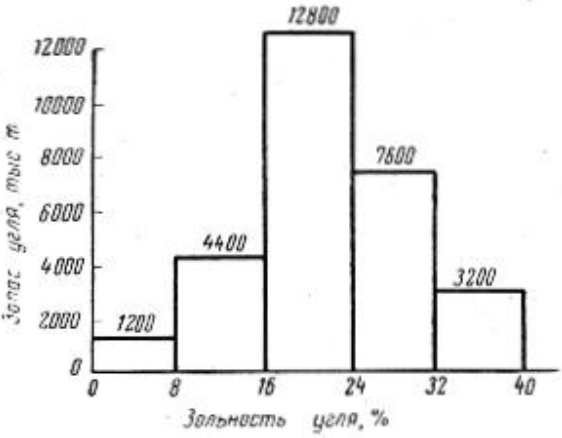


Рис. 26. Гистограмма распределения зольности

Гистограмма, построенная по этим данным без учета длин интервала (рис. 28), показывает значительное преобладание значений среднего интервала, но это впечатление обманчиво. Стоит нам только разбить средний интервал на шесть равных интервалов (таких же по длине, как и крайние интервалы), как это впечатление пропадает (рис. 29), и средним интервалам будут соответствовать минимальные значения частот.

В случае произвольного выбора неравных интервалов при построении гистограммы без учета длины интервала, а также в случае произвольного

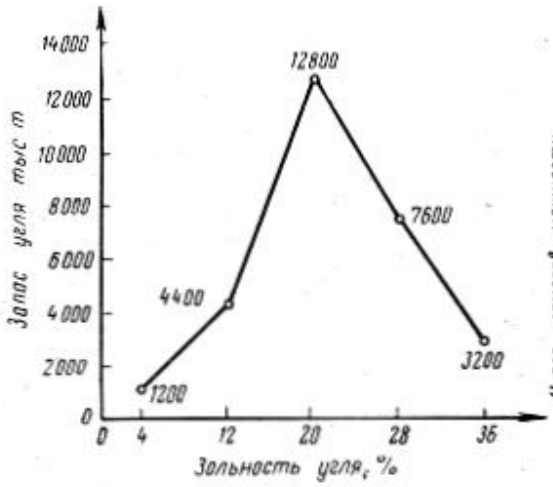


Рис. 27. Полигон распределения зольности угля

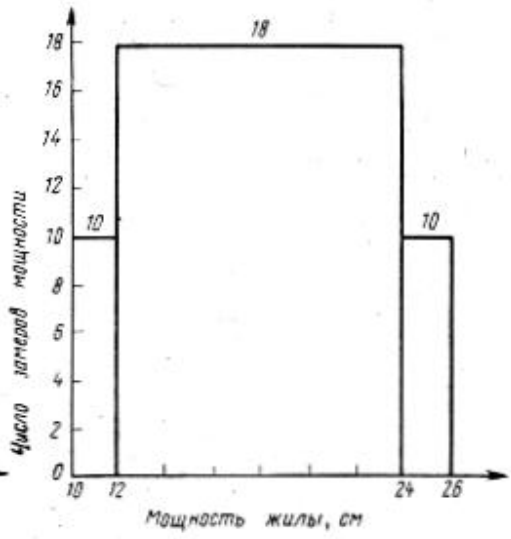


Рис. 28. Гистограмма распределения мощности жилы (при неравных интервалах)

дробления и объединения любую, даже самую малочисленную, группу значений (лишь бы число случаев в ней не равнялось нулю) можно (искусственно) превратить в преобладающую и, наоборот, любую самую много-

Таблица 31

Мощность жилы, см	10—12	12—24	24—26
Число пересечений жилы	10	18	10

численную группу сделать самой малочисленной. Это обстоятельство иногда упускают из виду, строя диаграмму с неравными интервалами без учета их длин и делая по ним, конечно, ошибочные выводы (о максимуме или минимуме групп).

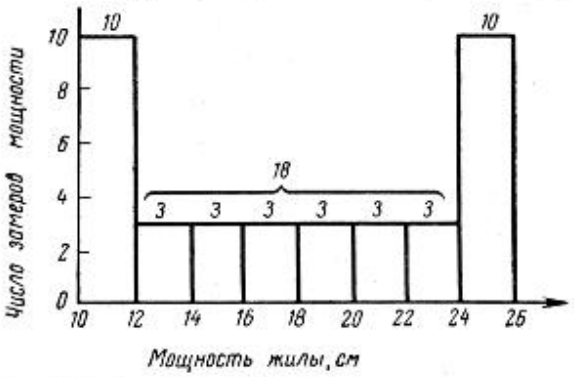


Рис. 29. Гистограмма распределения мощности жилы (при равных интервалах)

Иногда гистограмму заменяют кривой распределения. Для примера покажем такую кривую, построенную по данным табл. 32.

На рис. 30, изображающем распределение сурьмы, крайними показаны нулевые точки на расстоянии одного интервала вниз (к нулевому содержанию) и вверх (к содержанию 24—27%). Масштаб проставлен только по оси абсцисс. На оси же ординат масштаб в таких случаях не указывается. На кривой, показанной на рис. 7, отмечены исходные точки. По этой кривой нельзя определять частоту вне помеченных точек. Кривая только соединяет точки.

Представим себе, что в нашем распоряжении имеется очень большое число (n) результатов наблюдений, которые сгруппированы в некоторые интервалы длиной Δ_i. По значениям $\frac{n_i}{n\Delta_i}$ построена гистограмма. Если в условиях очень большого числа наблюдений сильно уменьшить длины интервалов, то полученная в результате гистограмма будет очень близка к непрерывной кривой.

Переход от гистограммы к кривой распределения показан на рис. 31.

Кумулята (кривая сумм), или кумулятивная кривая, представляет собой диаграмму, на оси абсцисс которой откладываются значения признака, а на оси ординат — накопленные частоты или частости каждого значения. Соседние точки с этими координатами соединяются прямыми линиями, в результате чего получается ломаная линия, в дальнейшем заменяемая кривой.

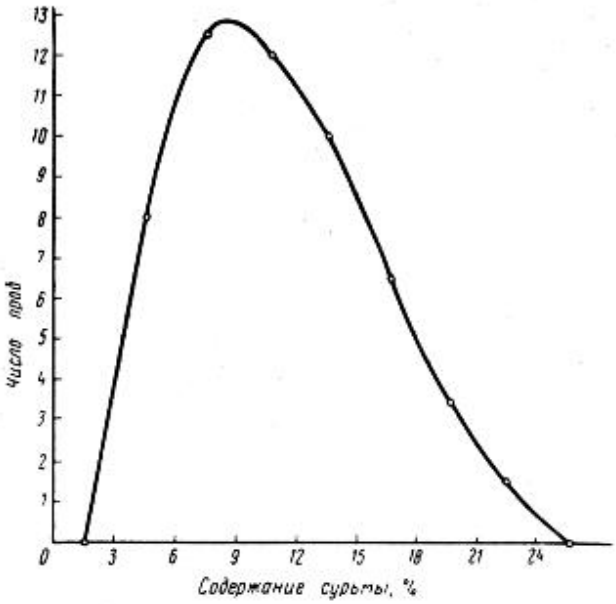


Рис. 30. Кривая распределения содержания сурьмы в пробах

Таблица 32

	Содержание сурьмы, %							Всего проб
	3—6	6—9	9—12	12—15	15—18	18—21	21—24	
Число проб . . .	16	25	24	20	13	7	3	108

Кумуляты можно построить как для восходящего, так и для нисходящего кумулирования. В первом случае необходимо на оси ординат откладывать соответствующие интервалам накопленные частоты или частости. В случае нисходящего кумулирования на оси абсцисс отмечают нижние границы интервалов, а на оси ординат — соответствующие им накопленные частоты или частости.

Необходимо заметить, что кумулята, которую обозначим $\overline{F(x_i)}$, является статистическим аналогом функции распределения $F(x) = P(\xi \leqslant x)$.

Для примера построения кумуляты воспользуемся данными по распределению сурьмы. Подсчет накопления показан в табл. 33.

По этим данным построена восходящая и нисходящая кумуляты (рис. 32).

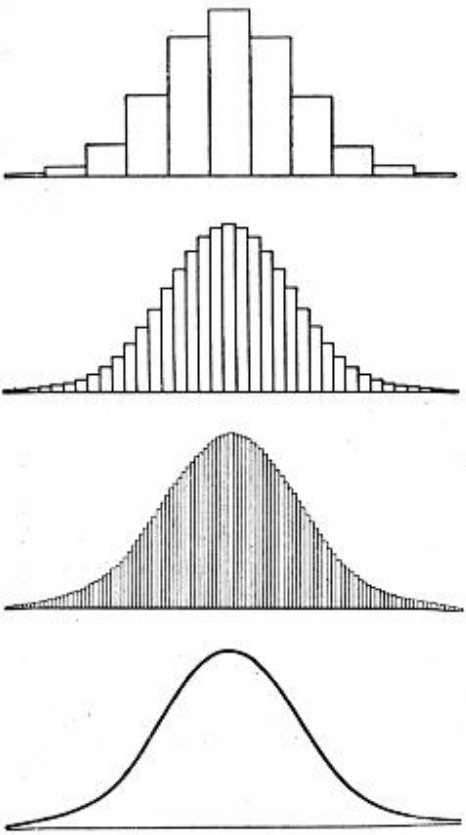


Рис. 31. Получение из гистограммы кривой нормального распределения (по А. К. Митропольскому)

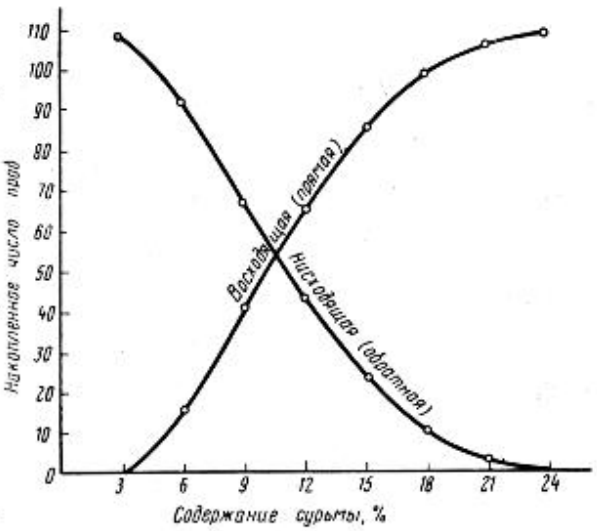


Рис. 32. Кумуляты (восходящая и нисходящая) содержания сурьмы в пробах

Кумулята является интегральным графиком, тогда как гистограмма и полигон — дифференциальные графики.

Кумуляты иногда используют для определения бортового минимума (Прокофьев, 1950) и для нахождения критического содержания в ураганных (выдающихся) пробах, учитываемого при подсчете запасов.

Таблица 33

Показатели	Содержание сурьмы, %							Всего проб
	3—6	6—9	9—12	12—15	15—18	18—21	21—24	
Число проб	16	25	24	20	13	7	3	108
Восходящая кумуляция	16	41	65	85	98	105	108	—
Нисходящая кумуляция	108	92	67	43	23	10	3	—

Пример. По геологически однородному участку золотой россыпи получены следующие результаты опробования:

Содержание золота, *г/м*	Число проб
0	13
Следы *	32
0,1—1,0	51
1,0—2,0	84
2,0—3,0	116
3,0—5,0	96
5,0—10,0	32
10,0—20,0	13
18,7	1
18,8	1
19,4	2
19,9	1
20,0	1
29,5	1
47,1	1
92,0	1
193,4	1
Всего	447

* Под «следами» понимается содержание, не превышающее предела чувствительности анализа (в данном случае не превышающее 0,1 *г/м*).

Необходимо определить критическое содержание золота в пробе, полностью учитываемое при подсчете запасов по этому участку (избыточное содержание будет отнесено к запасам всего месторождения). Есть много методов решения проблемы ураганных проб. Ураганной считается проба, показавшая содержание полезного компонента выше критического. Критическим называют содержание в пробе, в определенное число раз превышающее среднее содержание полезного компонента. Пусть в нашем случае это число будет равно 10. Для того чтобы показать, как «богатые» пробы влияют на среднее содержание, и вычислить критическое содержание, сделаем расчеты в табл. 34.

Таблица 34

№ п/п	Содержание золота, *г/м*	Среднее по классу, *г/м*	Число проб в классе	Сумма содержаний в классе	Нарастающая сумма	Нарастающая сумма %	Убывающая сумма	Убывающая сумма, %
1	0	0	13	0	0	0	1742,9	100,0
2	0—0,1	0,05	32	1,6	1,6	0,1	1742,9	100,0
3	0,1—1,0	0,55	51	28,1	29,7	1,7	1741,3	99,9
4	1—2	1,5	84	126,0	155,7	8,8	1713,2	98,3
5	2—3	2,5	116	290,0	445,7	25,6	1587,2	91,2
6	3—5	4,0	96	384,0	829,7	47,6	1297,2	74,4
7	5—10	7,5	32	240,0	1069,7	61,5	913,2	52,4
8	10—20	15,0	13	195,0	1264,7	72,5	673,2	38,5
9	18,7	18,7	1	18,7	1283,4	73,7	478,2	27,5
10	18,8	18,8	1	18,8	1302,2	74,6	459,5	26,3
11	19,4	19,4	2	38,8	1341,0	77,0	440,7	25,4
12	19,9	19,9	1	19,9	1360,9	78,1	401,9	23,0
13	20,0	20,0	1	20,0	1380,9	79,3	382,0	21,9
14	29,5	29,5	1	29,5	1410,4	81,0	362,0	20,7
15	47,1	47,1	1	47,1	1457,5	84,6	332,5	19,0
16	92,0	92,0	1	92,0	1549,5	88,9	285,4	15,4
17	193,4	193,4	1	193,4	1742,9	100,0	193,4	11,1
		3,90	447					

Из данных таблицы видно, что среднее содержание по всем 447 пробам равно 3,90 *г/т*. Критическим, следовательно, будет содержание 39,0 *г/т*. У нас имеется три пробы с содержанием, превышающим критическое, т. е. три ураганные пробы с содержанием 47,1; 92,0 и 193,4 *г/т*. На данных

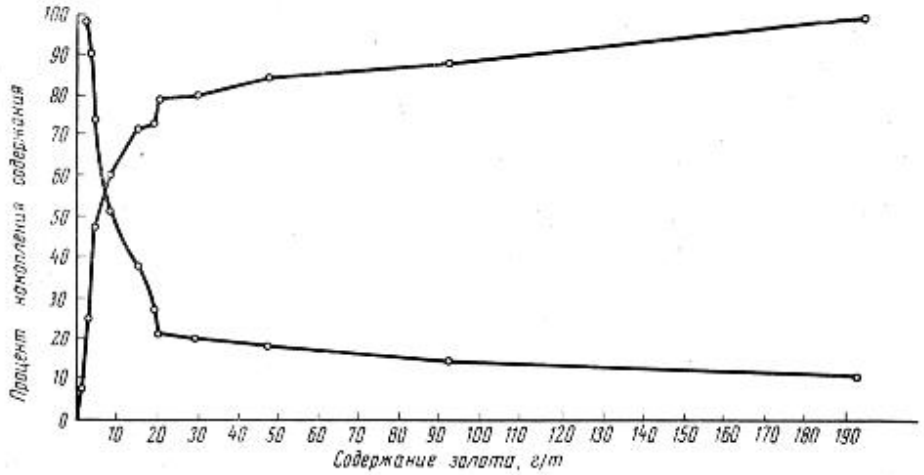


Рис. 33. Кумулята накопления содержания золота

этих трех проб основываются 19% всех запасов (при подсчете их методом среднего арифметического). Избыточным будем содержание 8,1; 53,0; 154,4 *г/т*, а в сумме — 215,5 *г/т*, или 12,4% от всех запасов.

Если же в соответствии с другим методом подсчета запасов мы условимся, что ураганные пробы должны составлять не более определенного процента запасов, например не более 20%, то по графику мы берем содержание, отвечающее 20%-ному накоплению нисходящей кумуляции, или 80%-ному накоплению восходящей кумуляции. Это содержание и будет критическим.

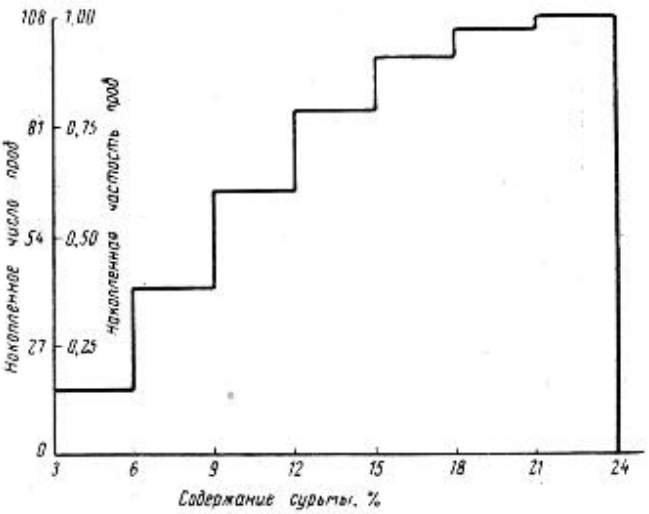


Рис. 34. Кумулятивная гистограмма накоплений содержания сурьмы

Все эти расчеты можно сделать как аналитическим, так и графическим путем (рис. 33).

Кумуляты нередко используются для решения вопроса о недостающих фракциях золота в пробах (Разумовский, 1939), для решения проблемы самородков (Шарапов, 1947), для отыскания бортового минимума (Прокофьев, 1950), для характеристики запасов (Косыгин, 1960) и т. д.

Подобно гистограмме, на графике можно отразить накопление частот или частостей, в результате чего получится график, который условно можно назвать кумулятивной гистограммой. Пример кумулятивной гистограммы показан на рис. 34.

Кумулятивные гистограммы составляются как для группированных, так и для негруппированных рядов.

Если при изображении кумуляты переставить оси (ось ординаты поставить на место оси абсцисс, а ось абсцисс — на место оси ординаты), то получится так называемая огива (иногда ее называют огивой Гальтона). Пример огивы показан на рис. 35.

При графическом изображении статистических данных масштабы по координатным осям следует выбирать таким образом, чтобы чертеж имел форму прямоугольника, высота которого относится к ширине, как 5 : 8. Однако это условие необязательное. Оно преследует цель наглядности изображения. Если строится много однотипных графиков, в отдельных случаях это требование невозможно выполнить. Для удобства их сравнения друг с другом масштаб всех графиков должен быть одинаковым (по соответствующим осям).

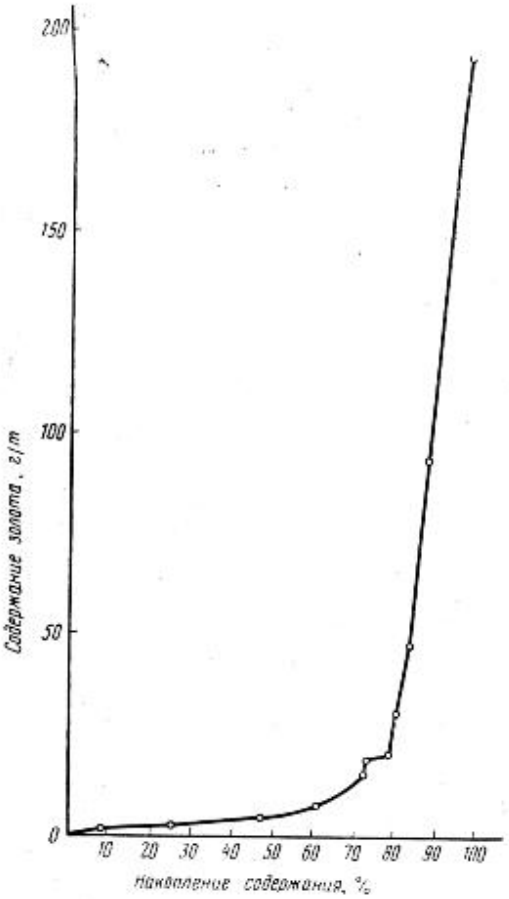


Рис. 35. Огива содержания золота в пробах

В эмпирических рядах всегда бывают случайные колебания значений признаков. Для того чтобы за этими случайными колебаниями увидеть закономерные изменения случайной величины, в практике статистических исследований пользуются приемом сглаживания ряда. Сущность этого приема покажем на следующем примере.

Пример. На разведочной линии, пересекающей россыпь, пройдено 23 шурфа, заданных через 10 *м*. Все шурфы добиты до плотика. Содержание золота на массу (россыпь будет отрабатываться драгой, поэтому содержание золота вычислено на массу) показано в табл. 35.

Изменения содержаний золота (x) по ширине россыпи, по данным опробования, показаны на рис. 36. По нему видно, что россыпь как бы делится на две струи. Для того чтобы закономерности строения распределения золота в россыпи были более рельефными, необходимо сгладить эмпирические данные, т. е. заменить реальные данные вычисленными данными. Для этого берем содержания золота по первым трем шурфам (1, 2, 3) и вычисляем по ним групповое среднее содержание. Оно оказывается равным

$$\frac{140 + 200 + 230}{3} = 190 \ мг/м^3.$$

Записываем содержание в средней из этих трех строк. Затем берем содержания по следующим трем шурфам (2, 3 и 4). Среднее содержание по ним 270 *мг/м*³. Записываем эту величину в средней строке для этой тройки. После этого берем содержания по шурфам 3, 4 и 5. Среднее для них 340 *мг/м*. Таким же способом вычисляем сглаженное содержание и для всех других строк, в результате чего получаем ряд сглаженных показателей x_1 (1-е сглаживание).

Таблица 35

№ шурфов	Содержание золота, мг/м³ (x)	Сглаженное содержание, мг/м³		
		после 1-го сглаживания, (x_1)	после 2-го сглаживания, (x_2)	после 3-го сглаживания, (x_3)
1	140			
2	200	190	(200)	(203)
3	230	270	267	(275)
4	380	340	357	352
5	400	460	433	432
6	600	500	507	492
7	500	560	537	538
8	590	550	570	560
9	560	600	573	569
10	650	570	563	540
11	510	520	483	483
12	400	360	403	409
13	170	330	341	369
14	420	333	364	368
15	410	430	390	399
16	460	433	435	420
17	430	443	425	422
18	440	400	405	397
19	330	373	361	361
20	350	310	318	314
21	250	270	262	(264)
22	210	207	(212)	(211)
23	160			

Если полученная кривая еще слишком извилиста, сглаживаем уже сглаженные показатели, в результате чего получаем ряд дважды сглаженных показателей x_2.

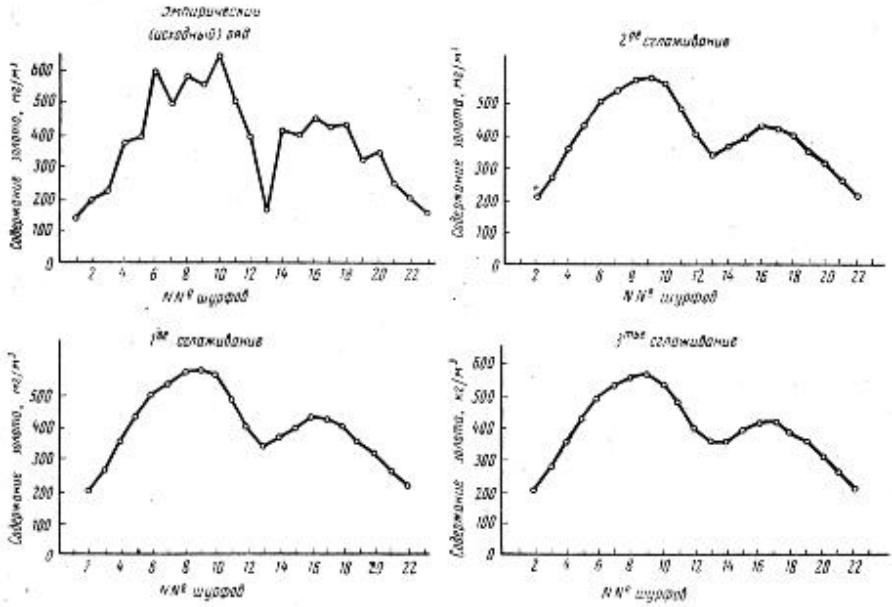


Рис. 36. Сглаживание эмпирических рядов

Таким же способом можем получить и трижды сглаженный ряд x_3. Он будет более плавным, чем первые ряды.

В начале и в конце рядов x_2 и x_3 имеются числа в скобках. Они вычислены с привлечением показателей из ряда x (т. е. показателей 140 и 160 мг/м³).

Сглаживание эмпирического ряда — операция чисто практическая. П. Л. Каллистов (1952) использовал ее для выявления эмпирической закономерности пространственного распределения золота в россыпи и для решения проблемы ураганных проб.

Более глубокое исследование эмпирических распределений проводится с помощью так называемого выравнивания.

Под выравниванием в практической статистике понимают подгонку эмпирического ряда, изображенного на полигоне, под теоретическую функцию, для которой имеется и формула, и соответствующая ей кривая.

Техника подбора кривой такова: эмпирические данные наносят в виде точек на кальку (по оси абсцисс — значение признака, по оси ординат — частоту или частость), которую затем накладывают то на одну, то на другую эталонную кривую и останавливаются на той кривой, которая меньше других отходит от эмпирических точек. Возможен и другой путь — эталонные кривые, вычерченные на кальке, прикладывают к эмпирическому графику.

Наиболее общим способом нахождения эмпирической или теоретической формулы распределения могло бы быть параболическое выравнивание по методу наименьших квадратов. Это выравнивание может быть весьма точным, если взять параболу достаточно высокой степени

$$I = a_0 + a_1 x + a_2 x^2 + \cdots + a_\nu x^\nu$$

и правильно найти параметры a_0, a_1, a_2, . . ., a_ν. Но объем вычислительной работы при этом будет настолько большим, что он окажется под силу только вычислительным машинам. Если же взять параболу невысокой степени, то и точность выравнивания может оказаться невысокой. Поэтому параболическое выравнивание применяется редко. Чаще поступают наоборот: сначала задаются законом распределения, потом по выборке оценивают его параметры. После этого строят теоретическое распределение и проверяют степень его близости к эмпирическому распределению.

V. ПРОВЕРКА ГИПОТЕЗ О ЗАКОНЕ РАСПРЕДЕЛЕНИЯ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

В процессе статистических исследований нередко возникают вопросы о близости эмпирических распределений к теоретическим, о принадлежности двух или большего числа выборок к одной общей совокупности, о допустимости тех или иных отклонений от среднего и др. Эти вопросы связываются с природой изучаемого признака и формулируются как статистические гипотезы, которые требуют проверки имеющимися эмпирическими данными.

Приведем примеры таких гипотез:

1. Собрав раковины спириферид в девоне Северо-Востока европейской части СССР, Д. В. Наливкин (1925) сформулировал гипотезу о принадлежности их к разным видам.

2. При изучении минерального состава гранитоидов Алтая, Д. А. Родионов (1961) проверил гипотезу о согласованности распределения содержаний пирита в этих породах с логарифмически нормальным законом.

3. Произведя некоторое число замеров мощности сильвинитового пласта в Соликамске, И. П. Шарапов сформулировал гипотезу о том, что статистическое распределение мощности не противоречит нормальному закону.

4. Изучив ошибки анализа геологических проб, сделанного в разные годы, мы проверяем гипотезу о принадлежности ошибок к одной общей статистической совокупности, т. е. об их независимости от времени анализа.

Статистическими гипотезами являются также предположения, что неизвестное среднее значение признака больше или равно заданной величине, что средние двух совокупностей равны, что неизвестные дисперсии двух совокупностей можно рассматривать как равные и т. д.

С каждой из этих гипотез связаны важные для практических геологических исследований выводы.

Проверка гипотез осуществляется с помощью статистических критериев, которые иногда называют критериями согласия.

Статистический критерий для проверки некоторой гипотезы H представляет собой случайную величину, для которой известен закон распределения при условии, что гипотеза H верна.

В «Комментариях редактора перевода» книги К. А. Браунли (1949) А. Н. Колмогоров формулирует понятие критерия следующим образом. «Для проверки гипотезы H измеряются величины: $\xi_1, \xi_2, \ldots, \xi_n$.

Из значений этих величин образуется комбинация

$$K = f(\xi_1, \xi_2, \ldots, \xi_n),$$

которая при гипотезе H является случайной величиной с вполне определенным законом распределения. Эта величина называется «критерием» для проверки гипотез (t-, F-, χ^2-критерием). Для каждой вероятности p устанавливаются такие границы K'_p и K''_p, что при гипотезе H

$$\text{вероятность} \quad \{K'_p < K < K''_p\} = 1 - p,$$

$$\text{вероятность} \quad \{K \leqslant K'_p \text{ или } K \geqslant K''_p\} = p.$$

Если следовать правилу — в каждом отдельном случае отбрасывать гипотезу H, когда наблюденное значение K не укладывается в пределы $K'_p < K < K''_p$, и считать гипотезу согласующейся с результатами наблюдений, когда K укладывается в эти пределы, то при массовом применении правила мы будем отбрасывать гипотезу H ошибочно в доле случаев, равной p.

Вероятность p, которую кладут в основу избранного правила проверки, называется уровнем значимости данного правила».

Ошибка, заключающаяся в ложном отклонении проверяемой гипотезы H, когда она в действительности верна, называется ошибкой первого рода. Таким образом, уровень значимости критерия является вероятностью появления ошибки первого рода.

Допустим, что гипотеза H, которую мы проверяем, неверна, а в действительности справедлива альтернатива H_1. Если в подобной ситуации в результате применения критерия будет принята ложная гипотеза H, а правильная гипотеза H_1 отклонена, то возникает так называемая ошибка второго рода.

Обозначим вероятность появления этой ошибки через β. Величина $1 - \beta$ выражает вероятность непоявления ошибки второго рода и называется мощностью критерия относительно альтернативной гипотезы H_1.

Вполне естественно, что исследователь заинтересован в выборе таких критериев, которые делают вероятности появления обеих ошибок достаточно малыми.

Статистика располагает очень большим числом критериев согласия. Одни из них применимы для проверки гипотез в условиях непрерывных распределений, другие — дискретных.

В геологических исследованиях нередко приходится проверять гипотезу о близости эмпирического распределения к теоретическому нормальному закону. Эта гипотеза может быть проверена разными способами, в том числе и очень простыми.

Есть два признака нормального распределения. Первый состоит в том, что третий центральный момент нормально распределенной случайной величины равен нулю, т. е. $\nu_3 = 0$.

Второй признак заключается в том, что четвертый центральный момент этой же случайной величины равен утроенному квадрату второго центрального момента, т. е. $\nu_4 = 3\nu_2^2$.

В связи с этими признаками необходимо охарактеризовать так называемые постоянные Пирсона (β_1 и β_2): первая определяется по формуле

$$\beta_1 = \frac{\nu_3^2}{\nu_2^3}.$$

Величина β_1 связана с асимметрией K равенством

$$K = \sqrt{\beta_1} = \frac{\nu_3}{\sqrt{\nu_2^3}}.$$

Обозначим оценку для ν_3 через $\hat{\nu}_3$, оценку для ν_2 — через s^2 (оценка дисперсии), а для величины K — через K^*. Тогда

$$K^* = \frac{\hat{\nu}_3}{s^3}.$$

Среднее квадратическое отклонение случайной величины K^* приблизительно равно

$$\sigma_{K^*} = \sqrt{\frac{6}{n}},$$

где n — число наблюдений в выборке.

Как уже говорилось, положительная асимметрия указывает на то, что кривая более крута в ее левой ветви, а отрицательная асимметрия свидетельствует о большой крутости правой ветви.

Проверяемую гипотезу о непротиворечивости нормального распределения выборочным данным можно сформулировать как требование одновременного выполнения следующих двух равенств:

$$\nu_3 = 0, \quad \text{или} \quad K = 0;$$

$$\nu_4 = 3\nu_2^2, \quad \text{или} \quad \frac{\nu_4}{\nu_2^2} - 3 = 0.$$

Если гипотеза $K = 0$ верна, то случайная величина $\frac{K^*}{\sigma_{K^*}}$ будет распределена асимптотически нормально с математическим ожиданием, равным нулю, и дисперсией, равной единице. Поэтому гипотеза $K = 0$ отвергается, если $\left|\frac{K^*}{\sigma_{K^*}}\right| > 3$.

Вторая постоянная Пирсона определяется по формуле

$$\beta_2 = \frac{\nu_4}{\nu_2^2}.$$

Обозначим через E величину $\frac{\nu_4}{\nu_2^2} - 3$, называемую эксцессом распределения.

Пусть ее статистическая оценка будет E^*, т. е.

$$E^* = \frac{\hat{\nu}_4}{s^4} - 3.$$

Среднее квадратическое отклонение величины E^* определяется по приближенной формуле

$$\sigma_{E^*} = \sqrt{\frac{24}{n}}.$$

Случайная величина $\frac{E^*}{\sigma_{E^*}}$ в случае, если верна гипотеза $E = 0$, распределена асимптотически нормально с математическим ожиданием, равным нулю, и дисперсией, равной единице. Поэтому гипотеза $E = 0$ отвергается, если $\left|\frac{E^*}{\sigma_{E^*}}\right| > 3$.

Таким образом, гипотеза о нормальном распределении может быть принята как непротиворечащая эмпирическим данным, если будут выполнены два неравенства:

$$\left|\frac{K^*}{\sigma_{K^*}}\right| < 3$$

и

$$\left|\frac{E^*}{\sigma_{E^*}}\right| < 3.$$

Если же хотя бы одно из этих неравенств не выполняется, то гипотезу о нормальном распределении следует отвергнуть.

Ввиду сложности вычисления оценки асимметрии и эксцесса иногда применяют другие, более простые, но менее надежные способы для характеристики особенностей формы кривой распределения.

Пирсон предложил следующую простую оценку коэффициента асимметрии:

$$\alpha = \frac{\bar{x} - Mo}{\sigma},$$

где α — коэффициент асимметрии Пирсона,

$\bar{x}$ — средняя арифметическая,

Mo — мода,

σ — среднее квадратическое отклонение.

Поскольку получение оценки моды затруднительно, последнюю формулу упрощают еще более, введя в нее оценку медианы вместо моды. Таким образом, имеем

$$\alpha = \frac{3(\bar{x} - Me)}{\sigma}.$$

Среднее квадратическое отклонение случайной величины α равно

$$\sigma_\alpha = \sqrt{\frac{3}{2n}}.$$

Гипотеза $M_\alpha = 0$ отвергается, если $\left|\frac{\alpha}{\sigma_\alpha}\right| > 3$, что основано на асимптотической нормальности распределения случайной величины $\frac{\alpha}{\sigma_\alpha}$.

Есть еще один способ измерения асимметрии, тоже довольно простой. Его предложил Линдберг. Заключается он в определении числа S, которое находят по следующей формуле:

$$S = 100\frac{r}{n} - 50,$$

где S — число, характеризующее асимметрию,

r — число тех значений в выборке, которые больше средних,

n — объем выборки.

Среднее квадратическое отклонение величины S равно

$$\sigma_S = \frac{30}{\sqrt{n}}.$$

Для измерения эксцесса Линдберг предлагает число e, определяемое по формуле

$$e = 100\frac{S}{n} - 38{,}29,$$

где S — число тех членов выборки, значения которых заключены в интервале от $\bar{x} - \frac{1}{2}\sigma$ до $\bar{x} + \frac{1}{2}\sigma$.

Среднее квадратическое отклонение величины e равно

$$\sigma_e = \frac{42}{\sqrt{n}}.$$

Как и в предыдущих случаях, гипотеза о нормальном распределении может быть проверена путем рассмотрения отношений $\left|\frac{S}{\sigma_S}\right|$ и $\left|\frac{e}{\sigma_e}\right|$.

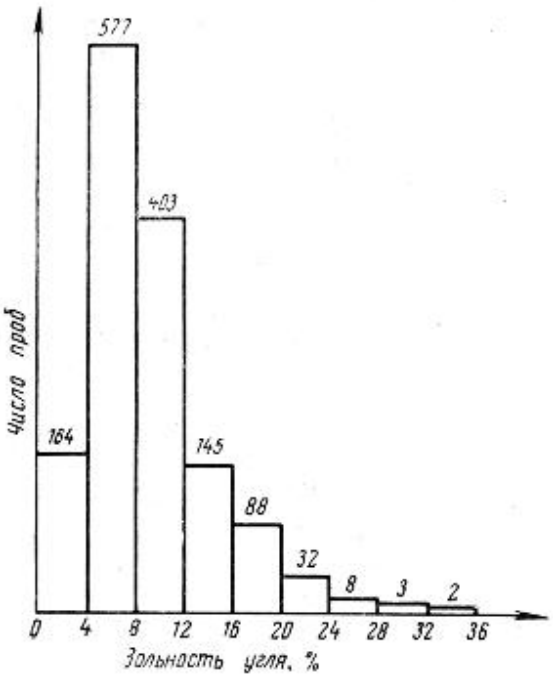


Рис. 37. Распределение зольности в каменном угле одного из месторождений Донбасса

Таблица 36

Зольность, %	Число проб, n	Середина интервала x, %
0—4	164	2
4—8	577	6
8—12	403	10
12—16	145	14
16—20	88	18
20—24	32	22
24—28	8	26
28—32	3	30
32—36	2	34
	1422	

Проверяемая гипотеза должна быть отвергнута, если хотя бы одна из этих величин больше 3.

Технику вычислений, связанных с проверкой гипотезы о нормальном законе и с определением критериев согласия, покажем на практическом примере.

Пример. Распределение зольности в каменном угле одного из месторождений Донбасса приведено в табл. 36 и на рис. 37.

Используя методы, приведенные в главе III, найдем оценки $\bar{x}$ и s^2 средней зольности и дисперсии:

$$\bar{x} = 8{,}79,$$
$$s^2 = 24{,}3.$$

Оценка стандартного отклонения $s = \sqrt{24{,}3} = 4{,}93$. Найдем также оценки асимметрии K^* и эксцесса E^* и соответствующие им стандартные отклонения σ_{K^*} и σ_{E^*}.

$$K^* = 1{,}19,$$
$$E^* = 2{,}03,$$
$$\sigma_{K^*} = 0{,}0649$$
$$\sigma_{E^*} = 0{,}13.$$

Для проверки гипотез $K = 0$ и $E = 0$, равносильных гипотезе о нормальном распределении зольности, нужно вычислить отношения $\left|\frac{K^*}{\sigma_{K^*}}\right|$ и $\left|\frac{E^*}{\sigma_{E^*}}\right|$:

$$\left|\frac{K^*}{\sigma_{K^*}}\right| = 18{,}36;$$

$$\left|\frac{E^*}{\sigma_{E^*}}\right| = 15{,}6.$$

Так как обе эти величины значительно превышают 3, то проверяемая гипотеза должна быть отвергнута.

Пример. Приведем еще один пример, связанный с проверкой гипотезы о нормальном законе. В табл. 37 показано распределение содержаний

Таблица 37

Содержание брома, сотые доли %	Середина интервала x	Число проб n
0—1	0,5	4
1—2	1,5	44
2—3	2,5	112
3—4	3,5	223
4—5	4,5	325
5—6	5,5	275
6—7	6,5	106
7—8	7,5	17
8—9	8,5	4
Всего		1110

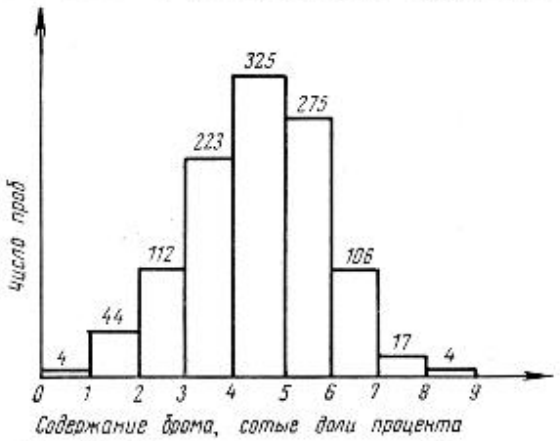


Рис. 38. Распределение содержаний брома в калийных солях

брома сильвинитового месторождения по данным химического анализа 1110 проб.

Данные таблицы графически изображены на рис. 38.

По данным, приведенным в этой таблице, вычислены:

$$\bar{x} = 4{,}363;$$

$$s^2 = 1{,}844;$$

$$s = 1{,}357;$$

$$K^* = -0{,}195;$$

$$E^* = 17{,}1;$$

$$\sigma_{K^*} = 0{,}0735;$$

$$\sigma_{E^*} = 0{,}147;$$

$$\frac{K^*}{\sigma_{K^*}} = 2{,}65;$$

$$\frac{E^*}{\sigma_{E^*}} = 116{,}0.$$

Так как одно из отношений $\frac{E^*}{\sigma_{E^*}}$ значительно превышает 3, то гипотезу о нормальном распределении содержаний брома в сильвинитах следует забраковать как неподтвердившуюся.

Пример. В табл. 38 и на рис. 39 показано распределение содержаний железа по данным анализа 275 проб железной руды.

По этим данным получены следующие характеристики:

$$\bar{x} = 46 - 0{,}035 = 45{,}965 \simeq 46;$$

$$s^2 = 31{,}86;$$

$$s = \sqrt{31{,}86} = 5{,}64;$$

$$K^* = -0{,}0766;$$

$$\sigma_{K^*} = 0{,}1526;$$

$$\frac{K^*}{\sigma_{K^*}} = -0{,}5036.$$

Так как $\left|\frac{K^*}{\sigma_{K^*}}\right| < 3$, то распределение можно рассматривать как симметричное.

$$E^* = -0{,}24;$$

$$\sigma_{E^*} = 0{,}306.$$

Отношение $\frac{E^*}{\sigma_{E^*}} = 0{,}77$.

Так как это отношение меньше трех, то гипотеза $E = 0$ не отвергается.

Общий вывод: в качестве модели распределения содержаний железа можно принять нормальный закон.

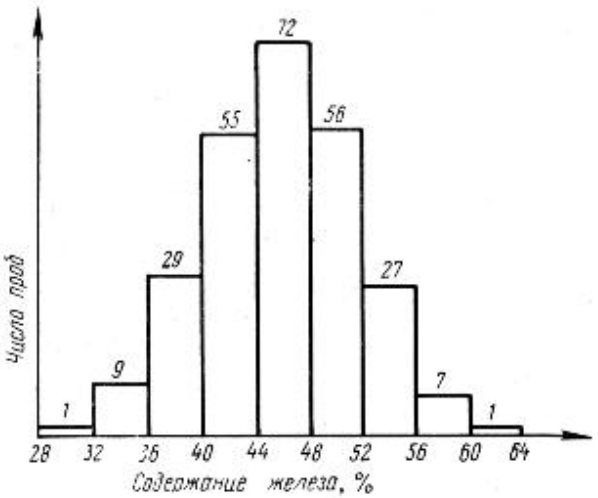


Рис. 39. Распределение содержаний железа

Гипотеза о согласованности эмпирического распределения с теоретическим может быть проверена с помощью так называемого критерия Пирсона χ^2, который представляет собой следующую сумму:

Таблица 38

Содержание железа, %	Середина интервала x	Число проб n
28—32	30	1
32—36	34	9
36—40	38	29
40—44	42	55
44—48	46	72
48—52	50	56
52—56	54	27
56—60	58	7
60—64	62	1
		257

$$\chi^2_H = \sum_{i=1}^{m} \frac{(n_i - n_i')^2}{n_i'},$$

где i — номер интервала группировки данных;

n_i — число наблюдений, попавших в i-тый интервал;

n_i' — теоретическое число наблюдений для i-того интервала, вычисленное исходя из распределения, согласованность с которым проверяется;

m — число интервалов.

Если гипотетическое распределение не противоречит эмпирическим данным, то случайная величина χ^2, значение которой χ^2_H мы наблюдаем,

будет распределена по закону Пирсона с $m - l$ степенями свободы, где l — число наложенных связей. Это распределение табулировано (приложение 14). Если χ^2_H окажется больше, чем допустимое значение $\chi^2_{q,\ m-l}$ при заданном уровне значимости q и $m - l$ степенях свободы, то проверяемая гипотеза отвергается.

Пример. Рассмотрим применение этого критерия на примере проверки гипотезы о нормальном распределении по данным предыдущего примера (см. табл. 38). Прежде всего нужно вычислить теоретические частоты n'_i при условии, что гипотеза о нормальном распределении верна и полученные оценки $\bar{x}$ и s^2 совпадают с истинными значениями параметров.

Первым шагом в вычислении теоретических частот является отыскание (по каждому интервалу) значения t случайной величины τ, которая, если гипотеза верна, распределена нормально со средним 0 и дисперсией, равной 1. Она определяется по формуле

$$t_i = \frac{x_i - \bar{x}}{s}.$$

Величину t_i записываем в отдельную графу (табл. 39). По этой величине в специальной таблице (приложение 3) находим соответствующее значение Z_{t_i} и записываем в следующую графу.

Теоретическое количество проб n'_i, какое у нас было бы, если гипотеза о нормальном распределении верна, вычисляется по формуле

$$n'_i = \frac{n\,\Delta_i}{s} Z_t,$$

где n — общее количество проб, в нашем примере равное 257,

Δ_i — длина интервала, в нашем случае равная 4,

Z_t и s — в прежнем значении.

Так как длина интервала Δ_i постоянна для всех i, то выражение $\frac{n\Delta_i}{s}$ равно 182,6.

Величину n'_i для каждого интервала записываем в следующей графе табл. 39.

Таблица 39

x_i	n_i	$x_i - \bar{x}$	t_i	Z_{t_i}	n'_i	Округленные значения n'_i
30	1	—16	—2,84	0,00707	1,3	1
34	9	—12	—2,13	0,04128	7,6	8
38	29	—8	—1,42	0,14556	26,6	27
42	55	—4	—0,71	0,31006	56,6	56
46	72	0	0	0,39894	72,8	73
50	56	4	0,71	0,31006	56,6	56
54	27	8	1,42	0,14556	26,6	27
58	7	12	2,13	0,04128	7,6	8
62	1	16	2,84	0,00707	1,3	1
	257				257,0	257

Таким образом, теоретические частоты получены (рис. 40).

Разница между фактической и теоретической частотами по каждому интервалу в отдельности видна из простого сопоставления соответствующих граф таблицы. Однако без применения соответствующего статистического критерия нельзя делать вывода о том, принимать эти расхождения как существенные или же пренебречь ими как случайными.

Вычисленные критерии согласия Пирсона для этого примера приведены в табл. 40.

Таким образом, для данного примера $\chi^2 = 0{,}43$. Число степеней свободы $k = 9 - 3 = 6$, так как в данном случае наложены три связи:

1) $\sum_{i=1}^{m} \frac{n_i}{n} = 1$,

2) $a = \bar{x}$,

3) $\sigma^2 = s^2$.

Так как допустимое значение χ^2 при уровне значимости 0,05 и 6 степенях свободы равно 12,59 (приложение 14), что значительно превышает

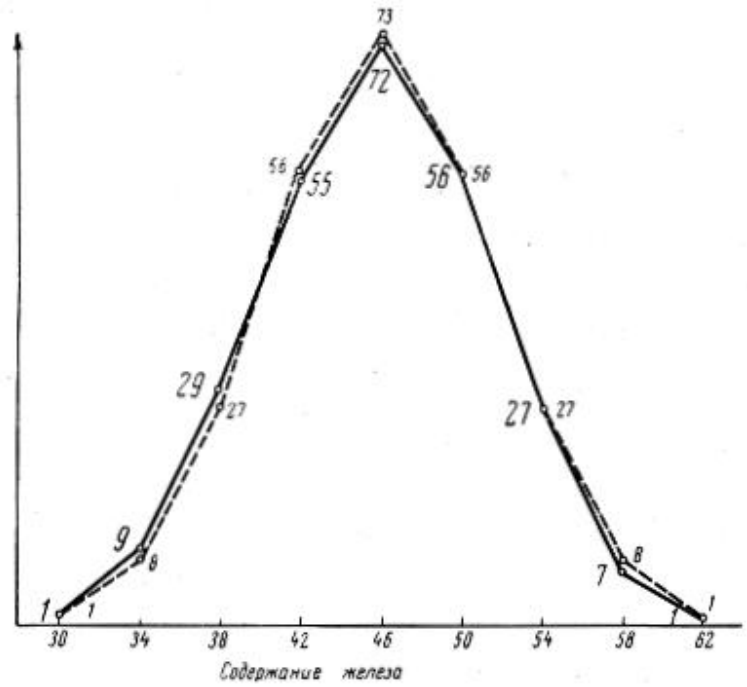


Рис. 40. Фактическое (сплошная линия) и теоретическое (пунктир) число проб с соответствующим содержанием железа в руде

вычисленное значение 0,43, то проверяемая гипотеза может быть принята как подтвердившаяся.

Если под руками нет соответствующих таблиц, то оценить величину χ^2 можно по методу В. И. Романовского, который предложил счи-

Таблица 40

x	n	n'	$n - n'$	$(n - n')^2$	$\frac{(n - n')^2}{n'}$
30	10	9	1	1	0,125
34					
38	29	27	2	4	0,148
42	55	56	—1	1	0,018
46	72	79	—1	1	0,014
50	56	56	0	0	0
54	27	27	0	0	0
58	8	9	—1	1	0,125
62					
	257	257			0,430

тать расхождение между теоретическим и фактическим распределением неслучайным (существенным), если

$$\frac{\chi^2 - k}{\sqrt{2k}} \geqslant 3,$$

где k — число степеней свободы.

В нашем примере

$$\frac{\chi^2 - k}{\sqrt{2k}} = \frac{0{,}430 - 6}{\sqrt{2 \cdot 6}} = \frac{5{,}570}{3{,}4} = 1{,}64.$$

Так как полученное значение меньше 3, то расхождение между теоретическим и эмпирическим распределением следует считать несущественным.

В связи с тем, что критерий χ^2 является широко используемым в статистических приложениях, полезно рассмотреть вопрос о возникновении распределения χ^2.

Пусть $\eta_1, \eta_2, \ldots, \eta_K$ — независимые нормально распределенные случайные величины, имеющие математические ожидания $a_1, \ldots, a_K$ и дисперсии $\sigma_1^2, \sigma_2^2, \ldots, \sigma_K^2$ соответственно. Образуем новую последовательность случайных величин $\xi_1, \xi_2, \ldots, \xi_K$, где $\xi_i = \frac{\eta_i - a_i}{\sigma_i}$, которые также будут независимы, но одинаково нормально распределены с математическим ожиданием, равным нулю, и дисперсией, равной единице.

Величина

$$\chi^2 = \sum_{i=1}^{k} \xi_i^2,$$

представляющая собой сумму квадратов k независимых одинаково нормально распределенных случайных величин, имеющих математическое ожидание, равное нулю, и дисперсию, равную 1, распределена как

$$P_k(x) = \frac{x^{\frac{k}{2}-1} e^{-\frac{x^2}{2}}}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)},$$

где $P_k(x)$ — плотность вероятности,

$\Gamma\left(\frac{k}{2}\right)$ — гамма-функция с параметром $\frac{k}{2}$,

k — число степеней свободы.

Плотность вероятностей $P_k(x)$ не имеет максимума при $k \leqslant 2$, а при $k > 2$ достигает максимума в точке $x = k - 2$.

При $k > 30$ кривая $P_k(x)$ становится очень близкой к нормальной кривой, и в этих условиях можно считать, что случайная величина χ^2 распределена асимптотически нормально с математическим ожиданием, равным k, и дисперсией $2k$. Тогда величина

$$\frac{\chi^2 - k}{\sqrt{2k}}$$

будет распределена асимптотически нормально с математическим ожиданием, равным нулю, и дисперсией, равной единице. Это свойство может быть использовано при проверке нулевой гипотезы, когда $k > 30$, или же когда нет таблиц распределения χ^2 при $k < 30$.

Кроме критериев согласия можно пользоваться и более простым, хотя и менее точным методом определения степени близости теоретического нормального распределения к эмпирическому — числами Вестергарда.

Вот эти числа: 0,3; 0,7; 1,1; и 3,0. Распределение считается нормальным или близким к нормальному, если:

1) в промежутке от $\bar{x}$—0,3σ до $\bar{x} + 0{,}3\sigma$ находится четвертая часть всей совокупности;

2) в промежутке от $\bar{x} - 0{,}7\sigma$ до $\bar{x} + 0{,}7\sigma$ находится половина всей совокупности;

3) в промежутке от $\bar{x} - 1{,}1\sigma$ до $\bar{x} + 1{,}1\sigma$ находится три четверти всей совокупности;

4) в промежутке от $\bar{x} - 3\sigma$ до $\bar{x} + 3\sigma$ находится 0,998 всей совокупности.

Заменив величину σ ее оценкой s в данном примере получим:

$\bar{x} = 45{,}965;$
$s = 5{,}640;$
$0{,}3s = 1{,}692;$
$0{,}7s = 3{,}948;$
$1{,}1s = 6{,}204;$
$3s = 16{,}920.$

В промежутке: от 44,273 до 47,657 имеем 25% n;
» » от 41,747 до 49,643 имеем 50% n;
» » от 39,491 до 51,899 имеем 75% n;
» » от 29,045 до 62,885 имеем 100% n.

Выше были рассмотрены аналитические методы определения степени близости эмпирического ряда к теоретическому (нормальному), но есть еще и графические методы.

Один из них предложен Анри (Henry). Сущность этого метода заключается в следующем.

Пусть ξ — случайная величина, распределенная нормально с параметрами a и σ^2. $F_\xi(x)$ — функция ее распределения. Образуем новую случайную величину

$$\tau = \frac{\xi - a}{\sigma},$$

которая также распределена нормально с математическим ожиданием, равным нулю, и дисперсией, равной 1. Нетрудно заметить, что зависимость

$$t = \frac{x - a}{\sigma}$$

представляет линейное уравнение, и график этой функции, построенный для всех значений x, выразится в виде прямой.

Пусть $x_1, x_2, \ldots, x_n$ — выборочные значения случайной величины ξ. Обозначим через $F(x_i)$ эмпирическую функцию распределения в точке x_i. Эту функцию нетрудно построить, расположив x_i в возрастающем порядке, т. е. $x_1 < x_2 < \cdots < x_n$.

Тогда

$$F(x_i) = \frac{i - 1}{n}.$$

Обозначим через t_i функцию, обратную $F(x_i)$ (таблицу значений этой функции можно найти в работе Ван дер Вардена, 1960). Каждому значению x_i будет соответствовать одно значение t_i, взятое из таблицы. Если гипотеза о нормальном распределении не противоречит эмпирическим данным, то значения t_i и x_i должны быть связаны линейной зависимостью, т. е. точки, нанесенные на график, должны располагаться близко к прямой линии (ось абсцисс — x_i, ось ординат — t_i).

Если же точки, нанесенные на график, нельзя рассматривать как близкие к прямой, то из этого следует, что гипотеза о нормальном распределении должна быть отвергнута.

Графический метод проверки этой гипотезы наиболее удобен, когда в распоряжении исследователя имеется так называемая нормальная вероятностная бумага (рис. 41). На этой бумаге по оси ординат откладываются значения x_i. Величины обратной функции t_i нанесены на ней по оси ординат, но вместо этих значений проставлены величины функции $F(x_i)$.

Таким образом, достаточно вычислить для всех x_i значения $F(x_i)$ и нанести их на бумагу. Если точки будут располагаться близко к прямой линии, то гипотезу о нормальном распределении можно принять как подтвердившуюся.

Следует отметить, что гипотеза о согласованности эмпирического распределения с логарифмическим нормальным законом, равносильна гипотезе о нормальном распределении логарифмов изучаемой случайной величины. Поэтому при проверке гипотезы о логнормальном распределении нужно значения признака заменить их логарифмами, и по новым данным проверить гипотезу о нормальном распределении каким-либо статистическим критерием. В случае применения графического метода для проверки этой гипотезы удобно пользоваться так называемой логнормальной вероятностной бумагой (рис. 42).

Пример. Приведем пример графической проверки гипотезы о нормальном распределении.

В Кизеловском угольном бассейне маркшейдером была замерена мощность угольного пласта № 5 в 370 точках. Результаты замера приведены в табл. 41.

Таблица 41

Мощность, *см*	Число точек замера	Накопление числа точек	Накопление числа точек, %
40—50	2	2	0,5
50—60	5	7	1,9
60—70	16	23	6,2
70—80	21	44	11,9
80—90	71	115	31,1
90—100	86	201	54,4
100—110	74	275	74,4
110—120	33	308	83,3
120—130	32	340	92,0
130—140	19	359	97,1
140—150	9	368	99,5
150—160	2	370	100,0

По этим данным составлена гистограмма (рис. 43).

На вероятностной бумаге (рис. 44) нанесены точки, отвечающие проценту накопления числа точек. Они оказались примерно на прямой линии, на основании чего можно сделать вывод о близости этого эмпирического распределения к нормальному. Точками удобно отмечать и индивидуальные (негруппированные) замеры. Группированные замеры лучше показывать в виде ступени.

Пример. Приведем пример проверки гипотезы о логнормальном распределении.

Распределение содержаний пирита в алтайских гранитах (Родионов, 1961) по 58 пробам приведено в табл. 42.

Здесь x_i — границы интервала группировки, $\frac{1}{2}\lg x_{i_1} + \frac{h}{2}$ — середина логарифмического интервала, n_i — число наблюдений в i-том интервале.

Содержание пирита x_i дается в граммах на тонну.

Среднее арифметическое логарифмов содержаний пирита

$$\overline{\lg x} = 0{,}296.$$

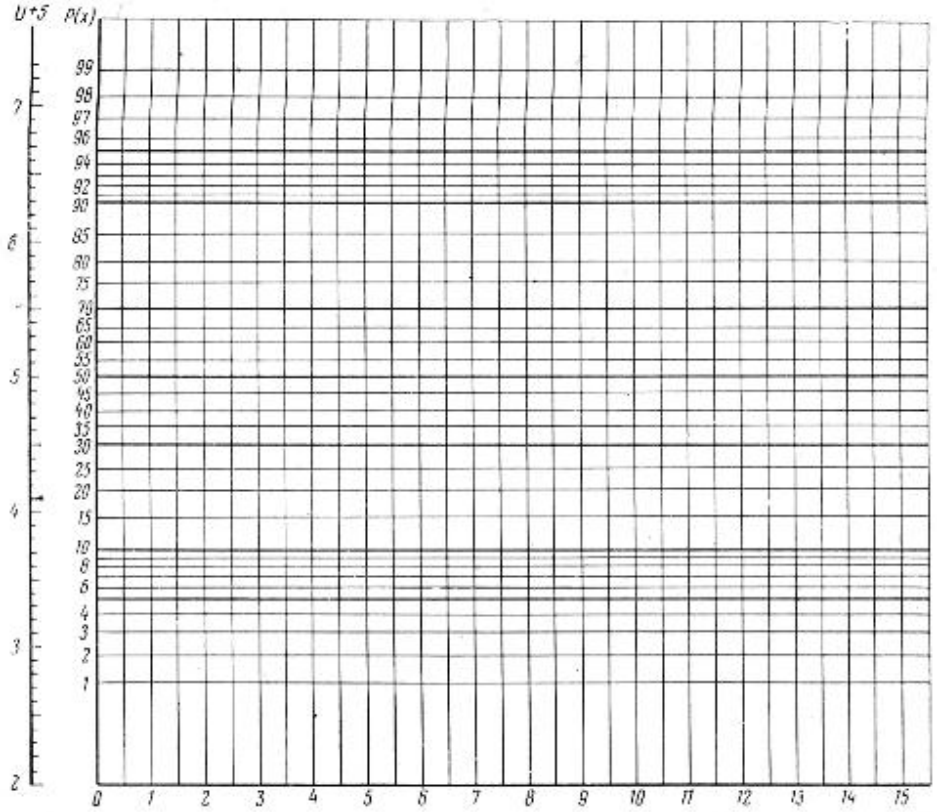


Рис. 41. Нормальная вероятностная бумага

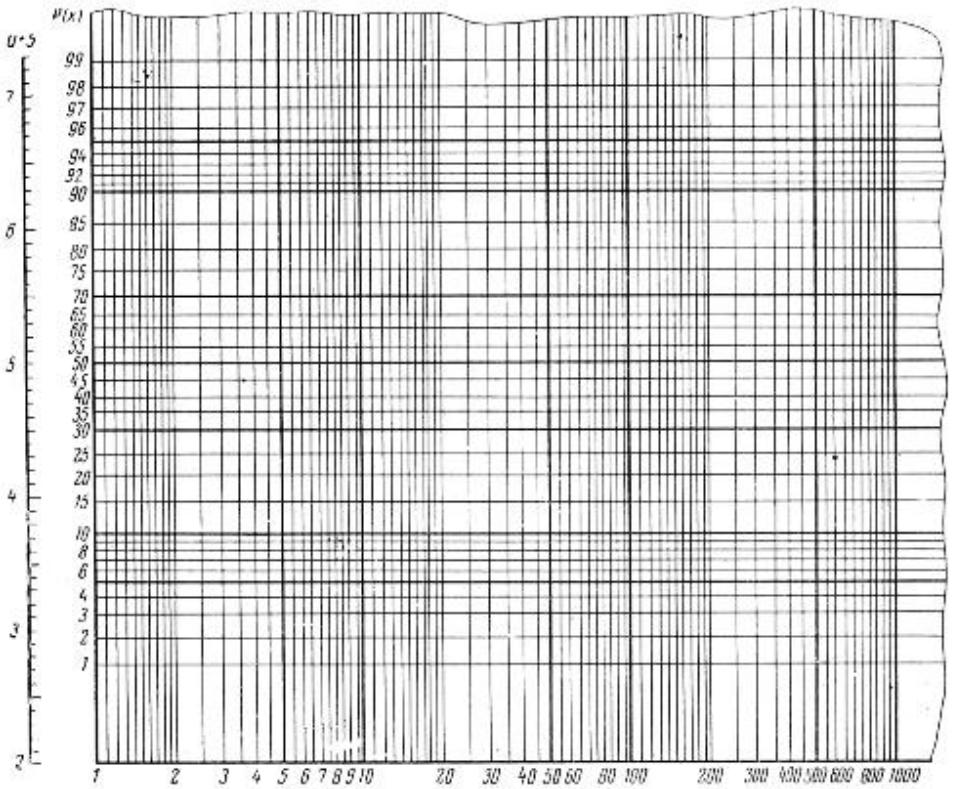


Рис. 42. Логнормальная вероятностная бумага

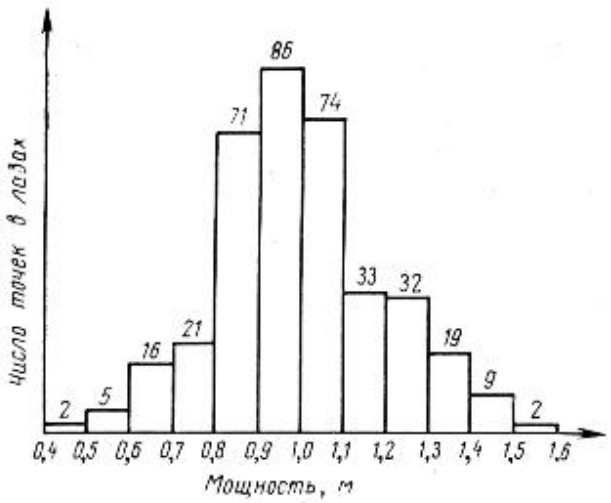


Рис. 43. Распределение мощности пласта № 5 по 370 замерам в шахтах им. 40-летия Октября и № 48

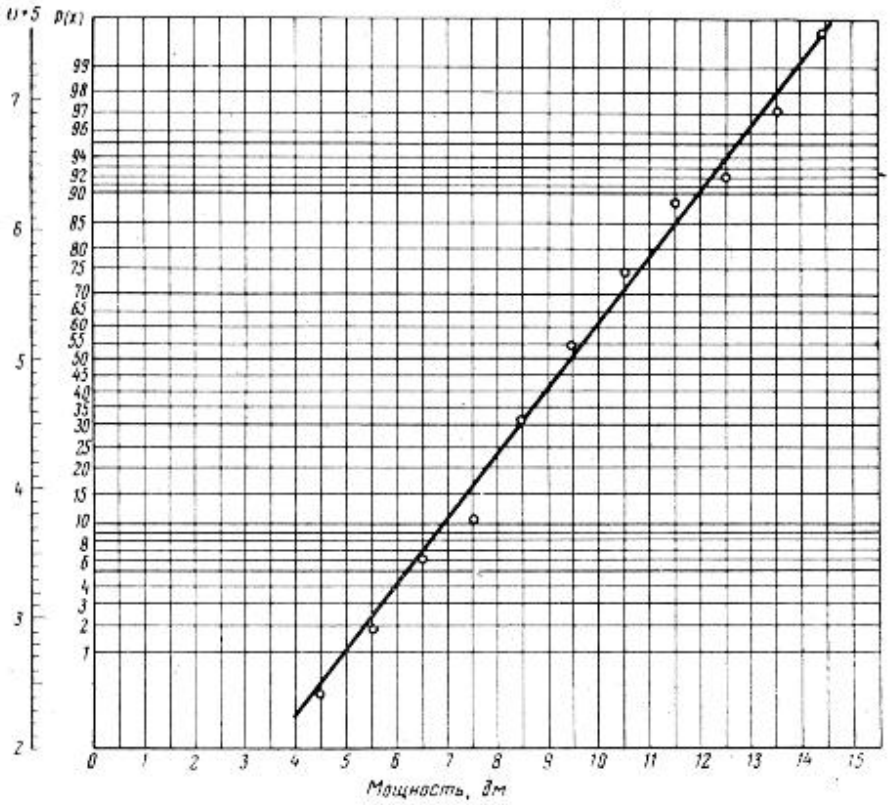


Рис. 44. Распределение мощности пласта № 5

Таблица 42

x_i	$\lg x_i$	$\lg x_i + \frac{h}{2}$	n_i
0,05	—1,301	—1,150	3
0,10	—1,000	—0,761	10
0,30	—0,523	—0,261	15
1,00	0,000	0,500	18
10,00	1,000	1,500	8
100,00	2,000	2,500	2
1000,00	3,000	2,349	1
5000,00	3,699	3,850	1
10000,00	4,000		
			58

Оценка дисперсии логарифмов содержаний:

$$s = 1{,}408.$$

Оценка асимметрии распределения логарифмов (S_K), оценка эксцесса этого распределения (E) и их средние квадратические отклонения (σ_{s_k}, σ_E):

$$S_K = -0{,}288;$$
$$E = -0{,}565;$$
$$\sigma_{s_k} = 0{,}322;$$
$$\sigma_E = 0{,}642.$$

Отношения оценок асимметрии и эксцесса к соответствующим стандартным отклонениям:

$$\frac{S_K}{\sigma_{s_k}} = 0{,}894;$$
$$\frac{E}{\sigma_E} = 0{,}877.$$

Так как эти отношения менее трех, можно сделать вывод о соответствии эмпирического распределения теоретическому (логнормальному).

Пример. Приведем пример графической проверки гипотезы о логнормальном распределении.

Мощность одного из пластов Кизеловского угольного бассейна по маркшейдерским замерам, сделанным в отрабатываемых лавах Коспашского района, имеет следующее эмпирическое распределение (Шарапов, 1964) (табл. 43).

Таблица 43

Мощность, м	Число замеров	Накопление	Накопление, %
0—0,3	8	8	0,9
0,3—0,6	44	52	5,9
0,6—0,9	168	220	24,8
0,9—1,2	317	537	47,5
1,2—1,5	173	710	80,2
1,5—1,8	84	794	89,6
1,8—2,1	61	855	96,5
2,1—2,4	18	873	98,8
2,4—2,7	8	881	99,4
2,7—3,0	3	884	99,8
3,0—3,3	2	886	100,0
	886		

По данным этой таблицы составлены два графика. На одном из них (см. рис. 15) мощность показана в равномерном, а на другом (см. рис. 16) — в логарифмическом масштабе. Если бы мы разбили общее число точек замера (886) на классы по равным интервалам логарифма (например, через каждые 0,1), то получили бы симметричное распределение, близкое к нормальному.

Если данные табл. 43 нанести на вероятностную логнормальную бумагу, то получим график распределения мощности (рис. 45). Близость

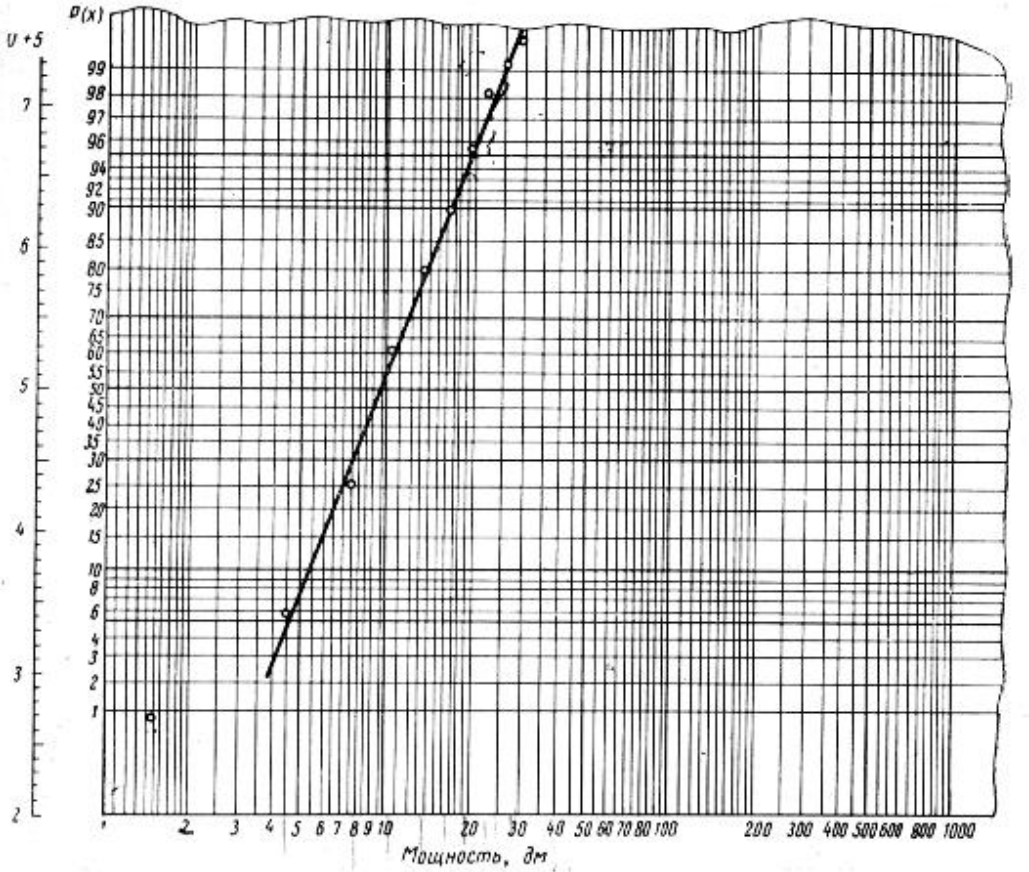


Рис. 45. Распределение мощности пласта № 11 на логнормальной бумаге

точек к прямой на этом графике свидетельствует о близости эмпирического распределения к логнормальному.

Мы показали, что правоасимметричное распределение иногда не противоречит логнормальному закону. Здесь же покажем, что и левоасимметричное распределение иногда может быть трансформировано в нормальное с помощью того же приема логарифмирования.

Отрицательно асимметричные эмпирические распределения иногда хорошо согласуются с так называемым зеркально отраженным логнормальным распределением. Если ξ — отрицательно асимметрично распределенная случайная величина, A — константа ($\xi \leqslant A$), то

$$P(\xi \leqslant x) = 1 - \Lambda(A - x;\ \mu, \sigma^2) = \Lambda\left(\frac{1}{A - x};\quad -\mu,\ \sigma^2\right),$$

где Λ — логнормальная функция с аргументом и параметрами, указанными в скобках.

Естественно, что гипотезе о согласованности распределения с зеркально отраженным логнормальным равносильны гипотезы

о согласованности с нормальным законом распределений случайных величин:

$$\ln \eta = \ln (A - \xi)$$

или

$$\ln \zeta = \ln \left(\frac{1}{A - \xi} \right).$$

Напомним, что при правоасимметричном распределении логарифмируется величина признака x_i. В случае же левоасимметричного распределения логарифмировать нужно дополнение до константы $A > \max x_i$, т. е. взять логарифмы разностей $(A - x_i)$ или $\ln \left(\frac{1}{A - x_i} \right)$.

По Д. А. Родионову (1963), проверка гипотезы о непротиворечивости выборочного распределения случайной величины ξ и функции $1 - \Lambda (A - x, \mu, \sigma^2)$ осуществляется при проверке равносильной гипотезы о нормальном распределении случайной величины $\ln \eta = \ln (A - \xi)$ по выборочным значениям $\ln y_1$, $\ln y_2$, . . ., $\ln y_n$, где $\ln y_i = \ln (A - x_i)$.

Пример (Родионов, 1963). По 32 образцам дунита Полярного Урала определено следующее содержание MgO (табл. 44).

Таблица 44

№ образца	Содержание MgO, % (x_i)	N	x_i	N	x_i	N	x_i
1	43,36	9	44,33	17	24,87	25	41,66
2	43,16	10	44,62	18	36,10	26	41,79
3	49,61	11	44,61	19	41,43	27	42,83
4	40,53	12	44,61	20	45,60	28	36,64
5	43,57	13	36,67	21	36,96	29	42,07
6	43,75	14	38,93	22	43,65	30	37,39
7	39,92	15	11,97	23	39,55	31	39,52
8	43,15	16	30,14	24	39,01	32	31,42

Таблица 45

m	n_m	Pm	n'_m	$n_m - n'_m$	$(n_m - n'_m)^2$	$\frac{(n_m - n'_m)^2}{n'_m}$
0	24	0,04979	24	0	0	0
1	75	0,14936	76	1	1	0,013
2	110	0,22404	109	1	1	0,009
3	113	0,22404	109	4	16	0,147
4	84	0,16803	86	2	4	0,047
5	50	0,10082	49	1	1	0,020
6	23	0,05041	25	2	4	0,160
7	10	0,02160	11	1	1	0,091
8	4	0,00810	4	0	0	0
9	1	0,00270	1			
	494		494			0,487

Распределение, показанное в этой таблице, имеет отрицательную асимметрию (—2,083) и положительный эксцесс (5,206). Выборочные значения (x_i) содержаний MgO были заменены величинами $y_i = 60 - x_i$, для которых и была проверена гипотеза о логнормальном распределении. В результате проверки установлено, что величины y_i распределены логнормально, так как отношения

$$\left| \frac{K^*}{\sigma_{K^*}} \right| = 1{,}898$$

и

$$\left|\frac{E^*}{\sigma_{E^*}}\right| = 1{,}532$$

в обоих случаях меньше трех.

Рассмотрим теперь использование критериев согласия при определении степени близости эмпирического распределения к распределению Пуассона. В качестве такого критерия используем критерий Пирсона.

Пример. В табл. 45 приводится расчет этого критерия по результатам определения содержаний свинца в медной руде одного уральского месторождения, выраженных в условных единицах (рис. 46).

Таким образом, $\chi^2_{II} = 0{,}487$, число степеней свободы $k = 10 - 2 = 8$. Поэтому

$$\frac{|\chi^2 - k|}{\sqrt{2k}} = \frac{|0{,}487 - 8|}{\sqrt{16}} = \frac{7{,}513}{4} = 1{,}878.$$

Так как полученное значение меньше трех, то можно сделать вывод о несущественности расхождения между эмпирическим и теоретическим распределениями. Следовательно, эмпирическое распределение не противоречит закону Пуассона.

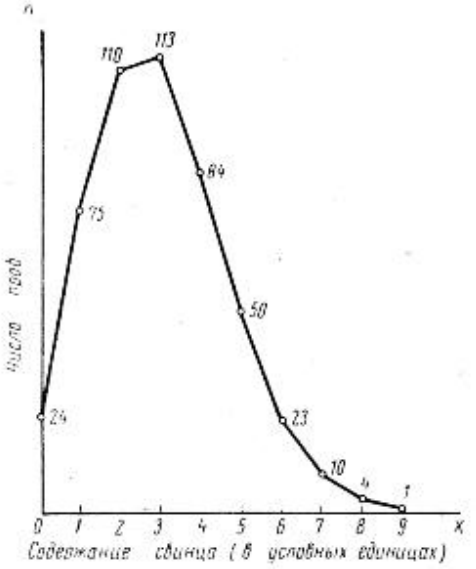


Рис. 46. Распределение содержаний свинца в пробах одного уральского меднорудного месторождения

VI. ВЫБОРКА И ПРОВЕРКА ГИПОТЕЗ О СРЕДНЕМ

Выше уже говорилось о том, что разведка месторождений полезных ископаемых представляет собой с точки зрения математической статистики не что иное, как взятие выборки из месторождения, которое рассматривается как генеральная совокупность. При этом объем выборки очень мал по сравнению с объемом генеральной совокупности.

Выборочный метод исследования, как замечает В. Романовский, единственно возможен, когда генеральная совокупность практически бесконечна или, по крайней мере, очень велика, а изучаемый признак имеет стохастический характер. Главная цель выборочного метода — найти важнейшие характеристики выбранной совокупности и перенести их на генеральную. В этом отношении выборочный метод — это умение по малому судить о многом, что является главным в статистическом исследовании вообще.

Удачно сделанная выборка (под выборкой понимают как процесс отбора, так и совокупность отобранных элементов) называется представительной или репрезентативной, если в такой выборке распределение и среднее значение изучаемого признака настолько близко к его распределению и среднему в генеральной совокупности, что этим расхождением можно пренебречь. Практически выборку считают представительной, если интересующие нас характеристики выборочной и генеральной совокупности отличаются друг от друга не более чем на допустимую, заранее заданную величину.

В основе теории выборки лежит представление о том, что распределение величины изучаемого признака в генеральной совокупности обусловливается некоторым законом распределения, и наша задача состоит в том, чтобы этот закон найти, а затем, используя вытекающие из него следствия,

вычислить основные характеристики изучаемого распределения. В связи с этим имеют место понятия о выборочном и теоретическом распределениях.

Выборочное распределение — это распределение в фактически полученной совокупности. Теоретическое распределение — распределение, предполагаемое в неизвестной генеральной совокупности, не противоречащее выборочным данным.

Исходя из предположенного теоретического распределения, мы можем найти характеристики неизвестной совокупности и сделать различные прогнозы.

Выборка является способом получения информации о неизвестной генеральной совокупности. Всякая выборка ограниченна. Информация, даваемая ею, никогда не может быть исчерпывающей.

Выборки бывают повторные и бесповторные. Первая производится по схеме возвращенного шара (шар, вынутый из урны, после его регистрации возвращается снова в урну и повторно участвует в игре, т. е. он еще раз может оказаться вынутым). Бесповторная выборка производится по схеме невозвращенного шара, при которой один и тот же член генеральной совокупности не может более чем один раз попасть в выборку. Будучи раз взятым, этот член «выбывает из игры». Если объем совокупности достаточно велик по сравнению с выборкой, то бесповторная выборка дает почти такие же результаты, что и повторная.

При повторной выборке вероятность, или доля членов генеральной совокупности, обладающих интересующим нас признаком, не изменяется на протяжении всего процесса отбора. При бесповторной выборке эта вероятность, или доля, непрерывно, от испытания к испытанию, изменяется. При этом объем генеральной совокупности все время сокращается, т. е. тоже изменяется.

Процесс отбора элементов выборки необходимо производить наудачу. Одним из признаков, по которым производится отбор, может служить номер членов совокупности, если на каждого члена генеральной совокупности (когда это возможно) завести карточки и пронумеровать их, не взирая на содержание записей в них. Карточки тщательно тасуются. Затем одна из них извлекается наудачу, ее номер записывается, после чего карточка возвращается в общую пачку или картотеку. После этого наудачу снова извлекается карточка и т. д. Исследованию по окончании процесса отбора могут подвергаться либо сами карточки, если в них записано все необходимое, либо те объекты (пронумерованные), представителями которых являются вынутые карточки.

Если обследуются не записи в карточках, а сами объекты, образовавшие выборочную совокупность, то на карточках никаких записей можно не делать; карточки следует лишь пронумеровать. В конечном итоге обследуются лишь объекты, имеющие те же номера, что и вынутые карточки. В этом случае можно обойтись и совсем без карточек. Делается это так: все объекты генеральной совокупности нумеруются, затем по таблице случайных чисел (приложение 1) берем на любой, случайно открытой странице столько номеров (подряд), сколько нам надо для целей выборки. Объекты с этими номерами считаются вошедшими в выборку. Если количество случайных чисел в таблице больше числа членов генеральной совокупности, то все номера, превышающие объем генеральной совокупности, можно просто пропустить. При этом один и тот же объект может дважды или большее число раз попасть в выборку.

Техника производства бесповторной выборки также должна обеспечивать одинаковую возможность каждому из еще невыбранных членов совокупности попасть в выборку.

Выборку можно производить различными способами. Кроме выше описанных способов, различают типическую, серийную (гнездовую), механическую и комбинированную выборки. Все они так или иначе при-

меняются в практике геологоразведочных работ. Все выборки, каким бы способом они ни делались, случайны, но характер этой случайности и способ ее использования разные.

Выборки собственно случайная, типическая и серийная могут быть как бесповторными, так и повторными. Механическая выборка — только бесповторная.

Типическая выборка производится в следующем порядке. Генеральная совокупность разделяется на несколько типических в отношении исследуемого признака групп, после чего из каждой группы делается выборка наудачу. Разбивку на группы нужно производить так, чтобы в выборке были представлены все типы изучаемых объектов. При типической выборке, например при отборе дубликатов для контрольного анализа, выделяют руды окисленные и неокисленные, сульфидные и карбонатные, богатые и бедные; разделение производится по участкам, по методам опробования полезного ископаемого и т. д. Правда, для такого разделения нужно знать многое о генеральной совокупности, в данном случае о месторождении. Без такого знания разделение на типы невозможно, а потому невозможна и типическая выборка. Таким образом, типические выборки тоже случайные, только их делается несколько из различных (частных) совокупностей.

При производстве типической выборки нужно иметь в виду, что выделяемые типические группы (типы, частные совокупности) почти всегда будут неодинакового объема. Поэтому возникает вопрос; как отбирать объекты из групповой совокупности? Тут возможны три метода отбора:

а) пропорционально объему группы;

б) непропорционально объему группы;

в) приблизительно пропорционально степени изменчивости признака в группе, т. е. в зависимости от величины дисперсии (чем больше дисперсия размера признака, тем больше объем выборки).

Серийная (гнездовая) выборка* в своей начальной стадии напоминает типическую: генеральная совокупность разбивается, по крайней мере мысленно, на ряд групп, причем разбивается по признаку, лишь косвенно связанному с изучаемым признаком. Далее серийная выборка отличается от типической. Если в последней из каждой группы отбираются объекты для выборочной совокупности, то в первой такой отбор совсем отсутствует. Вместо отбора отдельных членов генеральной совокупности в серийной выборке производится отбор целых групп для их сплошного обследования. Зато другие группы (неотобранные) остаются совершенно не затронутыми процессом отбора.

Пример серийной выборки можно привести из практики опробования. Для того чтобы проконтролировать метод взятия проб, иногда подвергают вторичному (контрольному) опробованию целый штрек или целый горизонт или три-четыре участка из общего, довольно значительного числа штреков, горизонтов, участков. Если по всему месторождению было взято, например, 15 000 проб, то в серийную выборку попадает 500—1000 проб.

Серийная выборка желательна там, где добываемое минеральное сырье смешивается в участковые потоки.

Механическая выборка осуществляется тоже по группам, но последние выделяются обязательно по признакам, не имеющим никакой связи с исследуемой совокупностью. Для осуществления такого способа отбора генеральную совокупность разбивают на группы чисто механически, например по порядковым номерам или по квадратной сетке, или в шахматном порядке; затем из каждой группы берется также меха-

---

* Термин гнездовая выборка автор использует не в том смысле, как его употребляют в статистике, где гнездовые, или иерархические, выборки применяются для изучения различных уровней изменчивости. (Прим. *Ред.*).

нически каждый первый или каждый второй или вообще каждый n-й член и включается в выборку. Отличие механической выборки от собственно случайной состоит в том, что в последней нет разбивки на группы. Поэтому при механической выборке выбранные объекты расположены более равномерно, чем при собственно случайной.

Механическая выборка иногда используется при сдаче на анализ оптического сырья: например, из каждого пятого ящика берется каждый четвертый кристалл. Несколько напоминает такую же выборку валовое опробование добытой руды, когда в пробу идет каждая десятая вагонетка; а из таких вагонеток берется по правилу конверта (по углам вагонетки и из середины) пять ковшей руды. Анализ пробы из каждой вагонетки делается отдельно, а иногда и в смеси с пробами из всех вагонеток сменной добычи. Это опробование похоже на серийную выборку, но отобранные объекты расположены более равномерно.

Разведка месторождения буровыми скважинами или шурфами, расположенными по треугольной, прямоугольной или квадратной сетке, внешне напоминает механическую выборку. И тут и там месторождение разбивается на n участков или ячеек, одинаковых по размерам, форме и ориентировке. Естественно, возникает вопрос, как в пределах такой ячейки наметить точку для заложения выработки. При механической выборке эта точка берется случайно. Она может оказаться в любом месте ячейки, причем положение точки в ячейке не зависит от положения такой же точки в другой ячейке. Механическая выборка, таким образом, является совокупностью n собственно случайных выборок, каждая из которых дает одну точку (один объект).

Иначе обстоит дело с разведочной сеткой. Положение точки в ней строго зафиксировано, т. е. зависит от положения всех других точек и каждой из них. Если сетка служит для геологического опробования, то рудный материал, взятый по всем ячейкам сетки, по существу является не пробами, а порциями одной сложной пробы.

Разведочная, или опробовательская сетка может быть ориентирована бесконечно большим числом способов. Так же бесконечно велико число положений начальной точки. Эта точка, а в связи с нею и все другие точки, может задаваться в любом месте ячейки. Поэтому совокупность возможных значений признака, определяемых по всей сетке для каждого из ее положений, беспредельно велика. Геолог же имеет дело только с одним положением сетки, т. е. единственным членом, случайно выбранным из бесконечной совокупности.

Статистическому исследованию подвергается всегда лишь выборочная совокупность. При этом чаще всего решаются две следующие задачи: 1) нахождение закона распределения изучаемой совокупности, 2) оценка параметров распределения.

Кроме этих задач ставятся и другие (проверка гипотез о связи признаков, планирование наблюдений и т. д.), но на них остановимся позже, а здесь рассмотрим лишь первые две.

Первой задачей статистического исследования выборочной совокупности является нахождение закона распределения, случайной величины в этой совокупности. Н. В. Смирнов и И. В. Дунин-Барковский (1959) пишут по этому поводу следующее: «. . .выяснение или оценка закона распределения по данным выборки (так называемая параметризация) и составляет существенную проблему математической статистики: только овладев законами распределения изучаемых величин, мы можем решать возникающие на практике задачи по анализу, сравнению и предсказанию результатов массового процесса».

Закон распределения мы можем иногда вывести из теоретических соображений. Так, например, если изучаемая случайная величина является суммой многих независимых равномерно малых слагаемых, то следует ожидать нормальное распределение суммы.

Очень важно и чисто феноменологическое исследование массовых явлений. Из массы фактического материала будут проступать контуры искомого закона распределения.

Так или иначе мы почти всегда получаем возможность предположить, что в интересующей нас совокупности действует тот или иной закон.

Второй задачей статистического исследования является вычисление параметров распределения. Их мы получаем, исходя из статистик изучаемой совокупности. Для этого мы вычисляем оценки параметров распределения по выборочной совокупности, исходя из заданной теоретической модели распределения. Оценка приемлема в том случае, если она удовлетворяет требованиям состоятельности, несмещенности и эффективности.

Состоятельной оценкой параметра ϑ называется статистика $\bar{\vartheta}$, сходящаяся по вероятности к ϑ при $n \to \infty$ (Крамер, 1948).

Несмещенная оценка — это «. . .выборочная характеристика, обладающая тем свойством, что ее математическое ожидание при любом объеме выборки равно оцениваемому параметру» (Смирнов и Дунин-Барковский, 1959).

Пусть $\bar{\vartheta}_1$ и $\bar{\vartheta}_2$ — две различные оценки неизвестного параметра ϑ, полученные по одной и той же выборке. Обозначим через $D(\bar{\vartheta}_1)$ и $D(\bar{\vartheta}_2)$ их дисперсии. Если $D(\bar{\vartheta}_1) < D(\bar{\vartheta}_2)$, то оценка $\bar{\vartheta}_1$ более эффективна, чем $\bar{\vartheta}_2$, так как она обеспечивает меньшую случайную погрешность.

Оценка, имеющая наименьшую дисперсию, называется эффективной оценкой. Эффективными являются так называемые максимально правдоподобные оценки, рассматриваемые ниже. Эффективностью оценки параметра считается отношение дисперсии эффективной оценки параметра к дисперсии рассматриваемой оценки.

Если случайная величина ξ распределена нормально с параметрами $M\xi$ и σ^2, то среднее арифметическое и выборочная медиана, по данным n наблюдений, распределены асимптотически нормально со средним, равным $M\xi$, и с дисперсиями, равными соответственно $\frac{\sigma^2}{n}$ и $\frac{\sigma^2\pi}{2n}$. Сравнивая эти дисперсии, можно отметить, что дисперсия среднего арифметического составляет 0,637 от дисперсии выборочной медианы. Иначе говоря, дисперсия среднего в выборке объема $n_1 = 637$ равна дисперсии медианы в выборке объема $n_2 = 1000$. Таким образом, в условиях нормального распределения среднее арифметическое является более эффективной оценкой среднего, чем выборочная медиана.

В связи с тем, что максимально правдоподобные оценки являются эффективными, коротко рассмотрим метод максимального правдоподобия, предложенный Р. Фишером.

Обозначим случайную величину через ξ, и пусть $f_\xi(x,\theta)$ — плотность ее распределения, зависящая от параметра θ. Обозначим через x_1, x_2, . . ., x_n выборочные значения величины ξ в выборке объема n. Эти значения можно рассматривать как n значений независимых, одинаково распределенных случайных величин ξ_1, ξ_2, . . ., ξ_i, . . ., ξ_n, где над каждой величиной ξ_i проведено по одному наблюдению. Пусть $f(x_1, \theta)$, $f(x_2, \theta)$, . . ., $f(x_n, \theta)$ — плотности вероятностей, соответствующие набору случайных величин ξ_1, ξ_2, . . ., ξ_n. Рассмотрим совместную плотность распределения этих случайных величин. В силу независимости она будет представлена произведением соответствующих одномерных плотностей, т. е.

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

Так как выборочные значения $x_1, x_2, \ldots, x_n$ заданы, мы будем рассматривать это произведение как функцию неизвестного параметра θ, которую обозначим $L_X(\theta)$, где $X = \{x_1, x_2, \ldots, x_n\}$, т. е.

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

Эта функция называется функцией правдоподобия.

Представляется естественным выбрать из всех возможных значений параметра θ то, при котором функция $L_X(\theta)$ достигает максимума. Для этого нужно решить уравнение

$$\frac{dL_X(\theta)}{d\theta} = 0$$

относительно θ. Полученное таким образом значение $\bar{\theta}$, являющееся функцией выборочных значений $x_1, x_2, \ldots, x_n$, будет представлять максимально правдоподобную оценку параметра θ. Вместо приведенного выше уравнения можно воспользоваться иногда более простым уравнением

$$\frac{d \ln L_X(\theta)}{d\theta} = 0.$$

В качестве примера рассмотрим процедуру определения максимально правдоподобных оценок в условиях нормального распределения.

Пусть ξ — случайная величина, которая распределена нормально с параметрами a и σ^2, т. е. плотность вероятности ее распределения дается выражением

$$f_\xi(x;\ a,\ \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}.$$

Пусть также над случайной величиной ξ проведено n наблюдений, результаты которых представлены $x_1, x_2, \ldots, x_n$. Построим функцию правдоподобия:

$$L_X(a,\ \sigma^2) = \prod_{i=1}^{n} f(x_i;\ a,\ \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-a)^2} =$$

$$= (\sigma^2)^{-\frac{n}{2}} (2\pi)^{-\frac{n}{2}} exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-a)^2\right\}.$$

Найдем логарифм этого выражения:

$$\ln L_X(a,\ \sigma^2) = -\frac{n}{2}\ln\sigma^2 - \frac{n}{2}\ln 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-a)^2.$$

Для определения максимально правдоподобной оценки параметра a нужно найти частную производную

$$\frac{\partial \ln L_X(a,\ \sigma^2)}{\partial a},$$

приравнять ее к нулю и решить уравнение относительно a:

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

Этому выражению равносильно (при $\sigma^2 > 0$)

$$\sum_{i=1}^{n}(x_i-a) = 0.$$

Откуда найдем

$$\sum_{i=1}^{n} x_i = na,$$

$$\bar{a} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Эту величину чаще обозначают через $\bar{x}$. Таким образом, мы нашли, что в условиях нормального распределения среднее арифметическое является максимально правдоподобной оценкой математического ожидания случайной величины ξ. Теперь найдем максимально правдоподобную оценку s^2 для второго неизвестного параметра, дисперсии σ^2:

$$\frac{\partial \ln L_X(a, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - a)^2 = 0.$$

После несложных преобразований получим

$$n\sigma^2 = \sum_{i=1}^{n} (x_i - a)^2,$$

откуда найдем оценку для σ^2:

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

Если точное значение параметра a неизвестно, то его можно заменить соответствующей оценкой $\bar{x}$. Тогда

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

Однако в последнем случае максимально правдоподобная оценка s^2 хотя и является эффективной, смещена относительно σ^2. Этот недостаток легко устраняется. Исправленное значение оценки будет

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

Если оценка неизвестного параметра распределения дается одним числом, то она называется точечной. Для того чтобы определить точность полученной оценки параметра при заданной надежности, используют так называемый доверительный интервал. Вычисление доверительного интервала особенно важно при малом объеме выборки, так как случайные погрешности оценки неизвестного параметра в этом случае могут оказаться недопустимо большими.

Сущность доверительного интервала заключается в следующем. Условимся, что $\bar{a}$ будет несмещенной оценкой параметра a, а q — заданное малое значение вероятности. Нам нужно найти такой интервал $\bar{a} - \alpha$, $\bar{a} + \alpha$, который бы с большой вероятностью $(1 - q)$ «накрывал» истинное неизвестное значение a; иначе говоря, чтобы соблюдалось следующее равенство:

$$P(\bar{a} - \alpha \leqslant a \leqslant \bar{a} + \alpha) = 1 - q.$$

Соответственно вероятность того, что a находится вне этого интервала, должна быть мала и равна q.

Интервал $\bar{a} - \alpha$, $\bar{a} + \alpha$ называется доверительным. Вероятность того, что он «накроет» точку a, называется доверительной вероятностью. Длина его равна 2α, а крайние значения $(\bar{a} - \alpha, \bar{a} + \alpha)$ называются доверительными границами.

Процедуру определения доверительных границ рассмотрим на примере среднего арифметического $\bar{x}$, вычисленного по наблюдениям и используемого в качестве оценки математического ожидания $M\xi = a$. Пусть случайная величина ξ распределена нормально с дисперсией σ^2. Пусть $1 - q$ — заданное значение вероятности того, что интервал $\bar{x} - \alpha$, $\bar{x} + \alpha$ «накрывает» истинное значение неизвестного параметра a, т. е.

$$P(\bar{x} - \alpha \leqslant a \leqslant \bar{x} + \alpha) = 1 - q.$$

Этому выражению равносильно

$$P(\bar{x} - a \leqslant \alpha) - P(\bar{x} - a \leqslant -\alpha) =$$

$$= P\left(\frac{\bar{x} - a}{\sigma_{\bar{x}}} \leqslant \frac{\alpha}{\sigma_{\bar{x}}}\right) - P\left(\frac{\bar{x} - a}{\sigma_{\bar{x}}} \leqslant -\frac{\alpha}{\sigma_{\bar{x}}}\right) = 1 - \frac{q}{2} - \frac{q}{2}.$$

Так как величина $\frac{\bar{x} - a}{\sigma_{\bar{x}}}$ распределена нормально с параметрами 0,1, то

$$P\left(\frac{\bar{x} - a}{\sigma_{\bar{x}}} \leqslant \frac{\alpha}{\sigma_{\bar{x}}}\right) = \Phi\left(\frac{\alpha}{\sigma_{\bar{x}}}\right) = 1 - \frac{q}{2}$$

и

$$P\left(\frac{\bar{x} - a}{\sigma_{\bar{x}}} \leqslant -\frac{\alpha}{\sigma_{\bar{x}}}\right) = \Phi\left(-\frac{\alpha}{\sigma_{\bar{x}}}\right) = \frac{q}{2},$$

где Φ — нормальная функция с параметрами 0,1.

Обозначим значение $\frac{\alpha}{\sigma_{\bar{x}}}$, соответствующее вероятности $\frac{q}{2}$, через $t_{\frac{q}{2}}$, а для $1 - \frac{q}{2}$ через $t_{1-\frac{q}{2}}$. Тогда

$$\frac{\alpha}{\sigma_{\bar{x}}} = t_{\frac{q}{2}},$$

$$\frac{\alpha}{\sigma_{\bar{x}}} = t_{1-\frac{q}{2}}.$$

Отсюда можно определить доверительные границы $\bar{x} - \alpha_{\frac{q}{2}}$ и $\bar{x} - \alpha_{1-\frac{q}{2}}$, которые обеспечивают вероятность $1 - q$ того, что значение a лежит между ними:

$$\alpha_{\frac{q}{2}} = t_{\frac{q}{2}} \sigma_{\bar{x}},$$

$$\alpha_{1-\frac{q}{2}} = t_{1-\frac{q}{2}} \sigma_{\bar{x}}.$$

Таким образом,

$$P\left(\bar{x} - t_{\frac{q}{2}} \sigma_{\bar{x}} \leqslant a \leqslant \bar{x} + t_{1-\frac{q}{2}} \sigma_{\bar{x}}\right) = 1 - q.$$

Так как $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, то

$$P\left(\bar{x} - \frac{t_{\frac{q}{2}} \sqrt{n}}{\sigma} \leqslant a \leqslant \bar{x} + \frac{t_{1-\frac{q}{2}} \sqrt{n}}{\sigma}\right) = 1 - q.$$

В тех случаях, когда точное значение параметра σ неизвестно, можно воспользоваться оценкой s этого параметра. Тогда

$$P\left(\bar{x} - \frac{t_{\frac{q}{2}}\sqrt{n}}{s} \leqslant a \leqslant \bar{x} + \frac{t_{1-\frac{q}{2}}\sqrt{n}}{s}\right) \simeq 1 - q.$$

Значения $t_{\frac{q}{2}}$, соответствующие заданной надежности $1 - q$, приведены в табл. 46.

Таблица 46

$1-q$	$t_{\frac{q}{2}}$	$1-q$	$t_{\frac{q}{2}}$
0,80	1,282	0,91	1,694
0,81	1,310	0,92	1,750
0,82	1,340	0,93	1,810
0,83	1,371	0,94	1,880
0,84	1,404	0,95	1,960
0,85	1,439	0,96	2,053
0,86	1,475	0,97	2,169
0,87	1,513	0,98	2,325
0,88	1,554	0,99	2,576
0,89	1,597	0,9973	3,000
0,90	1,643	0,9990	3,290

Каждое из чисел $t_{\frac{q}{2}}$ в этой таблице показывает, сколько раз среднее квадратическое отклонение надо отложить вправо и влево (по оси абсцисс) от величины $\bar{x}$, для того чтобы доверительный интервал «накрыл» значение a с вероятностью $1 - q$ (при условии нормального распределения величины ξ).

В тех случаях, когда распределение случайной величины отличается от нормального (при выборках большого объема), распределение среднего арифметического будет близко к нормальному, что позволяет для построения доверительного интервала пользоваться изложенными приемами.

Таблица 47

i	x_i	n_i
1	10	1
2	20	3
3	30	6
4	40	8
5	50	8
6	60	2
7	70	1
		29

Пример. Построим доверительный интервал по данным, приведенным в табл. 47. Замерена мощность 29 жил (i — номер класса; x_i — среднее значение мощности в классе, *см*; n_i — число жил в i-том классе). Найдем оценку $\bar{x}$ среднего значения мощности и оценку среднего квадратического отклонения s:

$$\bar{x} = \frac{1}{\sum_{i=1}^{n} n_i} \sum_{i=1}^{n} n_i x_i = \frac{1}{29} \cdot 1160 = 40,$$

$$s^2 = \frac{1}{\left(\sum_{i=1}^{m} n_i\right) - 1}\left[\sum_{i=1}^{m} n_i x_i^2 - \frac{\left(\sum_{i=1}^{m} n_i x_i\right)^2}{\sum_{i=1}^{m} n_i}\right] = 185.$$

Таким образом,

$$s = \sqrt{185} = 13{,}6;$$

$$s_{\bar{x}} = \frac{13{,}6}{\sqrt{29}} = 2{,}53.$$

Если нам нужен доверительный интервал, соответствующий надежности $1 - q = 0{,}95$, то по приведенной выше таблице находим $t_{0{,}025} = 1{,}96$. Далее определяем

$$t_{0{,}025} \cdot s_{\bar{x}} = 1{,}96 \cdot 2{,}53 = 4{,}96 \simeq 5.$$

Доверительные границы будут равны:

$$\bar{x} - \alpha_1 = 40 - 5 = 35,$$
$$\bar{x} + \alpha_2 = 40 + 5 = 45.$$

Доверительный интервал, отвечающий вероятности 0,95, составляет (35, 45). Это значит, что с вероятностью 0,95 истинное значение средней содержится в интервале (35, 45).

Построим доверительный интервал для среднего квадратического отклонения. При этом предполагается, что величина $\frac{ns^2}{\sigma^2}$ распределена по закону χ^2 с $n - 1$ степенями свободы (здесь σ^2 — неизвестная дисперсия, а s^2 — ее оценка, n — объем выборки). Выбираем пределы χ_1^2 и χ_2^2 так, чтобы

$$1 - P(\chi^2 > \chi_1^2) - P(\chi^2 < \chi_1^2) = \frac{q}{2};$$

$$P(\chi^2 > \chi_2^2) = \frac{q}{2},$$

где q — заданное малое значение вероятности.
Отсюда

$$P\left(\chi_1^2 < \frac{ns^2}{\sigma^2} < \chi_2^2\right) = 1 - P(\chi^2 - \chi_1^2) - P(\chi^2 > \chi_2^2) = 1 - q.$$

Таким образом, имеем

$$P\left(\frac{ns^2}{\chi_2^2} < \sigma^2 < \frac{ns^2}{\chi_1^2}\right) = 1 - q.$$

Преобразуем это выражение так, чтобы в него входила не σ^2, а σ:

$$P\left(\frac{\sqrt{n}\,s}{\chi_2} < \sigma < \frac{\sqrt{n}\,s}{\chi_1}\right) = 1 - q.$$

Из этого выражения видно, что нижней границей искомого интервала является $\frac{\sqrt{n}\,s}{\chi_2}$, а верхней — $\frac{\sqrt{n}\,s}{\chi_1}$.

Искомый интервал, который содержит σ с вероятностью $1 - q$, следовательно, будет равен

$$\left(\frac{\sqrt{n}\,s}{\chi_2}, \frac{\sqrt{n}\,s}{\chi_1}\right).$$

Пример. Построим доверительный интервал по выборочному среднему квадратическому отклонению s.

Исходные данные для расчета возьмем из предыдущего примера (см. табл. 47). Средняя арифметическая мощность жилы $\bar{x} = 40$ *см*,

$s = 13{,}6$, число степеней свободы $29 - 1 = 28$. Полагая, что $1 - q = 0{,}95$, q будет равной 0,05.

По числу степеней свободы 28 находим для вероятности 0,05 (приложение 14) значение $\chi_2^2 = 41{,}3$, а для вероятности 0,95 значение $\chi_1^2 = 16{,}9$.

Неравенство, вероятность которого нас интересует, можно теперь записать так:

$$\frac{29 \cdot 13{,}6^2}{16{,}9} < \sigma^2 < \frac{29 \cdot 13{,}6^2}{41{,}3}, \text{ или } 317 < \sigma^2 < 1300.$$

Переходя от дисперсии σ^2 к среднему квадратическому отклонению, получим

$$P\ (17{,}8 < \sigma < 36{,}0) = 0{,}95.$$

Наиболее распространенной группой статистических задач в геологии является проверка различных гипотез о среднем значении признака. Однако прежде чем перейти к формулировке этих гипотез с последующим рассмотрением соответствующих проверочных критериев, полезно ознакомиться с так называемым распределением Стьюдента, которое играет большую роль в статистических приложениях.

Пусть $\xi_1, \xi_2, \ldots, \xi_i, \ldots, \xi_n$ — последовательность независимых, одинаково нормально распределенных случайных величин, имеющих математическое ожидание, равное нулю, и дисперсию, равную единице.

Пусть также ξ_0 — нормально распределенная случайная величина с параметрами 0,1, которая не зависит от $\xi_1, \xi_2, \ldots, \xi_i, \ldots, \xi_n$. Образуем новую случайную величину

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

которая представляет частное от деления нормально распределенной случайной величины ξ_0 на корень квадратный из χ^2 — распределенной случайной величины $\sum_{i=1}^{n} \xi_i^2$, деленной на n. Распределение величины τ называется распределением Стьюдента и представлено следующим выражением:

$$S_n(t) = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{(n-1)\pi} \cdot \Gamma\left(\frac{n-1}{2}\right)} \int_{-\infty}^{t} \left(1 + \frac{t^2}{n-1}\right)^{-\frac{n}{2}} dt.$$

При $n \to \infty$ распределение $S_n(t)$ стремится как к своему пределу к распределению

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{t^2}{2}} dt,$$

т. е. к центрированному нормальному закону. Уже при $n = 20$ распределение t довольно близко к нормальному.

Плотность вероятности распределения Стьюдента графически напоминает нормальную кривую, которая при малой величине n заметно растянута по горизонтальной оси. Это означает, что большие отклонения при распределении Стьюдента встречаются чаще, чем при нормальном распределении, но по мере роста n кривая Стьюдента приближается к нормальной кривой. Э. Пирсон в 1929 г. экспериментально доказал, что распределением Стьюдента практически можно пользоваться не только при нормальном, но и вообще при одновершинном симметричном распределении признака в исследуемой генеральной совокупности, т. е. при любом эксцессе.

Рассмотрим теперь следующую ситуацию. Пусть из нормальной совокупности взята выборка $x_1, x_2, \ldots, x_n$, по которой вычислены оценки среднего $\bar{x}$ и дисперсии s^2. Обозначим неизвестное значение среднего через a. Пусть a_0 — заданное значение. Требуется проверить гипотезу, что $a = a_0$; иначе говоря, необходимо решить, можно ли пренебречь расхождением между оценкой среднего $\bar{x}$ и заданной величиной a_0 как случайным, или же этого делать нельзя, и расхождение следует считать существенным? Обозначим проверяемую гипотезу

$$H_0 : a = a_0$$

и будем называть ее нулевой гипотезой. Множество конкурирующих гипотез обозначим через

$$H_1 : a \neq a_0.$$

Для проверки нулевой гипотезы можно воспользоваться величиной t, вычисленной по формуле

$$t = \frac{\sqrt{n}(\bar{x} - a_0)}{\sqrt{\frac{n}{n-1} s^2}} = \frac{\sqrt{n-1}(\bar{x} - a_0)}{s},$$

и которая, если H_0 верна, распределена по закону Стьюдента с $n - 1$ степенью свободы.

Таким образом, гипотезу H_0 следует принять, если вычисленное значение t окажется меньше по абсолютной величине допустимого $t_{q, n-1}$, при уровне значимости q и $n - 1$ степени свободы. Если же $|t| > t_{q, n-1}$, то гипотезу H_0 следует забраковать и принять H_1. Значение $t_{q, n-1}$ можно найти в приложении 15.

Очень часто при геологических исследованиях требуется сравнить средние арифметические $\bar{x}_1$ и $\bar{x}_2$, полученные по n_1 и n_2 наблюдениям из двух совокупностей. Эту задачу можно сформулировать как проверку нулевой гипотезы о равенстве двух неизвестных средних a_1 и a_2 по их статистическим оценкам $\bar{x}_1$ и $\bar{x}_2$. Таким образом,

$$H_0 : a_1 = a_2,$$

при альтернативе

$$H_1 : a_1 \neq a_2.$$

Обозначим через s_1^2 и s_2^2 соответствующие оценки неизвестных дисперсий. Если распределения сравниваемых случайных величин нормальные, а гипотеза H_0 верна, то случайная величина

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

будет распределена по закону Стьюдента с $n_1 + n_2 - 2$ степенями свободы.

Таким образом, гипотеза H_0 должна отвергаться в тех случаях, когда вычисленное значение t превышает допустимое (t_{q, n_1+n_2-2}) при уровне значимости q и $n_1 + n_2 - 2$ степенях свободы, и приниматься как непротиворечащая выборочным данным, если $t \leqslant t_{q, n_1+n_2-2}$.

Однако при геологических исследованиях сравнения средних часто не ограничиваются двумя совокупностями. Очень часто приходится сталкиваться с ситуацией, когда рассматриваются m совокупностей одновременно, например несколько типов пород, несколько месторождений или несколько районов и т. п. Пусть в каждой из m совокупностей проведено по n_i наблюдений одного и того же признака, и по этим результатам вычислены средние арифметические $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_i, \ldots, \bar{x}_n$, являющиеся

оценками неизвестных средних $a_1, a_2, \ldots, a_i, \ldots, a_m$. Соответствующие оценки неизвестных дисперсий обозначим через $s_1^2, s_2^2, \ldots, s_i^2, \ldots, s_m^2$. По имеющимся данным требуется проверить гипотезу

$$H_0 : a_1 = a_2 = \ldots = a_i = \ldots = a_m = a_0$$

при конкурирующей гипотезе, которая заключается в том, что хотя бы для одного i $a_i \neq a_0$.

Допустим, что распределения рассматриваемых случайных величин во всех совокупностях нормальные и дисперсии их равны. Если эти условия выполнены, то при правильной нулевой гипотезе H_0 все m величин t_i, которые даются выражением

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

будут распределены по закону Стьюдента с $N-2$ степенями свободы. В выражении t_i

$$N = \sum_{i=1}^{m} n_i,$$

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

где

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{m} n_i \bar{x}_i,$$

$$s^2 = \frac{1}{N-m} \sum_{i=1}^{m} (n_i - 1) s_i^2.$$

Таким образом, гипотеза H_0 отвергается, если хотя бы одно из вычисленных значений t_i превысит допустимое $t_{q, N-2}$ при уровне значимости q и $N-2$ степенях свободы. Если же для всех i имеет место неравенство $t_i \leqslant t_{q, N-2}$, то средние во всех m совокупностях можно рассматривать как равные. В тех случаях, когда гипотеза H_0 отвергается, из набора средних арифметических $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m$ нужно исключить ту величину $\bar{x}_i$, для которой $|\bar{x}_i - \bar{x}| = \max$, а для оставшихся $m-1$ значений провести проверку, предварительно вычислив новые величины $\bar{x}$ и s^2. Этот процесс повторяется до тех пор, пока не будет принята H_0.

Пример. Приведем пример проверки гипотезы $H_0 : a = a_0$.

С 10 участков полиметаллического (в основном цинкового) месторождения взято по одной валовой пробе, причем пробы взяты одним и тем же способом.

Каждая проба имеет вес 1000 *кг*. Запасы всего месторождения в десятки тысяч раз превышают общий вес взятых проб, поэтому условно можно считать выборку повторной, хотя (если смотреть на это более строго) она, конечно, бесповторная. Содержание цинка в пробах:

Участок	Содержание, % x	Квадрат содержания x^2
1	0,6	0,36
2	2,4	5,76
3	2,1	4,41
4	1,4	1,96
5	1,2	1,44
6	4,8	23,04
7	0,9	0,81
8	1,1	1,21
9	3,5	12,25
10	3,0	9,00
Сумма	21,0	60,24

По этим данным требуется проверить гипотезу, заключающуюся в том, что неизвестное среднее содержание цинка (a) по месторождению можно рассматривать как равное 2,70%. Таким образом,

$$H_0 : a = 2{,}70\%.$$

при альтернативе

$$H_1 : a \neq 2{,}70\%.$$

Оценка среднего содержания цинка $\bar{x} = 2{,}10\%$, а оценка дисперсии $s^2 = 1{,}414$. Тогда $s = 1{,}19$.
Применяя отношение Стьюдента, получим

$$t = \frac{|(\bar{x}_1 - a_0)\sqrt{n-1}|}{s} = \frac{2{,}10 - 2{,}70\sqrt{10-1}}{1{,}19} = \frac{0{,}6 \cdot 3}{1{,}19} = 1{,}51.$$

Допустимое значение t при уровне значимости 0,05 и числе степеней свободы 9 равно 2,262. Так как вычисленное значение t много меньше 2,262, то гипотезу $H_0 : a = 2{,}70$ следует принять как подтвердившуюся. Полученный результат означает, что нет оснований считать среднее содержание цинка на месторождении существенно отличающимся от 2,7%.

Пример. На полиметаллическом руднике рудные жилы залегают в различных вмещающих породах, в частности в кварцевых кератофирах и известняках.

Для определения объемного веса было взято по 7 проб каждого типа руды. Результаты определений объемного веса оказались такими. В кератофирах: 2,50; 2,55; 2,60; 2,75; 2,80; 2,80 и 2,95; в известняках: 2,50; 2,80; 2,85; 2,90; 2,95; 2,95 и 3,40. Среднее арифметическое по первой группе проб $\bar{x}_1 = 2{,}71$ и по второй $\bar{x}_2 = 2{,}91$. Оценки дисперсий соответственно равны $s_1^2 = 0{,}0262$ и $s_2^2 = 0{,}0712$.

Спрашивается, можно ли пренебречь расхождением между $\bar{x}_1$ и $\bar{x}_2$ как случайным и рассматривать средние значения объемного веса двух типов руд как равные.

Подставив значения $\bar{x}_1$, $\bar{x}_2$, n_1, n_2, s_1^2 и s_2^2 в выражение

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

получим

$$t = \frac{|2{,}71 - 2{,}91|}{\sqrt{\frac{7 \cdot 0{,}0262 + 7 \cdot 0{,}0712}{7 + 7 - 2}\left(\frac{1}{7} + \frac{1}{7}\right)}} = 1{,}6.$$

Поскольку число степеней свободы $k = 7 + 7 - 2 = 12$, допустимое значение t при уровне значимости 0,05 равно 2,179. Таким образом, гипотезу о равенстве средних для обоих типов руд следует принять как подтвердившуюся. Это значит, что никаких выводов о разнице в средних делать нельзя.

Пример. Рассмотрим еще один пример сравнения двух выборочных средних.

На железорудном месторождении в сходных геологических условиях производилось рудничное опробование двумя методами: 1) пунктирной бороздой и 2) сплошной бороздой. Сечение борозд одинаковое. Содержание

железа в пробах, взятых первым методом (с округлением), x_{1_i} и квадраты этого содержания $x_{1_i}^2$ таковы:

Проба	x_{1_i}, %	$x_{1_i}^2$
1	45	2 025
2	53	2 809
3	48	2 304
4	59	3 481
5	44	1 936
6	60	3 600
7	41	1 681
8	43	1 849
9	57	3 249
Сумма	450	22 934

Содержание железа в пробах, взятых вторым методом (с округлением), x_{2_i} и квадраты этого содержания $x_{2_i}^2$:

Проба	x_{2_i}, %	$x_{2_i}^2$
1	51	2 601
2	50	2 500
3	42	1 764
4	44	1 936
5	39	1 521
6	40	1 600
7	48	2 304
8	38	1 444
9	59	3 481
10	55	3 025
11	51	2 601
Сумма	517	24 777

По данным этого опробования необходимо решить вопрос: можно ли рассматривать эти методы как равноценные, т. е. обеспечивающие в результате их применения одни и те же значения среднего? Иначе говоря, мы проверяем гипотезу $a_1 = a_2$.

Среднее арифметическое по пробам первой группы

$$\bar{x}_1 = \frac{450}{9} = 50,$$

а по пробам второй группы

$$\bar{x}_2 = \frac{517}{11} = 47.$$

Вычисленные по приведенным выше данным оценки дисперсий равны соответственно

$$s_1^2 = 54{,}25,$$

$$s_2^2 = 47{,}8.$$

Подставив эти данные в выражение

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{(n_1 s_1^2 + n_2 s_2^2)(n_1 + n_2)}{(n_1 + n_2 - 2)\, n_1 n_2}}},$$

получим

$$t = \frac{|50 - 47|}{\sqrt{\frac{(9 \cdot 54{,}25 + 11 \cdot 47{,}80)\ (9 + 11)}{(9 + 11 - 2)\, 9 \cdot 11}}} = 0{,}89.$$

Допустимое значение t при уровне значимости 0,05 и 18 степенях свободы равно 2,101. Таким образом, проверяемую гипотезу нужно принять как подтвердившуюся и оба метода опробования рассматривать как равноценные.

Пример. В геологической практике нередко приходится сталкиваться со следующей задачей. Два пласта сильвинита опробованы одним и тем же способом. Пробы обработаны и проанализированы одним методом. Содержание хлористого калия по первому пласту x_{1i} в процентах и квадраты этого содержания (с округлением) x_{1i}^2 приводятся ниже:

Проба	x_{1i}	x_{1i}^2
1	35	1225
2	31	961
3	35	1225
4	36	1296
5	33	1089
6	34	1156
7	34	1156
Сумма	238	8108

Содержание хлористого калия по второму пласту x_{2i} и соответствующие им квадраты (с округлением) x_{2i}^2 таковы:

Проба	x_{2i}	x_{2i}^2
1	21	441
2	28	784
3	30	900
4	22	484
5	23	529
6	29	841
7	28	784
8	28	784
9	21	441
10	22	484
11	31	961
12	29	841
Сумма	312	8274

По проведенным данным опробования необходимо решить, существенна ли разница между пластами по среднему содержанию в них хлористого калия или же она случайна и ее влиянием можно пренебречь. Для этого проделаем следующие вычисления.

Оценим сначала среднее содержание по первому пласту. Среднее арифметическое по нему равно

$$\bar{x}_1 = \frac{238}{7} = 34.$$

Аналогичная оценка среднего содержания для второго пласта

$$\bar{x}_2 = \frac{312}{12} = 26.$$

Воспользовавшись приемом, применявшимся в предыдущем примере, вычислим значение $t = 4{,}98$. Так как t превышает допустимое (3,965) при уровне значимости 0,001 и степенях свободы 17, то гипотезу о равенстве средних содержаний хлористого калия в пластах можно уверенно забраковать. Этот результат позволяет сделать вывод о том, что один из пластов существенно богаче хлористым калием по сравнению с другим.

Проверка таких гипотез по выборочным средним очень важна для геологии. Она позволяет узнать, не противоречат ли факты тому или иному выводу, построенному на разности оценок средних. Однако (заметим еще раз) для успешного применения критерия надо сначала убедиться в нормальности или, по крайней мере, в симметричности распределения. В нашем примере распределение более или менее симметричное.

Если под руками нет таблиц допустимых значений t, то можно воспользоваться упрощенным способом проверки гипотезы о равенстве средних.

Расхождение между средними считают случайным и несущественным, если

$$t < 3\sigma_t,$$

и неслучайным, существенным, если

$$t > 3\sigma_t.$$

Величина σ_t вычисляется следующим образом:

$$\sigma_t = \sqrt{\frac{n_1 + n_2 - 2}{n_1 + n_2 - 4}}.$$

Применим этот способ к данным последних двух примеров. Первый пример. Имеем $\bar{x}_1 = 50$, $n_1 = 9$, $\bar{x}_2 = 47$, $n_2 = 11$, $t = 0{,}89$.

$$\sigma_t = \sqrt{\frac{9 + 11 - 2}{9 + 11 - 4}} = \sqrt{\frac{18}{16}} = 1{,}06,$$

$$3\sigma_t = 3 \cdot 1{,}06 = 3{,}18.$$

Так как $0{,}89 < 3{,}18$, т. е. $t < 3\sigma_t$, то средние можно считать равными.

В последнем примере также $\sigma_t = 1{,}06$, а $t = 4{,}97$. Так как $t > 3\sigma_t = 3{,}18$, то гипотезу о равенстве средних следует считать неприемлемой.

VII. ПРОВЕРКА ГИПОТЕЗ О ДИСПЕРСИЯХ И ДИСПЕРСИОННЫЙ АНАЛИЗ

Как и математическое ожидание, дисперсия является важной характеристикой распределения случайной величины. Во многих практических задачах сравнения выборочных данных, помимо проверки гипотезы о равенстве средних, иногда нужно сделать выводы о дисперсиях, т. е. определить, одинакова ли степень рассеяния отдельных результатов наблюдения вокруг средних.

Таким образом, задачу можно сформулировать как проверку гипотезы о равенстве дисперсий двух случайных величин по выборочным данным.

Пусть ξ и η — две независимые нормально распределенные случайные величины с дисперсиями σ_ξ^2 и σ_η^2 соответственно. Обозначим через x_1, x_2, . . ., x_m выборочные значения случайной величины ξ, а через y_1, y_2, . . ., y_n — значения величины η. Оценки для σ_ξ^2 и σ_η^2 будут:

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

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Известно, что величины

$$\chi_1^2 = \frac{(m-1)\,s_x^2}{\sigma_\xi^2},$$

$$\chi_2^2 = \frac{(n-1)s_y^2}{\sigma_\eta^2}$$

распределены как χ^2 с $m-1$ и $n-1$ степенями свободы соответственно.

Если гипотеза $H_0: \sigma_\xi^2 = \sigma_\eta^2$ верна, то

$$w = \frac{\chi_1^2}{\chi_2^2} = \frac{(m-1)\,s_x^2}{(n-1)\,s_y^2},$$

и величина w будет иметь функцию распределения $H(w)$, зависящую только от числа степеней свободы $f_1 = m-1$ и $f_2 = n-1$:

$$H(w) = \frac{\Gamma\left(\frac{f_1+f_2}{2}\right)}{\Gamma\left(\frac{f_1}{2}\right)\Gamma\left(\frac{f_2}{2}\right)} \int_0^w z^{\frac{1}{2}f-1}\,(z+1)^{-\frac{1}{2}(f_1+f_2)}\,dz.$$

Обозначив выражение $\frac{s_x^2}{s_y^2}$ через F, получим

$$w = \frac{f_1}{f_2} F.$$

Используя это соотношение, нетрудно вычислить значения F, соответствующие заданному уровню значимости при данном числе степеней свободы $f_1 = m-1$ и $f_2 = n-1$.

Таким образом, в качестве критерия для проверки гипотезы $H_0: \sigma_\xi^2 = \sigma_\eta^2$ используется отношение F статистических оценок дисперсий s_x^2 и s_y^2. Гипотеза $\sigma_\xi^2 = \sigma_\eta^2$ отвергается, если вычисленное значение F превысит допустимое $(F_{q,\,f_1,\,f_2})$ при уровне значимости q, f_1, f_2 степенях свободы. Наоборот, если $F \leqslant F_{q,\,f_1,\,f_2}$, то гипотеза $\sigma_\xi^2 = \sigma_\eta^2$ принимается как непротиворечащая эмпирическим данным (значения $F_{q,\,f_1,\,f_2}$ приведены в приложениях 24 и 25).

Пример. Пробы на месторождении анализировались в химической лаборатории двумя методами, причем каждая проба анализировалась только один раз. Содержание интересующего нас элемента x (в %) и квадраты этого содержания приведены в табл. 48 и 49.

Первая партия проб, т. е. пробы, проанализированные одним методом, дали следующие результаты (табл. 48).

Таблица 48

Номера проб	$x_{1,\,i}$	$x_{1,\,i} - \bar{x}_1$	$(x_{1,\,i} - \bar{x}_1)^2$
1	1	−2	4
2	0	−3	9
3	10	7	49
4	2	−1	1
5	3	0	0
6	2	−1	1
Сумма	18		64

Выборочное среднее и оценка дисперсии по этой партии равны:

$$\bar{x}_1 = \frac{18}{6} = 3; \quad s_1^2 = \frac{64}{6-1} = 12{,}8.$$

Вторая партия проб, т. е. пробы, проанализированные другим методом, показали следующие результаты (табл. 49).

Таблица 49

Номера проб	$x_{2,i}$	$x_{2,i}-\bar{x}_2$	$(x_{2,i}-\bar{x}_2)^2$
1	4	1	1
2	3	0	0
3	2	—1	1
4	2	—1	1
5	4	1	1
6	2	—1	1
7	5	2	4
8	6	3	9
9	0	—3	9
10	2	—1	1
Сумма	30		28

Оценка среднего и выборочная дисперсия имеют следующие значения:

$$\bar{x}_2 = \frac{30}{10} = 3; \quad s_2^2 = \frac{28}{10-1} = 3{,}1.$$

По этим данным требуется выяснить, можно ли рассматривать случайные погрешности применяемых аналитических методов как равнозначные. Этому равносильна проверка гипотезы о равенстве неизвестных дисперсий σ_1^2 и σ_2^2, обусловленных сравниваемыми методами. Для проверки гипотезы вычислим отношение

$$F = \frac{s_1^2}{s_2^2} = \frac{12{,}8}{3{,}1} = 4{,}14.$$

Допустимое значение F при уровне значимости 0,05 и степенях свободы 5 и 9 равно 4,77. Так как вычисленное значение меньше 4,77, то для вывода о существенном различии между дисперсиями методов нет оснований.

Методы проверки гипотезы о равенстве дисперсий можно с успехом использовать и для проверки гипотезы о равенстве нескольких средних, соответствующих различным уровням изменения некоторого неслучайного фактора. Статистические приемы, предназначенные для решения подобных задач, объединены под общим названием дисперсионного анализа.

В геологии очень часто приходится встречаться с влиянием нескольких причин, сил или факторов на какое-либо явление. Допустим, что мы изучаем влияние глубины на содержание металла в руде. Фактор глубины можно представить его различными уровнями, например горизонтами 1, 2, 3, . . ., p. На каждом горизонте берется некоторое число проб.

Обозначим случайную величину (содержание металла в пробах), соответствующую i-тому горизонту, через ξ_i и допустим, что она распределена нормально с математическим ожиданием a_i и дисперсией σ^2. Величину дисперсии σ^2 для всех горизонтов будем считать постоянной. Гипотезу о том, что фактор глубины не влияет на содержание металла в пробах, можно сформулировать как равенство $a_1 = a_2 = \ldots = a_p$. Пусть над каждой случайной величиной ξ_i произведено по n_i наблюдений, которые мы обозначим $x_{i1}, x_{i2}, \ldots, x_{ij}, x_{in_i}$. Таким образом, x_{ij} означает j-тое значение в i-той группе. Среднее арифметическое, являющееся оценкой для a_i, полученное по n_i наблюдениям в i-той группе, составит

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

Обозначим через n сумму всех n_i, т. е.

$$n = \sum_{i=1}^{p} n_i.$$

Тогда среднее арифметическое всех n значений

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{n_i} x_{ij} = \frac{1}{n} \sum_{i=1}^{p} n_i \bar{x}_{i\cdot}.$$

В данном случае будет иметь место тождество

$$\sum_{i,\,j} (x_{ij} - \bar{x})^2 = \sum_{i,\,j} (x_{ij} - \bar{x}_{i\cdot})^2 + \sum_{i,\,j} (\bar{x}_{i\cdot} - \bar{x})^2.$$

Изменив порядок слагаемых, получим

$$\sum_{i=1}^{p} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \sum_{i=1}^{p} n_i (\bar{x}_{i\cdot} - \bar{x})^2 + \sum_{i=1}^{p} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i\cdot})^2.$$

Запишем это тождество короче:

$$Q = Q_1 + Q_2.$$

Если гипотеза $a_1 = a_2 = \ldots = a_p$ верна, то отношения $\frac{Q}{\sigma^2}$, $\frac{Q_1}{\sigma^2}$ и $\frac{Q_2}{\sigma_2}$ будут распределены как χ^2 с $n - 1$, $p - 1$ и $n - p$ степенями свободы соответственно. Положив,

$$s^2 = \frac{Q}{n-1} = \frac{1}{n-1} \sum_{i=1}^{p} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2,$$

$$s_1^2 = \frac{Q_1}{p-1} = \frac{1}{p-1} \sum_{i=1}^{p} n_i (\bar{x}_{i\cdot} - \bar{x})^2,$$

$$s_2^2 = \frac{Q_2}{n-p} = \frac{1}{n-p} \sum_{i=1}^{p} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i\cdot})^2,$$

получим, что в условиях нулевой гипотезы математические ожидания оценок дисперсий s^2, s_1^2 и s_2^2 равны σ^2, т. е.

$$M(s^2) = M(s_1^2) = M(s_2^2) = \sigma^2.$$

Тогда отношение

$$\frac{s_1^2}{s_2^2} = \frac{\frac{Q_1}{p-1}}{\frac{Q_2}{n-p}}$$

будет распределено по закону Фишера (F — распределения) с $p - 1$ и $n - p$ степенями свободы. Таким образом, гипотеза $a_1 = a_2 = \ldots = a_p = a$ отвергается, если вычисленное значение этого отношения окажется больше допустимого ($F_{q,\,p-1,\,n-p}$) при уровне значимости q и $p - 1$ и $n - p$ степенях свободы. Если же

$$\frac{s_1^2}{s_2^2} \leqslant F_{q,\,p-1,\,n-p},$$

то гипотезу о равенстве средних можно принять как не противоречащую эмпирическим данным, а исследуемый фактор рассматривать как не оказывающий влияния на изучаемую случайную величину.

Пример. Для иллюстрации применения методов дисперсионного анализа рассмотрим пример.

На одном полиметаллическом руднике рудные жилы залегают в семи различных типах вмещающих пород: в кварцевых кератофирах, глинистых сланцах, туффитах, песчаниках, альбитофирах, известняках и гранитах.

Для того чтобы выяснить вопрос о влиянии боковых пород на объемный вес руды, было взято семь серий проб, причем каждая серия состояла из семи проб (по одной для каждого типа вмещающей породы). Всего проб, таким образом, взято 49. Результаты анализа проб показаны в табл. 50 (в данном случае x — объемный вес).

Таблица 50

j	i							$\sum_{i=1}^{7} x_{ij}$	$\bar{x}_{\cdot j}$
	1	2	3	4	5	6	7		
1	2,95	2,60	2,65	2,55	2,75	2,80	2,60	18,90	2,70
2	2,50	2,95	2,75	2,85	2,45	2,50	2,55	18,55	2,65
3	2,55	2,70	2,80	2,60	2,90	2,85	2,70	19,10	2,73
4	2,80	2,90	2,75	2,65	3,00	2,95	2,70	19,75	2,82
5	2,80	2,65	2,60	3,10	2,50	2,95	2,95	19,55	2,79
6	2,60	3,25	3,00	2,70	3,00	2,90	2,80	20,25	2,89
7	2,75	2,50	3,40	3,10	3,60	3,40	3,15	21,90	3,19
$\bar{x}_{i\cdot}$	2,71	2,76	2,85	2,79	2,89	2,91	2,78	138,00	$\bar{x} = 2{,}82$

В этой таблице типы вмещающих пород обозначены номерами $i = 1, 2, \ldots, 7$ (в порядке перечисления этих типов, данного выше), а номера проб — $j = 1, 2, \ldots, 7$. Таким образом, в данном примере $n_1 = n_2 = \ldots = n_p = q$.

Общая схема дисперсионного анализа для решения поставленной задачи дается в табл. 51.

Таблица 51

Вид дисперсии	Сумма квадратов	Число степеней свободы	Средний квадрат
Между группами	$q\sum_i (\bar{x}_{i\cdot} - \bar{x})^2 = Q_1$	$p - 1$	$s_1^2 = \frac{Q_1}{p-1}$
Внутри групп	$\sum_{j,i} (x_{ij} - \bar{x}_{i\cdot})^2 = Q_2$	$p(q-1)$	$s_2^2 = \frac{Q_2}{p(q-1)}$
Сумма	$\sum_{j,i} (x_{ij} - \bar{x})^2 = Q$	$pq - 1$	$s^2 = \frac{Q}{pq-1}$

Суммы квадратов Q и Q_1 вычисляются непосредственно, а Q_2 получается путем вычитания второй из первой. Знаком $\bar{x}_{i\cdot}$ обозначен средний удельный вес руды по тому типу породы, который обозначен данным индексом i. Число p — количество типов породы, число q — количество серий. При этом $p = q = 7$.

Вычисленные по этой схеме значения величин могут быть представлены в следующем виде (табл. 52).

Таблица 52

Дисперсия	Сумма квадратов	Число степеней свободы	Средний квадрат
Между группами	0,2121	6	$s_1^2 = 0{,}03535$
Внутри групп	2,3579	42	$s_2^2 = 0{,}05614$
Сумма	2,5700	48	$s^2 = 0{,}05354$

Разделив s_1^2 на s_2^2, вычислим значение критерия Фишера F:

$$F = \frac{0{,}03535}{0{,}05614} = 0{,}63.$$

Так как вычисленное значение значительно меньше, чем допустимое $F_{0{,}05;\,6;\,42} = 2{,}34$, то гипотезу об отсутствии влияния типа вмещающих пород на объемный вес руды следует принять.

Выше был рассмотрен простейший случай дисперсионного анализа (однофакторный), который изучает влияние только одного фактора. В нем таким фактором является состав вмещающих рудные жилы пород. Однако в ряде случаев требуется рассмотреть результаты одновременного действия двух факторов, например мощности жил и глубины их залегания, на содержание в них свинца.

Обозначим изучаемые признаки или факторы через A и B. Пусть по признаку A все наблюдения делятся на p групп (A_1, A_2 . . ., A_p), а по признаку B — на q групп (B_1, B_2, . . ., B_q). Таким образом, всего будет pq групп.

Рассмотрим наиболее простой случай двухфакторного анализа, когда на каждую группу приходится только по одному наблюдению. Результаты наблюдений x_{ij} удобно записывать в виде следующей таблицы (табл. 53).

Таблица 53

	B_1	B_2	. . . B_j	. . . . B_q	Среднее, $\bar{x}_{i\cdot}$
A_1	x_{11}	x_{12}	. . . x_{1j}	. . . . x_{1q}	$\bar{x}_{1\cdot}$
A_2	x_{21}	x_{22}	. . . x_{2j}	. . . . x_{2q}	$\bar{x}_{2\cdot}$
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
A_i	x_{i1}	x_{i2}	. . . x_{ij}	. . . . x_{iq}	$\bar{x}_{i\cdot}$
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
A_p	x_{p1}	x_{p2}	. . . x_{pj}	. . . . x_{pq}	$\bar{x}_{p\cdot}$
Среднее, $\bar{x}_{\cdot j}$	$\bar{x}_{\cdot 1}$	$\bar{x}_{\cdot 2}$	. . . $\bar{x}_{\cdot j}$	. . . . $\bar{x}_{\cdot q}$	$\bar{x}$

Средние арифметические по строкам $(\bar{x}_{i\cdot})$ и по столбцам $(\bar{x}_{\cdot j})$, а также общая средняя $(\bar{x})$ подсчитываются по формулам:

$$\bar{x}_{i\cdot} = \frac{1}{q}\sum_{j=1}^{q} x_{ij},$$

$$\bar{x}_{\cdot j} = \frac{1}{p}\sum_{i=1}^{p} x_{ij}, \qquad \bar{x} = \frac{1}{pq}\sum_{i=1}^{p}\sum_{j=1}^{q} x_{ij}.$$

Как и в случае однофакторного анализа, получим тождество

$$Q = \sum_{i=1}^{p}\sum_{j=1}^{q}(x_{ij}-\bar{x})^2 = q\sum_{i=1}^{p}(\bar{x}_{i\cdot}-\bar{x})^2 + p\sum_{j=1}^{q}(\bar{x}_{\cdot j}-\bar{x})^2 +$$

$$+\sum_{i=1}^{p}\sum_{j=1}^{q}(x_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}+\bar{x})^2 = Q_1 + Q_2 + Q_3.$$

Если нулевая гипотеза утверждает, что средние значения для всех групп равны при условии, что все рассматриваемые случайные величины распределены нормально с одной и той же дисперсией σ^2, то отношения

$$\frac{Q}{\sigma^2},\ \frac{Q_1}{\sigma^2},\ \frac{Q_2}{\sigma^2}\ \text{и}\ \frac{Q_3}{\sigma^2}$$

будут распределены по закону χ^2 с $pq-1$, $p-1$, $q-1$ и $(p-1)(q-1)$ степенями свободы соответственно. Тогда оценками для общей дисперсии и ее компонентов будут служить выражения:

$$s^2 = \frac{Q}{pq-1} = \frac{1}{pq-1}\sum_{i=1}^{p}\sum_{j=1}^{q}(x_{ij}-\bar{x})^2,$$

$$s_1^2 = \frac{Q_1}{p-1} = \frac{q}{p-1}\sum_{i=1}^{p}(\bar{x}_{i\cdot}-\bar{x})^2,$$

$$s_2^2 = \frac{Q_2}{q-1} = \frac{p}{q-1}\sum_{j=1}^{q}(\bar{x}_{\cdot j}-\bar{x})^2,$$

$$s_3^2 = \frac{Q_3}{(p-1)(q-1)} = \frac{1}{(p-1)(q-1)}\sum_{i=1}^{p}\sum_{j=1}^{q}(x_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}+\bar{x})^2.$$

Для проверки гипотез равенства средних по строкам или по столбцам исходных данных таблицы (табл. 53) применяются следующие критерии:

$$F_A = \frac{\frac{1}{p-1}Q_1}{\frac{1}{(p-1)(q-1)}Q_3} = \frac{s_1^2}{s_3^2},$$

$$F_B = \frac{\frac{1}{q-1}Q_2}{\frac{1}{(p-1)(q-1)}Q_3} = \frac{s_2^2}{s_3^2}.$$

Если оба фактора не оказывают существенного влияния на изучаемый признак, то обе величины лишь с малой вероятностью должны принимать значения, превосходящие допустимые, при соответствующем уровне значимости и числе степеней свободы.

Общую схему двухфакторного дисперсионного анализа, когда в каждую группу входит только по одному наблюдению (бесповторный анализ), можно представить в виде следующей таблицы (табл. 54).

Более сложная схема двухфакторного дисперсионного анализа возникает в тех случаях, когда для каждой комбинации A_iB_j произведено по n наблюдений (анализ с повторением). Это значит, что в каждой клетке таблицы исходных данных находится не один, а n результатов наблюдений, которые мы обозначим x_{ijk}. В данном случае $i = 1, 2, \ldots, p$; $j = 1, 2, \ldots, q$ сохраняют тот же смысл, что и раньше, а $k = 1, 2, \ldots, n$— это номер наблюдения в ij-той группе.

Таблица 54

Вид дисперсии	Сумма квадратов	Число степеней свободы	Оценка дисперсии
Между средними по строкам	$Q_1 = q\sum_{i=1}^{p}(\bar{x}_{i.} - \bar{x})^2$	$p-1$	$s_1^2 = \frac{Q_1}{p-1}$
Между средними по столбцам	$Q_2 = p\sum_{j=1}^{q}(\bar{x}_{.j} - \bar{x})^2$	$q-1$	$s_2^2 = \frac{Q_2}{q-1}$
Остаточная	$Q_3 = \sum_{i=1}^{p}\sum_{j=1}^{q}(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2$	$(p-1)(q-1)$	$s_3^2 = \frac{Q_3}{(p-1)(q-1)}$
Общая (сумма) . . .	$Q = \sum_{i=1}^{p}\sum_{j=1}^{q}(x_{ij} - \bar{x})^2$	$pq-1$	$s^2 = \frac{Q}{pq-1}$

При дисперсионном анализе с повторением используются следующие оценки средних:

$$\bar{x}_{ij.} = \frac{1}{n}\sum_{k=1}^{n} x_{ijk},$$

$$\bar{x}_{i..} = \frac{1}{q}\sum_{j=1}^{q}\bar{x}_{ij.} = \frac{1}{qn}\sum_{j=1}^{q}\sum_{k=1}^{n} x_{ijk},$$

$$\bar{x}_{.j.} = \frac{1}{p}\sum_{i=1}^{p}\bar{x}_{ij.} = \frac{1}{pn}\sum_{i=1}^{p}\sum_{k=1}^{n} x_{ijk},$$

$$\bar{x} = \frac{1}{npq}\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n} x_{ijk}.$$

Общую схему двухфакторного дисперсионного анализа с повторением можно представить в виде следующей таблицы (табл. 55).

Таблица 55

Вид дисперсии	Сумма квадратов	Число степеней свободы	Средний квадрат
Между средними по строкам	$Q_1 = nq\sum_{i=1}^{p}(\bar{x}_{i..} - \bar{x})^2$	$p-1$	$s_1^2 = \frac{Q_1}{p-1}$
Между средними по столбцам	$Q_2 = np\sum_{j=1}^{q}(\bar{x}_{.j.} - \bar{x})^2$	$q-1$	$s_2^2 = \frac{Q_2}{q-1}$
Смешанная	$Q_3 = n\sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x})^2$	$(p-1)(q-1)$	$s_3^2 = \frac{Q_3}{(p-1)(q-1)}$
Остаточная	$Q_4 = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n}(x_{ijk} - \bar{x}_{ij.})^2$	$pq(n-1)$	$s_4^2 = \frac{Q_4}{pq(n-1)}$
Общая	$Q = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n}(x_{ijk} - \bar{x})^2$	$npq-1$	$s^2 = \frac{Q}{npq-1}$

Используя данные таблицы, можно построить отношения оценок дисперсий для проверки гипотез о влиянии факторов как в отдельности, так и совместно:

$$F_A = \frac{s_1^2}{s_4^2},$$

$$F_B = \frac{s_2^2}{s_4^2},$$

$$F_{AB} = \frac{s_3^2}{s_4^2}.$$

Если влияние рассматриваемых факторов отсутствует, то распределение величин F_A, F_B и F_{AB} будет совпадать с F-распределением Фишера, что позволяет выбирать критические значения F для заданного уровня значимости и соответствующего числа степеней свободы (приложения 24, 25).

Для иллюстрации применения двухфакторного дисперсионного анализа с повторением рассмотрим следующий пример.

Пример. При разведке полиметаллического месторождения возник вопрос о влиянии мощности жилы и глубины ее залегания на среднее содержание свинца.

Обозначим фактор мощности через A, а фактор глубины через B. Выделим следующие уровни (классы) фактора A по мощности жилы: от 5 до 15 *см* — A_1, от 15 до 25 *см* — A_2, от 25 до 35 *см* — A_3, от 35 до 45 *см* — A_4. Фактор B представлен тремя эксплуатационными горизонтами: B_1 (первый сверху), B_2 (средний, или второй) и B_3 (третий сверху, или нижний).

Исходные (эмпирические) данные содержания свинца в процентах приведены в табл. 56.

Таблица 56

Фактор	x_{ijk}		
	B_1	B_2	B_3
A_1	1, 5, 6	1, 3, 5	2, 3, 7
A_2	2, 2, 5	2, 5, 8	3, 8, 10
A_3	1, 4, 10	2, 2, 11	6, 7, 8
A_4	5, 10, 12	2, 10, 15	4, 7, 10

Общее число проб npq — 36. Для каждой комбинации фактора A с фактором B имеются три пробы. Для комбинации A_1B_1, например, имеем три пробы; содержание свинца в них 1, 5 и 6%.

Вычисляем сумму содержаний для каждой комбинации, для каждой строки и столбца таблицы. Результаты вычисления приведены в табл. 57.

Таблица 57

Фактор	$\sum_{k=1}^{n} x_{ijk}$			Сумма $\sum_{i=1}^{q}\sum_{k=1}^{n} x_{ijk}$
	B_1	B_2	B_3	
A_1	12	9	12	33
A_2	9	15	21	45
A_3	15	15	21	51
A_4	27	27	21	75
Сумма $\sum_{i=1}^{p}\sum_{k=1}^{n} x_{ijk}$	63	66	75	204

Вычислим среднее арифметическое по каждой клетке $(\bar{x}_{ij.})$, а также по каждой строке $(\bar{x}_{i..})$ и по каждому столбцу $(\bar{x}_{.j.})$, для чего разделим полученные суммы (см. табл. 57) на соответствующее количество проб. Результат вычислений приводится в табл. 58.

Таблица 58

Фактор	B_1	B_2	B_3	Среднее
A_1	4	3	4	3,67
A_2	3	5	7	5,00
A_3	5	5	7	5,67
A_4	9	9	7	8,33
Среднее . . .	5,25	5,50	6,25	5,67

Общее среднее содержание свинца $\bar{x} = 5{,}67\%$. Таким образом, создается впечатление, что среднее содержание свинца растет с глубиной и с увеличением мощности жил.

Таким образом, простая группировка проб по глубине и мощности позволяет сделать предположение о существенном влиянии факторов A и B на содержание свинца в руде. Но подтвердится ли это дисперсионным анализом? Возможно, что изменение оценок средних с глубиной окажется случайным.

Чем ближе величина отношения дисперсии групповых средних к остаточной дисперсии, тем больше основания считать, что расхождение между средними является существенным, а не вызванным лишь действием неучтенных факторов или случайных причин.

Факторы A и B действуют изолированно (A, B) и совместно (AB). Рассмотрим эти действия, но прежде найдем отклонение содержания в каждой пробе от общей средней $(x_{ijk} - \bar{x})$, т. е. от 5,67%. Эти отклонения показаны в табл. 59.

Таблица 59

Фактор	$x_{ijk} - \bar{x}$		
	B_1	B_2	B_3
A_1	—4,67; —0,67; 0,33	—4,67; —2,67; —0,67	—3,67; —2,67; 1,33
A_2	—3,67; —3,67; —0,67	—3,67; —0,67; 2,33	—2,67; 2,33; 4,33
A_3	—4,67; —1,67; 4,33	—3,67; —3,67; 5,33	0,33; 1,33; 2,33
A_4	—0,67; 4,33; 6,33	—3,67; 4,33; 9,33	—1,67; 1,33; 4,33

Квадрат каждого отклонения $(x_{ijk} - \bar{x})^2$ приведен в табл. 60.

Сумма квадратов $\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n}(x_{ijk} - \bar{x})^2$ равна 459,94.

Найдем теперь сумму квадратов отклонений групповых средних $\bar{x}_{i..}$, условно характеризующих влияние фактора A от общей средней, т. е. от 5,67 (табл. 61).

Число проб в каждой из этих четырех строк 9. Сумму квадратов отклонений, условно связываемых с влиянием фактора A, мы получим, если умножим каждую из величин $(\bar{x}_{i..} - \bar{x})^2$ на число 9 и сложим эти произведения. Для нашего примера получим

$$nq\sum_{i=1}^{p}(\bar{x}_{i..} - \bar{x})^2 = 4{,}00 \cdot 9 + 0{,}45 \cdot 9 + 0 \cdot 9 + 7{,}08 \cdot 9 = 103{,}77.$$

Таблица 60

Фактор	$(x_{ijk}-\bar{x})^2$		
	B_1	B_2	B_3
A_1	21,81; 0,45; 0,11	21,81; 7,13; 0,45	13,47; 7,13; 1,77
A_2	13,47; 13,47; 0,45	13,47; 0,45; 5,43	7,13; 5,43; 18,75
A_3	21,81; 2,79; 18,75	13,47; 13,47; 28,41	0,11; 1,77; 5,43
A_4	0,45; 18,75; 40,07	13,47; 18,75; 87,05	2,79; 1,77; 18,75
Всего	152,28	223,36	84,30

Таблица 61

Фактор	Средняя групповая $\bar{x}_{i..}$	Отклонение от общей средней $\bar{x}_{i..}-\bar{x}$	Квадрат отклонения $(\bar{x}_{i..}-\bar{x})^2$
A_1	3,67	—2,00	4,00
A_2	5,00	—0,67	0,45
A_3	5,67	0	0
A_4	8,33	2,66	7,08

Определим сумму квадратов отклонений $np\sum_{j=1}^{q}(\bar{x}_{.j.}-\bar{x})^2$, условно характеризующих влияние фактора B, от общей средней 5,67 (табл. 62).

Таблица 62

Фактор	Средняя $\bar{x}_{.j.}$	Отклонение $\bar{x}_{.j.}-\bar{x}$	Квадрат отклонений $(\bar{x}_{.j.}-\bar{x})^2$
B_1	5,25	—0,42	0,18
B_2	5,50	—0,17	0,03
B_3	6,25	0,58	0,34

Число проб, в которых действует фактор B, в каждой группе равно 12. Сумма квадратов отклонений

$$np\sum_{j=1}^{q}(\bar{x}_{.j.}-\bar{x})^2 = 0{,}18\cdot 12 + 0{,}03\cdot 12 + 0{,}34\cdot 12 = 6{,}60.$$

Рассмотрим отклонения, вызванные комбинированным действием факторов AB.

Средние величины $\bar{x}_{ij.}$, возникшие в условиях комбинированного действия факторов AB, приведены в табл. 58. Если бы на этих средних величинах не отражалось действие неучтенных факторов, то об эффекте, возникшем под изолированным действием факторов, например A_1 и B_1, можно было бы судить по тому, что средняя при A_1 равна 3,67, а общая средняя 5,67. Эффект изолированного действия фактора A_1 равен 5,25— — 5,67 = — 0,42. Сумма действия этих факторов (—2,00) + (—0,42) = = — 2,42 равна эффекту изолированного действия факторов A и B, между тем как в случае комбинированного действия факторов A и B отклонение от общей средней (от 5,67) 4 — 5,67 = — 1,67. Отклонения от общей средней в случае комбинированного действия факторов A и B $(\bar{x}_{ij.}-\bar{x})$ сведены в табл. 63.

Таблица 63

Фактор	B_1	B_2	B_3
A_1	—1,67	—2,67	—1,67
A_2	—2,67	—0,67	1,33
A_3	—0,67	—0,67	1,33
A_4	3,33	3,33	1,33

Эффекты изолированного действия факторов A и B $(\bar{x}_{i..} - \bar{x}) + (\bar{x}_{.j.} - \bar{x})$ приведены в табл. 64.

Таблица 64

Фактор	B_1 (—0,42)	B_2 (—0,17)	B_3 (0,58)
A_1 (—2,00)	—2,42	—2,17	—1,42
A_2 (—0,67)	—1,09	—0,84	—0,09
A_3 (0)	—0,42	—0,17	0,58
A_4 (2,66)	2,24	2,49	3,24

В табл. 64 около символа каждого фактора в скобках показана величина двух средних — групповой $\bar{x}_{i..}$ или $\bar{x}_{.j.}$ и общей $\bar{x}$.

Сравним теперь друг с другом данные последних двух таблиц (табл. 63, 64). В первой из них указан эффект комбинированного действия факторов A и B, а во второй — эффект изолированного действия тех же факторов. Если из первого (эффекта комбинированного действия) отнять второй (эффект изолированного действия), то получим дополнительный эффект комбинированного действия соответствующих факторов. Например, для факторов A_1B_1 дополнительный эффект комбинированного действия составит

$$\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x} = (-1{,}67) - (-2{,}42) = 0{,}75.$$

Так же поступим и в отношении других комбинаций факторов; результаты вычислений сведены в табл. 65.

Таблица 65

Фактор	B_1	B_2	B_3
A_1	0,75	—0,50	—0,25
A_2	—1,58	0,17	1,42
A_3	—0,25	—0,50	0,75
A_4	1,09	0,84	—1,91

Квадраты этих чисел $(\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x})^2$ приведены в табл. 66.

Таблица 66

Фактор	B_1	B_2	B_3
A_1	0,5625	0,2500	0,0625
A_2	2,4964	0,0289	2,0164
A_3	0,0625	0,2500	0,5625
A_4	1,1881	0,7056	3,6481

Умножив эти величины на число проб, т. е. на 3, получим значения $n(\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x})^2$, сведенные в табл. 67.

Таблица 67

Фактор	B_1	B_2	B_3	Всего
A_1	1,6875	0,7500	0,1875	2,6250
A_2	7,4892	0,0867	6,0492	13,6251
A_3	0,1875	0,7500	1,6785	2,6250
A_4	3,5643	2,1168	10,9443	16,6254
Сумма . . .	12,9285	3,7035	18,8685	35,5005

Полученную в итоге сумму $n\sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x})^2$ округляем до 35,5.

Силу действия неучтенных (в том числе и случайных) факторов можно определить как разность между суммой квадратов отклонений всех проб от общей средней, с одной стороны, и суммой квадратов отклонений групповых средних, возникших под влиянием факторов A и B, — с другой.

Вычислим эту, интересующую нас разность:

$$459{,}94 - (103{,}77 + 6{,}60 + 35{,}50) = 314{,}07 \simeq 314.$$

Это очень важная величина. Для контроля подсчитаем ее другим, прямым путем по формуле

$$\sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{n}(x_{ijk} - \bar{x}_{ij.})^2,$$

исходя из отклонения отдельных проб от средней величины (см. табл. 57 и 58). Результаты вычисления отклонений сведены в табл. 68.

Таблица 68

Фактор	B_1	B_2	B_3
A_1	—3, 1, 2	—2, 0; 2	—2, —1, 3
A_2	—1, —1, 2	—3, 0, 3	—4, 1, 3
A_3	—4; —1, 5	—3, —3, 6	—1, 0, 1
A_4	—4, 1, 3	—7, 1, 6	—3, 0, 3

Возведем в квадрат эти отклонения (табл. 69).

Таблица 69

Фактор	B_1	B_2	B_3	Всего
A_1	9, 1, 4	4, 0, 4	4, 1, 9	36
A_2	1, 1, 4	9, 0, 9	16, 1, 9	50
A_3	16, 1, 25	9, 9, 36	1, 0, 1	98
A_4	16, 1, 9	49, 1, 36	9, 0, 9	130
Сумма . . .	88	166	60	314

Сумма 314 сошлась с ранее вычисленной.

Вычислим теперь соответствующие числа степеней свободы.

Общее число степеней свободы

$$k = npq - 1 = 36 - 1 = 35.$$

Число степеней свободы фактора A

$$k_A = p - 1 = 4 - 1 = 3.$$

Число степеней свободы фактора B

$$k_B = q - 1 = 3 - 1 = 2.$$

Число степеней свободы совместного действия факторов AB

$$k_{AB} = (p - 1)(q - 1) = 3 \cdot 2 = 6.$$

Число степеней свободы неучтенных факторов и случайных причин

$$pq(n - 1) = 4 \cdot 3(3 - 1) = 24.$$

Оценим величину каждой дисперсии:

общая дисперсия: $s^2 = 459{,}94 : 35 = 13{,}141$;

дисперсия фактора A : $s_1^2 = 103{,}77 : 3 = 34{,}590$;

дисперсия фактора B : $s_2^2 = 6{,}60 : 2 = 3{,}300$;

смешанная дисперсия факторов AB: $s_3^2 = 35{,}50 : 6 = 5{,}917$;

остаточная дисперсия: $s_4^2 = 314 : 24 = 13{,}083$.

Каждая из этих дисперсий является следствием своей причины изменчивости.

Сравнение групповых выборочных дисперсий с остаточной позволит проверить предположение о том, что соответствующие факторы (A, B и их комбинированное действие) не оказывает существенного влияния на среднее содержание свинца в пробах.

Выше было сказано, что чем больше отношение частной дисперсии к остаточной, тем больше вероятность предположения о существенном действии соответствующего фактора.

Отношение групповых дисперсий к остаточной:

$$F_A = \frac{34{,}590}{13{,}089} = 2{,}645,$$

$$F_B = \frac{3{,}300}{13{,}083} = 0{,}252,$$

$$F_{AB} = \frac{5{,}917}{13{,}083} = 0{,}453.$$

Сравним эти величины с допустимыми значениями $F_{0{,}05}$ при уровне значимости 0,05 и соответствующем числе степеней свободы (табл. 70) (см. приложение 25).

Таблица 70

Дисперсия	Число степеней свободы		F	$F_{0.05}$
	числителя	знаменателя		
Фактора A	3	24	2,615	3,01
Фактора B	2	24	0,252	3,40
Фактора AB	6	24	0,453	2,51

В этой таблице F в двух случаях меньше единицы, так как сравниваются не дисперсии двух выборок, а две разные дисперсии — групповая и остаточная. Поэтому сделанное выше замечание о том, что для вычисления F надо делить бóльшую дисперсию на меньшую, к данному случаю (многофакторный анализ) не относится.

Так как значения F оказались меньше допустимого, мы можем сделать следующий вывод: предположение об отсутствии влияния фактора глубины и фактора мощности на содержание свинца в руде не противоречит эмпирическим данным.

Это очень важный вывод. Без него мы непременно считали бы, что глубина и мощность влияют на распределение содержания. Однако факторный анализ показал, что различие в групповых средних вполне может быть вызвано чисто случайными колебаниями содержания.

Рассмотренные приемы как однофакторного, так и двухфакторного дисперсионного анализа основаны на предположении, что распределение изучаемой случайной величины согласуется с нормальным законом. Если в изучаемой совокупности нет оснований для принятия априорного предположения о нормальном распределении значений признака, то по имеющимся выборочным данным до применения дисперсионного анализа следует проверить гипотезу о возможности подобного допущения.

Пример. Одно ртутное месторождение имеет сложный штокверковкрапленный характер оруденения. Киноварь заполняет систему мельчайших жилок и образует неправильные вкрапления в песчаниках. Песчаники смяты в асимметричные брахиантиклинали. Южные крылья последних крутые, северные — пологие.

Опробование месторождения сопряжено с большими трудностями, так как киноварь легко выкрашивается, а оруденелый песчаник даже в пределах забоя имеет неодинаковую плотность. К тому же он трещиноват. Отборщикам проб не удается получить борозду опробования правильного сечения, а это приводит к большим ошибкам. Для того чтобы найти наиболее подходящий к условиям этого месторождения метод опробования, необходимо прежде всего выяснить, какие факторы существенно влияют на величину содержания ртути в пробах. По геологическим предположениям, факторы, влияющие на содержание киновари в пробе, следующие:

1. Стратиграфическая глубина (месторождение многоярусное, при этом среднее содержание ртути по разным ярусам как будто различное, хотя детально изучен пока только один ярус).

2. Крепость руды (руды более мягкие, по-видимому, богаче руд более крепких, но эмпирический материал по этому фактору не собран).

3. Крупность кусков руды после отпалки (руда мелкокусковатая, возможно, богаче руды крупнокусковой, но и по этому фактору эмпирический материал не собран).

4. Тип оруденения; оруденение представлено жилками или вкраплениями (в пространстве эти типы не отделены друг от друга, хотя в отдельных местах и наблюдается преобладание жилок над вкраплением или преобладание вкраплений над жилками, при этом жилки значительно богаче вкраплений, хотя точных данных по этому фактору не имеется).

5. Положение участка, т. е. места взятия проб относительно оси брахиантиклинали (структурный фактор).

6. Способ опробования.

Из-за отсутствия фактического материала, который характеризовал бы первые четыре фактора, проверить гипотезы относительно них не представляется возможным. По последним же двум факторам удалось собрать 60 проб, обработанных и проанализированных одним и тем же способом в одной и той же лаборатории.

Содержание ртути в условных единицах по этим пробам приведено в (исходной) табл. 71.

Таблица 71

Фактор	B_1	B_2	B_3
A_1	0, 1, 5, 6	0, 3, 4, 9	3, 4, 5, 20
A_2	0, 6, 7, 7	9, 10, 10, 11	8, 12, 14, 14
A_3	6, 7, 9, 10	7, 8, 11, 11	10, 13, 13, 16
A_4	6, 9, 10, 11	8, 9, 12, 15	8, 9, 10, 13
A_5	9, 9, 11, 11	8, 11, 12, 13	7, 9, 14, 18

В этой таблице буквой A обозначен фактор метода опробования (A_1 — бороздовый метод, A_2 — точечный, A_3 — горстевой I, A_4 — горстевой II, A_5 — валовый). Борозды имели сечение 5×10 *см*, длину около 2 *м*. Точечная проба состояла из 10 кусков руды весом 300—500 *г* каждый, отбитых в забое по двумя линиям, в каждой из которых было 5 точек (линии ориентированы перпендикулярно направлению жилок киновари). При горстевом I методе опробования каждая проба состояла из 30 порций руды весом приблизительно по 400 *г* каждая (порции брались из отбитой руды около забоя). Метод горстевой II характеризуется также 30 порциями, но вес порции 200 *г*. В валовую пробу отбиралась каждая пятая вагонетка с рудой из числа 10—25 всех вагонеток с участка опробования; затем проба сокращалась до 10—15 *кг*.

Структурный фактор обозначен B.

В связи с тем, что какие-либо сведения о характере распределения содержаний ртути отсутствовали, была проверена гипотеза о непротиворечивости нормального распределения выборочным данным. Проверка показала, что гипотеза приемлема. Этот результат дает возможность уверенного применения дисперсионного анализа.

Применив описанные выше приемы, вычислим суммы содержаний, которые приведены в табл. 72

Таблица 72

Фактор	B_1	B_2	B_3	Всего
A_1	12	16	32	60
A_2	20	40	48	108
A_3	32	36	52	120
A_4	36	44	40	120
A_5	40	44	48	132
Сумма . . .	140	180	220	540

Разделив каждое из чисел, приведенных в этой таблице, на соответствующее количество проб, получим среднее содержание для каждой из 15 комбинаций факторов A и B, а также групповые средние и общую среднюю. Результаты этих вычислений приведены в табл. 73.

Как видно из данных табл. 73, оценка среднего содержания ртути для всех 60 проб (общее среднее) равна 9.

Определим отклонение содержания каждой пробы от общей средней, т. е. от 9 (результаты вычислений сведены в табл. 74).

Возведем каждое из этих отклонений в квадрат (табл. 75)

Таблица 73

Фактор	B_1	B_2	B_3	Среднее
A_1	3	4	8	5
A_2	5	10	12	9
A_3	8	9	13	10
A_4	9	11	10	10
A_5	10	11	12	11
Среднее	7	9	11	9

Таблица 74

Фактор	B_1	B_2	B_3
A_1	−9, −8, −4, −3	−9, −6, −5, 0	−6, −5, −4, 11
A_2	−9, −3, −2, −2	0, 1, 1, 2	−1, 3, 5, 5
A_3	−3, −2, 0, 1	−2, −1, 1, 2	1, 4, 4, 7
A_4	−3, 0, 1, 2	−1, 0, 3, 6	−1, 0, 1, 4
A_5	0, 0, 2, 2	−1, 2, 3, 4	−2, 0, 5, 9

Таблица 75

Фактор	B_1	B_2	B_3	Всего
A_1	81, 64, 16, 9	81, 36, 25, 0	36, 25, 16, 121	510
A_2	81, 9, 4, 4	0, 1, 1, 4	1, 9, 25, 25	164
A_3	9, 4, 0, 1	4, 1, 1, 4	1, 16, 16, 49	106
A_4	9, 0, 1, 4	1, 0, 9, 36	1, 0, 1, 16	78
A_5	0, 0, 4, 4	1, 4, 9, 16	4, 0, 25, 81	148
Сумма . .	304	234	468	1006

Вычислим сумму квадратов отклонений групповых средних, характеризующих влияние фактора A, от общей средней, т. е. от 9 (табл. 76).

Таблица 76

Фактор	$\bar{x}_{i..}$	$\bar{x}_{i..} - \bar{x}$	$(\bar{x}_{i..} - \bar{x})^2$	nq	$nq\,(\bar{x}_{i..} - \bar{x})^2$
A_1	5	−4	16	12	192
A_2	9	0	0	12	0
A_3	10	1	1	12	12
A_4	10	1	1	12	12
A_5	11	2	4	12	48
Сумма . . .					264

Сделаем такое же вычисление для фактора B (табл. 77).

Определим отклонение от общей средней для комбинации факторов AB, т. е. для $\bar{x}_{ij.} - \bar{x}$.

Таблица 77

Фактор	$\bar{x}_{.j.}$	$\bar{x}_{.j.} - \bar{x}$	$(\bar{x}_{.j.} - \bar{x})^2$	np	$np\,(\bar{x}_{.j.} - \bar{x})^2$
B_1	7	—2	4	20	80
B_2	9	0	0	20	0
B_3	11	2	4	20	80
Сумма . . .					160

Средняя для каждой из 15 комбинаций $\bar{x}_{ij.}$ приведена была выше (см. табл. 73), а общая средняя $\bar{x} = 9$. Разности этих двух средних приведены в табл. 78.

Таблица 78

Фактор	B_1	B_2	B_3
A_1	—6	—5	—1
A_2	—4	1	3
A_3	—1	0	4
A_4	0	2	1
A_5	1	2	3

Далее найдем эффект совместного действия факторов A и B. Для этого сложим разности между соответствующими групповыми средними и общей средней, т. е.

$$(\bar{x}_{i..} - \bar{x}) + (\bar{x}_{.j.} - \bar{x}).$$

Для комбинации A_1B_1, например, имеем

$$(5 - 9) + (7 - 9) = (-4) + (-2) = -6.$$

Эти разности $(\bar{x}_{i..} - \bar{x}$ и $\bar{x}_{.j.} - \bar{x})$ были вычислены раньше (см. табл. 76 и 77). Результаты сложения их приведены в табл. 79

Таблица 79

Фактор	B_1	B_2	B_3
A_1	—6	—4	—2
A_2	—2	0	2
A_3	—1	1	3
A_4	—1	1	3
A_5	0	2	4

Вычисляем дополнительный эффект комбинированного действия факторов A и B (см. табл. 78 и 79). Для комбинации A_1B_2, например, имеем $(-5) - (-4) = -1$.

Результаты вычисления приведены в табл. 80.

Возведем в квадрат числа дополнительного эффекта факторов A и B (табл. 81).

Умножим эти квадраты на число проб для каждой комбинации факторов A и B (табл. 82).

Таблица 80

Фактор	B_1	B_2	B_3
A_1	0	—1	1
A_2	—2	1	1
A_3	0	—1	1
A_4	1	1	—2
A_5	1	0	—1

Таблица 81

Фактор	B_1	B_2	B_3
A_1	0	1	1
A_2	4	1	1
A_3	0	1	1
A_4	1	1	4
A_5	1	0	1

Таблица 82

Фактор	B_1	B_2	B_3	Всего
A_1	0	4	4	8
A_2	16	4	4	24
A_3	0	4	4	8
A_4	4	4	16	24
A_5	4	0	4	8
Сумма . . .	24	16	32	72

Определим силу действия неучтенных факторов B, как сумму квадратов отклонения под влиянием этих факторов. Для этого из суммарного эффекта выявленных и невыявленных факторов, представленного суммой квадратов отклонений каждой пробы от общего среднего, которая равна 1006, вычтем суммарный эффект одних только выявленных факторов, т. е. сумму 264 + 160 + 72. Получим

$$1006 - (264 + 160 + 72) = 510.$$

Сумму квадратов отклонений под влиянием неучтенных факторов вычислим для контроля другим путем, исходя из отклонений отдельных проб (см. табл. 74) от средней величины (см. табл. 78). Поскольку в одной и той же комбинации факторов разница в содержании не может быть объяснена действием данных, скомбинированных факторов, мы ее можем объяснить чем-то иным, т. е. действием неучтенных факторов (отклонения от собственной средней приведены в табл. 83).

Возведем эти отклонения в квадрат (табл. 84).

Получим число 510, как и в первом случае.

Определим число степеней свободы для каждого фактора и для их комбинации.

Общее число степеней свободы

$$k = qn - 1 = 60 - 1 = 59.$$

Таблица 83

Фактор	B_1	B_2	B_3
A_1	—3, —2, 2, 3	—4, —1, 0, 5	—5, —4, —3, 12
A_2	—5, 1, 2, 2	—1, 0, 0, 1	—4, 0, 2, 2
A_3	—2, —1, 1, 2	—2, —1, 1, 2	—3, 0, 0, 3
A_4	—3, 0, 1, 2	—3, —2, 1, 4	—2, —1, 0, 3
A_5	—1, —1, 1, 1	—3, 0, 1, 2	—5, —3, 2, 6

Таблица 84

Фактор	B_1	B_2	B_3	Всего
A_1	9, 4, 4, 9;	16, 1, 0, 25	25, 16, 9, 144	262
A_2	25, 1, 4, 4	1, 0, 0, 1	16, 0, 4, 4	60
A_3	4, 1, 1, 4	4, 1, 1, 4	9, 0, 0, 9	38
A_4	9, 0, 1, 4	9, 4, 1, 16	4, 1, 0, 9	58
A_5	1, 1, 1, 1	9, 0, 1, 4	25, 9, 4, 36	92
Сумма . .	88	98	324	510

Число степеней свободы фактора A

$$k_A = p - 1 = 5 - 1 = 4.$$

Число степеней свободы фактора B

$$k_B = q - 1 = 3 - 1 = 2.$$

Число степеней свободы совместного действия факторов AB

$$k_{AB} = (p - 1)(q - 1) = 4 \cdot 2 = 8.$$

Число степеней свободы неучтенных факторов и случайных причин

$$pq(n - 1) = 5 \cdot 3(4 - 1) = 45.$$

Вычислим величины дисперсий.

Общая дисперсия

$$s^2 = 1006 : 59 = 17{,}06.$$

Дисперсия фактора A

$$s_1^2 = 264 : 4 = 66{,}00.$$

Дисперсия фактора B

$$s_2^2 = 160 : 2 = 88{,}00.$$

Смешанная дисперсия факторов AB

$$s_3^2 = 72 : 9 = 9{,}00.$$

Остаточная дисперсия

$$s_4^2 = 566 : 45 = 12{,}58.$$

Отношение общей дисперсии и групповых дисперсий к остаточной дисперсии:

$$F_A = \frac{66{,}00}{12{,}58} = 5{,}25;$$

$$F_B = \frac{80{,}00}{12{,}58} = 6{,}36;$$

$$F_{AB} = \frac{9{,}00}{12{,}58} = 0{,}72.$$

Сопоставим полученные отношения эмпирических дисперсий с соответствующими допустимыми значениями, которые возьмем из таблицы (приложения 24 и 25) для данного числа степеней свободы (табл. 85)

Таблица 85

Дисперсия	Число степеней свободы		F	$F_{0,05}$	$F_{0,01}$
	числителя	знаменателя			
A	4	45	5,25	2,58	3,77
B	2	45	6,36	3,21	5,11
AB	8	45	0,72	2,15	2,94

Ввиду того что F_A и F_B больше, чем $F_{0,01}$, влияние этих факторов следует считать доказанным. Следовательно, содержание ртути в пробах существенно зависит от структурного фактора (от положения участка опробования относительно оси складки) и метода опробования. Бороздовое опробование показывает весьма заниженное, по сравнению с результатами валового опробования, содержание ртути. Горстевое опробование также занижает содержание, но незначительно.

Примеры двухфакторного дисперсионного анализа заняли очень много места и потребовали составления большого числа таблиц, но на практике все расчеты обычно помещают в одной большой таблице довольно сложного вида. Новичку в области математической статистики трудно разобраться в большой и громоздкой таблице факторного анализа, поэтому мы и расчленили ее на множество малых и простых таблиц.

Трехфакторный, четырехфакторный анализ еще более сложен. Его описание дано в книге К. Браунли (1949).

VIII. РАЗМАХ ВЫБОРКИ И КРАЙНИЕ ЧЛЕНЫ ВАРИАЦИОННОГО РЯДА

Имеется ряд замеров, наблюдений или анализов $x_1, x_2, x_3, \ldots, x_n$, расположенных в порядке возрастания значений, т. е. $x_1 \leqslant x_2 \leqslant x_3 \ldots \ldots \leqslant x_n$. В этом ряду x_1 имеет наименьшее значение (иногда таких наименьших значений бывает два, три или больше, причем все они равны между собой), x_n — наибольшее значение (иногда таких наибольших результатов также бывает несколько, причем все они тоже равны друг другу). Величина x_1 называется наименьшим, или минимальным, членом вариационного ряда, а величина x_n — наибольшим, или максимальным, членом ряда. Оба они называются крайними, или экстремальными, членами ряда. Иногда вместо x_n и x_1 пишут соответственно x_{max} и x_{min}.

Разность $x_n - x_1$ называется выборочным размахом или просто размахом (обозначим его через w).

Значение размаха и крайних членов ряда в статистике очень велико (Лукомский, 1955; Гумбель, 1958). Большое значение эти величины играют и при решении практических задач. Например, при проектировании горных комбайнов для Кизеловского угольного бассейна конструкторы столкнулись с необходимостью определить типоразмеры высоты рабочего органа комбайна, которые зависят от размахов мощности пластов, угля по лавам. Размах мощности угольных пластов в Кизеловском бассейне очень велик. В пределах одной, произвольно взятой лавы мощность пласта может колебаться от 0 до 5 *м*, а чаще всего от 0,5 до 2,0 *м*. Расчеты показали, что высота рабочего органа для одного типоразмера комбайна не может изменяться более чем в два раза. Поэтому в большинстве лав не

сможет работать ни один комбайн. Рассчитывать же на то, что в одной и той же лаве будет работать несколько комбайнов, нельзя по ряду важных причин.

В Кизеловском бассейне наблюдаются угольные пласты с относительно малыми колебаниями мощности, но такие именно пласты, в каких районах и как часто они встречаются — это предстоит выяснить. Таким образом, изучение размахов мощности пластов угля в Кизеловском бассейне имеет важное значение.

При опробовании и подсчете запасов месторождений драгоценных, редких и цветных металлов, алмазов и некоторых других полезных ископаемых геологи испытывают большие затруднения с учетом выдающихся, очень богатых (ураганных) содержаний. Известно около 20 методов учета таких содержаний при подсчете запасов, но ни один из них не дает хороших результатов. Между тем проблема ураганных содержаний могла бы решаться с помощью теории распределения максимального члена вариационного ряда.

Оценка качества руды ряда месторождений во многом зависит от содержаний вредных примесей (серы и фосфора в железной руде, золы и серы в угле и т. д.), а закономерности распределения этих примесей, по-видимому, можно изучать с помощью статистики минимальных членов вариационного ряда.

С крайними значениями признака в выборке геолог сталкивается и тогда, когда вариационный ряд построен по результатам определения ошибок анализа проб, по размерам драгоценных камней, золотых самородков, кристаллов пьезокварца, слюды, на нагрузкам, испытываемым буровыми снарядами, по температурам горной породы на глубине, по результатам измерения длины волокна асбеста и т. п.

Если взять очень большое число случайных выборок равного объема из генеральной совокупности, подчиняющейся закону нормального распределения, и выписать по каждой выборке $x_{\min}$, $x_{\max}$ и ω, то получим три новые совокупности: минимальных членов, максимальных членов и размахов. Для каждой из этих совокупностей в статистике теоретически выведены законы распределения (они отличаются от нормального распределения), а по этим законам можно решать ряд практических задач.

В общем виде закон распределения размахов в выборке из n наблюдений над случайной величиной ξ можно представить следующей формулой:

$$P(\omega_n \leqslant t) = n \int_{-\infty}^{+\infty} \varphi(x) [F(x+t) - F(x)]^{n-1} dx.$$

Здесь левая часть равенства представляет собой вероятность того, что при объеме выборочной совокупности, равном n, размах ω будет меньше некоторого числа t, выраженного в единицах среднего квадратического отклонения σ случайной величины ξ, $\varphi(x)$ — дифференциальная функция распределения случайной переменной ξ, а $F(x+t)$ и $F(x)$ — значения функции распределения той же случайной переменной в точках x и $x+t$.

В данном случае величина ξ заменена случайной величиной ξ^1, которая распределена нормально с параметрами 0,1.

Такая замена позволяет, используя плотность

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

с любой вероятностью P, найти p — процентные нормы (квантили) для отношения $\frac{\omega_n}{\sigma}$ при объеме выборки, равном n. Эти нормы вычислены для случая, когда значение признака ξ распределено нормально (приложение 16).

Поскольку число выборок может быть довольно большим, имеет смысл говорить о среднем размахе, о средней величине отношения размаха к среднему квадратическому отклонению изучаемой случайной переменной и о соответствующих им оценках.

В приложении 16 приведено математическое ожидание отношения $\frac{w_n}{\sigma_x}$, обозначенное через α_n, среднее квадратическое отклонение величины α, обозначенное как β_n, отношение $\frac{\beta_n}{\alpha_n} = \gamma_n$, являющееся коэффициентом вариации величины α, и p-процентные пределы для отношения $\frac{w_n}{\sigma}$ при разных уровнях вероятности.

Все эти показатели даны для объемов выборки n из нормальной совокупности.

Пример. При исследовании мощности угольных пластов Кизеловского бассейна (Шарапов, 1964) были собраны данные маркшейдерских замеров по очень большому числу лав. По одному из четырех исследованных пластов (пласт № 5) замеры сделаны в 42 лавах. Число замеров в лаве колебалось от одного до 34, а в среднем составляло 8—9. В четырех лавах было сделано по 8 замеров. В одной из них (лава 51-бис были получены следующие результаты измерения мощности в *м*: 0,70; 0,70; 0,75; 0,80; 0,80; 0,85; 1,00; 1,10. По этим данным вычислено, что средняя мощность пласта 0,84 *м*, минимальная 0,70 *м*, максимальная 1,10 *м*; размах $w = 0{,}40$ *м*. Оценка среднего квадратического отклонения, вычисленная по формуле

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

равна 0,143. Теоретические значения (приложение 16): $\alpha = 2{,}847$, $\beta = 0{,}820$, $\gamma = 0{,}288$.

Размах w, как и среднее квадратическое отклонение σ, является характеристикой степени рассеяния величины x. Поэтому между w и σ имеется связь; по величине w можно оценить величину σ.

Частное $\frac{w}{\alpha}$ является несмещенной оценкой величины σ, т. е. $\sigma = \frac{w}{\alpha}$. При $n \leqslant 10$ эта оценка довольно точна. При $n \leqslant 20$ ее точность снижается, а при $n > 20$ точность недостаточна для решения большинства практических задач.

Расхождение в значениях оценок для σ, вычисленных прямым способом (по классической формуле) и по размаху, зависит от особенностей распределения.

Определение оценки $\hat{\sigma}$ по величине w значительно проще, чем по классической формуле, но для отдельных выборок при этом возможны большие ошибки. Обоснование этого заключения приведено А. Хальдом (1956).

Оценка $\hat{\sigma}$, полученная по серии выборок, исходя из размаха, дает более точные результаты. Для этого вычисления необходимо проделать следующее.

Всю совокупность наблюдений N разобьем случайным образом на k групп по n наблюдений в каждой группе ($kn = N$), вычислим по каждой группе размах w и определим их среднее арифметическое $\bar{w}$. После чего найдем оценку $\hat{\sigma}$ для совокупности N наблюдений по формуле

$$\hat{\sigma} = \frac{\bar{w}}{\alpha_n}.$$

Объем групп в некоторых случаях может быть и неодинаковым, но эти различия не должны быть большими.

Этот метод применим для n, находящегося в пределах от 5 до 10, лучше всего равного 8.

Пример. Для примера возьмем 64 замера мощностей, сделанные в шахте № 6 «Капитальная» Кизеловского района в 8 лавах, по пласту № 11. Разобьем эти замеры на 8 групп по 8 замеров в каждой. Результаты этой разбивки приведены в табл. 86.

Таблица 86

№ группы	Мощность, м	Размах w, м
1	1,30; 1,10; 1,00; 0,73; 1,20; 1,05; 0,90; 1,05	0,57
2	1,30; 1,00; 0,90; 0,87; 1,10; 1,02; 1,10; 0,95	0,43
3	0,96; 0,90; 0,95; 1,00; 0,90; 1,20; 1,00; 0,80	0,40
4	1,00; 0,98; 0,82; 1,00; 1,08; 1,07; 1,02; 1,00	0,26
5	1,00; 1,10; 0,80; 0,88; 0,80; 1,00; 0,90; 0,87	0,30
6	1,00; 0,95; 1,00; 0,90; 0,90; 1,20; 1,14; 1,00	0,30
7	1,02; 1,10; 0,90; 0,95; 0,90; 0,80; 0,85; 0,80	0,30
8	0,97; 0,95; 0,95; 1,03; 0,95; 1,00; 0,90; 0,90	0,13

Сумма размахов равна 2,69. Среднее арифметическое размахов

$$\overline{w} = \frac{2{,}69}{8} = 0{,}371.$$

По приложению 16 для $n = 8$ находим

$$\alpha_8 = 2{,}847.$$

По упрощенной формуле получаем

$$\hat{\sigma} = \frac{0{,}371}{2{,}847} = 0{,}130.$$

С целью сравнения вычислим оценку s для σ обычным способом:

$$\frac{\sum_{i=1}^{n}(x-\overline{x})^2}{n-1} = \sqrt{\frac{0{,}906}{63}} = 0{,}120.$$

Величина $\hat{\sigma} = 0{,}130$, вычисленная по размахам, очень близка к величине $s = 0{,}120$, вычисленной обычным способом.

Приемлемость упрощенного метода для вычисления оценки $\hat{\sigma}$ (по размахам) обусловлена нормальным распределением признака в генеральной совокупности.

Пример. Распределение мощностей угольных пластов на многих шахтах Кизеловского бассейна довольно близко к нормальному. Так, выборочное распределение по данным 64 замеров приближается к нормальному распределению. В табл. 87 показано распределение мощностей в этой выборке.

Нанесение квантилей на вероятностную бумагу позволяет констатировать близость данного распределения к нормальному (рис. 47, линия *1*). Попутно покажем еще один способ нанесения квантилей на вероятностную бумагу. На этой бумаге очень малые (меньше 1%) и очень большие (более 99%) вероятности показать нельзя. Поэтому мы можем кумулировать частости не только в прямом (снизу вверх), но и в обратном (сверху вниз) порядке. Для обратного порядка на графике получены точки, по которым проведена линия *2*.

Таблица 87

Мощность, м	Средняя мощность по интервалу, м	Число замеров	Накопленное число замеров	Накопленная частость, %	Квантили, %	
					прямая	обратная
0,65—0,75	0,7	1	1	1,56	2,845	—
0,75—0,85	0,8	7	8	12,5	3,850	98,4
0,85—0,95	0,9	21	29	45,4	4,884	87,6
0,95—1,05	1,0	22	51	79,5	5,824	54,7
1,05—1,15	1,1	8	59	92,3	6,426	20,3
1,15—1,25	1,2	3	62	96,9	6,866	7,81
1,25—1,35	1,3	2	64	100,0	—	3,13
Всего		64				

Эта линия с осью абсцисс составляет угол $90° + \varphi$, где φ — угол между линией *1* и осью абсцисс.

Обратную кумуляцию можно показать и иначе. Для этого на оси абсцисс нанесем показания в обратном порядке (числа в скобках). Тогда получим линию *3*, параллельную линии *1*.

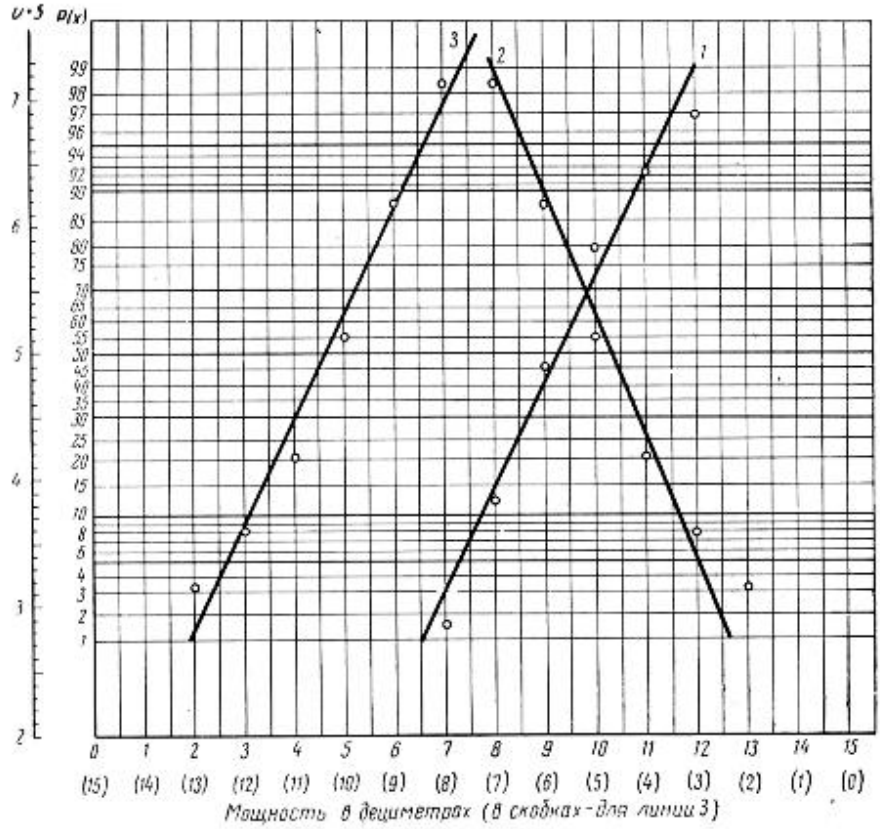


Рис. 47. Распределение мощности угольных пластов Кизеловского бассейна по 64 замерам

Линии *2* или *3* (кроме линии *1*) заставляют «работать» крайнюю (максимальную) квантиль (в данном случае квантиль, отвечающую мощности 1,3 *м*). При обычном способе применения вероятностной бумаги эта квантиль не используется.

Таким образом, при проведении линии *1* «не работает» самая нижняя квантиль, а у линий *2* и *3* — самая верхняя. При проведении обеих

линий, т. е. линий *1* и *2* или *1* и *3*, все квантили «работают». В тех случаях, когда с крайними квантилями связаны большие частости, проведение двух линий предохранит нас от значительной ошибки.

Проанализировав изображенное на рис. 47, можно сделать вывод, что распределение мощности угольных пластов близко к нормальному, и поэтому определение оценки среднего квадратичного отклонения мощности пластов по размаху оказывается для Кизеловского бассейна возможным.

Геолога иногда интересует, с какой вероятностью можно встретить величину размаха в заданных пределах. В равной степени ему часто необходимо узнать величину размаха, которая встретится с заданной вероятностью. Ответ на этот вопрос можно найти, использовав таблицу квантилей (приложение 16), составленную, как уже отмечалось, для случая нормального распределения признака в генеральной совокупности. Покажем это на материале приведенного выше примера (пример с лавой 51-бис).

Пример. Если нас удовлетворяет вероятность 95%, то p-процентный предел (квантиль) равен 4,29, а максимальный размах для $n = 8$ составит $4{,}29\sigma$, т. е. $4{,}29 \cdot 0{,}143 = 0{,}613$ *м*. Для вероятности 99% размах не превзойдет величины $4{,}99\sigma$, или 0,714 *м*, и т. д.

Можно решать и обратную задачу — по величине размаха для данного n найти соответствующую ему вероятность.

В приведенном примере размах w равен 0,4 *м*. Он в 2,8 раза больше величины σ. Вероятность этого события (приложение 16) для $n = 8$ близка к 50%.

В геологической практике приходится решать важные задачи, связанные с вероятностью встречи максимального или минимального члена в выборке. Так, например, сейсмолога интересует вероятность наиболее сильного землетрясения в том или ином районе, рудничного геолога — вероятность появления максимальной ошибки в анализах проб или вероятность максимально высокого содержания металла в руде суточного потока, геолога-разведчика — вероятность максимально высокого содержания полезного компонента в серии проб и т. д.

В оценке качества слюды играет роль наименьшее пробивное напряжение электрического тока, в оценке качества кристаллов, предназначенных для производства оптических изделий, — наименьший (критический) размер бездефектной области, в изучении подземных вод иногда важно знать наименьшую глубину их залегания и т. д.

Знание этих вероятностей дает возможность делать научные прогнозы изучаемых явлений, а иногда и управлять последними.

Закон распределения как максимального, так и минимального членов можно сформулировать следующим образом.

Пусть объем выборки равен n. При этом будем считать, что члены выборки не зависят друг от друга. Величина x_i любого i-того члена выборки случайная. Функцию распределения для любого i-того члена выборки обозначим как $P(\xi_i < x) = P(x)$.

Предположим, что

$$\zeta_n = \min(\xi_1, \xi_2, \ldots, \xi_n).$$

Нужно найти вероятность того, что ζ_n будет не более некоторой заданной величины x, т. е. нужно найти вероятность неравенства $\zeta_n \geqslant x$.

Так как мы имеем n членов, то для того чтобы величина ζ_n была не меньше x при любом i ($i = 1, 2, \ldots, n$), необходимо выполнить неравенство $\xi_i \geqslant x$. При этом

$$P(\xi_n \geqslant x) = 1 - P(\xi_i < x) = 1 - P(x).$$

По теореме умножения вероятность одновременного выполнения всех неравенств $\xi_i \geqslant x$ может быть выражена в виде

$$P(\zeta_n \geqslant x) = P(\xi_1 \geqslant x;\ \xi_2 \geqslant x, \ldots,\ \xi_n \geqslant x) = \\ = [1 - P(x)]^n.$$

Отсюда

$$P(\zeta_n < x) = 1 - [1 - P(x)]^n.$$

Если обозначить минимальный член в выборке объема n через $x_{\min} = \zeta_n$, то его распределение будет соответствовать приведенной формуле. Это и есть общий вид закона распределения минимального члена в выборке объема n.

Подобно распределению минимального члена $x_{\min}$, максимальный член $x_{\max}$ выборки из n членов имеет распределение, определяемое равенством

$$P(x_{\max} < x) = [P(x)]^n.$$

Последнее выражение показывает вероятность совместного выполнения неравенств $\xi_1 < x$; $\xi_2 < x$; . . .; $\xi_n < x$, что равносильно неравенству $x_{\max} < x$.

Для функции распределения $x_{\max}$

$$P(x_{\max} < x) = [P(x)]^n$$

можно найти квантили $x_{\max, p}$ из уравнения

$$P(x_{\max} < x_{\max, p}) = P,$$

или $[P(x)]^n = P$ при $x = x_{\max, p}$.

В результате получаем

$$P(x) = P^{\frac{1}{n}} = P_1 \text{ при } x = x_{\max, p}$$

и

$$x_{\max, p} = x_{p_1}.$$

Если взять $P = 0{,}95$, то значения P_1 в зависимости от n будут следующими:

n	$P_1 = 0{,}95^{\frac{1}{n}}$
1	0,95
2	0,975
5	0,990
10	0,995
51	0,999
102	0,9995

Эта таблица показывает, что, например, 95%-ный квантиль для наибольшего из 51 члена выборки равен 99,9%-ному квантилю для распределения каждого из наблюдений.

В приложении 17 приведены значения u p-процентных норм для отклонения максимального члена $x_{\max}$ выборки из нормальной совокупности от центра распределения.

Для $u_{(51)\,0{,}95} = u_{0,\,999} = 3{,}09$ (в таблице есть значения и для $n = 50$ и $n = 60$, но нет для $n = 51$; путем интерполяции получаем $u_{0{,}999} = 3{,}09$). Это означает, что с вероятностью 0,95 максимальный член выборки объема $n = 51$ будет меньше, чем $\bar{x} + 3{,}09\sigma$.

Пример. Среднее арифметическое содержание металла в 80 пробах $\bar{x} = 2{,}34\%$, а оценка среднего квадратического отклонения $s = 1{,}17$. С вероятностью 0,99 мы хотим узнать отклонение максимального члена, т. е. самого высокого содержания, от среднего арифметического. В при-

ложении 17 находим, что величина этого отклонения не превысит $3{,}67\sigma$, т. е., заменив σ оценкой s, получим $3{,}67 \cdot 1{,}17 = 4{,}29\%$. Так как среднее арифметическое равно 2,34%, то самое высокое содержание в пробе будет $4{,}29 + 2{,}34 = 6{,}63\%$. Если же фактическое содержание в пробе составило, например, 7,14%, то это означает, что (если распределение нормальное) подобное содержание может встретиться реже, чем в одной пробе из ста (при $P = 1 - 0{,}99$). В противном случае распределение нельзя считать нормальным.

По приведенной здесь степенной формуле можно вычислить математическое ожидание максимального члена и его стандарт. Для некоторых значений n эти величины будут следующими (табл. 88).

Таблица 88

Объем выборки n	Среднее значение $M_{x_{max}}$	Среднее квадратическое отклонение $\sigma_{x_{max}}$
2	0,56	0,83
5	1,16	0,67
10	1,54	0,59
20	1,87	0,53
60	2,32	0,45
100	2,51	0,13
200	2,75	0,40
500	3,04	0,37
1000	3,24	0,35

Данные таблицы говорят о том, что если мы возьмем очень много выборок равного объема ($n = 2$; $n = 5$ и т. д.) из нормальной генеральной совокупности и выпишем для каждой выборки значение одного самого большого члена, то получим совокупность максимальных членов. В этой совокупности мы можем вычислить среднее значение максимального члена и его среднее квадратическое отклонение. При этом обе величины берутся в единицах среднего квадратического отклонения σ признака в генеральной совокупности.

Во всех предшествующих рассуждениях заданное значение являлось критическим, т. е. служило то верхней, то нижней границей. В зависимости от того, ограничена она или нет, законы распределения крайних членов будут иметь некоторые различия.

Рассмотрим случай с ограниченной критической величиной.

Допустим, что величина каждого члена выборки будет больше величины a, но меньше $a + h$ (здесь $h > 0$). Тогда

$$P(x_{min} < a) = P_a = 0, \text{ а } P(a + h) > 0.$$

Пусть далее положительная величина h мала. Тогда

$$P(a + h) = (c + \varepsilon_h)\, h,$$

где $c > 0$ и $\varepsilon_h \to 0$ при $h \to 0$.

Вероятность же

$$P(x_{min} < a + h) = 1 - [1 - P(a + h)]^n$$

будет близка к единице при достаточно большом n, так как

$$1 - P(a + h) < 1 \text{ и } [1 - P(a + h)]^n \to 0 \text{ (при } n \to \infty).$$

Возьмем $h = \frac{t}{n}$, где t — постоянное положительное число. Для этого случая

$$P\left(a + \frac{t}{n}\right) = (c + \varepsilon_h)\frac{t}{n},$$

здесь $\varepsilon_h = \varepsilon_{\frac{t}{n}} \to 0$ (при $n \to 0$)

Из последних формул получим

$$P\left(x_{\min} < a + \frac{t}{n}\right) = 1 - [1 - (c + \varepsilon_h)\, t]^n \simeq 1 - e^{-ct}.$$

Приняв $a + \frac{t}{n} = x$ и $cn = \frac{1}{\nu}$ $(\nu > 0)$ при достаточно большом n, получим

$$P(x_{\min} < x) \simeq 1 - e^{-\frac{(x-a)^\alpha}{\nu}} \text{ (при } x > a)$$

и

$$P(x_{\min} < x) \simeq 0 \qquad \text{при } (x < 0).$$

Таким образом, при больших выборках будет иметь место показательный закон распределения.

Если вместо $P(a + h) = (c + \varepsilon_h)\, h$ взять

$$P(a + h) = (c + \varepsilon_h)\, h^\alpha,$$

где $c > 0$, $\alpha > 0$ и $\varepsilon_h \to 0$ (при $h \to 0$),
то закон распределения минимального члена будет следующим:

$$P(x_{\min} < x) = 1 - e^{-\frac{(x-a)^\alpha}{\nu}} \text{ (при } x > a)$$

и

$$P(x_{\min} < x) = 0 \quad (\text{при } x < a).$$

Если $b = -a$, то для максимального члена получим следующий закон распределения:

$$P(x_{\max} < x) = 1 - P(x_{\min} < -x) \simeq e^{-\frac{(b-x)^\alpha}{\nu}} (\text{при } x < b)$$

и

$$P(x_{\max} < x) = 1 \quad (\text{при } x > b).$$

Напомним, что приведенные здесь законы распределения минимального и максимального членов относятся лишь к случаю, когда значения величины x лежат в отрезке a, b. Очень важно, что эти законы действительны для любой функции $P(x)$. Последняя имеет ограничения, приведенные выше, только лишь в крайних точках распределения.

Случай с ограниченной величиной x может наблюдаться при изучении содержания металла в пробе. Максимум этой величины определяется содержанием металла в рудном минерале. Для свинцовой руды это будет, например, содержание свинца в галените, для железной — в магнетите и т. д.

В других задачах геологии возможен, однако, случай, когда значения крайних членов в выборке ничем не ограничены. Например, относительная ошибка в определении содержания металла в пробе теоретически может быть любой.

Для этого случая закон распределения максимального члена (вывод этого закона можно найти у Н. В. Смирнова и И. В. Дунина-Барковского, 1959) будет таким:

$$P(x_{\max} < \gamma \ln n + z) \to e^{-e^{-\frac{z}{\gamma}}},$$

где $\gamma \ln n + z = x$ имеют прежнее значение.

В более общей форме этот закон имеет следующий вид:

$$P(x) = e^{-e^{-y}},$$

где $y = \alpha(x - q)$, $\alpha > 0$, а q — некоторая константа.

Это двойной показательный закон, применимый как для максимального, так и для минимального членов (с различными значениями y).

Зависимость функции $e^{-e^{-y}}$ от аргумента y показана в приложении 18. Практически, однако, удобнее определить аргумент y в зависимости от значения функции $e^{-e^{-y}}$ (приложение 19).

Пример. Предположим, на одном каменноугольном месторождении нарезано 70 лав, и в каждой из них сделано одно и то же число маркшейдерских замеров мощности пласта. Один из этих замеров максимальный. В результате этого получена совокупность 70 максимальных членов. Расположим их в порядке возрастания и попробуем узнать значение, например y_2 и y_{69} — нормированного отклонения для 2-го и 69-го членов этого ряда.

$$\varphi(y_2) = \frac{m}{N+1} = \frac{2}{70+1} = 0{,}0282,$$

$$\varphi(y_{69}) = \frac{m}{N+1} = \frac{69}{70+1} = 0{,}972,$$

где величина $N + 1$ принимается потому, что $N + 1$ членов совокупности дают N интервалов, а число m — номер исследуемого члена.

В приложении 19 для этих значений $\varphi(y)$ находим (при помощи интерполяции) $y_2 = -1{,}27$ и $y_{69} = 3{,}56$. Это означает, что 2-й член будет на $1{,}27\sigma$ меньше среднего величины в совокупности максимальных членов, а 69-й член на $3{,}56\sigma$ больше того же среднего.

Закон распределения, выраженный формулой $P(x) = e^{-e^{-y}}$, можно представить и в другом виде.

При $n \to \infty$ для любого значения u получим

$$\begin{cases} P\left(x_{\max} - Mx_{\max} < u\sqrt{Dx_{\max}}\right) \to e^{-e^{-t}}, \\ P\left(x_{\min} - Mx_{\min} < u\sqrt{Dx_{\min}}\right) \to e^{-e^{-t}}, \end{cases}$$

где $Mx_{\max}$ и $Mx_{\min}$ — математические ожидания крайних членов, а $Dx_{\max}$ и $Dx_{\min}$ — дисперсии тех же величин.

Величина u показывает, во сколько раз теоретический стандарт (среднее квадратическое отклонение) превышает фактический.

Необходимо иметь в виду, что последние две асимптотические формулы справедливы лишь тогда, когда n достаточно велико (стремится к бесконечности). В противном случае значительно точнее будут предшествующие степенные формулы.

В вариационном ряду иногда встречаются резко выделяющиеся наблюдения. Если проверка записи наблюдений не дает возможности найти описку или механическую ошибку («грубый промах», как иногда говорят), то необходимо оценить это выдающееся наблюдение статистическим методом. Такая оценка может дать некоторое обоснование для сомнения в правильности данного наблюдения, но она возможна только в том случае, если вид функции распределения изучаемой случайной величины известен.

Сущность статистической оценки такого наблюдения состоит в решении вопроса о том, принадлежит ли это наблюдение к одной, общей для всех других, генеральной совокупности или же оно попало сюда из какой-то другой совокупности. Иначе говоря, речь идет о том, не сделано ли это подозрительное наблюдение при каких-то других условиях, чем исследуемые наблюдения. Постоянство условий производства наблюдений приводит к тому, что все наблюдения будут относиться к одной совокупности; нарушение этих условий означает, что мы имеем дело с разными совокупностями.

Прямая проверка условий позволяет убедиться в ошибочности выдающегося наблюдения, но такая проверка далеко не всегда возможна. В таком случае остается статистический путь проверки выдающегося наблюдения.

Статистический метод возможен в двух вариантах. Один применяется в случае, когда имеется две или большее число выборок, а выдающееся наблюдение входит в одну из них; второй — когда имеет место только одна выборка, и выдающееся наблюдение входит в нее.

Первый вариант. Отбрасывается, как явно ошибочное, такое наблюдение, которое выходит за границы допустимого размаха (при нормальном распределении). Допустимый же размах определяется по таблице (приложение 16) для выбранной величины вероятности. Для этой же цели можно использовать критерий Стьюдента, вычисляемый дважды: один раз для всех наблюдений, другой — с исключением выдающегося наблюдения.

При этом обычно оказывается, что критерий Стьюдента, определенный для всех наблюдений(способ вычисления описан в главе V), свидетельствует о принадлежности двух выборок к разным генеральным совокупностям, а тот же критерий, вычисленный с исключением подозрительного выдающегося члена, показывает, что обе выборки (одна из них содержала выдающийся член) вполне могли быть взяты из одной общей генеральной совокупности.

Второй вариант. По этому варианту можно проверить как одно выдающееся наблюдение (минимальное или максимальное), так и два (минимальное и максимальное).

Проверка по второму варианту заключается в вычислении критерия значимости.

Для случая одного выдающегося наблюдения вычисление критерия значимости (в случае нормального распределения) производится следующим образом:

1. Границы значимости для x_{max} можно найти по формуле

$$x_{max,\,P} \simeq a + \sigma u_{P_1} \text{ (при } P_1 = P^{\frac{1}{n}}),$$

где a и σ — среднее значение признака и среднее квадратическое отклонение — характеризуют генеральную совокупность.

В геологической практике параметры генеральной совокупности обычно неизвестны. Поэтому мы условно принимаем $a = \bar{x}$ (среднее выборочное), а $\sigma = \hat{\sigma}$ (выборочный стандарт).

2. Далее по таблицам Граббса (Grubbs, 1950) получаем следующий критерий:

$$x_{max,\,P} = \bar{x} + su_{P_1}.$$

3. Если реальная величина выдающегося наблюдения x_{max} больше вычисленной по таблицам Граббса для вероятности P величины $x_{max,\,P}$, то это наблюдение считается ошибочным и отбрасывается. Если же оно равно или меньше величины, получаемой по таблицам Граббса, то наблюдение следует считать приемлемым.

Для случая одного выдающегося наблюдения есть еще один критерий. Его предложил И. Ирвин (Irwin, 1925).

Критерий λ Ирвина основан на разности между x_n и $x_{(n-1)}$. Его формула имеет следующий вид:

$$\lambda = \frac{x_{(n)} - x_{(n-1)}}{\sigma}.$$

Ирвин составил таблицу для функции распределения величины λ при заданных уровнях вероятности. В результате получены следующие значения квантилей:

n	$\lambda_{0,95}$	$\lambda_{0,99}$
2	2,8	3,7
3	2,2	2,9
10	1,5	2,0
20	1,3	1,8
30	1,2	1,7
50	1,1	1,6
100	1,0	1,5
400	0,9	1,3
1000	0,8	1,2

Пример. В рудном блоке взято 10 проб, показавших следующее содержание молибдена в %; 0,2; 0,4; 0,0; 0,9; 0,3; 0,1; 0,0; 0,2; 0,2; 0,1.

Для этого ряда оценено среднее квадратическое отклонение $s = 0{,}25$. Заметим, что в этом ряду $x_{(n)} = 0{,}9$, а $x_{(n-1)} = 0{,}4$.

По формуле Ирвина находим

$$\lambda = \frac{0{,}9 - 0{,}4}{0{,}25} = 2{,}0.$$

Сравниваем эту величину с теоретическими значениями критерия Ирвина по приведенной выше таблице. Для вероятности 0,95 и $n = 10$ имеем $\lambda = 1{,}5$, а вероятности 0,99 соответствует $\lambda = 2{,}0$. Таким образом, если нас удовлетворяет вероятность 0,95, то содержание 0,9% мы должны считать ошибочным, а потому подлежащим исключению. Если же мы хотим решить этот вопрос с вероятностью 0,99, то содержание 0,9 оказывается критическим в верхнем пределе, а потому допустимым.

Пример. В другом блоке того же рудника получены следующие содержания молибдена в %: 0,5; 0,0; 0,2; 0,0; 0,3; 1,5; 0,2; 0,2; 0,1; 0,2; 0,1; 0,2; 0,2; 0,1; 0,1.

Здесь $x_{(n)} = 1{,}5$, а $x_{(n-1)} = 0{,}4$. Величина s для этого примера равна 0,36.

По формуле Ирвина имеем

$$\lambda = \frac{1{,}5 - 0{,}5}{0{,}36} = 2{,}8.$$

В таблице Ирвина нет значений для $n = 15$, а есть $n = 10$ и $n = 20$. Для них имеем $\lambda_{0,95} = 1{,}5$ и 1,3 и $\lambda_{0,99} = 2{,}0$ и 1,8. Фактическое значение $\lambda = 2{,}8$ превышает эти нормы, поэтому содержание 1,5% считаем ошибочным и исключаем.

И. Ирвин (Irwin, 1925) составил таблицы критических вероятностей того, что как первый ($x_{(n)}$), так и второй ($x_{(n-1)}$) члены упорядоченного выборочного ряда принадлежат одной общей совокупности и что как второй ($x_{(n-1)}$), так и третий ($x_{(n-2)}$) члены того же ряда входят в общую совокупность. На русском языке эти две таблицы публикуются впервые.

Вопрос о применимости критерия Ирвина к определению нижней границы ураганной пробы нуждается в более детальном рассмотрении. При этом надо учитывать характер распределения особо богатых проб.

Пример. В геологической практике могут встретиться случаи, когда выдающимися (сомнительными) наблюдениями являются минимальные члены ряда. Так, например, анализ 10 проб гранита показал следующие содержания SiO_2 в %: 72,5; 59,4; 75,6; 68,0; 63,0; 70,1; 72,9; 68,5; 54,5; 78,0.

Для того ряда получаем $s = 6{,}7$. Сомнительным содержанием здесь является $x_{(1)} = 54{,}5$. Следующее за ним содержание $x_{(2)} = 59{,}4$.

Критерий Ирвина вычисляем по формуле

$$\lambda = \frac{x_{(2)} - x_{(1)}}{s} = \frac{59{,}4 - 54{,}5}{6{,}7} = \frac{4{,}9}{6{,}7} = 0{,}7.$$

По таблице Ирвина для $P = 0{,}95$ и $n = 10$ получаем $\lambda = 1{,}5$. Следовательно, содержание SiO_2, равное 54,5, приемлемо.

Пример. На одном из участков Артемовского месторождения каменной соли взято 10 проб, показавших следующие содержания натрия в % (с округлением): 40,35, 39, 40, 40, 25, 38, 36, 39, 39. Для этого ряда $s = 4{,}7$, $x_{(1)} = 25$, $x_{(2)} = 35$ и $\lambda = 2{,}1$.

Для $n = 10$ по таблице Ирвина $\lambda_{0,99} = 2{,}0$. Поскольку фактическое значение $\lambda = 2{,}1$ больше теоретического, мы можем с вероятностью 0,99 считать, что содержание натрия, равное 25%, ошибочное и подлежит исключению (при этом мы можем предположить, что возможно тут сделана описка: вместо 35 написано 25). Возможно также другое предположение, а именно, что распределение ненормальное (это надо проверить).

Приведенная выше таблица Ирвина составлена только для двух значений вероятности: $P(\lambda) = 0{,}95$ и $P(\lambda) = 0{,}99$. В приложениях же 20 и 21 приводятся различные вероятности при определенных n и λ. По этим приложениям можем найти вероятность того, что при данном n будет та или иная величина λ. Так, например, если при $n = 10$ величина λ будет равна 2,0, вероятность этого будет равна 0,011.

Способ Ирвина дает возможность проверять не только самый крайний, но и соседний с ним член упорядоченного ряда, т. е. смежный член, расположенный в порядке возрастания значений (см. приложение 21). Причем принцип проверки тот же. Так, если второй член будет больше третьего в λ-кратное число стандартных отклонений, т. е. в λs раз, то вероятность можно найти по таблице (приложение 21). Для $n = 300$ и $\lambda = 0{,}6$ вероятность будет равна 0,029. Что же касается того, при какой вероятности проверяемый член отбрасывается, то это зависит от конкретных условий практической задачи.

Все эти примеры относились к случаю, когда мы имеем одно сомнительное наблюдение — максимальное или минимальное.

В статистике (Хальд, 1956) существует также критерий для случая, когда имеется два сомнительных наблюдения — одно минимальное и другое максимальное. Здесь эти критерии не описываются ввиду того, что геологам еще не приходилось сталкиваться с такими случаями.

Иногда в серии проб имеется две сомнительные, причем обе они находятся в конце одного ряда, т. е. относятся к самым богатым или самым бедным пробам. В этом случае можно пользоваться критерием Ирвина последовательно — сначала для одной (крайней) пробы, потом для другой. Можно также проверять минимальный из сомнительных членов, т. е. $x_{(n-1)}$, так как в случае его неприемлемости автоматически бракуется и $x_{(n)}$.

IX. ПОСЛЕДОВАТЕЛЬНЫЙ АНАЛИЗ

Все описанные в предыдущих главах методы статистического анализа применяются после того, как сбор исходных данных закончен. При этом иногда оказывается, что для обоснования сделанных в результате анализа выводов вполне хватило бы и меньшего числа данных, что мы собрали излишний материал — взяли лишние пробы, сделали лишние анализы, пробили лишние шурфы и т. д.

Для того чтобы не делать лишних наблюдений статистический анализ необходимо производить не после сбора исходных данных, а в процессе этого сбора. Пробурить небольшое число скважин, подсчитать результат и проверить его обоснованность: если он окажется обоснованным — больше скважин не бурить, если окажется необоснованным, то пробурить еще одну-две скважины, снова подсчитать результат. Правда, при этом увеличивается объем вычислительной работы, но легче и дешевле подсчитать данные по ста скважинам, чем пробурить хотя бы одну из них.

Метод последовательного анализа был предложен Вальдом (1960). Этот анализ в то время применялся в связи с контролем качества массовой продукции, выпускаемой военными заводами. Минимум наблюдений для максимума информации — такова была цель нового метода математической статистики (Гнеденко, 1949).

Вальд (1960) доказал, что последовательный анализ зачастую позволяет сократить число наблюдений в среднем в два раза по сравнению с тем, что требуется при использовании классических методов проверки гипотез. Но достоинства нового метода этим не исчерпываются. При последовательном анализе можно задать одновременно как уровень значимости, так и мощность критерия. В статистике же с фиксированным объемом наблюдений это оказывается невозможным.

Из военного дела последовательный анализ проник в технологию, физику и в другие дисциплины, а затем и в геологию. В геологии этот метод применялся пока очень мало.

Ниже дается краткое описание метода последовательного анализа. Более полное описание можно найти у А. Вальда (1960) или у А. Е. Башаринова и Б. С. Флейшмана (1962).

Башаринов и Флейшман (1962) считают, что последовательный анализ может быть с успехом применен в следующих областях исследования:

1. В двухальтернативных ситуациях (выбор между двумя решениями):
а) контроль качества продукции и материалов;
б) испытание приборов и систем на надежность;
в) поиск неисправностей в сложных системах;
г) обнаружение сигналов при наличии шумов.

2. В многоальтернативных ситуациях (выбор между более чем двумя решениями):
а) многосортная классификация выборочных данных (изделий) из одной генеральной совокупности (партии);
б) сравнительный анализ выборочных данных (изделий) из нескольких генеральных совокупностей (партий);
в) многоканальное обнаружение сигналов при наличии шумов.

3. В сложных кибернетических системах:
а) процедуры поиска при наличии помех;
б) информационные системы с обратной связью;
в) кибернетические устройства для различения информационных потоков и др.

Для каждой из этих областей можно найти ее аналоги в геологии. Так, аналогом изделия является проба, аналогом шума — фон геофизического явления, мешающий выявить анамалию (гравитационную, сейсмическую, магнитную и пр.) *.

Конкретной целью последовательного анализа заинтересовавшего нас явления ставится проверка какой-либо гипотезы относительно этого явления. Последнюю называют начальной или нулевой гипотезой (H_0). Ей противопоставляется конкурирующая гипотеза (H_1).

Для проверки нулевой гипотезы проводятся последовательно (одно за другим) статистические испытания, т. е. берутся, например, пробы, или фотографируются обнажения, или же делаются отсчеты на геофизическом приборе и т. д. При этом после каждого испытания делаются вычисления по сумме всех сделанных испытаний и выносится одно из следующих трех решений.

1. Проверяемая гипотеза принимается и дальнейшие испытания не производятся.

---

* Под шумом в статистике понимают вообще помехи. Разумовский (1962) разработал методику выделения аномалий на фоне шума в разведке месторождений полезных ископаемых.

2. Проверяемая гипотеза отвергается (принимается альтернативная, конкурирующая гипотеза) и испытания прекращаются.

3. Проверяемая гипотеза не принимается и не отвергается, поэтому испытания продолжаются.

Когда вычисления, сделанные после какого-то (по счету) испытания, позволят принять первое или второе решение, анализ считается законченным. Вальд (1960) доказал, что процесс проверки не может быть бесконечным. Обычно он сравнительно быстро заканчивается.

Таким образом, область возможных значений результата статистических испытаний здесь разделена на три части. Деление на две части (без выделения «зоны сомнения») не может удовлетворить нас, хотя на практике такое двухчленное деление и делается почти всегда (таким путем определяется норма выхода керна, норма точности химических анализов, нижний предел для ураганных проб и пр.). Выделение «зоны сомнения» делает исследование гибким и более точным.

Трехчленное деление возможных значений результата статистических испытаний — важнейшая особенность последовательного анализа.

Проверяемая гипотеза и ее альтернатива могут выглядеть по-разному, в частности так:

1. Новый метод опробования руд можно считать приемлемым, если разница в содержании свинца в серии проб между этим и эталонным методами не превысит в среднем 0,2 абсолютного процента.

2. Новый метод опробования руд можно считать неприемлемым, если разница в содержании свинца в серии проб между этим и эталонным методами превысит в среднем 0,3 абсолютного процента.

3. Если же эта разница лежит где-то между указанными пределами (0,2 и 0,3%), то решение не выносится, новый метод опробования остается под сомнением, а испытания продолжаются, т. е. берется еще проба или серия проб.

Таким образом, при последовательном анализе на основании испытаний (проб, наблюдений, замеров) мы делаем выбор между двумя взаимно исключающими друг друга гипотезами, первую из которых относительно параметра μ обозначим $H_0 : \mu = \mu_0$, а вторую $H_1 : \mu = \mu_1$ или же продолжаем испытания.

Необходимо заметить, что обе эти гипотезы относятся к гипотезам о среднем значении случайной величины.

Принимая одну из двух гипотез, мы в большинстве случаев поступаем правильно, но иногда можем и ошибиться. В статистике вообще нет ничего абсолютно безошибочного, и с этим вполне можно мириться. Важно лишь одно: чтобы вероятность появления ошибки в выводе была не более, чем заданная малая величина.

Если мы примем неверную гипотезу $\mu = \mu_1$, тогда как верной является гипотеза $\mu = \mu_0$, то мы совершим ошибку первого рода, вероятность которой условимся считать меньшей, чем некоторая величина α, (эту величину называют уровнем значимости — см. главу V). Если же мы примем гипотезу $\mu = \mu_0$, тогда как в действительности верна гипотеза $\mu = \mu_1$, то совершим ошибку второго рода, вероятность которой условимся считать меньшей, чем β.

Пусть над изучаемой случайной величиной ξ проведено m наблюдений. Обозначим полученные значения $x_1, x_2, \ldots, x_m$. Эти результаты можно рассматривать как m значений m независимых одинаково распределенных случайных величин $\xi_1, \xi_2, \ldots, \xi_i, \ldots, \xi_m$. Пусть $f(x_i, \mu)$ — плотность распределения случайной величины с номером i в последовательности. Допустим, что верна гипотеза H_0 и построим при этом условии функцию правдоподобия

$$L_m(H_0) = \prod_{i=1}^{m} f(x_i, \mu_0).$$

Если предположить, что верна гипотеза H_1, функция правдоподобия примет следующий вид:

$$L_m(H_1) = \prod_{i=1}^{m} f(x_i, \mu_1).$$

Для простоты записи обозначим $L_m(H_0) = P_{0,m}$ и $L_m(H_1) = P_{1,m}$.

Вполне естественно, что отношение $\frac{P_{1,m}}{P_{0,m}}$ при условии, что верна гипотеза H_1, должно превышать единицу, а если верна гипотеза H_0, наоборот, должно быть меньше 1. Это отношение называется отношением правдоподобия.

Для того чтобы построить последовательный критерий отношения вероятностей, необходимо сначала выбрать две положительные величины A и B, являющиеся критическими, причем $A > B$. После каждого испытания, имеющего номер m, вычисляется отношение $\frac{P_{1,m}}{P_{0,m}}$. Если это отношение окажется в интервале (A, B), т. е. если осуществится неравенство

$$B < \frac{P_{1,m}}{P_{0,m}} < A,$$

то исследование продолжается (делается следующее испытание). Если же отношение

$$\frac{P_{1,m}}{P_{0,m}} \geqslant A,$$

то исследование прекращается, гипотеза H_0 отклоняется и принимается гипотеза H_1. Если

$$\frac{P_{1,m}}{P_{0,m}} \leqslant B,$$

то исследование прекращается, но при этом принимается гипотеза H_0.

Если для какой-либо выборки окажется, что $P_{1,m} = P_{0,m} = 0$, то условно будем считать, что отношение $\frac{P_{1,m}}{P_{0,m}}$ равно единице. Если в другой выборке получим $P_{1,m} > 0$, но $P_{0,m} = 0$, то неравенство $\frac{P_{1,m}}{P_{0,m}} \geqslant A$ считается выполненным и H_0 отвергается.

Числа A и B Вальд выбирает так, чтобы критерий имел заданную силу. Он принимает

$$A = \frac{1-\beta}{\alpha}, \quad B = \frac{\beta}{1-\alpha},$$

где α и β — вероятности ошибок первого и соответственно второго рода.

Описанный здесь критерий применим при соблюдении следующих трех условий (Хальд, 1956):

1. Математический вид функции распределения известен.

2. Проверяемая и альтернативная гипотезы (H_0 и H_1) о значении неизвестного параметра μ в функции распределения фиксированы.

3. Выбраны значения уровня значимости α для H_0 и функции мощности $1 - \beta$ относительно гипотезы H_1.

В том случае, когда неизвестная генеральная совокупность, из которой мы берем выборку, подчиняется закону нормального распределения,

а произведенные n испытаний независимы одно от другого, то справедливо будет следующее неравенство. *

$$2{,}3\frac{\sigma^2}{\delta}\log\frac{\beta}{1-\alpha}+n\bar{\mu}<T_n<2{,}3\frac{\sigma^2}{\delta}\log\frac{1-\beta}{\alpha}+n\bar{\mu},$$

где $T_n=\sum_{i=1}^{n}x_i$; $\delta=\mu_1-\mu_0$; $\bar{\mu}=\frac{\mu_1+\mu_0}{2}$.

Неизвестной здесь является только σ^2 — дисперсия изучаемой случайной величины (величина σ^2 принимается условно).

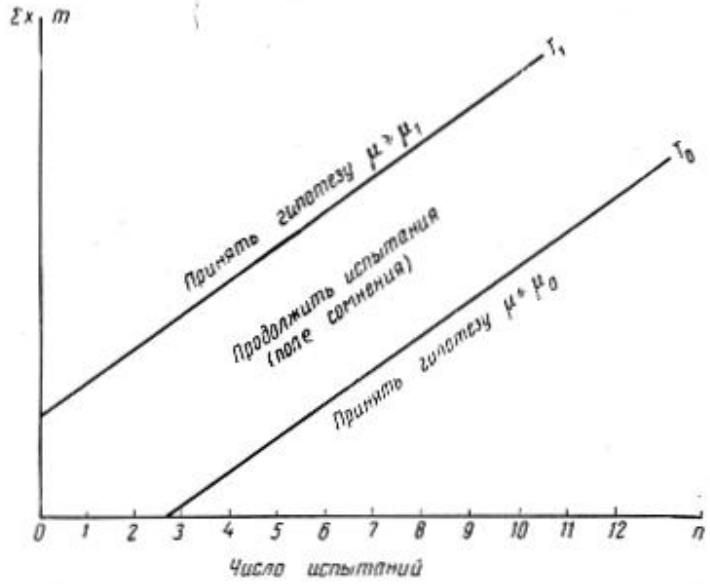


Рис. 48. Схема применения последовательного критерия

Для практического использования последней формулы строим по ней график в прямоугольных координатах (рис. 48). По оси абсцисс откладываем число n (номер и число сделанных испытаний), а по оси ординат $\sum_{i=1}^{n}x_i$. На графике проводим две прямые линии (иногда эти линии — кривые; их формулы дает Вальд). Одна из них — линия T_0 — определяется уравнением

$$T_0=a_0+bn,$$

другая линия T_1 — уравнением

$$T_1=a_1+bn,$$

где

$$a_0=2{,}3\frac{\sigma^2}{\delta}\log\frac{\beta}{1-\alpha},$$

$$a_1=2{,}3\frac{\sigma^2}{\delta}\log\frac{1-\beta}{\alpha},$$

$$b=\bar{\mu}.$$

* Теория последовательного анализа для неоднородных зависимых выборок, как отмечают А. Е. Башаринов и Б. С. Флейшман (1962), еще не разработана. Это ограничивает область применения последовательного анализа в геологии.

После каждого испытания находим $\sum_{i=1}^{n} x_i$ и ставим на графике точку с координатами: по горизонтальной оси n и по вертикальной оси $\sum_{i=1}^{n} x_i$ или в других случаях среднее значение признака $\bar{x}$.

Эти точки затем соединяем друг с другом, что дает нам ломаную линию.

Испытания надо продолжать до тех пор, пока последняя точка не выйдет из «зоны сомнения» — вверх или вниз от нее. Заметим, что «зоной сомнения» мы условно называем поле между линиями T_1 и T_0. Эту зону можно также назвать «полем безразличия или непринятия решений».

Рассмотрим задачу, связанную с проверкой нулевой гипотезы о равенстве среднего $\mu = \mu_0$ при альтернативе $\mu \neq \mu_1$.

При этом мы будем исходить из того, что вид функции распределения изучаемого признака заранее известен, что нулевая гипотеза и ее альтернатива (т. е. значения μ_0 и μ_1) фиксированы и что вероятности ошибочного принятия нулевой гипотезы или ее альтернативы выбраны равными соответственно α и β.

Характер случайной переменной (ее дискретность или непрерывность) и закон распределения могут быть различными. Поэтому и критерий может принимать различный вид. Хальд (1956) приводит примеры, когда признак распределен по нормальному, биноминальному и Пуассонову законам.

Для иллюстрации сказанного приведем следующий пример последовательного анализа, заимствованный у Налимова (1960). «При ускоренном определении SiO_2 в шлаке весовым методом получаются заниженные результаты из-за неучета SiO_2 в фильтрате. Перед аналитиками была поставлена задача — повысить полноту выделения SiO_2 путем применения специальных приемов (например, использование желатина для коагуляции и пр.). Для проведения исследования был выбран стандартный образец, содержащий 24,5% SiO_2. На основании обработки предыдущего экспериментального материала была оценена ошибка воспроизводимости $\sigma \simeq 0{,}25\%$. Исходя из практических соображений, обусловленных требованиями производства, было принято, что осаждение можно считать достаточно полным, если результаты анализа стандартного образца будут больше чем 24,25%».

«Таким образом, проверяемая и альтернативная гипотезы были сформулированы следующим образом: $\mu \leqslant \mu_0 = 24{,}25\%$, $\mu \geqslant \mu_1 = 24{,}50\%$. При оценке результатов было принято $\alpha = \beta = 0{,}05$».

«Результаты вычислений дают:

$$\delta = 24{,}50 - 24{,}25 = 0{,}25;$$

$$\bar{\mu} = \frac{24{,}25 + 24{,}50}{2} = 24{,}375;$$

$$a_0 = \frac{2{,}30 \cdot 0{,}25^2}{0{,}25} \log \frac{0{,}05}{0{,}95} = -0{,}735;$$

$$a_1 = \frac{2{,}30 \cdot 0{,}25^2}{0{,}25} \log \frac{0{,}95}{0{,}05} = 0{,}735;$$

$$b = 24{,}375''.$$

«Область продолжающихся испытаний будет ограничена прямыми:

$$T_0 = -0{,}735 + 24{,}375n,$$

$$T_1 = 0{,}735 + 24{,}375n».$$

«Результаты последовательных определений для одного из изучаемых вариантов приведены на рис. 49. По оси абсцисс здесь отложено число последовательных определений, а по оси ординат — сумма результатов последовательных определений, причем для удобства построения графика из каждого результата вычиталось число, равное 24,0%. Поэтому при построении прямых вместо написанных выше уравнений использовались уравнения:

$$T_0 = -0{,}735 + 0{,}375n,$$

$$T_1 = 0{,}735 + 0{,}375n».$$

«Испытание закончилось после четвертого определения *. Четыре последовательных определения дали следующие результаты: 24,50; 23,78; 24,12; 23,90».

«Вычитая из каждого результата число 24,00, получаем следующие накопленные суммы: 0,50; 0,28; 0,40; 0,30, которые нанесены на график».

Таким образом, принимается гипотеза $\mu \leqslant 24{,}25\%$.

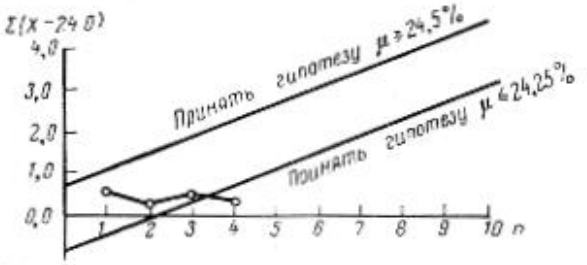


Рис. 49. Применение последовательного критерия при изучении одного из вариантов определения SiO_2 в шлаке (по Налимову)

В том случае, когда проверяемая гипотеза относится не к среднему значению какого-либо признака (например, среднее содержание SiO_2 в шлаке), а к определению вероятности (частости) некоторого события, последовательный анализ проводится несколько иначе.

Допустим, что встреча одного или нескольких зерен ценного минерала в пробе речного песка зависит только от случая. Назовем такую пробу продуктивной. Пусть геолог, ведущий поиски россыпи этого минерала, сформулировал нулевую гипотезу так. Поиски можно считать перспективными, если частость продуктивных проб, т. е. их доля в общей массе промытых проб, будет больше или равна 5%. Альтернативная гипотеза будет следующая: если доля таких проб окажется меньше или равна 1%, поиски будем считать бесперспективными.

Обозначим через α вероятность появления ошибки первого рода, а через β — вероятность ошибки второго рода и дадим им следующие значения: $\alpha = 0{,}20$, $\beta = 0{,}16$.

Вероятность ошибки нашего заключения относительно нулевой гипотезы не должна превышать эти границы. По формулам, приведенным выше, найдем:

$$A = \frac{1 - 0{,}16}{0{,}20} = 4{,}2,$$

$$B = \frac{0{,}16}{1 - 0{,}2} = 0{,}2.$$

Условимся, что в случае встречи хотя бы одного зерна интересующего нас ценного минерала в пробе будем считать такую пробу продуктивной и положим, что случайная величина — содержание — приняла значение, равное единице ($x = 1$). С другой стороны, в случае отсутствия зерен интересующего нас минерала будем считать пробу пустой ($x = 0$).

Вероятность взятия пустой пробы обозначим через q, вероятность противоположного события — через p, причем $p + q = 1$.

* Здесь Налимов под испытанием подразумевает серию испытаний, а под определением — отдельное испытание.

Таким образом, принимаем $p_0 = 0{,}05$ и $p_1 = 0{,}01$, $q_0 = 0{,}95$ и $q_1 = 0{,}99$.

Если для некоторого числа взятых проб гипотеза $p \geqslant p_0$ не отвергается, то мы сочтем поиски перспективными. Если же эта гипотеза отвергается, а $p \leqslant p_1$ не отвергается, то поиски сочтем бесперспективными. Если же ни та, ни другая гипотеза не отвергается, то никакого вывода о перспективности или о бесперспективности делать не будем и продолжим опробование.

Взяв n проб и установив, что m из них являются продуктивными, мы можем вычислить вероятность появления этого результата по биномиальному закону:

если верна нулевая гипотеза,

$$P_{0,\,n} = p_0^m q_0^{n-m},$$

если верна гипотеза $p \leqslant p_1$,

$$P_{1,\,n} = p_1^m q_1^{n-m}.$$

Отношение правдоподобия определим так:

$$\frac{P_{0,\,n}}{P_{1,\,n}} = \left(\frac{p_0}{p_1}\right)^m \left(\frac{q_0}{q_1}\right)^{n-m}.$$

Подставим в эту формулу значения переменных величин:

$$\frac{P_{0,\,n}}{P_{1,\,n}} = \left(\frac{0{,}05}{0{,}01}\right)^m \left(\frac{0{,}95}{0{,}99}\right)^{n-m} = \left(\frac{0{,}95}{0{,}99}\right)^n \left(\frac{0{,}0495}{0{,}0095}\right)^m = 0{,}96^n \cdot 5{,}21^m.$$

Допустим, что эта величина будет лежать в пределах между A и B, т. е. $4{,}2 > 0{,}96^n \cdot 5{,}21^m > 0{,}2$.

Возьмем два крайних случая:

1. $0{,}96^n \cdot 5{,}21^m = 4{,}2$.

2. $0{,}96^n \cdot 5{,}21^m = 0{,}2$.

Задавая значения величины n, мы можем из этих неравенств найти пределы для колебаний значения m, и, наоборот, задавая m, найдем пределы для n.

Найдем, например общее число проб n, если исходить из заданного значения m. Для этого логарифмируем обе части того и другого равенства:

1. $n \lg 0{,}96 + m \lg 5{,}21 = \lg 4{,}2$

2. $n \lg 0{,}96 + m \lg 5{,}21 = \lg 0{,}2$

Отсюда находим:

1. $n = \dfrac{\lg 4{,}2 - m \lg 5{,}21}{\lg 0{,}96}$.

2. $n = \dfrac{\lg 0{,}2 - m \lg 5{,}21}{\lg 0{,}96}$.

Произведя вычисления, получим:

1. $n = -34{,}6 + 39{,}8m$.

2. $n = 38{,}8 + 39{,}8m$.

В этих пределах будет лежать n при фиксированном (заданном) значении m.

Найдем теперь число продуктивных проб m, если исходить из заданного общего числа проб n:

1. $m = 0{,}870 + 0{,}025n$.
2. $m = -0{,}975 + 0{,}025n$.

Значения m, нанесенные на график (рис. 50), покажут нам границы «поля сомнения», вверх от которого лежит поле принятия нулевой гипотезы, а вниз — поле принятия альтернативной гипотезы.

На этот же график нанесем эмпирические данные, полученные от промывки 12 серий шлиховых проб. Здесь роль отдельного испытания играет взятие и промывка серии проб, так как продуктивные пробы встречаются очень редко. Эти данные приведены в табл. 89.

12-я серия проб (12-е испытание) показывает, что нулевая гипотеза неверна. Ее надо отвергнуть, а принять альтернативную гипотезу. Иначе говоря, мы можем сделать вывод о бесперспективности поисков ввиду малой вероятности встречи продуктивных проб. В этом примере вероятность ошибки первого рода $\alpha = 0{,}2$, а $\beta = 0{,}16$.

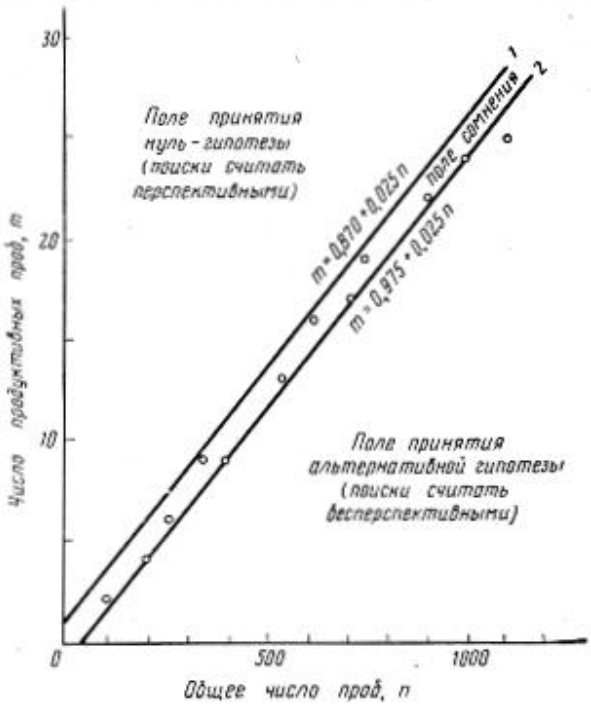


Рис. 50. Схема последовательного анализа с целью проверки гипотезы о вероятности (доли признака)

Таблица 89

№ испытания (серии)	Число всех проб в серии	Число продуктивных проб	Нарастающий итог	
			общего числа проб	числа продуктивных проб
1	100	2	100	2
2	100	2	200	4
3	61	2	261	6
4	78	3	339	9
5	61	0	400	9
6	138	4	538	13
7	81	3	619	16
8	91	2	710	17
9	40	2	750	19
10	148	3	898	22
11	101	2	999	24
12	102	1	1101	25

X. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ (ЛИНЕЙНАЯ КОРРЕЛЯЦИЯ)

Во всех предшествующих главах речь шла о таких статистических совокупностях, которые содержат лишь одну случайную величину (например, мощность пласта).

Такие совокупности называются одномерными, но в математической статистике рассматриваются и другие совокупности — такие, в которых

объединяются различные значения двух, трех или большего числа признаков. Такие совокупности называются многомерными. В зависимости от числа признаков различают двухмерные, трехмерные и т. д. совокупности.

В геологии многомерные совокупности играют особенно важную роль, так как нередко по данным буровых скважин или в геологических пробах мы измеряем величину не одного, а нескольких признаков.

Связь между признаками бывает функциональная и стохастическая. Первая заключается в том, что какому-либо значению признака A, например x_0, соответствует одно и только одно значение y_0 признака B. Приведем пример такой связи. Вес буровой трубы определенного диаметра практически функционально связан с ее длиной, если отвлечься от возможного непостоянства толщины стенок и диаметра труб. Вторая, т. е. стохастическая связь, наоборот, состоит в том, что одному, произвольно взятому, значению признака A может соответствовать несколько разных значений признака B с различными вероятностями или более строго в том, что изменение величины A влечет за собой изменение закона распределения величины B.

Стохастическая связь вызывается тем, что как на A, так и на B действуют общие факторы, например C, D, E, но кроме них есть факторы F, G, H и другие, действующие только на A, и факторы I, J, K и другие, действующие только на B. В результате получается связь между A и B, проявляющаяся в виде тенденции.

Первой задачей совместного исследования нескольких признаков является выявление их стохастической сопряженности. Если эта сопряженность имеется, то тогда можно делать прогноз значений одного признака в связи с определенным изменением другого, измерять силу их связи и делать другие расчеты. Научную основу таких расчетов дает теория корреляции или корреляционный анализ.

Прежде чем перейти к методам корреляционного анализа, коротко остановимся на рассмотрении распределений многомерных случайных величин.

Пусть $\Xi = \{\xi_1, \xi_2, \ldots, \xi_m\}$ m-мерная случайная величина, образованная одномерными случайными величинами $\xi_1, \xi_2, \ldots, \xi_m$. Функция распределения $F_\Xi(X)$ m-мерной случайной величины Ξ есть вероятность совместного выполнения неравенств $\xi_1 \leqslant x_1, \xi_2 \leqslant x_2, \ldots, \xi_m \leqslant x_m$, т. е. $F_\Xi(X) = P(\xi_1 \leqslant x_1, \xi_2 \leqslant x_2, \ldots, \xi_m \leqslant x)$, где $X = \{x_1, x_2, \ldots, x_m\}$.

Одной из важных характеристик распределения m-мерной случайной величины Ξ является ее математическое ожидание $M\Xi$, которое представляет вектор, образованный математическими ожиданиями величин ξ_i, т. е.

$$M\Xi = \begin{pmatrix} M\xi_1 \\ M\xi_2 \\ \vdots \\ M\xi_m \end{pmatrix}.$$

Второй характеристикой m-мерного распределения является так называемая ковариационная матрица, элементы которой представляют собой математические ожидания произведений $(\xi_i - M\xi_i)(\xi_j - M\xi_j)$, для всех $i = 1, 2, \ldots, m$ и $j = 1, 2, \ldots, m$. Таким образом,

$$\operatorname{cov}(\xi_i, \xi_j) = M[(\xi_i - M\xi_i)(\xi_j - M\xi_j)] = \sigma_{ij}.$$

Ковариационная матрица $\operatorname{cov}[\Xi]$ (обозначим ее для простоты $\boldsymbol{\Sigma}$) будет иметь следующий вид:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11}\sigma_{12} \ldots \sigma_{1j} \ldots \sigma_{1m} \\ \sigma_{21}\sigma_{22} \ldots \sigma_{2j} \ldots \sigma_{2m} \\ \cdots\cdots\cdots\cdots\cdots \\ \sigma_{i1}\sigma_{i2} \ldots \sigma_{ij} \ldots \sigma_{im} \\ \cdots\cdots\cdots\cdots\cdots \\ \sigma_{m1}\sigma_{m2} \ldots \sigma_{mj} \ldots \sigma_{mm} \end{pmatrix}.$$

Диагональные элементы этой матрицы $\sigma_{11}, \sigma_{22}, \ldots, \sigma_{ii}, \ldots, \sigma_{mm}$ — есть дисперсии одномерных случайных величин $\xi_1, \xi_2, \ldots \xi_i, \ldots \xi_m$, которые мы будем обозначать $\sigma_1^2, \sigma_2^2, \ldots, \sigma_i^2, \ldots, \sigma_m^2$, а те элементы, которые не лежат на главной диагонали (σ_{ij} при $i \neq j$), характеризуют силу зависимости между парами величин ξ_i и ξ_j. Если ξ_i и ξ_j независимы, соответствующая им ковариация (σ_{ij}) равна нулю. Нередко в качестве меры зависимости между ξ_i и ξ_j используется безразмерная величина

$$\varrho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} = \frac{M[(\xi_i - M\xi_i)(\xi_j - M\xi_j)]}{\sqrt{M[(\xi_i - M\xi_i)^2] M[(\xi_j - M\xi_j)^2]}},$$

которая называется коэффициентом корреляции.

Как пример многомерного распределения можно привести m-мерный нормальный закон. Пусть $\Xi = (\xi_1, \xi_2, \ldots, \xi_m)$ m-мерный случайный вектор-столбец, который распределен нормально с математическим ожиданием $\mu = \mu_1, \mu_2, \ldots, \mu_m$ и ковариационной матрицей Σ. Функция распределения этого вектора дается выражением

$$F_\Xi(X) = P(\xi_1 \leqslant x_1, \xi_2 \leqslant x_2, \ldots, \xi_m \leqslant x_m) =$$

$$= \frac{1}{(2\pi)^{\frac{m}{2}} |\Sigma|^{\frac{1}{2}}} \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_m} e^{-\frac{1}{2}(X-\mu)' \Sigma^{-1} (X-\mu)} dX,$$

где $X = (x_1, x_2, \ldots, x_m)$, $|\Sigma|$ — детерминант матрицы Σ, Σ^{-1} — матрица, обратная матрице Σ.

Соответствующая m-мерная плотность вероятности $f(X)$ будет равна

$$f(X) = \frac{1}{(2\pi)^{\frac{m}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(X-\mu)' \Sigma^{-1} (X-\mu)}.$$

Рассмотрим наиболее простой случай, когда $m = 2$, т. е. $\Xi = (\xi_1, \xi_2)$. Обозначим математические ожидания ξ_1 и ξ_2 через μ_1 и μ_2, дисперсии — через σ_1^2 и σ_2^2, а ковариацию как σ_{12}. Последнюю можно выразить через ϱ, σ_1 и σ_2:

$$\sigma_{12} = \varrho \sigma_1 \sigma_2.$$

Тогда ковариационную матрицу можно представить в следующем виде:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \varrho\sigma_1\sigma_2 \\ \varrho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

и совместная плотность распределения случайных величин будет

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\varrho^2}} e^{-\frac{1}{2(1-\varrho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{2\varrho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right]}.$$

Если случайные величины ξ_1 и ξ_2 независимы, то $\varrho = 0$ и функция $f(x, y)$ примет следующий вид:

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2} e^{-\frac{1}{2}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right]},$$

т. е. является произведением одномерных плотностей $f(x)$ и $f(y)$

$$f(x, y) = f(x) \cdot f(y).$$

Следует отметить, что коэффициент корреляции ϱ может принимать значения в интервале от —1 до 1. Если $\varrho = 1$, то ξ_1 и ξ_2 связаны линейной

функциональной зависимостью и с увеличением значений одной из величин возрастают значения другой. Если же $\varrho = -1$, то зависимость носит обратный характер, т. е. с увеличением значений одной из величин уменьшаются значения другой.

Если над двухмерной случайной величиной произведено n наблюдений $x_1, y_1; x_2, y_2; \ldots; x_n, y_n$, то оценкой коэффициента корреляции ϱ является величина r, определяемая по формуле

$$r = \frac{\sum_{i=1}^{n} [(x_i - \bar{x})(y_i - \bar{y})]}{(n-1) s_x s_y},$$

где $\bar{x}$ и $\bar{y}$ — средние арифметические;

n — число парных измерений x_i и y_i;

s_x и s_y — оценки стандартных отклонений для x_i и y_i в выборке.

Оценкой для ковариации будет величина

$$\operatorname{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

Оценку коэффициента корреляции можно вычислить другим, более удобным способом, который дает значительную экономию во времени при использовании клавишных электрических вычислительных машин.

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

После этого теоретического вступления перейдем к эмпирическому, более наглядному описанию корреляции. Для того чтобы установить связь с предыдущим текстом, некоторые формулы будут повторены.

Рассмотрим следующий пример. Ниже приведены результаты определения удельного веса и зольности в 18 пробах угля (табл. 90). Можно

Таблица 90

№ пробы	Удельный вес x_i	Зольность, % y_i
1	1,5	25
2	1,2	4
3	1,7	30
4	1,4	20
5	1,8	36
6	1,3	7
7	1,3	5
8	1,5	24
9	1,7	33
10	1,3	4
11	1,5	17
12	1,5	24
13	1,6	25
14	1,4	6
15	1,6	26
16	1,5	24
17	1,4	20
18	1,4	9

Таблица 91

x_i	y_i
1,2	4
1,3	4
1,3	5
1,3	7
1,4	6
1,4	9
1,4	20
1,4	20
1,5	17
1,5	24
1,5	24
1,5	24
1,5	25
1,6	25
1,6	25
1,7	30
1,7	33
1,8	36

ли рассматривать значения этих двух признаков как независимые или же такой вывод делать нельзя, и оба признака следует рассматривать как зависящие один от другого?

Эта таблица называется исходной или первичной перечневой таблицей.

Для того чтобы скрытая здесь закономерность стала заметной, расположим пробы в неубывающем порядке величины x, а внутри групп с одинаковым значением x расположим пробы в неубывающем порядке величины y. В результате этого получим таблицу упорядоченных данных (табл. 91). Номера проб можно опустить, так как они никакой роли в расчетах играть не будут.

Эту таблицу можно еще более упростить, если выписать только неповторяющиеся значения x и соответствующие им значения y и вычислить среднее значение y, т. е. величину $\bar{y}$ (табл. 92)

Таблица 92

x	y	$\bar{y}$
1,2	4	4,00
1,3	4, 5, 7	5,33
1,4	6, 9, 20, 20	13,75
1,5	17, 24, 24, 24, 25	22,80
1,6	25, 26	22,50
1,7	30, 33	31,50
1,8	36	36,00

Таким же путем мы можем распределить все пробы по группам в порядке неубывания y, а внутри этих групп расположить значения x в порядке его неубывания, а затем вычислить $\bar{x}$ (табл. 93).

Таблица 93

y	x	$\bar{x}$
4	1,2; 1,3	1,25
5	1,3	1,30
6	1,4	1,40
7	1,3	1,30
9	1,4	1,40
17	1,5	1,50
20	1,4; 1,4	1,40
24	1,5; 1,5; 1,5	1,50
25	1,5; 1,6	1,55
26	1,6	1,60
30	1,7	1,70
33	1,7	1,70
36	1,8	1,80

Две последние таблицы мы можем соединить в одну «шахматную» таблицу (табл. 94).

В этой таблице приведены все конкретные значения x и y из исходной таблицы. Числа, стоящие в клетках, на пересечении того или иного x с тем или иным y, означают количество проб с такими показателями. Число 3, например, стоящее на пересечении $x = 1{,}5$ с $y = 24$, означает, что у нас имеется 3 пробы с удельным весом угля 1,5 и зольностью его 24%. $\bar{x}_i$ и $\bar{y}_j$ — средние арифметические по строкам и столбцам соответственно.

Таблица 94

x	y													Всего проб	$\bar{y}_j$
	4	5	6	7	9	17	20	24	25	26	30	33	36		
1,2	1													1	4,00
1,3	1	1		1										3	5,33
1,4			1		1		2							4	13,75
1,5						1		3	1					5	22,80
1,6									1	1				2	25,50
1,7											1	1		2	31,50
1,8													1	1	36,00
Итого проб	2	1	1	1	1	1	2	3	2	1	1	1	1	18	
$\bar{x}_i$	1,25	1,30	1,40	1,30	1,40	1,50	1,40	1,50	1,55	1,60	1,70	1,70	1,80		

Общие средние таковы: $\bar{x} = 1{,}41$, $\bar{y} = 18{,}83$.

Данные, приведенные в «шахматной» таблице, легко вынести на график (рис. 51). На последнем точками показаны все 18 проб. Линии проведены через средние значения $\bar{x}$ и $\bar{y}$. Линия зависимости среднего значения y от конкретных значений x более плавная, чем среднего значения x от конкретных значений y. Линии средних значений x и y ломаные. В ходе корреляционного анализа необходимо найти плавные линии, которые заменили бы эти ломаные линии, так чтобы средние показатели ряда x и средние показатели ряда y остались неизменными.

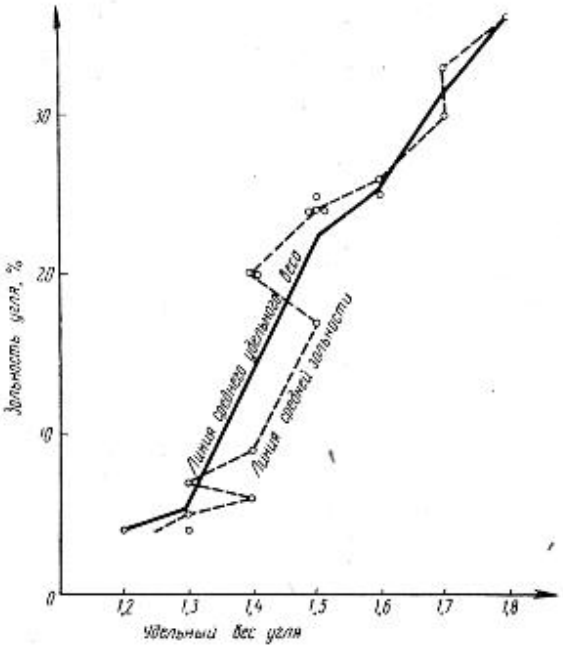


Рис. 51. Связь удельного веса и зольности угля

В данном примере такими линиями будут прямые. Уравнения этих прямых будут найдены ниже.

Такая таблица становится очень громоздкой и неудобной при большом числе значений x и y. Поэтому ее можно преобразовать так, чтобы вместо конкретных x и y брать их интервалы. Последние для x, так же как и для y, могут быть как равными, так и неравными. При этом интервалы по x вообще не равны интервалам по y. Лишь иногда интервалы по x оказываются такими же, как и по y.

Эту таблицу обычно называют корреляционной. Для примера с удельным весом и зольностью угля можно составить корреляционную таблицу (табл. 95).

В этой таблице между соседними интервалами нет перерывов. В других корреляционных таблицах они могут быть, но величина каждого «перерыва» должна быть кратной длине интервала, хотя иногда можно и отступить от этого правила (только методика расчетов в таких случаях изменится).

Изучение корреляционной связи является технической операцией, которая должна дополняться выявлением причинно-следственных отношений в изучаемых процессах или состояниях. Открыв связь между зольностью и удельным весом угля, мы еще не знаем, что здесь является причиной, а что следствием или, если эти оба признака являются параллельными следствиями чего-то третьего, то мы, не нашедши этого третьего, еще не знаем причины соответствия данных признаков друг другу. Так,

Таблица 95

x	y				Всего
	0—10	10—20	20—30	30—40	
1,0—1,2	1				1
1,2—1,4	5	2			7
1,4—1,6		1	6		7
1,6—1,8			1	2	3
Итого . . .	6	3	7	2	18

на удельный вес угля кроме величины зольности влияет состав золы и многие другие факторы, в частности степень метаморфизма или степень углефикации органического вещества, влажность и пр.

Выше говорилось, что корреляционная связь бывает положительной и отрицательной, иногда ее называют прямой и обратной. Прямая связь такая, когда с ростом одного показателя вообще растет и другой, или, что то же самое, когда с уменьшением одного показателя вообще уменьшается и другой. Пример прямой корреляции — связь зольности угля с его удельным весом. При этом отдельные пробы отклоняются от этого правила, но большинство проб подчиняется ему. Обратная связь, наоборот, показывает уменьшение одного признака с ростом другого или, что то же самое, рост одного признака с уменьшением другого.

Пример обратной корреляционной связи: с увеличением стратиграфической глубины залегания угля вообще падает выход летучих (правило Хильта), или второй пример: с уменьшением зольности обогащенного угля в большинстве проб повышается его калорийность.

Используя данные примера с зольностью и удельным весом угля, приведенного выше, оценим коэффициент корреляции (табл. 96).

Таблица 96

x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})\times(y_i-\bar{y})$
1,2	4	—0,28	—14,8	0,0784	219,04	4,144
1,3	4	—0,18	—14,8	0,0324	219,04	2,664
1,3	5	—0,18	—13,8	0,0324	190,44	2,484
1,3	7	—0,18	—11,8	0,0324	139,24	2,124
1,4	6	—0,08	—12,8	0,0064	163,84	1,024
1,4	9	—0,08	—9,8	0,0064	96,04	0,784
1,4	20	—0,08	1,2	0,0064	1,44	—0,096
1,4	20	—0,08	1,2	0,0064	1,44	—0,096
1,5	17	0,02	—1,8	0,0004	3,24	—0,036
1,5	24	0,02	5,2	0,0004	27,04	0,104
1,5	24	0,02	5,2	0,0004	27,04	0,104
1,5	24	0,02	5,2	0,0004	27,04	0,104
1,5	25	0,02	6,2	0,0004	38,44	0,124
1,6	25	0,12	6,2	0,0144	38,44	0,744
1,6	26	0,12	7,2	0,0144	51,84	0,864
1,7	30	0,22	11,2	0,0484	125,44	2,464
1,7	33	0,22	14,2	0,0484	201,64	3,124
1,8	36	0,32	17,2	0,1024	295,84	5,504
Всего 26,6	339			0,4312	1866,52	26,132

В этой таблице x означает удельный вес, а y — зольность. Средний удельный вес $\bar{x}=\frac{26,6}{18}\approx 1,48$, а средняя зольность $\bar{y}=\frac{339}{18}\approx 18,8$.

Вычислим сначала оценки средних квадратичных отклонений:

$$s_x=\sqrt{\frac{0,4312}{18}}=0,1548\approx 0,155,$$

$$s_y=\sqrt{\frac{1866,52}{18}}=10,128\approx 10,13.$$

Теперь вычислим оценку коэффициента корреляции

$$r=\frac{26,132}{18\cdot 0,155\cdot 10,13}=0,94.$$

При вычислении оценки коэффициента корреляции по сгруппированным данным удобно пользоваться следующей формулой:

$$r = \frac{\sum_{i=1}^{m}\sum_{j=1}^{k} n_{ij}x_i y_j - \frac{1}{N}\left(\sum_{i=1}^{m} n_{i\cdot}x_i\right)\left(\sum_{j=1}^{k} n_{\cdot j}y_j\right)}{\sqrt{\left[\sum_{i=1}^{m} n_{i\cdot}x_i^2 - \frac{1}{N}\left(\sum_{i=1}^{m} n_{i\cdot}x_i\right)^2\right]\left[\sum_{j=1}^{k} n_{\cdot j}y_j^2 - \frac{1}{N}\left(\sum_{j=1}^{k} n_{\cdot j}y_j\right)^2\right]}},$$

где $x_1, x_2, \ldots, x_i, \ldots, x_m$ — середины интервалов, на которые разделена область значений x;

$y_1, y_2, \ldots, y_j, \ldots, y_k$ — аналогичные величины для y;

n_{ij} — число наблюдений, соответствующих x_i, y_j;

$n_{i\cdot} = \sum_{j=1}^{k} n_{ij}$ — число наблюдений, соответствующих значению x_i;

$n_{\cdot j} = \sum_{i=1}^{m} n_{ij}$ — число наблюдений, соответствующих y_j;

$N = \sum_{i=1}^{m}\sum_{j=1}^{k} n_{ij}$ — объем выборки.

Оценить коэффициент корреляции можно и другим способом, не используя в вычислениях значений середин интервалов группировки. Рассмотрим этот способ на следующем примере.

На одном из коренных месторождений колумбита было взято (из керна буровых скважин и из опробовательских борозд в горных выработках) 834 пробы. Анализом определялось содержание пятиокиси ниобия, двуокиси циркония и ряда других компонентов. По результатам анализа этих проб автором была составлена корреляционная таблица (табл. 97).

В этой таблице x означает содержание ZrO_2, а y — содержание Nb_2O_5 (и то, и другое в процентах).

В данной таблице — 9 строк и 9 столбцов. Всего, следовательно, имеется 81 клетка. В некоторых клетках стоят числа, означающие количество проб. Число 14, например, стоящее в первой слева вверху клетке, означает, что из 834 проб 14 проб с содержанием ZrO_2 от 0,02 до 0,06% и Nb_2O_5 — от 0,01 до 0,02% (в тех же пробах).

Если же в какой-либо пробе содержание Nb_2O_5 будет больше 0,02%, но не больше 0,03%, то она попадет в следующую строку.

При составлении корреляционных таблиц необходимо так выбирать величину интервалов x и y, т. е. так выбрать количество строк и столбцов, чтобы, во-первых, относительно наибольшее число проб не оказалось в самой крайней строке или в самом крайнем столбце, а во-вторых, придерживаются того правила, которое выше было описано для выбора числа классов в вариационном ряду (как для x, так и для y). Однако от этого правила можно отступать, когда точность определения величины признака x или y (в приведенном примере — точность химического анализа проб на цирконий и ниобий) настолько мала, что фактическая ошибка анализа оказывается меньше размера интервала x или y.

В классическом корреляционном анализе интервалы величины x должны быть одинаковы для всей таблицы. Также одинаковы должны быть интервалы величины y, но интервалы по x вообще не равны интервалам по y, хотя в отдельных случаях они и могут совпасть.

Если в какой-либо клетке нет ни одной пробы, то нуль в ней обычно не ставится, хотя если бы он и был проставлен, на расчетах это не отразится.

В корреляционной таблице справа и снизу примыкают однотипные расчетные графы. Порядок заполнения этих граф, т. е. порядок расчетов таков.

Таблица 97

y	x										n_l	$n_l l$	$n_l l^2$	$n_{kl}k$	$\sum_k n_{kl}k$	$l\sum_k n_{kl}k$
	0,02—0,06	0,06—0,10	0,10—0,14	0,14—0,18	0,18—0,22	0,22—0,26	0,26—0,30	0,30—0,34	0,34—0,38	l						
0,01—0,02	14	19	1							—3	34	—102	306	—42, —38, —1	—81	243
0,02—0,03	32	117	25	3					1	—2	177	—354	708	—96, —234, —25	—355	710
0,03—0,04	17	129	98	15	4	1			1	—1	265	—265	265	—51, —258, —98, 4, 2, 5	—396	396
0,04—0,05	3	52	73	60	16	5	2			0	211	—	—	—	—	—
0,05—0,06	2	6	30	44	22	7	2			1	113	113	113	—6, —12, —30, 22, 14, 6	—6	—6
0,06—0,07			3	2	6	8	2	1	1	2	23	46	92	—3, 6, 16, 6, 4, 5	34	68
0,07—0,08				2	4	2				3	8	24	72	4, 4	8	24
0,08—0,09							1	1		4	2	8	32	3, 4	7	28
0,09—0,10		1								5	1	5	25	—2	—2	—10
k	—3	—2	—1	0	1	2	3	4	5		831	—525	1613			1453
n_k	68	324	230	126	52	23	7	2	2	834						
$n_k k$	—240	—648	—230	—	52	46	21	8	10	—945						
$n_k k^2$	612	1296	230	—	52	92	63	32	50	2427						
$n_{lk}l$	—42 —64 —17 2	—57 —234 —129 6	—3 —50 —98 30		—4 22 12 12	—1 7 16 6	2 4 4	2 4	—1 2							
$\sum_l n_{lk}l$		5	6													
$k\sum_l n_{lk}l$	—121	—409	—115	—	42	28	10	6	1							
	363	818	115	—	42	56	30	24	5	1453						

В столбце l приведены номера строк, а в строке k — номера столбцов. При этом нумеруются и те строки (и те столбцы), в которых нет ни одной пробы. Нулевой номер по l обычно ставится в промежутке между средней строкой и строкой с наибольшим числом проб. Там же ставится нуль и по k.

Вверх от нуля по l идут последовательно отрицательные номера интервалов, т. е. —1, —2 и т. д. Вниз от этого нуля идут последовательно положительные номера интервалов, т. е. 1, 2 и т. д.

Подобным образом и по k влево от нуля идут отрицательные, а вправо от нуля — положительные номера. Номера интервалов, т. е. величина k и величина l называются также рангом величины x (в первом случае) и y (во втором). От места нуля результат вычислений не зависит, но при удачном положении нуля расчеты облегчаются. В нашем примере нулевое значение x и y, а также длины интервалов (d) по x и по y таковы: $x_0 = 0{,}160$; $y = 0{,}045$; $d_x = 0{,}040$; $d_y = 0{,}010$.

Столбец n_l и строка n_k означают число проб в данной строке или в данном столбце. Итог по n_l равен итогу по n_k, т. е. $\sum n_l = \sum n_k = n$.

Столбец $n_l l$ и строка $n_k k$ — произведение числа проб и номера строки в первом случае и столбца — во втором. При записи этих величин нужно следить за знаком минус или плюс, хотя последний можно и не проставлять, но нужно подразумевать.

Столбец $n_l l^2$ и строка $n_k k^2$ — произведение чисел предыдущего столбца (строки) и номера столбца (строки). По этим двум столбцам и строкам подсчитываются итоги. В нашем примере это будут числа: —525, 1613, —945 и 2427.

Столбец $n_{kl}k$ заполняется так: число проб в каждой клетке строки l, т. е. число n_{kl}, стоящее в строке l и в столбце k умножается на номер столбца k, например, 14 умножается на —3, а результат —42 записывается в этом широком столбце. Далее 19 умножаем на —2 (получаем —38), число 1 умножаем на —1 (получаем —1). Сумма этих чисел $(-42) + (-38) + (-1) = -81$ записывается в следующем столбце, обозначенном $\sum n_{kl}k$.

Соответствующим образом заполняются и графы $n_{lk}l$ и $\sum n_{lk}l$, только число проб в каждой клетке графы умножается не на номер столбца, а на номер строки (т. е. на l).

Далее идет столбец $l\sum n_{kl}k$. В нем записано произведение числа, стоящего в предыдущем столбце, и номера строки, например, произведение $(-81)\cdot(-3) = 243$ или $(-355)\cdot(-2) = 710$. Итоги по этим графам обязательно будут одинаковы. В нашем примере это число 1453.

Если эти итоги будут отличаться друг от друга, то это укажет нам на наличие ошибки в расчетах. Ее надо найти и устранить.

Номер класса величины x (в равноинтервальной корреляционной таблице) определяется по формуле

$$k = \frac{x - x_0}{d_x},$$

где k — (номер класса) величины x;

x — значение первого признака (середина интервала);

x_0 — место нуля величины x;

d_x — длина интервала для x.

Номер класса величины y (в равноинтервальной корреляционной таблице) определяется по аналогичной формуле

$$l = \frac{y - y_0}{d_y},$$

где l — номер класса величины y;

y — значение второго признака (середина интервала);

y_0 — место нуля величины y;

d_x — длина интервала по y.

Оценку среднего содержания ZrO_2 для всей совокупности проб (т. е. для 834 проб), обозначенную через x, вычисляем по формуле

$$\bar{x} = \frac{\sum n_k k}{n} d_x + x_0 = \frac{-945}{834} 0{,}04 + 0{,}16 = 0{,}115.$$

Оценку среднего содержания Nb_2O_5 вычисляем подобным образом:

$$\bar{y} = \frac{\sum n_l l}{n} d_y + y_0 = \frac{-525}{834} 0{,}01 + 0{,}045 = 0{,}039.$$

Далее вычисляем оценку среднего квадратичного отклонения номеров k и l, т. е. вычислим s_k и s_l:

$$s_k = \sqrt{\frac{\sum n_k k^2}{n} - \left(\frac{\sum n_k k}{n}\right)^2} = \sqrt{\frac{2427}{834} - \left(\frac{-945}{834}\right)^2} = 1{,}275,$$

$$s_l = \sqrt{\frac{\sum n_l l^2}{n} - \left(\frac{\sum n_l l}{n}\right)^2} = \sqrt{\frac{1613}{834} - \left(\frac{-525}{834}\right)^2} = 1{,}240.$$

По этим данным можно вычислить эмпирическое среднее квадратичное отклонение величин x и y:

$$s_x = s_k d_x = 1{,}275 \cdot 0{,}04 = 0{,}0510,$$

$$s_y = s_l d_y = 1{,}240 \cdot 0{,}01 = 0{,}0124.$$

Оценка коэффициента корреляции r вычисляется по формуле

$$r = \frac{\hat{\mu}_{2xy} - \hat{\mu}_{1x}\hat{\mu}_{1y}}{\sqrt{\hat{\mu}_{2x} - \hat{\mu}_{1x}^2}\sqrt{\hat{\mu}_{2y} - \hat{\mu}_{1y}^2}},$$

где $\hat{\mu}_{1x}$ — оценка первого момента номеров для x;
$\hat{\mu}_{1y}$ — » » » » » y;
$\hat{\mu}_{2x}$ — » второго » » » x;
$\hat{\mu}_{2y}$ — » » » » » y;
$\hat{\mu}_{xy}$ — » смешанного » » » xy;
или по формуле

$$r = \frac{\hat{\mu}_{2xy} - \hat{\mu}_{1x}\hat{\mu}_{1y}}{s_k s_l}.$$

Оценки моментов номеров для величин x и y определяются по формулам:

$$\hat{\mu}_{1x} = \frac{\sum n_k k}{n},$$

$$\mu_{1y} = \frac{\sum n_l l}{n},$$

$$\hat{\mu}_{2x} = \frac{\sum n_k k^2}{n},$$

$$\hat{\mu}_{2y} = \frac{\sum n_l l^2}{n},$$

$$\hat{\mu}_{xy} = \frac{\sum (l \sum n_{kl} k)}{n}$$

или

$$\hat{\mu}_{xy} = \frac{\sum (k \sum n_{kl} l)}{n}.$$

Если не вычислять моментов, то выражение для коэффициента корреляции можно представить так:

$$r = \frac{\sum (l \sum n_{kl} k) - \sum n_k k \cdot \sum n_l l}{n s_k s_l}.$$

При большом значении n величина r по последней формуле вычисляется значительно легче, чем по ранее приведенной.

В нашем примере имеем:

$$\hat{\mu}_{1x} = \frac{-945}{834} = -1{,}136,$$

$$\hat{\mu}_{1y} = \frac{-525}{834} = -0{,}630,$$

$$\hat{\mu}_{2x} = \frac{2427}{834} = 2{,}914,$$

$$\hat{\mu}_{2y} = \frac{1613}{834} = 1{,}937,$$

$$\hat{\mu}_{xy} = \frac{1453}{834} = 1{,}746.$$

Подставив значения этих пяти моментов в формулу коэффициента корреляции, получим

$$r = \frac{1{,}746 - (-1{,}136)(-0{,}630)}{\sqrt{2{,}914 - (-1{,}136)^2}\sqrt{1{,}937 - (-0{,}630)^2}} = 0{,}6527 \simeq 0{,}65.$$

По вычисленному значению выборочного коэффициента корреляции (r) требуется проверить гипотезу $H_0 : \varrho = 0$ при альтернативе $H_1 : \varrho \neq 0$.

Если нулевая гипотеза будет принята, то делать вывод о наличии корреляционной зависимости нельзя. Если же H_0 будет отвергнута и принята H_1, то из этого следует, что рассматриваемая пара случайных величин связана корреляционной зависимостью.

В качестве статистического критерия для проверки гипотезы $H_0 : \varrho = 0$ обычно используется величина

$$t = \frac{|r|}{\sqrt{1 - r^2}}\sqrt{n-2},$$

(где n — число наблюдений), которая в условиях H_0 распределена по закону Стьюдента с $n - 2$ степенями свободы. Таким образом, H_0 отвергается (т. е. зависимость считается установленной), если t превысит допустимое значение $t_{q, n-2}$ при уровне значимости q и $n - 2$ степенях свободы.

Используя приведенное выше выражение, можно при заданных величинах $t_{q, n-2}$ и n вычислить допустимые значения $r_{q, n-2}$. Гипотеза H_0 отвергается, если вычисленное значение r превысит $r_{q, n-2}$. Таблицы значений $r_{q, n-2}$ можно найти в книге Ван дер Вардена (1960).

Для примера связи удельного веса угля с зольностью имеем

$$t = \frac{0{,}94}{\sqrt{1 - 0{,}94^2}}(18 - 2) = 44.$$

Допустимое значение t при уровне значимости 0,001 и 16 степенях свободы равно 4,015. Из этого делаем вывод, что отличие r от нуля существенное, т. е. связь реальная.

Для примера с пятиокисью ниобия и циркония соответственно получаем

$$t = \frac{0{,}65}{\sqrt{1 - 0{,}65^2}}(834 - 2) = 715.$$

Для этого случая связь еще более реальная.

Другой критерий для проверки гипотезы $\varrho = 0$ предложен Фишером, который ввел новую переменную

$$z = \frac{1}{2}\ln\frac{1+r}{1-r} = 1{,}1513\lg\frac{1+r}{1-r},$$

причем

$$-1 \leqslant r \leqslant 1$$

$$-\infty < z < \infty.$$

Таким образом, если $\varrho = 0$, то случайная величина

$$u = \frac{z}{\sqrt{\frac{1}{n-3}}} = z\sqrt{n-3}$$

будет распределена приблизительно нормально с параметрами 0,1. Это обстоятельство можно использовать при выборе критической области для отклонения гипотезы $\varrho = 0$. Так, если вычисленное значение u превысит 1,96, гипотеза $\varrho = 0$ может быть отклонена при уровне значимости 0,05.

Преобразование (z) Фишера можно использовать при построении доверительного интервала для ϱ. Используя нормальное приближение, можно вычислить

$$z_1 = z + u_q\sigma_z$$
$$z_2 = z - u_q\sigma_z,$$

где u_q — значения нормальной функции с параметрами 0,1, соответствующие доверительной вероятности $1 - q$.

Значения r_1 и r_2 вычисляются путем подстановки z_1 и z_2 в уравнение

$$r = \frac{e^{2z} - 1}{e^{2z} + 1}.$$

Распределение величины z можно использовать также для проверки гипотезы о равенстве двух неизвестных коэффициентов корреляции по их оценкам.

Распределение z даже при малых значениях n довольно близко к нормальному со средним

$$Mz \simeq \frac{1}{2}\ln\frac{1+\varrho}{1-\varrho} + \frac{\varrho}{2(n-1)}$$

и дисперсией

$$Dz \simeq \frac{1}{n-3}.$$

Следовательно, распределение величины

$$u = \frac{z - Mz}{\sqrt{\frac{1}{n-3}}} = (z - Mz)\sqrt{(n-3)}$$

близко к нормальному с параметрами (0,1).

Если, например, для двухмерных совокупностей объема n_1 и n_2 оценки коэффициентов корреляции соответственно равны r_1 и r_2, то мы можем проверить гипотезу $\varrho_1 = \varrho_2 = \varrho$, заменив ее равносильной $\zeta_1 = -\zeta_2 = \zeta$. Если гипотеза верна, то величины

$$z_1 = \frac{1}{2}\ln(1 + r_1) - \ln(1 - r_1)$$

и

$$z_2 = \frac{1}{2}\ln(1 + r_2) - \ln(1 - r_2)$$

распределены нормально со средней ζ и дисперсиями, соответственно равными $\frac{1}{n_1 - 3}$ и $\frac{1}{n_2 - 3}$.

Таким образом, величина

$$u = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

должна иметь нормальное распределение с параметрами (0,1). Если осуществляется неравенство $|u| < u_q$, где u_q равно, например, 1,96, то гипотеза не отклоняется.

Гипотеза $H_0 : \varrho_1 = \varrho_2 = \cdots = \varrho_k = \varrho$ может быть проверена по оценкам $r_1, r_2, \ldots, r_k$ с помощью χ^2 — распределения; так как

$$\sum_{i=1}^{k} u_i^2 = \sum_{i=1}^{k} (z_i - \zeta)^2 (n_i - 3) = \chi^2$$

распределена в условиях H_0 по закону χ^2 с k степенями свободы.

Величину ζ обычно заменяют оценкой

$$\bar{z} = \frac{\sum_{i=1}^{k} (n_i - 3) z_i}{\sum_{i=1}^{k} (n - 3)},$$

т. е. получают

$$\chi^2 = \sum_{i=1}^{k} (z_i - \bar{z})^2 (n_i - 3).$$

Поэтому величина χ^2 здесь распределена с $k - 1$ степенями свободы.

Гипотеза H_0 отвергается, если вычисленное значение χ^2 превысит допустимое при заданном уровне значимости и $k - 1$ степени свободы.

Доверительный интервал для ϱ можно построить и упрощенным способом. Мерой отклонения r от ϱ служит оценка среднего квадратичного отклонения величины r, определяемая по формуле

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

С помощью этого отклонения строят критическую область вида

$$|r| > u_p s_r$$

с уровнем значимости p.

Доверительный интервал величины ϱ с p-процентным уровнем значимости определяется формулой

$$r - u_p \frac{1 - r^2}{\sqrt{n}} < \varrho < r + u_p \frac{1 - r^2}{\sqrt{n}}.$$

Если величина r не попадает в эту область, то гипотезу $\varrho = 0$ отвергают, а связь считают реальной.

Есть еще более простой способ проверки значимости связи.

При небольшом объеме выборки (n менее 30—50) для этого можно воспользоваться критерием Романовского (R_r)

$$R_r = |r| \sqrt{n-1} \geqslant 3.$$

Если это неравенство выполнено, то полученную оценку коэффициента корреляции считают существенной. В противном случае это уклонение от нуля оценки следует признать случайным.

Для приведенного примера связи удельного веса с зольностью угля

$$|r| \sqrt{n-1} = 0{,}94 \sqrt{18-1} = 3{,}83 > 3.$$

Из этого следует, что оценка коэффициента корреляции существенно отличается от нуля.

Для большого объема выборки существенность или надежность связи проверяется обычно по другой, тоже упрощенной формуле, в которой участвует среднее квадратичное отклонение вычисленного значения коэффициента корреляции. Это отклонение определяется по формуле

$$\sigma_r = \frac{1-r^2}{n}.$$

В примере с Nb_2O_5 и ZrO_2 имеем

$$\sigma_r = \frac{1-0{,}65^2}{\sqrt{834}} = 0{,}02.$$

Поэтому мы считаем, что ϱ содержится в интервале $0{,}65 \pm 0{,}02$.

Условием надежности этой оценки (по Романовскому) является формула

$$r > 3\sigma_r.$$

В данном случае $r = 0{,}65 > 3 \cdot 0{,}02 = 0{,}06$, т. е. связь выявлена надежно и она тесная, так как $|r| > 0{,}5$, положительная, так как величина r положительная.

Поскольку каждому значению x соответствует среднее значение y и каждому значению y соответствует среднее значение x, мы можем, исходя из меры силы связи, т. е. из коэффициента корреляции, составить уравнение связи x с y или y с x. Эти уравнения называются также уравнениями регрессии x на y и y на x.

Формула уравнения регрессии x на y такова:

$$x = r \frac{s_x}{s_y} (y - \bar{y}) + \bar{x},$$

а y на x:

$$y = r \frac{s_y}{s_x} (x - \bar{x}) + \bar{y}.$$

Термин «уравнение регрессии» впервые был применен в биологических исследованиях (в области наследования признаков), а затем получил обобщающее значение в статистике, хотя он и представляет некоторые неудобства. Вместо «уравнения регрессии» часто говорят «уравнение связи».

Так, если $r = 0{,}65$, $x = 0{,}115$, $y = 0{,}039$, $s_x = 0{,}0510$, $s_y = 0{,}0124$, получим уравнения регрессии

$$x = 0{,}65\,\frac{0{,}0510}{0{,}0124}(y - 0{,}039) + 0{,}115;$$

$$y = 0{,}65\,\frac{0{,}0124}{0{,}0510}(x - 0{,}115) + 0{,}039$$

или в окончательном виде

$$x = 2{,}683y + 0{,}010$$
$$y = 0{,}159x + 0{,}021$$

Уравнениями связи, если коэффициент корреляции реален или достаточно надежен, т. е. если $\varrho \neq 0$, можно воспользоваться для определения величины одного признака, исходя из величины другого, в частности, можно определять содержание элемента-примеси в руде, исходя из содержания главного элемента.

Из уравнения регрессии y на x можно найти значение x, если задавать значения y, а из уравнения регрессии x на y можно найти значения y, если задавать значения x. Получается два уравнения регрессии. Они совпадают друг с другом только при $r = \pm 1$. Во всех других случаях это разные уравнения.

Статистики-теоретики (Слуцкий, 1912; Чупров, 1926) считают эту двойственность неизбежной, лежащей в основе корреляционного анализа.

Статистики-практики не могут примириться с этой двойственностью. Они предложили несколько методов единого или срединного уравнения регрессии (Обухов, 1923; Срезневский, 1928; H. E. Jones, 1937; C. Gini, 1939; Конюс, 1949; Шарапов, 1957; Мирчинк, Бухарцев, 1959).

Теоретическое рассмотрение этого вопроса («Ортогональная средняя квадратическая регрессия») имеется у Г. Крамера (1948).

В формуле уравнения регрессии x на y величина $r\frac{s_x}{s_y}$ называется эмпирическим коэффициентом регрессии. То же название носит и величина $r\frac{s_y}{s_x}$ в уравнении регрессии y на x. Первый из них — коэффициент регрессии x на y, а второй y на x. Коэффициент регрессии можно обозначить через b_x и b_y.

В рассмотренном примере коэффициенты регрессии таковы:

$$b_x = 0{,}65\,\frac{0{,}0510}{0{,}0124} = 2{,}683;$$

$$b_y = 0{,}65\,\frac{0{,}0124}{0{,}0510} = 0{,}159.$$

Простой способ вычисления эмпирических коэффициентов регрессии дает Э. Вуллей (E. Wooley, 1941).

Применение этого способа в геологической практике описали Раевский и Шурубор (1958).

Коэффициенты регрессии необходимо проверить, а для этого требуется сначала оценить среднее квадратичное отклонение величин b_x и b_y, что делается по следующей формуле:

$$s_{b_x} = \frac{s_x}{s_y}\sqrt{\frac{1 - r^2}{n - 3}}.$$

Романовский (1947) говорит, что это равенство точно выполняется для выборок из нормальной генеральной совокупности.

В приведенном примере распределение величин x и y отклоняется от нормального, поэтому последняя формула даст лишь приближенный результат.

Подставив вычисленные ранее значения s_x, s_y, r и заданную величину n в последнюю формулу, получим

$$s_{b_x} = \frac{0{,}0510}{0{,}0124}\sqrt{\frac{1-0{,}65^2}{834-3}} = 0{,}108.$$

Для вычисления среднего квадратичного отклонения коэффициента регрессии b_y, т. е. величины s_{b_y} служит подобная формула

$$s_{b_y} = \frac{s_y}{s_x}\sqrt{\frac{1-r^2}{n-3}}.$$

Она, как и предыдущая формула, точна для нормальных коллективов. В нашем примере имеем

$$s_{b_y} = \frac{0{,}0124}{0{,}0510}\sqrt{\frac{1-0{,}65^2}{834-3}} = 0{,}064.$$

Оценка коэффициентов регрессии делается путем сравнения с утроенной дисперсией, т. е. по следующим формулам:

$$b_x > 3s_{b_x};$$

$$b_y > 3s_{b_y}.$$

Здесь также проверяются гипотезы $Mb_x = 0$ и $Mb_y = 0$.

В нашем примере по первой из этих двух формул имеем $2{,}683 > 3 \cdot 0{,}108 = 0{,}324$, а это означает, что вычисленный коэффициент регрессии не случайно отклоняется от нуля, а имеет существенное значение.

Другой коэффициент регрессии по второй из приведенных формул будет меньше утроенного значения дисперсии, т. е. $0{,}159 < 3 \cdot 0{,}064 = 0{,}192$; это означает, что он ненадежен, так как величины 0,159 и 0,192 мало отличаются друг от друга.

Описанная здесь проверка гипотезы о равенстве коэффициентов регрессии нулю пригодна только в случае выборок большого объема. В случае малого объема выборок гипотеза проверяется иначе.

Выводы о зависимости, сделанные по выборке из одной совокупности, нельзя механически распространять на другие статистические коллективы.

Каждую совокупность необходимо изучать глубоко и всесторонне. Лучшим средством для этого является разбивка коллектива на части, т. е. группировка по дополнительным признакам, и вычисление оценки коэффициента корреляции для каждой части отдельно. Наряду с разбивкой совокупности на части хорошие результаты иногда дает объединение нескольких частных коллективов в один общий.

Принципы группировки, т. е. дополнительные признаки, могут быть самыми различными. Если, например, мы изучаем ошибки анализов руды, то общую совокупность анализов мы можем разбить на частные группы по лабораториям, в которых производились основные анализы, и получим столько частных совокупностей, сколько было лабораторий. Общий коллектив анализов мы можем разбить также по лабораториям, в которых производился контрольный анализ, по годам анализа, по месту взятия проб, по содержанию в них того или иного элемента, по методу опробования и по многим другим признакам.

Нередко в коллективе, где связь изучаемых признаков (x и y) отсутствует ($\varrho = 0$), можно выделить частные группы, где связь будет заметной. Надо только уметь найти признак, по которому эти частные совокупности можно выделить.

Любую группу числовых данных, характеризующих какой-либо признак, необходимо анализировать статистическими методами, руководствуясь гипотезами о причинно-следственных связях. Так, совокупность чисел, показывающих процентное содержание главного, сопутствующего и других химических элементов в геологических пробах изучаемого место-

рождения полезного ископаемого, необходимо анализировать статистическими методами, руководствуясь данными о генезисе месторождения, о причинах возможных ошибок в химическом анализе проб, о влиянии метода опробования на минералогический состав проб и т. д.

Применение метода группировок для открытия замаскированных связей между фактами, явлениями или признаками, а также между факторами, определившими интересующие нас свойства изучаемых событий или состояний, можно показать на следующем примере.

Пусть x — содержание главного элемента в геологической пробе из одного месторождения (в процентах), а y — содержание элемента-примеси в той же пробе (в граммах на тонну). Пусть далее пробы брались на различных участках, разными методами, а анализировались они в четырех лабораториях (каждая проба анализировалась в одной какой-либо лаборатории). В журнале опробования записано время взятия, обработки и анализа пробы, содержание различных элементов, в том числе главного элемента и элемента-примеси, указан метод взятия пробы, назван участок опробования, проставлены координаты точки опробования и т. д. Анализ всех этих данных может дать очень много интересного, но обратим внимание только на три признака: содержание главного элемента (x), содержание элемента-примеси (y) и место взятия пробы (участок).

Допустим, выборка содержит следующие данные (см. табл. 98).

Если надо узнать, имеется ли связь x с y, то мы должны будем вычислить коэффициент корреляции между ними.

Таблица 98

№ п/п	x_i	y_i	Участок	№ п/п	x_i	y_i	Участок
1	12,5	1,7	Восточный	37	16,0	0,2	Новый
2	8,1	0,9	Кварцевый	38	15,1	1,0	Кварцевый
3	16,8	1,2	»	39	9,0	1,0	»
4	18,0	1,0	»	40	11,1	1,1	Западный
5	10,9	1,4	Новый	41	11,3	1,1	Новый
6	7,2	1,1	Кварцевый	42	15,9	0,7	»
7	14,5	1,5	Западный	43	16,0	1,9	Западный
8	7,3	0,1	»	44	13,0	0,0	Восточный
9	12,4	0,7	Восточный	45	13,1	0,8	Новый
10	12,4	0,6	»	46	13,4	0,9	Кварцевый
11	16,8	1,0	Кварцевый	47	8,7	0,9	»
12	18,1	0,2	Новый	48	11,5	1,0	Восточный
13	7,9	1,3	»	49	14,0	1,2	Кварцевый
14	17,9	0,6	Восточный	50	11,0	0,7	Западный
15	17,8	0,7	»	51	9,1	0,5	»
16	10,1	0,1	Кварцевый	52	12,1	0,8	Восточный
17	8,5	0,7	Восточный	53	15,0	1,4	Западный
18	9,9	0,3	Кварцевый	54	9,3	1,6	Новый
19	14,2	0,6	Новый	55	9,6	0,9	Кварцевый
20	7,3	0,5	»	56	12,7	0,4	Восточный
21	10,9	1,7	Кварцевый	57	14,9	1,3	Западный
22	11,1	1,8	»	58	11,6	0,9	Кварцевый
23	11,2	1,9	»	59	10,4	1,2	»
24	7,9	1,9	Новый	60	12,1	1,6	Восточный
25	10,7	0,8	Западный	61	10,0	0,5	Западный
26	13,0	2,0	Восточный	62	10,9	0,8	»
27	13,0	1,9	»	63	14,2	0,5	Новый
28	16,8	1,9	Западный	64	14,5	2,0	Западный
29	11,4	0,9	Восточный	65	14,4	0,6	»
30	11,1	1,0	Западный	66	14,5	1,5	Новый
31	12,4	1,2	Кварцевый	67	13,3	1,0	Кварцевый
32	13,3	0,7	Новый	68	10,6	1,3	Новый
33	14,3	1,6	Западный	69	11,2	0,9	»
34	14,5	0,3	Новый	70	10,6	0,7	Западный
35	14,9	1,5	Западный	71	11,9	1,3	Восточный
36	10,8	1,3	Новый				

Ниже приводится корреляционная таблица с вычислением коэффициента корреляции (табл. 99).

Вычисление показало, что в этом статистическом коллективе связь x с y отсутствует, но это для опытного глаза видно и без вычисления.

Разложим теперь общий статистический коллектив на частные коллективы (по участкам, где брались пробы) и представим эти частные коллективы в виде корреляционных таблиц.

Для Восточного участка исходные данные и вычисление коэффициента корреляции приведены в табл. 100.

Для Кварцевого участка соответствующие данные приведены в табл. 101.

В этом коллективе связь x с y также отсутствует. Если на Восточном участке x почти не изменяется в связи с изменением y, то на Кварцевом участке x изменяется в широких пределах, а y остается приблизительно одним и тем же.

По Западному участку корреляционная таблица имеет уже иной вид (табл. 102).

Связь x с y явная, так как оценка коэффициента корреляции здесь равна 0,906, а ее стандарт только 0,022.

Мы можем построить уравнение регрессии этой зависимости. Для этого сначала оценим стандарты x и y.

$$s_x = s_k d_x = 1{,}168 \cdot 2 = 2{,}336;$$
$$s_y = s_l d_y = 1{,}242 \cdot 0{,}4 = 0{,}497.$$

Уравнение регрессии x на y таково:

$$x = r\frac{s_x}{s_y}(y - \bar{y}) + \bar{x} = 0{,}906\,\frac{2{,}336}{0{,}497}(y - 1{,}044) + 12{,}333,$$

т. е. $x = 4{,}25y + 7{,}90$.

Уравнение регрессии y на x

$$y = r\frac{s_y}{s_x}(x - \bar{x}) + \bar{y} = 0{,}906\,\frac{0{,}497}{2{,}336}(x - 12{,}333) + 2{,}044,$$

т. е. $y = 0{,}193x - 1{,}34$.

Для Нового участка исходные данные и вычисления приводятся в табл. 103.

Связь x с y ярко выраженная, поэтому мы можем вывести уравнения регрессии.

Оценки стандартов x и y таковы:

$$s_x = 1{,}460 \cdot 2 = 2{,}920;$$
$$s_y = 1{,}228 \cdot 0{,}4 = 0{,}491.$$

Уравнения регрессии:

$$x = -0{,}756\,\frac{2{,}920}{0{,}491}(y - 0{,}911) + 12{,}333 = 4{,}49y + 16{,}43;$$

$$y = -0{,}756\,\frac{0{,}491}{2{,}920}(x - 12{,}333) + 0{,}911 = -0{,}127x + 2{,}481.$$

Сопоставим теперь характеристики рассмотренных статистических коллективов друг с другом (табл. 104):

Таким образом, группировки по участкам оказалось достаточно для того, чтобы обнаружить связь признака x с признаком y на двух участках (Западном и Новом). На Восточном и Кварцевом участках связь между x и y отсутствует, но зато есть некоторые особенности в распределении

Таблица 99

y \ x	7–9	9–11	11–13	13–15	15–17	17–19	l	n_l	$n_l l$	$n_l l^2$	$n_{kl}k$	$\sum_k n_{kl}k$	$l\sum_k n_{kl}l$
0—0,4	1	2	2	1	1	1	−2	8	−16	32	−2, −2, 1, 2, 3	2	−4
0,4—0,8	2	7	3	4	1	2	−1	19	−19	19	−4, −7, 4, 2, 6	1	−1
0,8—1,2	4	2	8	3	3	1	0	21	0	0	—	—	—
1,2—1,6	1	5	2	5			1	13	13	13	−2, −5, 5	−2	−2
1,6—2,0	1	3	3	1	2		2	10	20	40	−2, −3, 1, 4	0	0
k	−2	−1	0	1	2	3		71	−2	104			−7
n_k	9	19	18	14	7	4	71						
$n_k k$	−18	−19	0	14	14	12	3						
$n_k k^2$	36	19	0	14	28	36	133						
$n_{lk}l$	−2	−4	—	−2	−2	−2							
	−2	−7		−4	−1	−2							
	1	5		5	4								
	2	6		2									
$\sum_l n_{lk}l$	−1	0	—	1	1	−4							
$k\sum_l n_{lk}l$	2	0	—	1	2	−12	−7						

$$x_0 = 12;$$
$$y_0 = 1{,}0;$$
$$d_x = 2;$$
$$d_y = 0{,}4;$$

$$\mu_{1x} = \frac{\sum n_k k}{n} = \frac{3}{71} = 0{,}0422;$$

$$\mu_{1y} = \frac{\sum n_l l}{n} = \frac{-2}{71} = -0{,}0282;$$

$$\mu_{2x} = \frac{\sum n_k k^2}{n} = \frac{133}{71} = 1{,}874;$$

$$\mu_{2y} = \frac{\sum n_l l^2}{n} = \frac{104}{71} = 1{,}464;$$

$$\mu_{2xy} = \frac{k\sum n_{lk}l}{n} = \frac{-7}{71} = -0{,}0986.$$

$$r = \frac{\mu_{2xy} - \mu_{1x}\mu_{1y}}{s_k s_l} = -0{,}059;$$

$$s_r = \frac{1 - (-0{,}059)}{\sqrt{71}} = 0{,}126;$$

$$\bar{x} = \mu_{1x} d_x + x_0 = 12{,}0844;$$
$$\bar{y} = \mu_{1y} d_y + y_0 = 0{,}9887;$$
$$s_k = \sqrt{\mu_{2x} - \mu_{1x}^2} = 1{,}37;$$
$$s_l = \sqrt{\mu_{2y} - \mu_{1y}^2} = 1{,}21.$$

Таблица 100

y \ x	7–9	9–11	11–13	13–15	15–17	17–19	n_l	l	$n_l l$	$n_l l^2$	$n_{kl}k$	$\sum n_{kl}k$	$l\sum n_{kl}k$
0—0,4			2				2	−1	−2	2			
0,4—0,8	1		3			2	6	0	0	0	−2,6	4	0
0,8—1,2			2				2	1	2	2			
1,2—1,6			2				2	2	4	8			
1,6—2,0			3				3	3	9	27			
n_k	1	0	12	0	0	2	15		13	39			0

y \ x	7—9	9—11	11—13	13—15	15—17	17—19	n_l	l	$n_l l$	$n_l l^2$	$n_{kl}k$	$\sum n_{kl}k$	$l\sum n_{kl}k$
k	−2	−1	0	1	2	3							
$n_k k$	−2	0	0	0	0	6	4						
$n_k k^2$	4	0	0	0	0	18	22						
$n_{lk}l$			−2 2 4 9										
$\sum_l n_{lk}l$			13										
$k\sum_l n_{lk}l$			0				0						

$x_0 = 12;\ y_0 = 0{,}6;\ d_x = 2;\ d_y = 0{,}4$

$$\mu_{1x} = \frac{4}{71} = 0{,}0564$$

$$\mu_{1y} = \frac{13}{71} = 0{,}1831$$

$$\mu_{2x} = \frac{22}{71} = 0{,}3095$$

$$\mu_{2y} = \frac{39}{71} = 0{,}5490$$

$$\mu_{2xy} = \frac{0}{71} = 0$$

$$\bar{x} = 0{,}0564 \cdot 2 + 12 = 12{,}11$$

$$\bar{y} = 0{,}1831 \cdot 0{,}4 + 0{,}6 = 0{,}68$$

$$s_k = \sqrt{0{,}3095 - 0{,}0564^2} = 0{,}554$$

$$s_l = \sqrt{0{,}5490 - 0{,}1831^2} = 0{,}716$$

$$r = \frac{0 - 0{,}0564 \cdot 0{,}1831}{0{,}554 \cdot 0{,}716} = \frac{-0{,}001}{0{,}397} = -0{,}00252$$

$$s_r = \frac{1 - (-0{,}00252)}{\sqrt{15}} = 0{,}26$$

Таблица 101

y \ x	7—9	9—11	11—13	13—15	15—17	17—19	n_l	l	$n_l l$	$n_l l^2$	$n_{kl}k$	$\sum n_{kl}k$	$l\sum n_{kl}k$
0—0,4		2					2	−2	−4	8	−2	−2	4
0,8—1,2	4	2	2	3	3	1	15	0	0	0	−8, −2, 3, 6, 3	2	0
1,6—2,0		1	2				3	2	6	12	−1	−1	−2
n	4	5	4	3	3	1	20		2	20			2
k	−2	−1	0	1	2	3							
$n_k k$	−8	−5	0	3	6	3	−1						
$n_k k^2$	16	5	0	3	12	9	45						
$n_{lk}l$	0	−4	4	0	0	0							
		2											
$\sum n_{lk}l$	0	−2	4	0	0	0							
$k\sum n_{lk}l$	0	2	0	0	0	0	2						

$x_0 = 12$

$y_0 = 1{,}0$

$d_x = 2$

$d_y = 0{,}4$

$$\mu_{1x} = \frac{-1}{20} = -0{,}05$$

$$\mu_{1y} = \frac{2}{20} = 0{,}10$$

$$\mu_{2x} = \frac{45}{20} = 2{,}25$$

$$\mu_{2y} = \frac{20}{20} = 1{,}00$$

$$\mu_{2xy} = \frac{2}{20} = 0{,}10$$

$$\bar{x} = -0{,}05 \cdot 2 + 12 = 11{,}90$$

$$\bar{y} = 0{,}1 \cdot 0{,}4 + 1{,}0 + 1{,}04$$

$$s_k = \sqrt{2{,}25 - (-0{,}05)^2} = 1{,}50$$

$$s_l = \sqrt{1{,}00 - (0{,}1)^2} = 0{,}99$$

$$r = \frac{0{,}1 - (-0{,}05) \cdot 0{,}1}{1{,}5 \cdot 0{,}99} = 0{,}07$$

$$s_r = \frac{1 - 0{,}07}{\sqrt{20}} = 0{,}21$$

$$3s_r = 0{,}63$$

Таблица 102

y	x											
	7—9	9—11	11—13	13—15	15—17	n_l	l	$n_l l$	$n_l l^2$	$n_{kl}k$	$\sum n_{kl}k$	$l\sum n_{kl}k$
0—0,4	1					1	—2	—2	4	—2	—2	4
0,4—0,8		6		1		7	—1	—7	7	—6,1	—5	5
0,8—1,2			2			2	0	0	0	0	0	0
1,2—1,6				5		5	1	5	5	5	5	5
1,6—2,0				1	2	3	2	6	12	1,4	5	10
n_k	1	6	2	7	2	18		2	28			24
k	—2	—1	0	1	2							
$n_k k$	—2	—6	0	7	4	3						
$n_k k^2$	4	6	0	7	8	25						
$n_{lk}l$	—2	—6	0	—1	4							
				5								
				2								
$\sum n_{kl}l$	—2	—6	0	6	4							
$k\sum n_{kl}l$	4	6	0	6	8	24						

$$x_0 = 12$$

$$y_0 = 1{,}0$$

$$d_x = 2{,}0$$

$$d_y = 0{,}4$$

$$\mu_{1x} = \frac{3}{18} = 0{,}167$$

$$\mu_{1y} = \frac{2}{18} = 0{,}111$$

$$\mu_{2x} = \frac{25}{18} = 1{,}390$$

$$\mu_{2y} = \frac{28}{18} = 1{,}555$$

$$\mu_{2xy} = \frac{24}{18} = 1{,}335$$

$$\bar{x} = 0{,}167 \cdot 2 + 12 = 12{,}333$$

$$\bar{y} = 0{,}111 \cdot 0{,}4 + 1{,}0 = 1{,}044$$

$$s_k = \sqrt{1{,}390 - (-0{,}167)^2} = 1{,}168$$

$$s_l = \sqrt{1{,}555 - (-0{,}111)^2} = 1{,}242$$

$$r = \frac{1{,}335 - 0{,}167 \cdot 0{,}111}{1{,}168 \cdot 1{,}242} = 0{,}906$$

$$s_r = \frac{1 - 0{,}906}{\sqrt{18}} = 0{,}022$$

$$3s_r = 0{,}066$$

Таблица 103

y	x												
	7—9	9—11	11—13	13—15	15—17	17—19	n_l	l	$n_l l$	$n_l l^2$	$n_{lk}k$	$\sum n_{lk}k$	$l\sum n_{lk}k$
0—0,4				1	1	1	3	—2	—6	12	1, 2, 3	6	—12
0,4—0,8	1			4	1		6	—1	—6	6	—2, 4, 2	4	—4
0,8—1,2			2				2	0	0	0	0	0	0
1,2—1,6	1	4		1			6	1	6	6	—2, —4, 1	—5	—5
1,6—2,0	4						1	2	2	4	—2	—2	—4
n_k	3	4	2	6	2	1	18		—4	28			—25
k	—2	—1	0	1	2	3							
$n_k k$	—6	—4	0	6	4	3	3						
$n_k k^2$	12	4	0	6	8	9	39						
$n_{lk}l$	—1	4	0	—2	—2	—2							
	1			—4	—1								
	2			1									
$\sum n_{lk}l$	2	4	0	—5	—3	—2							
$k\sum n_{lk}l$	—4	—4	0	—5	—6	—6	—25						

$$x_0 = 12$$

$$y_0 = 1{,}0$$

$$d_x = 2$$

$$d_y = 0{,}4$$

$$\mu_{1x} = \frac{3}{18} = 0{,}167$$

$$\mu_{1y} = \frac{-4}{18} = -0{,}222$$

$$\mu_{2x} = \frac{39}{18} = 2{,}165$$

$$\mu_{2y} = \frac{28}{18} = 1{,}556$$

$$\mu_{2xy} = \frac{-25}{18} = -1{,}390$$

$$\bar{x} = 0{,}167 \cdot 2 + 12 = 12{,}333$$

$$\bar{y} = -0{,}222 \cdot 0{,}4 + 1{,}0 = 0{,}911$$

$$s_k = \sqrt{2{,}165 - 0{,}167^2} = 1{,}460$$

$$s_l = \sqrt{1{,}556 - (-0{,}222)^2} = 1{,}228$$

$$r = \frac{-1{,}39 - 0{,}167\,(-0{,}222)}{1{,}460 \cdot 1{,}228} = -0{,}756$$

$$s_r = \frac{1 - (+0{,}756)}{\sqrt{18}} = 0{,}058$$

$$3s_r = 0{,}174.$$

Таблица 104

Коллектив	n	$\bar{x}$	$\bar{y}$	r	$\bar{s}_r$	R_r
Общий	71	12 084	0,989	—0,059	0,126	0,5
Восточный участок	15	12,110	0,680	—0,003	0,260	0,0
Кварцевый »	20	11,900	1,040	0,070	0,210	0,3
Западный »	18	12,333	1,044	0,906	0,022	3,7
Новый »	18	12,333	0,911	—0,756	0,058	3,1

этих величин. На Восточном участке x почти постоянен, но y изменяется в широких пределах, а на Кварцевом участке y почти постоянен, а x изменяется в широких пределах. Природа этих закономерностей нуждается в объяснении, но для этого нужны специальные исследования.

Причины связи x с y по Западному и Новому участкам (в одном случае прямой, в другом — обратной) необходимо выяснить, но это уже не статистическая, а геологическая задача.

Если говорить о более сложных формах связи (криволинейная и множественная корреляция), то группировка может дать любой, даже самый неожиданный результат. Поэтому нельзя успокаиваться на открытии связи. Ее еще следует проверить с помощью группировки по разным факторам, оценить ее реальность, т. е. проверить, не случайна ли она.

XI. МНОЖЕСТВЕННАЯ КОРРЕЛЯЦИЯ

В практике разведки нередко встречаются случаи, когда интересующий нас признак существенно зависит не от одного, а от двух и большего числа других признаков. Корреляция таких величин называется множественной. В зависимости от числа определяющих элементов различают связь и корреляцию двухмерную, трехмерную, четырехмерную и т. д.

Примером двухмерной связи может служить связь кобальта с серой и железом в железной руде контактово-метасоматического (скарнового) типа. Тройной является связь брома с магнием, калием и натрием в калийных солях. Четверная связь имеется у монацита с ильменитом, рутилом, цирконом и магнетитом в песках. Еще более многосторонняя связь имеется у рассеянных элементов (индий, галлий, таллий, германий, теллур, селен) в медноколчеданной руде, содержащей кроме меди и названных элементов серу, железо, цинк, свинец, мышьяк, сурьму, висмут, молибден, кобальт, кадмий, серебро, золото и другие элементы.

Множественная связь в разведке встречается не только при изучении состава руды, но и при подсчете запасов, точность которого связана с множеством условий (степенью разведанности и степенью изменчивости месторождения, генезисом тел полезного ископаемого, методом опробования, особенностями анализа проб и т. д.). Такая же связь ставит нам свои загадки при поисках слепых рудных тел, при параллелизации угольных пластов, при изучении водоносности тех или иных горизонтов и во многих других случаях.

Механизм связи в разных случаях может быть различным. Он определяется главным образом генезисом месторождений и геохимией элементов, входящих в состав руды. Так, например, кобальт в железной руде скарнового типа связан с зонами пиритизации. Его основная масса сконцентрирована в кристаллах пирита. Возможно не весь пирит содержит в себе кобальт. Есть, по-видимому, и некобальтоносная генерация пирита, как есть и внепиритный кобальт. Изучение генетического механизма связи кобальта с пиритом и вообще связи элементов очень важно для разведки. Статистический анализ состава руды должен опираться на геологические закономерности. Он может подтвердить или не подтвердить правильность геологических гипотез, а иногда может дать толчок к появлению новых направлений исследования.

Связь брома с магнием, калием и натрием в калийных солях определена закономерностями выпадения солей из сгущающегося рассола. Бром выделяется непрерывно в течение всего процесса садки солей. У него наблюдается обратная связь с хлором. По-видимому, хлор частично замещается бромом в различных хлоридах, особенно в карналлите. При этом содержание брома в солях первой фазы кристаллизации, когда отлагается почти один только галит, невысокое, в солях второй фазы кристаллизации (сильвинит) — более высокое, а в солях третьей фазы (карналлит и сильвин с примесью галита) — наибольшее.

Форма связи брома с другими элементами в карналлите изучена мало, а в сильвините и галите совсем не изучена. Однако эта неизученность не мешает практически пользоваться эмпирическими закономерностями в количественных отношениях элементов. По теоретически пока необъяснимым, но фактически существующим связям элементов можно, например, подсчитать запасы элементов-примесей. Поиски теоретических объяснений необходимы, но это не должно мешать практическому использованию эмпирических обобщений, сделанных с помощью статистики.

Выбор вида связи (двойная, тройная, четверная и т. д.) зависит от того, какие компоненты нас интересуют и какая степень точности результата допускается. Эту степень можно приближенно определить по оценочным критериям, которые будут рассмотрены ниже.

Множественная корреляция любого вида (двойная, тройная и т. д.) может быть в одних случаях линейной, в других нелинейной (кривые второй, третьей и еще более высоких степеней).

Рассмотрим двойную линейную связь. Для конкретности изложения будем иметь в виду, что речь идет о связи элемента-примеси (его содержание обозначим через z) с двумя другими элементами (их содержание обозначим через x и y). Убедиться в наличии этой связи, а также измерить ее силу можно с помощью оценки сводного коэффициента корреляции R, определяемого по следующей формуле (Романовский, 1947):

$$R = + \sqrt{\frac{r_{xz}^2 - 2r_{xy}r_{xz}r_{yz} + r_{yz}^2}{1 - r_{xy}^2}},$$

где r_{xz} — оценка коэффициента корреляции для x и z;
r_{xy} — такой же выборочный коэффициент для x и y;
r_{yz} — такой же коэффициент для y и z.

Сводный коэффициент корреляции R всегда положителен. Его численное значение колеблется от 0 до 1. Если линейная связь z с x и y отсутствует, то математическое ожидание оценки R равно нулю ($MR = 0$) (нелинейная же связь может быть). При точной (функциональной) линейной связи z с x и y $MR = 1$. Во всех других случаях $0 < R < 1$.

Вычисление R можно показать на примере, взятом из практики.

По двум участкам, названным аномалиями (№ 1 и Малотаскинская) Теченского железорудного месторождения в 1956 г. были проанализированы 342 керновые пробы на различные элементы, в том числе на кобальт, серу и железо.

Содержание кобальта, выраженное в условных единицах, обозначим через z, содержание серы в процентах — через x и содержание железа в процентах через y. Вычисление оценок среднего содержания, среднего квадратического отклонения и парных коэффициентов корреляции дало следующие результаты:

$$\bar{z} = 11{,}7; \qquad s_y = 11{,}2;$$

$$\bar{x} = 1{,}61; \qquad r_{xz} = 40{,}684;$$

$$\bar{y} = 31{,}4; \qquad r_{yz} = +0{,}479;$$

$$s_z = 9{,}6; \qquad r_{xy} = +0{,}420.$$

$$s_x = 1{,}40;$$

Отсюда

$$R = + \sqrt{\frac{0{,}684^2 - 2 \cdot 0{,}420 \cdot 0{,}684 \cdot 0{,}479 + 0{,}479^2}{1 - 0{,}420^2}} = 0{,}716.$$

Близость R к единице говорит о наличии сильной линейной связи z с x и y.

Проверить гипотезу $H_0 : MR = 0$ можно с помощью следующего критерия (Романовский, 1947):

$$\tau = \frac{R^2 (n-5)}{2(1-R^2)} \sqrt{\frac{n-7}{n-3}}.$$

Случайная величина τ, если гипотеза $MR = 0$ верна, распределена асимптотически нормально с параметрами 0,1. Поэтому гипотезу H_0 можно уверенно отвергнуть, т. е. считать наличие зависимости доказанным, если вычисленное значение τ окажется больше 3. Для приведенного выше примера $R = 0{,}716$, $n = 342$, $\tau = 175$, что доказывает высокую надежность вывода о наличии сильной корреляционной зависимости содержания кобальта от содержания серы и железа.

Необходимо заметить, что есть и другие методы проверки гипотез о коэффициентах корреляции, а в равной степени и о коэффициентах регрессии, но приведенные формулы наиболее удобны *.

На z влияет как x, так и y. Силу влияния одного x или одного y на z при двойной связи нельзя точно измерить методами парной корреляции. Приведенные парные коэффициенты корреляции r_{xz}, r_{yz}, r_{xy} лишь приближенно измеряют связь. Более точно силу этой связи в случае линейной множественной корреляции можно измерить при помощи так называемых частных коэффициентов корреляции. Сущность частного коэффициента корреляции — исключение влияния зависимости между другими величинами или «очищение» связи (Ван дер Варден, 1960).

Частный коэффициент корреляции x и z при исключении влияния на них y оценивается по формуле

$$r_{xz\,(y)} = \frac{r_{xz} - r_{xy} \cdot r_{yz}}{\sqrt{\left(1 - r_{xy}^2\right)\left(1 - r_{yz}^2\right)}}.$$

Корень в знаменателе всегда берется со знаком плюс. Заключение в скобки индекса y при коэффициенте корреляции говорит о том, что этот коэффициент вычислен при условии, что влияние на связь между x и z величины y исключено.

Частный коэффициент корреляции в своих свойствах подобен обыкновенному коэффициенту корреляции: он не может быть меньше -1 и больше $+1$. Он также не зависит от размерности величин, и сам не имеет размерности. Этот коэффициент равен нулю, когда линейная связь между x и y отсутствует. Если же связь функциональная, то этот коэффициент будет равным $+1$ или -1.

Для примера с кобальтом, серой и железом имеем

$$r_{xz\,(y)} = \frac{0{,}684 - 0{,}420 \cdot 0{,}479}{\sqrt{(1 - 0{,}420^2)(1 - 0{,}479^2)}} = 0{,}606.$$

Подобным образом оценивается частный коэффициент корреляции между z и y:

$$r_{zy\,(x)} = \frac{r_{yz} - r_{xy} r_{xz}}{\sqrt{\left(1 - r_{xy}^2\right)\left(1 - r_{xz}^2\right)}}.$$

В нашем примере имеем

$$r_{zy\,(x)} = \frac{0{,}479 - 0{,}420 \cdot 0{,}684}{\sqrt{(1 - 0{,}420^2)(1 - 0{,}684^2)}} = 0{,}662.$$

---

* Автор рассматривает только вопрос о зависимости одной величины от двух других. Более подробно вопрос о множественной корреляции для случая более чем трех переменных рассмотрен в книге Т. Андерсона «Введение в многомерный статистический анализ», Ф. М. 1963. (Прим. ред.).

Итак, мы получили частный коэффициент корреляции кобальта с серой (при исключении влияния железа) равным 0,606 и частный коэффициент корреляции кобальта с железом (при исключении влияния серы) равным 0,662. Парные коэффициенты корреляции для этих пар элементов, как было показано выше, равны соответственно 0,684 и 0,479.

Чем объясняется расхождение между обычным парным и частным коэффициентом корреляции? Оно объясняется искажающим влиянием третьего признака. Так, на простой коэффициент корреляции кобальта с железом влияет содержание серы. С серой связана некоторая часть кобальта. Если в руде много серы, то и кобальта будет больше, а это отразится на связи его с железом. Если же в руде мало серы, то и кобальта будет меньше, а это также повлияет на корреляцию его с железом.

Возьмем вместо кобальта, серы и железа какие-либо другие элементы и обозначим их через a, b и c. Парная корреляция a с b может быть очень тесной, а частная корреляция a с b при исключении влияния c может быть даже нулевой или обратной по знаку, так как теснота и положительный характер связи a с b могут быть вызваны тем, что a связан с c, а c с b, тогда как a с b непосредственно не связан. Вот почему необходимо всесторонне анализировать связи признаков. Методы множественной корреляции дают более полные результаты, чем методы простой корреляции.

Оценив сводный и частные коэффициенты корреляции и убедившись в наличии множественной линейной связи, необходимо далее построить эмпирическое уравнение регрессии.

Последнее имеет вид $z - \bar{z} = A\,(x - \bar{x}) + B\,(y - \bar{y})$.

Величины A и B, входящие в эту формулу, являются оценками коэффициентов регрессии. Их можно вычислить по следующим формулам:

$$A = \frac{r_{xz} - r_{yz}r_{xy}}{1 - r_{xy}^2} \cdot \frac{s_z}{s_x};$$

$$B = \frac{r_{yz} - r_{xz}r_{xy}}{1 - r_{xy}^2} \cdot \frac{s_x}{s_y},$$

где r_{xy}, r_{xz} и r_{yz} — коэффициенты простой линейной корреляции; s_z, s_x и s_y — оценки стандартов величин z, x и y (соответственно).

Для приведенного выше примера имеем

$$A = \frac{0{,}684 - 0{,}479 \cdot 0{,}42}{1 - 0{,}420^2} \cdot \frac{9{,}6}{1{,}40} = 4{,}007;$$

$$B = \frac{0{,}479 - 0{,}684 \cdot 0{,}420}{1 - 0{,}420^2} \cdot \frac{9{,}6}{11{,}2} = 0{,}200.$$

Подставим полученные величины в уравнение регрессии $z - 11{,}7 = 4{,}007\,(x - 1{,}61) + 0{,}200\,(y - 31{,}4)$, откуда получаем (с округлением) окончательное уравнение

$$z = 4{,}007x + 0{,}2y - 1{,}3.$$

Это уравнение выражает плоскость, делящую совокупность точек с переменными координатами x, y, z, причем сумма квадратов отклонений этих точек от плоскости — минимальная из всех возможных.

Глядя на полученное уравнение связи кобальта с серой и железом, можно подумать, что влияние железа (y) ничтожно мало по сравнению с влиянием серы (x), так как коэффициент при y в 20 раз меньше коэффициента при x, но это мнение будет ошибочным, так как среднее содержание железа (31,4%) почти в 20 раз выше среднего содержания серы (1,61%). В результате оказывается, что влияние железа почти такое же, как и серы.

Свободный член уравнения (—1,3) указывает на неполноту учета определяющих факторов. На содержание кобальта кроме серы и железа

влияют некоторые другие факторы, причем влияют отрицательно, но сила этого влияния невелика. Так, если $x = 1{,}6$, а $y = 31{,}0$, то $z = 6{,}41 + 6{,}20 - 1{,}3 = 11{,}31$. От этого итогового содержания кобальта величина 1,3 составляет всего лишь 11%. При очень низком содержании серы и железа (первой около 0,2% второго — 3%) содержание кобальта будет в среднем почти нулевым (если выведенное уравнение связи действительно для таких содержаний серы и железа).

Для приведенного примера были построены уравнения регрессии также и путем простой линейной корреляции. Они таковы:

$$z = 4{,}609x + 4{,}1;$$

$$z = 0{,}376y + 0{,}1.$$

Вычисленное разными способами и для разных случаев содержание кобальта приводится в табл. 105.

Таблица 105

Исходное содержание, %		Вычисленное содержание кобальта, усл. ед.		
серы (x)	железа (y)	Множественная корреляция	Простая корреляция	
			с серой	с железом
1	10	3,3	8,7	3,9
2	10	10,1	13,3	3,9
1	20	5,3	8,7	7,6
2	20	12,1	13,3	7,6
3	30	21,0	17,9	11,4
7	30	48,4	36,4	11,4
1,61	31,4	11,7	11,5	11,9
4	40	29,8	20,5	15,1
9	40	64,1	45,6	15,1
2	50	18,1	13,3	18,9
5	50	38,7	27,1	18,9
3	60	27,0	17,9	22,7

Сравнение содержания кобальта, вычисленного по уравнениям простой и множественной регрессии, показывает значительные различия. Более точные результаты дает множественная регрессия.

Точность корреляционного метода вычисления содержания того или иного элемента можно определить следующим образом.

По одному пласту калийной соли было взято и проанализировано (на $MgCl_2$, NaCl, KCl, Br и другие компоненты) 733 пробы. По этим пробам составлена корреляционная таблица (табл. 106).

Таблица 106

NaCl, %	$MgCl_2$, %									Всего проб n	Среднее содержание $MgCl_2$, %
	6—9	9—12	12—15	15—18	18—21	21—24	24—27	27—30	30—33		
8—16								39	2	41	28,6
16—24							106	83		189	26,8
24—32						112	59			171	23,5
32—40					98	85	1	1		185	20,9
40—48				47	58					105	18,1
48—56			16	17						33	15,0
56—64		2	4							6	12,5
64—72	1	2								3	9,5
Итого	1	4	20	64	156	197	166	123	2	733	18,8

Затем по каждой клетке этой таблицы была составлена корреляционная таблица содержания брома (в сотых долях процента) и KCl (в про-

центах), на основании чего было вычислено среднее фактическое содержание брома (табл. 107).

Таблица 107

NaCl	$MgCl_2$								
	6—9	9—12	12—15	15—18	18—21	21—24	24—27	27—30	30—33
8—16								13,08	14,95
16—24							13,11	13,85	
24—32						11,68	12,47		
32—40					10,63	11,07	8,50	13,31	
40—48				9,38	10,03				
48—56			9,94	8,65					
56—64		7,50	9,00						
64—72	5,00	6,75							

Эта таблица — не корреляционная. В ее клетках проставлено не число проб, а оценки среднего содержания брома для тех проб, которые проставлены в тех же клетках предшествующей таблицы.

В клетке с 27—30% $MgCl_2$ и 8—16% NaCl стоит 39 проб (см. табл. 106). По каждой из этих 39 проб химики определили содержание брома, колеблющееся в каких-то пределах, а в среднем оно составляет 13,08%, что и показано в табл. 107. Таким же путем заполнены и другие клетки этой таблицы.

По методу множественной корреляции затем было выведено уравнение связи содержания брома с содержанием $MgCl_2$ и NaCl. Вот это уравнение: $Br = 0{,}2064\ MgCl_2 - 0{,}06866\ NaCl + 9{,}06$.

Здесь содержание Br выражено в сотых долях процента, а $MgCl_2$ и NaCl — в процентах.

По каждой из клеток последней таблицы затем было вычислено теоретическое содержание Br и сопоставлено со средним фактическим, показанным в этой таблице. Расхождение между средним фактическим и теоретическим содержанием, выраженное в относительных процентах (к среднему фактическому), показано в табл. 108.

Таблица 108

NaCl	$MgCl_2$								
	6—9	9—12	12—15	15—18	18—21	21—24	24—27	27—30	30—33
8—16								+8,0	—1,4
16—24							—1,2	—2,0	
24—32						+0,8	—0,6		
32—40					—0,2	+1,4	+39,4	—6,3	
40—48				+0,6	+0,3				
48—56			—16,7	+2,8					
56—64		—5,2	—14,1						
64—72	+18,8	—2,8							

В клетках этой таблицы проставлена относительная ошибка (в процентах к среднему фактическому), вычисленная по формуле содержания брома в пробах, имеющих определенное содержание $MgCl_2$ и NaCl. Так, например, в 39 пробах с содержанием $MgCl_2$ от 27 до 30% и NaCl от 8 до 16% теоретически вычисленное (по приведенной выше формуле) содержание брома на 8% выше, чем среднее (для всех 39 проб) фактическое содержание брома.

Распределение этих отклонений покажем в табл. 109.

Таблица 109

Отклонение	Число проб
—16,7	16
—14,1	4
—6,3	1
—5,2	2
—2,8	2
—2,0	83
—1,4	2
—1,2	116
—0,6	59
—0,2	58
0,3	58
0,6	47
0,8	112
1,4	85
2,8	17
8,0	39
18,8	1
39,4	1
Всего	733

Эта таблица показывает, что ошибка вычисления содержания брома по уравнению связи только в одной из 733 проб превышает допустимый для химических анализов предел, равный ±30%.

Приведенный здесь метод расчета ошибки корреляционного определения содержания элементов — приближенный. Он основан на средних показателях по каждой клетке.

Значительно точнее вычисление по группам проб, лежащих в узких интервалах содержания брома. Покажем и этот способ на том же примере.

Вычисление ошибки корреляционной формулы по узким интервалам брома (6—8, 8—10, 10—12% и т. д.) производится следующим образом.

Всю совокупность из 733 проб разобьем на группы по интервалам содержания компонентов. По каждой из этих групп покажем распределение числа проб (n) по среднему фактическому содержанию брома ($Br_ф$), среднее теоретическое (вычисленное по уравнению регрессии, выведенному выше) содержание брома по всей данной группе ($Br_т$) и ошибку, т. е. разницу в содержании — абсолютную (Δ) и относительную (d) — в процентах (табл. 110).

Таблица 110

$MgCl_2$	NaCl	$Br_т$	$Br_ф$	n	Δ	d
1	2	3	4	5	6	7
27—30	8—16	14,1	10	1	—4,1	—41,0
			11	12	—3,1	—28,2
			12	2	—2,1	—17,5
			13	11	—1,1	—8,5
			15	9	0,9	6,0
			17	4	2,9	17,1
30—33	8—16	14,7	12,9	1	—1,8	—14,0
			17	1	2,3	13,5

$MgCl_2$	NaCl	$Br_т$	$Br_ф$	n	Δ	d
1	2	3	4	5	6	7
24—27	16—24	13,0	9	3	—4,0	—44,4
			10	1	—3,0	—30,0
			11	23	—2,0	—18,2
			13	46	0	0
			14	1	1,0	7,1
			15	30	2,0	13,3
			17	2	4,0	23,5
			19	1	6,0	31,6
27—30	16—24	13,0	9	3	—4,0	—44,4
			10	1	—3,0	—30,0
			11	23	—2,0	—18,2
			13	46	0	0
			14	1	1,0	7,1
			15	30	2,0	13,3
			17	2	4,0	23,5
			19	1	6,0	31,6
27—30	16—24	13,6	11	13	—2,6	—29,6
			13	29	—0,6	—4,6
			15	32	1,4	9,3
			17	8	3,4	20,0
21—24	24—32	11,8	9	10	—2,8	—31,1
			10	2	—1,8	—18,0
			11	48	—0,8	—7,3
			12	2	0,2	1,7
			13	46	1,2	9,2
			15	4	3,2	21,3
24—27	24—32	12,4	7	1	—5,4	—77,1
			9	3	—3,4	—37,8
			10	1	—2,4	—24,0
			11	15	—1,4	—12,7
			13	29	0,6	4,6
			15	10	2,6	17,3
18—21	32—40	10,6	5,5	1	—5,1	—92,7
			7	3	—3,6	—51,4
			8	1	—2,6	—32,5
			9	23	—1,6	—17,8
			10,18	1	—0,42	—4,1
			10,9	1	0,3	2,8
			11	51	0,4	3,6
			13	16	2,4	18,5
			14	1	3,4	24,3
21—24	32—40	11,2	8	3	—3,2	—40,0
			9	14	—2,2	—24,4
			11	51	—0,2	—1,6
			13	16	1,8	13,8
			22	1	10,8	49,2
24—27	32—40	11,8	8,5	1	—3,3	—38,8
27—30	32—40	12,5	13,31	1	0,81	6,2

Продолжение табл. 110

$MgCl_2$	NaCl	$Br_т$	$Br_ф$	n	Δ	d
1	2	3	4	5	6	7
15—18	40—48	9,4	8	6	—1,4	—17,5
			9	27	—0,4	—4,5
			10	1	0,6	6,0
			11	11	1,6	14,5
			12,5	1	3,1	24,8
			13	1	3,6	27,7
18—21	20—48	10,1	7	6	—3,1	—44,3
			9	18	—26,1	—23,3
			11	32	0,9	8,2
			12,6	1	2,5	19,8
			13	1	2,9	22,3
12—15	48—56	8,3	7,4	2	—0,9	—12,2
			8	2	—0,3	—3,7
			9	10	0,7	7,8
			10,1	1	1,8	17,8
			12	1	3,7	30,2
15—18	48—56	8,9	7	2	—1,9	—27,2
			8	2	—0,9	—11,2
			9	13	0,1	1,1
9—12	56—64	7,1	7	1	—0,1	—1,4
			8	1	0,9	11,2
12—15	56—64	7,7	8	2	0,3	3,7
			9	1	1,3	14,5
			11	1	3,3	20,9
6—9	64—72	5,9	5	1	—0,9	—18,0
9—12	64—72	6,6	6	1	—0,6	—10,0
			7,5	1	0,9	12,0

По данным этой таблицы можно составить следующее распределение относительных ошибок (отклонений фактического содержания от теоретического) (табл. 111).

Таблица 111

Знак ошибки	Интервал отн. ошибки, %	Число проб
—	50—40	1
—	40—30	2
—	30—20	11
—	20—10	101
—	10—0	230
	0	46
+	0—10	171
+	10—20	77
+	20—30	61
+	30—40	18
+	40—50	10
+	50—60	3
+	60—70	—
+	70—80	1
+	80—90	—
+	90—100	1

Эта таблица показывает, что распределение близко к нормальному (рис. 52)*. Некоторая асимметрия вызвана влиянием неучтенного здесь хлористого калия.

В нормах ошибок химического анализа проб на бром знак ошибки не учитывается, поэтому дадим группировку полученных относительных ошибок по их абсолютной величине независимо от знака (табл. 112).

Эта таблица показывает, что в 95,08% всех случаев наибольшая ошибка по отдельно взятой пробе не превзойдет допустимой ошибки анализа (±30%). Если же брать сразу две, три или большее число соседних проб и вычислить теоретическое содержание брома по среднему для этих проб значению хлористого магния и хлористого натрия, то ошибка будет значительно меньше. С увеличением числа проб, т. е. объема выборки, она будет стремиться к нулю. Можно ожидать, что если теоретическое содержание брома определять по средним показателям $MgCl_2$ и NaCl из 3—5 проб, то

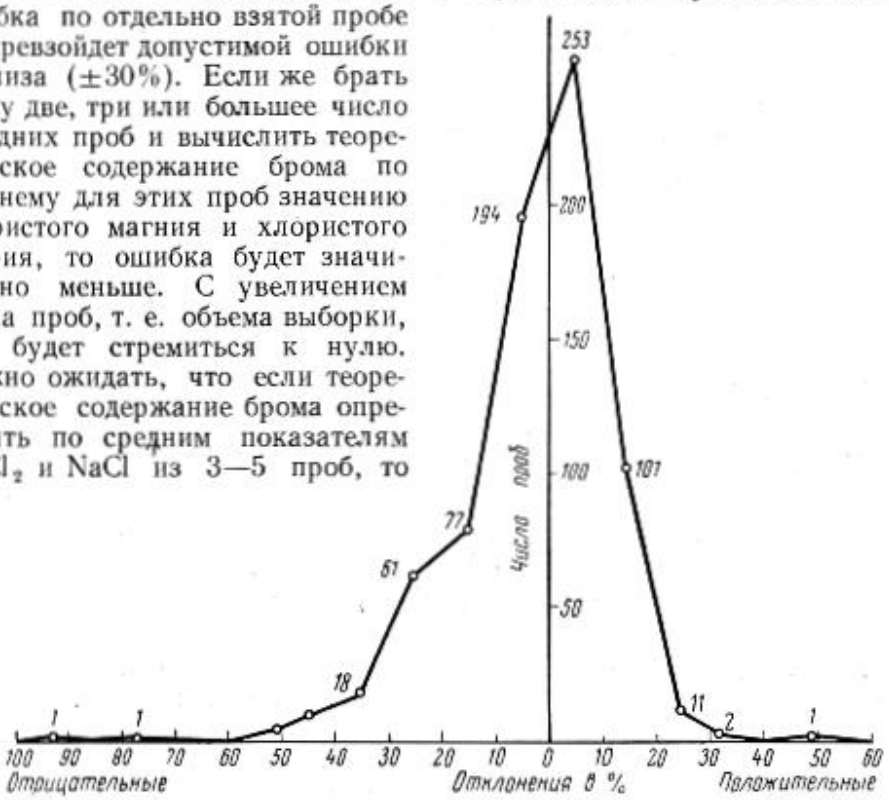


Рис. 52. Распределение отклонений теоретического содержания брома от фактического его содержания по пласту В в Соликамске

вероятность ошибки, большей, чем ошибка анализа, практически будет очень мала. Таким образом, корреляционный способ определения содержания брома оказывается практически вполне приемлемым.

Таблица 112

Интервал отн. ошибки, %	Число проб	Доля (частость), %	Нарастающая доля, %	Убывающая доля, %
0—10	466	60,84	60,84	100,00
10—20	179	24,42	85,26	39,16
20—30	72	9,82	95,08	14,74
30—40	20	2,73	97,81	4,92
40—50	11	1,50	99,31	2,19
50—60	3	0,41	99,72	0,69
60—70	—	—	99,72	0,28
70—80	1	0,14	99,86	0,28
80—90	—	—	99,86	0,14
90—100	1	0,14	100,00	0,00
Всего . . .	733	100,00		

* На рисунке знак каждой ошибки заменен на обратный, т. е. взято отклонение теоретического содержания от фактического.

XII. НЕЛИНЕЙНАЯ КОРРЕЛЯЦИЯ

Если кривые регрессии значительно отличаются от прямых линий, то в качестве меры зависимости между двумя признаками нередко используется так называемое корреляционное отношение.

Как было показано выше, дисперсию признака можно разложить на составные части. Общая дисперсия представляет собой сумму межгрупповой и внутригрупповой дисперсии. Межгрупповая дисперсия характеризует изменчивость, вызываемую тем фактором, который представлен признаком группировки, а внутригрупповая дисперсия является следствием изменчивости, вызываемой другими факторами.

Межгрупповая дисперсия групповых средних значений $\bar{y}_x$ вокруг общей средней $\bar{y}$ оценивается по формуле

$$s^2_{\bar{y}_x} = \frac{1}{n}\sum n_x(\bar{y}_x - \bar{y})^2,$$

где $s^2_{\bar{y}_x}$ — межгрупповая дисперсия групповых средних $\bar{y}_x$ вокруг общей средней $\bar{y}$;

n — объем выборки;

n_x — объем групп в выборке, выделяемых по величине x;

$\bar{y}_x$ — групповая средняя;

$\bar{y}$ — общая средняя.

Если потребуется найти межгрупповую дисперсию средних для x вокруг общей средней $\bar{x}$, то эта формула преобразуется так:

$$s^2_{\bar{x}_y} = \frac{1}{n}\sum n_y(\bar{x}_y - \bar{x})^2.$$

Здесь всюду вместо x поставлен y и вместо $\bar{y}$ поставлен $\bar{x}$.

Приведенные формулы неудобны для вычислений. Поэтому их следует преобразовать так:

$$\bar{s}^2_{\bar{y}_x} = \frac{\sum n_x \bar{y}^2_x}{n} - \bar{y}^2$$

и

$$\bar{s}^2_{\bar{x}_y} = \frac{\sum n_y \bar{x}^2_y}{n} - \bar{x}^2.$$

Внутригрупповая дисперсия оценивается по формулам

$$\bar{s}^2_y = \frac{\sum s_i^2 m_i}{\sum m_i}$$

и

$$\bar{s}^2_x = \frac{\sum s_h^2 b_h}{\sum b_h},$$

где $\bar{s}^2_i$ и $\bar{s}^2_h$ — внутригрупповые выборочные дисперсии (вокруг групповых средних);

m_i и b_h — численности групп.

По правилу сложения дисперсий имеем

$$s^2_y = s^2_{\bar{y}_x} + \bar{s}^2_y$$

и

$$s^2_x = s^2_{\bar{x}_y} + \bar{s}^2_x,$$

где s^2_y и s^2_x — общие дисперсии.

Так как все дисперсии в последних двух формулах — положительные числа, каждая из частных дисперсий, указанных в правой части равенств, меньше общей дисперсии, указанной в левой части тех же равенств.

Из этих дисперсий получаем выборочные стандарты (корни квадратные из дисперсий).

Доля межгруппового выборочного стандарта в общем выборочном стандарте называется оценкой корреляционного отношения y на x или x на y. Формула эмпирического корреляционного отношения y на x такова:

$$\eta_{y_x} = \frac{s_{\bar{y}_x}}{s_y},$$

где η_{y_x} — корреляционные отношения y на x;

$s_{\bar{y}_x}$ — выборочный стандарт групповых средних y вокруг общей средней y (группы выделяются по величине x).

s_y — общий стандарт y.

Для корреляционного отношения x на y соответственно имеем такую формулу:

$$\eta_{x_y} = \frac{s_{\bar{x}_y}}{s_x},$$

где η_{x_y} — корреляционное отношение x на y;

$s_{\bar{x}_y}$ — выборочный стандарт групповых средних x вокруг общей средней x (группы выделяются по величине y).

Величины η_{y_x} и η_{x_y} в общем случае, конечно, не равны друг другу.

Приведем примеры вычисления корреляционного отношения. На одном из месторождений Урала было проанализировано 35 проб медной руды на различные элементы, в том числе на мышьяк и теллур. Содержание мышьяка (x) теллура (y) в тысячных долях процента показано в табл. 113.

Таблица 113

№ п/п	x	y	№ п/п	x	y
1	2	3	1	2	3
1	33	5	19	6	0
2	30	7	20	5	0
3	70	9	21	55	5
4	50	10	22	55	4
5	17	6	23	70	4
6	35	8	24	25	4
7	35	8	25	20	3
8	50	7	26	75	6
9	18	4	27	55	4
10	40	5	28	50	5
11	40	6	29	45	4
12	12	0	30	17	14
13	7	0	31	30	6
14	10	0	32	12	20
15	12	0	33	15	2
16	15	0	34	25	3
17	45	6	35	15	0
18	22	0			

Для того чтобы выяснить вопрос о том, связаны ли друг с другом x и y, необходимо прежде всего систематизировать эти данные. Сгруппируем все анализы по интервалам содержания мышьяка, т. е. по интервалам величины x (табл. 114).

Таблица 114

№ интервалов	Интервал содержания мышьяка	Середина интервала x	Содержание теллура по отдельным пробам y	$\sum y$	Число проб n_x	Среднее содержание теллура $\bar{y}_x$
1	0—10	5	0, 0, 0, 0	0	4	0
2	10—20	15	6, 4, 0, 0, 0, 3, 2, 2, 14, 0	31	10	3,1
3	20—30	25	7, 0, 4, 6, 3	20	5	4,0
4	30—40	35	5, 8, 8, 5, 4	32	5	6,4
5	40—50	45	10, 7, 6, 5, 4	32	5	6,4
6	50—60	55	5, 4, 4	13	3	4,3
7	60—70	65	9, 4	13	2	6,5
8	70—80	75	6	6	1	6,0
Всего . .				147	35	4,2

Среднее содержание теллура по всем пробам $\bar{y} = 4{,}2$.

Среднее содержание теллура по классам содержания мышьяка можно нанести на график (рис. 53). Связь x с y криволинейная и видна на глаз, но графический метод обнаружения связи и определение ее характера часто бывает ненадежен. При большом числе проб и значительной дисперсии признака график может ввести в заблуждение, так как на нем не отмечается число проб (по каждой точке), а отмечается только средняя величина показателя. Если, например, для низких значений x на этом графике будет очень много анализов, а для высоких значений x очень мало (по 1—2 пробы), то изломы линии окажутся построенными на случайных данных, и вывод о криволинейности окажется ошибочным. В таком случае (т. е. в случае большого числа проб) следует пользоваться аналитическим методом, который заключается в оценке корреляционного отношения.

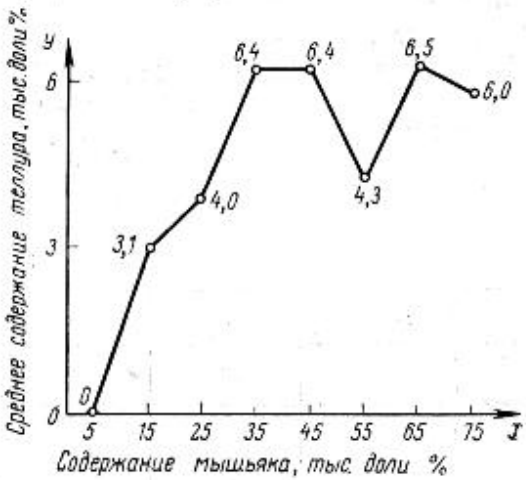


Рис. 53. Связь среднего содержания теллура с содержанием мышьяка в медной руде Уральского месторождения

Для вычисления эмпирического корреляционного отношения y на x составим таблицу (табл. 115).

Здесь i означает номер класса (группы) по мышьяку, x — середина интервала содержания мышьяка, n_x — число проб по классам, $\bar{y}_x$ — оценка среднего содержания теллура по классам.

Величину $s_{\bar{y}_x}$ вычислим по формуле

$$s_{\bar{y}_x} = \sqrt{s^2_{\bar{y}_x}},$$

т. е. как корень квадратный из дисперсии, а дисперсию определим так:

$$s^2_{\bar{y}_x} = \frac{761{,}7}{35} - 4{,}2^2 = 4{,}12,$$

откуда

$$s_{\bar{y}_x} = \sqrt{4{,}12} = 2{,}03.$$

Таблица 115

i	x	n_x	$\bar{y}_x$	$\bar{y}_x^2$	$n_x\bar{y}_x^2$
1	5	4	0	0	0
2	15	10	3,1	9,61	96,1
3	25	5	4,0	16,00	80,0
4	35	5	6,4	40,96	204,8
5	45	5	6,4	40,96	204,8
6	55	3	4,3	18,49	55,5
7	65	2	6,5	42,25	84,5
8	75	1	6,0	36,00	36,0
Всего		35	4,2		761,7

Общий выборочный стандарт s_y вычислим с помощью следующей таблицы (табл. 116).

Таблица 116

i	x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$y-\bar{y}$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
1	2	3	4	5	6	7	8
1	5	0	—26,9	723,61	—4,2	17,64	113,0
2	6	0	—25,9	670,81	—4,2	17,64	108,8
3	7	0	—24,9	620,01	—4,2	17,64	104,6
4	10	0	—21,9	479,61	—4,2	17,64	92,0
5	12	0	—19,9	396,01	—4,2	17,64	83,6
6	12	0	—19,9	396,01	—4,2	17,64	83,6
7	15	0	—16,9	285,61	—4,2	17,64	71,0
8	15	0	—16,9	285,61	—4,2	17,64	71,0
9	22	0	— 9,9	98,01	—4,2	17,64	41,6
10	12	2	—19,9	396,01	—2,2	4,84	43,8
11	15	2	—16,9	285,61	—2,2	4,84	37,2
12	20	3	—11,9	141,61	—1,2	1,44	13,3
13	25	3	— 6,9	47,61	—1,2	1,44	8,3
14	18	4	—13,9	193,21	—0,2	0,04	2,8
15	25	4	— 6,9	47,61	—0,2	0,04	1,4
16	45	4	13,1	171,61	—0,2	0,04	— 2,6
17	55	4	23,1	533,61	—0,2	0,04	— 4,6
18	55	4	23,1	533,61	—0,2	0,04	— 4,6
19	70	4	38,1	1 451,61	—0,2	0,04	— 7,6
20	33	5	1,1	1,21	0,8	0,64	0,9
21	40	5	8,1	65,61	0,8	0,64	6,5
22	50	5	18,1	327,61	0,8	0,64	14,5
23	55	5	23,1	533,61	0,8	0,64	18,5
24	17	6	—14,9	222,01	1,8	3,24	— 25,8
25	30	6	— 1,9	3,61	1,8	3,24	— 3,4
26	40	6	8,1	65,61	1,8	3,24	14,6
27	45	6	13,1	171,61	1,8	3,24	23,6
28	75	6	43,1	1 875,61	1,8	3,24	77,6
29	30	7	— 1,9	3,61	2,8	7,84	— 5,3
30	50	7	18,1	327,61	2,8	7,84	50,7
31	35	8	3,1	9,61	3,8	14,44	11,8
32	35	8	3,1	9,61	3,8	14,44	11,8
33	70	9	38,1	1 451,61	4,8	23,04	182,9
34	50	10	18,1	327,61	5,8	33,64	105,0
35	17	14	—14,9	222,01	9,8	96,04	—146,0
Всего . . .	1116	147		13357,55		387,60	1194,5

По итогам первых трех столбцов этой таблицы определены средние значения величин x и y, т. е.

$$\bar{x} = \frac{1116}{35} = 31{,}9;$$

$$\bar{y} = \frac{147}{35} = 4{,}2.$$

Далее можно оценить общие стандарты:

$$s_x = \sqrt{\frac{13357{,}55}{35}} = 19{,}54;$$

$$s_y = \sqrt{\frac{387{,}60}{35}} = 3{,}33.$$

Выше было определено, что для этого примера

$$s_{\bar{y}_x} = 2{,}03.$$

Корреляционное отношение y на x равно

$$\eta_{x_y} = \frac{2{,}03}{3{,}33} = 0{,}610.$$

Корреляционное отношение x на y вычислим таким же путем. Сначала сгруппируем все значения x по значениям y (табл. 117).

Здесь m — номер класса по порядку.

Таблица 117

m	Интервал содержания теллура	y	x	Σx	n_y	$\bar{x}_y$
1	0—2	1	5, 6, 7, 10, 12, 12, 15, 22, 12, 15	131	11	11,9
2	2—4	3	20, 25, 18, 25, 45, 55, 55, 70	313	8	39,1
3	4—6	5	33, 40, 50, 55, 17, 30, 40, 45, 75	385	9	42,8
4	6—8	7	30, 50, 35, 35	150	4	37,5
5	8—10	9	70, 50	120	2	60,0
6	10—12	11		—	—	—
7	12—14	13	17	17	1	17,0
Всего				1116	35	31,9

Таблица 118

m	y	n_y	$\bar{x}_y$	$\bar{x}_y^2$	$n_y \bar{x}_y^2$
1	1	11	11,9	141,61	1 557,71
2	3	8	39,1	1528,81	12 230,48
3	5	9	42,8	1841,84	16 576,56
4	7	4	37,5	1406,25	5 625,00
5	9	2	60,0	3600,00	7 200,00
6	11	—	—	—	—
7	13	1	17,0	289,00	289,00
Всего		35	31,9		43 478,75

Среднее содержание мышьяка $(\bar{x}_i)$ по классам содержания теллура можно отнести на график (рис. 54). Криволинейность можно заметить простым обозрением графика.

Для вычисления корреляционного отношения x на y составим таблицу (табл. 118).

Здесь m — номер класса по теллуру, y — середина интервала содержания теллура, n_y — число проб в классе, $\bar{x}_y$ — среднее содержание мышьяка в классе.

Величина $s^2_{\bar{x}_y}$ определяется так:

$$s^2_{\bar{x}_y} = \frac{43478{,}75}{35} - 31{,}9^2 = 224{,}64,$$

откуда

$$s_{\bar{x}_y} = \sqrt{224{,}64} = 14{,}93.$$

Общий стандарт s_x был выше определен равным 19,54.

Корреляционное отношение x на y равно

$$\eta_{x_y} = \frac{14{,}93}{19{,}54} = 0{,}765.$$

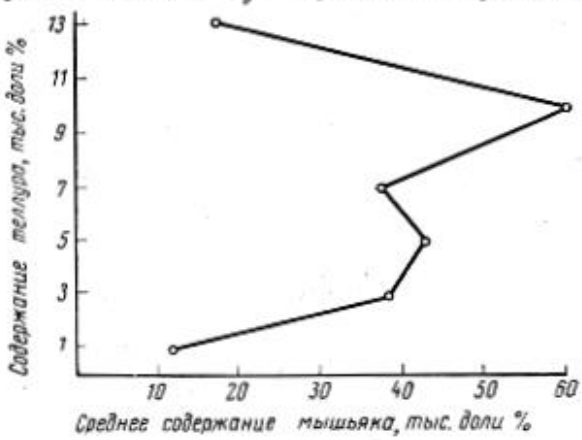


Рис. 54. Связь среднего содержания мышьяка с содержанием теллура в медной руде Уральского месторождения

Коэффициент корреляции для этого же примера

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{n s_x s_y} = \frac{1194{,}5}{35 \cdot 19{,}45 \cdot 3{,}33} = 0{,}525.$$

Оба корреляционных отношения, т. е. η_{y_x} и η_{x_y} — больше коэффициента корреляции r. Это указывает на некоторую криволинейность связи x с y и y с x. При этом $\eta_x > \eta_y$, что указывает на бо́льшую реальность кривизны связи x с y, чем y с x.

Таков метод вычисления эмпирического корреляционного отношения для малого объема выборки.

Для большого объема статистического коллектива метод расчета корреляционного отношения иной. Вычисление корреляционного отношения можно делать в той расчетной таблице, в какой вычисляется коэффициент корреляции. Для этого добавляется два столбца справа и две строки снизу. Добавочные столбцы используются для вычисления корреляционного отношения x на y, а добавочные строки — для вычисления корреляционного отношения y на x.

В столбце $\left(\sum n_{lk} k\right)^2$ записываются квадраты чисел столбца $\sum n_{lk} k$. Подобным образом заполняется и строка $\left(\sum n_{kl} l\right)^2$. Столбец $\frac{\left(\sum n_{lk} k\right)^2}{n_l}$ — частное от деления чисел предыдущего столбца на число проб. Так же заполняется и строка $\frac{\left(\sum n_{kl} l\right)^2}{n_k}$.

Корреляционное отношение определяется из равенства

$$\eta^2_{x_y} = \frac{1}{\hat{\mu}_{2x} - \hat{\mu}_{1x}} \left\{ \frac{1}{n} \sum \frac{\left(\sum n_{kl} k^2\right)}{n_l} - \hat{\mu}^2_{1x} \right\},$$

где все обозначения старые.

Отсюда находим

$$\eta_{x_y} = \sqrt{\eta^2_{x_y}}.$$

Подобным же образом находим η_{y_x}.

Таблица 119

y	x																	
	0—4	4—8	8—12	12—16	16—20	20—24	24—28	28—32	32—36	36—40	40—44	44—48	48—52	56—60	72—76	76—80	96—100	132—136
0—3	6	38	6	8	3	1						1						
3—6	7	39	15	10	7	1												
6—9	6	28	13	7	3	2	1		2					1				
9—12	12	23	6	7	8	2	5	3	4		2				1		1	
12—15	13	15	8	4	7	2	1	2	2	2	2	1		1				
15—18	13	15	1	4	1	2	2	2		1			1			1		
18—21	9	5	2	1			2											1
21—24	14	2	2			1	1		1									
24—27	14	1									1							
27—30		1					1			1								
30—33					1													
33—36										1								
k	—4	—3	—2	—1	0	1	2	3	4	5	6	7	8	10	14	15	20	29
n_k	94	167	53	41	30	11	13	7	9	5	5	2	1	2	1	1	1	1
$n_k k$	—376	—501	—106	—41	0	11	36	21	36	25	30	14	8	20	14	15	20	29
$n_k k^2$	1492	1503	212	41	0	11	72	63	144	125	180	98	64	200	196	225	400	841
$n_{xy}l$	—24 —21 —12 —12 13 18 42 56	—152 —177 —56 —23 15 10 6 4 5	—24 —45 —26 —6 1 4 6	—32 —30 —14 —7 4 2	—12 —21 —6 —8 1 6	—4 —3 —4 —2 2 3	—2 —5 2 4 3 5	—3 2	—4 —4 3	1 5 7	—2 4	—4	1	—2	—1	1	—1	2
$\sum n_{xy}l$	60	—308	—90	—77	—32	—8	7	—1	—5	13	2	—4	1	—2	—1	1	—1	—2
$k\sum n_{xy}l$	—240	924	180	77	0	8	14	—3	—20	65	12	—28	8	—20	—14	15	—20	58
$(\sum n_{xy}l)^2$	3600	94864	8100	5929	1024	64	49	1	25	169	4	16	1	4	1	1	1	4
$\frac{(\sum n_{xy}l)^2}{n_k}$	38,29	568,04	152,83	144,61	34,13	5,82	3,77	0,14	2,77	33,80	0,80	8	1	2	1	1	1	4

Продолжение табл. 119

y	l	n_l	$n_l l$	$n_l l^2$	$n_{xy}k$	$\sum n_{xy}k$	$l\left(\sum n_{xy}k\right)$	$\left(\sum n_{xy}k\right)^2$	$\frac{\left(\sum n_{xy}k\right)^2}{n_l}$
0—3	—4	63	—252	1008	—24, —114, —12, —8, 1,7	—150	600	22500	357,14
3—6	—3	79	—237	711	—28, —117, —30, —10,1	—184	552	33856	428,56
6—9	—2	63	—126	252	—24, —84, —26, —7, 2, 2, 8, 10	—119	238	14161	224,78
9—12	—1	74	—74	74	—48, 69, —12, —7, 2, 10, 9, 16, 12, 14, 20	—53	53	2809	37,96
12—15	0	60	0	0	—52, —45, —16, —4, 2, 3, 6, 8, 10, 12, 7, 10	—60	0	3600	60
15—18	1	43	43	43	—52, —45, —2, —4, 2, 4, 6, 5, 8, 15	—63	—63	3969	92,30
18—21	2	20	40	80	—36, —15, —4, —1, 4, 29	—23	—46	529	26,45
21—24	3	21	63	189	—56, —6, —4, 1, 2, 4	—59	—177	3481	182,91
24—27	4	16	64	256	—56 —3, 6	—53	—212	2809	175,56
27—30	5	3	15	75	—3, 2, 5	4	20	16	5,33
30—33	6	1	6	36	0	0	0	0	0
33—36	7	1	7	49	5	5	35	25	25
k		444	—451	2773			1000		1615,99

n_k	444
$n_k k$	—775
$n_k k^2$	5867
$n_{xy}l$	
$\sum n_{xy}l$	2
$k\sum n_{xy}l$	1000
$\left(\sum n_{xy}l\right)^2$	
$\frac{\left(\sum n_{xy}l\right)^2}{n_k}$	1003

Проверка гипотез $M\eta_{x_y} = 0$ или $M\eta_{y_x} = 0$ осуществляется с помощью следующего критерия:

$$\Theta_x = \frac{\left(\eta^2_{x_y} - r^2\right)(n_x - m_x - 2)}{\left(1 - \eta^2_{x_y}\right)(m_x - 2)} \sqrt{\frac{(m_x - 2)(n_x - m_x - 4)}{2(n_x - 4)}},$$

где m_x — число столбцов в таблице (считая и те, которые не содержат проб).

Для проверки гипотезы $M\eta_{y_x}$ вычисляется аналогичная величина Θ_y.

Если проверяемая гипотеза верна, то величина Θ_x или Θ_y распределена нормально с параметрами 0,1. Таким образом, гипотеза $M\eta_{y_x} = 0$ или $M\eta_{x_y} = 0$ должна быть отвергнута, если Θ_y или Θ_x превысят 3, и, следовательно, наличие зависимости можно считать установленным.

Приведем пример вычисления оценки корреляционного отношения и проверки гипотезы о равенстве его нулю.

На Теченском железорудном месторождении скарнового типа (Урал) исследовалась связь содержания фосфора с содержанием железа в руде. С двух участков этого месторождения (Малотаскинский и Первоаномальный) были взяты и проанализированы (в 1956 г. в лаборатории Уральского геологического управления) пробы. По результатам этих анализов составлена корреляционная таблица (табл. 119).

В ней учтено 444 пробы. Содержание фосфора в сотых долях процента обозначено через x, а содержание железа в процентах — через y.

По итогам таблицы можно сделать следующие вычисления:

$$x_0 = 18;$$

$$d_x = 4;$$

$$\bar{x} = \frac{-752}{444} \cdot 4 + 18 = 11{,}16;$$

$$s_u^2 = \frac{5824}{444} - \frac{-752^2}{444} = 10{,}32;$$

$$s_u = 3{,}21;$$

$$s_x = 3{,}19 \cdot 4 = 12{,}84;$$

$$y_0 = 33;$$

$$d_y = 6;$$

$$\bar{y} = \frac{-420}{444} \cdot 6 + 33 = 27{,}27;$$

$$s_v^2 = \frac{2402}{444} - \left(\frac{-420}{444}\right)^2 = 4{,}57;$$

$$s_v = 2{,}13;$$

$$s_y = 2{,}13 \cdot 6 = 12{,}78;$$

$$r = \frac{\frac{736}{444} - \left[\left(\frac{-752}{444}\right)\left(\frac{-420}{444}\right)\right]}{3{,}21 \cdot 2{,}13} = 0{,}0058;$$

$$\eta^2_{x_y} = \frac{1}{3{,}21}\left[\frac{1718{,}90}{444} - \left(\frac{-752}{444}\right)^2\right] = 0{,}309;$$

$$\eta_{x_y} = 0{,}556;$$

$$\eta^2_{y_x} = \frac{1}{2{,}13}\left[\frac{658{,}56}{444} - \left(\frac{-420}{444}\right)^2\right] = 0{,}486;$$

$$\eta_{y_x} = 0{,}697;$$

$$\Theta_x = 20{,}2.$$

Эти вычисления показывают, что линейная связь фосфора с железом отсутствует ($r = 0{,}0058$), а криволинейная связь имеется, так как $\eta_{x_y} = 0{,}556$ и $\Theta > 3$.

Корреляционное отношение обладает следующими основными свойствами:

1) оно всегда положительно, но не превышает единицы;

2) оно не может быть меньше абсолютного значения коэффициента корреляции;

3) если корреляционное отношение равно абсолютному значению коэффициента корреляции, то связь точно линейная;

4) если корреляционное отношение равно нулю, то никакой связи нет;

5) если корреляционное отношение равно единице, то связь функциональная;

6) чем ближе корреляционное отношение к единице, тем теснее связь;

7) чем ближе корреляционное отношение к нулю, тем слабее связь.

В том случае, когда связь оказывается криволинейной, необходимо вывести параболическое уравнение связи

$$\bar{y}_x = a + bx + cx^2,$$

где a, b, c — постоянные коэффициенты;

$\bar{y}_x$ — частные средние значения y, соответствующие различным заданным значениям x.

Если исходные данные систематизированы в виде корреляционной таблице, то коэффициенты a, b и c можно вычислить, решая систему уравнений:

$$\left.\begin{aligned} an + b\sum n_x x + c\sum n_x x^2 &= \sum n_x \bar{y}_x \\ a\sum n_x x + b\sum n_x x^2 + c\sum n_x x^3 &= \sum n_x x \bar{y}_x \\ a\sum n_x x^2 + b\sum n_x x^3 + c\sum n_x x^4 &= \sum n_x x^2 \bar{y}_x \end{aligned}\right\}.$$

Технику вычислений можно показать на примере с теллуром. Исходные данные взяты из табл. 114, где x означает содержание мышьяка, $\bar{y}_x$ — частное среднее содержание теллура (по классам содержания мышьяка). Все вычисления даются в табл. 120.

Таблица 120

x	l	n	nl	nl^2	nl^3	nl^4	$\bar{y}$	$n\bar{y}$	$n\bar{y}l$	$n\bar{y}l^2$
5	—3	4	—12	36	—108	324	0	0	0	0
15	—2	10	—20	40	—80	160	3,1	31	—62	124
25	—1	5	—5	5	—5	5	4,0	20	—20	20
35	0	5	0	0	0	0	6,4	32	0	0
45	1	5	5	5	5	5	6,4	32	32	32
55	2	3	6	12	24	48	4,3	13	26	52
65	3	2	6	18	54	162	6,5	13	39	117
75	4	1	4	16	64	256	6,0	6	24	96
Всего. .		35	—16	132	—46	960		147	39	441

В этой таблице x означает содержание мышьяка в тысячных долях процента, l — номер класса величины x, n — число проб, $\bar{y}$ — частное среднее содержание теллура в тысячных долях процента.

Подставив в систему уравнений итоги последней таблицы, получим:

$$\left.\begin{aligned} 35a - 16b + 132c &= 147 \\ -16a + 132b - 46c &= 39 \\ 132a - 46b + 960c &= 441 \end{aligned}\right\}.$$

Для решения этой системы уравнений нужно сначала разделить коэффициенты первого уравнения на 35, коэффициенты второго — на 16 и коэффициенты третьего — на 132. В результате этого деления получим систему уравнений:

$$\left.\begin{aligned} a - 0{,}46b + 3{,}77c &= 4{,}20 \\ a - 8{,}25b + 2{,}87c &= 2{,}44 \\ a - 0{,}35b + 7{,}27c &= 3{,}34 \end{aligned}\right\}. \qquad \text{(I)}$$

Вычтем первое уравнение из второго, а второе из третьего, в результате чего получим следующие два уравнения:

$$-7{,}79b - 0{,}90c = -6{,}64;$$
$$7{,}90b + 4{,}40c = 5{,}78.$$

Разделим коэффициенты первого из этих двух уравнений на —7,79 и коэффициенты второго уравнения на 7,90, в результате чего получим уравнения

$$\left.\begin{aligned} b + 0{,}12c &= 0{,}85 \\ b - 0{,}56c &= 0{,}73 \end{aligned}\right\} \qquad \text{(II)}$$

Вычтем из второго уравнения первое, в результате чего получим уравнение

$$0{,}44c = -0{,}12,$$

откуда

$$c = \frac{-0{,}12}{0{,}44} = -0{,}27.$$

Подставив это значение c в первое из уравнений (II), получим

$$b = 0{,}85 - 0{,}12\ (-0{,}27) = 0{,}88.$$

Наконец, подставив значения b и c в первое из уравнений (I), получим

$$a - 0{,}46 \cdot 0{,}88 + 3{,}77\ (-0{,}27) = 4{,}20$$

или

$$a = 5{,}60.$$

Таким образом, мы нашли, что

$$a = 5{,}60,$$
$$b = 0{,}88,$$
$$c = -0{,}27.$$

Параболическое уравнение регрессии теперь принимает следующий вид:

$$\bar{y}_т = 5{,}6 + 0{,}88l - 0{,}27l^2.$$

Здесь $\bar{y}_т$ — теоретическое содержание теллура в тысячных долях процента.

Сопоставим теоретическое содержание теллура с результатами наблюдений (табл. 121)

В этой таблице и ниже символ $\bar{y}_ф$ — это то же, что и $\bar{y}_x$. Индекс ф означает — фактическое (рис. 55).

Уравнение регрессии можно преобразовать так, чтобы в него входило значение x, а не l. Для этого подставим в него значение l:

$$l = 0{,}1x - 3{,}5,$$

в результате чего получим

$$\bar{y}_т = -0{,}753 + 0{,}276x - 0{,}00268x^2,$$

где $\bar{y}_т$ и x — в тысячных долях процента.

Вычисление $\bar{y}_т$ по этому уравнению дает почти те же результаты, что и по уравнению с участием l (разница лежит в пределах точности вычисления и объясняется округлением чисел).

Эта формула справедлива, по-видимому, для любых значений x в пределах от 5 до 75.

Ранжирование величины x, т. е. замена ее величиной l, во много раз облегчает вычисление, не уменьшая его точности.

Таблица 121

x	l	$\bar{y}_ф$	$\bar{y}_т$
5	—3	0	0,5
15	—2	3,1	2,8
25	—1	4,0	4,4
35	0	6,4	5,6
45	1	6,4	6,2
55	2	4,3	6,3
65	3	6,5	5,8
75	4	6,0	4,8

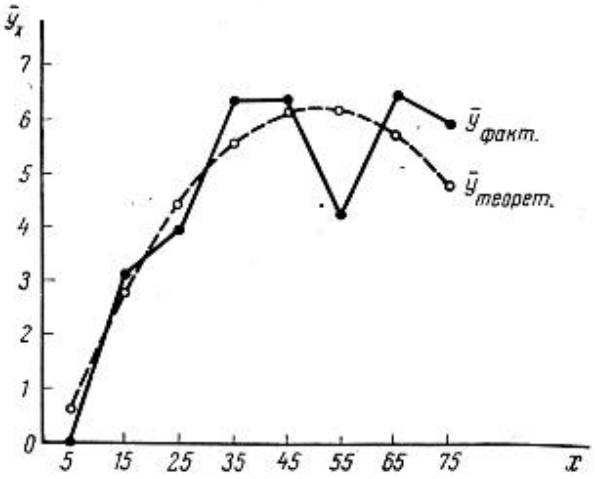


Рис. 55. Среднее содержание теллура «$\bar{y}_x$» по интервалам содержания мышьяка x (тыс. доли %)

Приведем еще пример нелинейной корреляции.

В калийных солях Верхнекамского месторождения есть примесь рубидия. Для использования этой примеси необходимо знать корреляционную связь рубидия с главными компонентами соли (Шарапов, 1957). По литературным данным и по пробам, отобранным Пермским политехническим институтом на одном пласте (данные любезно представлены Н. А. Косициной), было установлено, что рубидий сильнее всего связан с хлористым магнием.

Вывод уравнения регрессии рубидия и $MgCl_2$ таков:

1. Составлена корреляционная таблица по данным 170 проб (табл. 122). В ней фактическое содержание рубидия в пробах (в условных единицах) обозначено через y.

Таблица 122

$MgCl_2$ (x), %	l	n	nl	nl^2	nl^3	nl^4	$\bar{y}$, г/т	$n\bar{y}$	$n\bar{y}l$	$n\bar{y}l^2$
9—11	—5	4	—20	100	—500	2 500	35,0	140,0	—700,0	3 500,0
11—13	—4	7	—28	112	—448	1 792	34,0	238,0	—952,0	3 808,0
13—15	—3	13	—39	117	—351	1 053	87,7	1 140,1	—3 420,3	10 260,9
15—17	—2	20	—40	80	—160	320	72,0	1 440,0	—2 880,0	5 760,0
17—19	—1	16	—16	16	—16	16	67,5	1 080,0	—1 080,0	1 080,0
19—21	0	27	0	0	0	0	78,5	2 119,5	0	0
21—23	1	20	20	20	20	20	101 0	2 020,0	2 020,0	2 020,0
23—25	2	18	36	72	144	288	115,6	2 080,8	4 161,6	8 323,2
25—27	3	23	69	207	621	1 863	157,4	3 620,2	10 860,6	32 581,8
27—29	4	19	76	304	1216	4 864	158,9	3 019,1	12 076,4	48 305,6
29—31	5	3	15	75	375	1 875	193,3	579,9	2 899,5	14 497,5
Сумма		170	73	1103	901	14 591		17 477,6	22 985,8	130 096,0

2. По итогам этой таблицы составляем систему уравнений:

$$\left.\begin{aligned} 170a + 73b + 1103c &= 17477{,}6 \\ 73a + 1103b + 901c &= 22985{,}8 \\ 1103a + 901b + 14591c &= 130096{,}0 \end{aligned}\right\}. \qquad \text{(I)}$$

3. Разделим каждое уравнение на коэффициент при a, в результате чего получим:

$$\left.\begin{aligned} a + 0{,}4294b + 6{,}4882c &= 102{,}8094 \\ a + 15{,}1096b + 12{,}3425c &= 314{,}8740 \\ a + 0{,}8169b + 13{,}2285c &= 117{,}9474 \end{aligned}\right\}. \qquad \text{(II)}$$

4. Вычтем первое уравнение из второго, а второе из третьего:

$$\left.\begin{aligned} 14{,}6802b + 5{,}8543c &= 212{,}0646 \\ -14{,}2927b + 0{,}8860c &= -196{,}9266 \end{aligned}\right\}. \qquad \text{(III)}$$

5. Разделим каждое уравнение на коэффициент при b:

$$\left.\begin{aligned} b + 0{,}3988c &= 14{,}4737 \\ b - 0{,}0612c &= 13{,}7781 \end{aligned}\right\}. \qquad \text{(IV)}$$

6. Вычтем второе уравнение из первого:

$$0{,}4600c = 0{,}6956.$$

7. Определим c:

$$c = \frac{0{,}6956}{0{,}4600} = 1{,}5122.$$

8. Подставим значение c в первое уравнение (IV):

$$b + 0{,}3988 \cdot 1{,}5122 = 14{,}4737,$$

откуда

$$b = 14{,}4737 - 0{,}3988 \cdot 1{,}5122 = 13{,}8706.$$

9. Подставим b и c в первое уравнение (II):

$$a + 0{,}4294 \cdot 13{,}8706 + 6{,}4882 \cdot 1{,}5122 = 102{,}8094,$$

откуда

$$a = 102{,}8094 - 5{,}9560 - 9{,}8114 = 87{,}0420.$$

10. Таким образом, имеем:

$$a = 87{,}0420,$$

$$b = 13{,}8706,$$

$$c = 1{,}5122.$$

11. Уравнение регрессии теперь принимает вид:

$$\overline{y}_t = 87{,}0420 + 13{,}8706l + 1{,}5122l^2.$$

12. Но

$$l = \frac{x - 20}{2} = 0{,}5x - 10,$$

$$l^2 = (0{,}5x - 10)^2 = 0{,}25x^2 - 10x + 100.$$

13. Поэтому

$$\bar{y}_т = 87{,}0420 + 13{,}8706\,(0{,}5x - 10) + 1{,}5122\,(0{,}25x^2 - 10x + 100) =$$
$$= 87{,}0420 + 6{,}9353x - 138{,}706 + 0{,}37805x^2 - 15{,}122x + 151{,}22 =$$
$$= 99{,}556 - 8{,}1867x + 0{,}37805x^2.$$

14. Определяем область применения: $10 < x < 30$.

15. Вычислим теоретическое содержание рубидия $\bar{y}_т$ для фиксированных значений $MgCl_2$, т. е. для x.

x	10	12	14	16	18	20	22	24	26	28	30
$\bar{y}_т$	55,5	55,8	59,0	65,4	74,7	81,0	102,4	120,8	142,3	166,7	194,2

16. Нанесем на график фактическое $(\bar{y}_ф)$ и теоретическое $(\bar{y}_т)$ значения содержания рубидия по интервалам величины x (рис. 56).

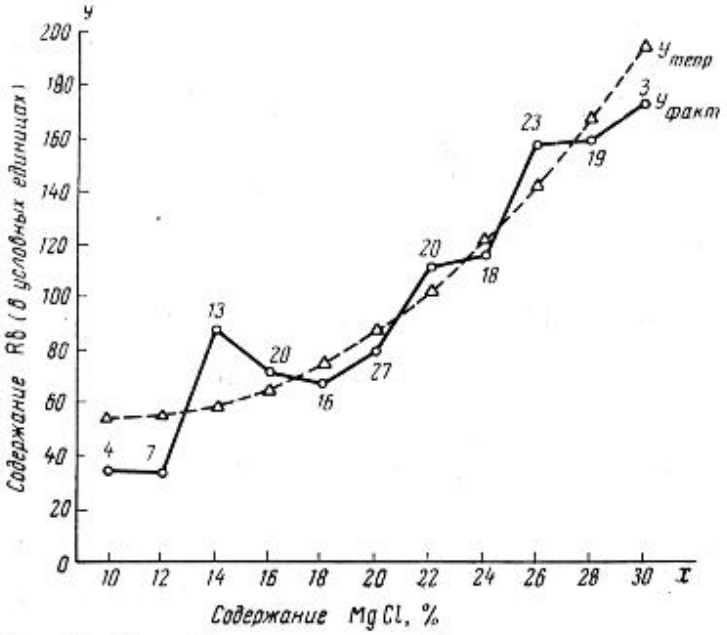


Рис. 56. Нелинейная корреляция рубидия и хлористого магния (числа на ломаной линии показывают количество проб)

17. Глядя на график (рис. 56), убеждаемся, в том, что вычисления сделаны без существенной ошибки и окончательно принимаем:

$$\bar{y}_т = 99{,}56 - 8{,}19x + 0{,}378x^2,$$

где $\bar{y}_т$ — расчетное (теоретическое) содержание рубидия в условных единицах;

x — фактическое содержание $MgCl_2$ в процентах.

Изучение нелинейной связи не заканчивается выводом уравнения параболической регрессии. Необходимы также оценка надежности этой регрессии, ее геологическое истолкование и использование для практических целей.

Оценка параболической регрессии дается по коэффициентам регрессии:

$$B = \frac{s_{\bar{y}_т}}{s_y},$$

где B — оценка коэффициента регрессии,

s_y — среднее квадратическое отклонение величины y, а величина $s_{\bar{y}_т}$ — вычисляется по формуле

$$s_{\bar{y}_т} = \sqrt{\frac{1}{n}\sum n_x \left(\bar{y}_т - \bar{y}_ф\right)^2}.$$

Величина B удовлетворяет условиям:

$$0 \leqslant B \leqslant 1,$$

$$B \leqslant \eta_y.$$

Величина B равна η_y только тогда, когда регрессия y на x представляется точно полученным уравнением.

Гипотезу $MB = 0$ можно проверить с помощью следующего критерия:

$$\vartheta = \frac{(n-m-2)\left(\eta_y^2 - B^2\right)}{(m-3)\left(1-\eta_y^2\right)} \sqrt{\frac{(m-3)(n-m-4)}{2(n-5)}},$$

где n — общее число проб;

m — число классов величины x, в том числе и те классы, в которых нет ни одной пробы, если они находятся не на самом краю таблицы;

η_y — оценка корреляционного отношения y на x;

B — эмпирический коэффициент параболической регрессии.

Если гипотеза $MB = 0$ верна, то величина ϑ распределена асимптотически нормально с параметрами 0,1. Поэтому гипотеза $MB = 0$ отвергается, если $\vartheta > 3$.

Применение этих формул проиллюстрируем на примере с теллуром и мышьяком (табл. 123).

Таблица 123

x	n_x	$\bar{y}_ф$	$\bar{y}_т$	$\bar{y}_т - \bar{y}_ф$	$(\bar{y}_т - \bar{y}_ф)^2$	$n(\bar{y}_т - \bar{y}_ф)^2$
5	4	0	0,5	0,5	0,25	1,00
15	10	3,1	2,8	—0,3	0,09	0,90
25	5	4,0	4,4	0,4	0,16	0,80
35	5	6,4	5,6	—0,8	0,64	3,20
45	5	6,4	6,2	—0,2	0,04	0,20
55	3	4,3	6,3	2,0	4,00	12,00
65	2	6,5	5,8	—0,7	0,49	0,98
75	1	6,0	4,8	—1,2	1,44	1,44
Всего . . .	35					20,52

Выборочное среднее квадратическое отклонение величины $\bar{y}_т$ равно

$$s_{\bar{y}_т} = \sqrt{\frac{1}{35} \cdot 20{,}52} = 0{,}766.$$

Среднее квадратическое отклонение величины y было вычислено ранее. Оно равно $s_y = 3{,}33$.

Коэффициент параболической регрессии B на основании этих данных определяется так:

$$B = \frac{0{,}766}{3{,}33} = 0{,}230.$$

Он значительно меньше корреляционного отношения, равного 0,610. Это указывает на то, что уравнение параболической регрессии неточно отражает интересующую нас связь теллура с мышьяком. Вычисленное значение ϑ равно 2,14.

Уравнение второй степени оказывается недостаточно надежным, поскольку $2{,}14 < 3$. В этом случае следовало бы использовать уравнение высшей степени, но цель этого примера состояла только в иллюстрации техники расчета. Поэтому от поисков уравнения третьей или четвертой степени мы воздержимся.

Геологическое истолкование выявленной статистическим путем связи дается с учетом всех конкретных условий и прежде всего с учетом минераграфических, петрографических и минералогических данных.

Минераграфические наблюдения показали, что мышьяк в медной руде находится в форме теннантита и что теннантит концентрируется главным образом в центральной части (по мощности) меднорудных линз.

Теллур в какой-то степени связан с теннантитом, но для выяснения причины криволинейности связи необходимо более детальное изучение месторождения.

По-видимому, условия кристаллизации теннантита и условия связи с ним теллура существенно изменились при содержании мышьяка в руде около 0,05%.

Других данных для геологического истолкования криволинейной связи теллура с мышьяком нет.

XIII. ВЫЯВЛЕНИЕ СВЯЗИ МЕЖДУ ДВУМЯ КАЧЕСТВЕННЫМИ ПРИЗНАКАМИ ПРИ ДВУХРАЗРЯДНОЙ ГРУППИРОВКЕ

Факт отсутствия связи между теми или иными признаками не менее интересен для геолога, чем факт ее наличия. Так, на коренных месторождениях алмазов в Якутии (Бобриевич, Бондаренко и др., 1959) было установлено (правда, неточными методами) отсутствие связи между крупностью алмазов и степенью их близости к краям кимберлитовых трубок. Вблизи краев этих трубок, так же как и в середине их (по горизонтальному сечению), встречаются как мелкие, так и крупные алмазы. Следовательно, алмазы выросли не в трубке, а гораздо глубже. С большой глубины алмазы были выжаты вверх вместе с кимберлитовой магмой. С этим выводом затем были связаны некоторые положения методики поисков алмазных месторождений и прогнозы оруденения на глубину.

Нахождение россыпных алмазов на Урале не связано с возрастом пород плотика, но отсутствие этой связи сужает круг ответов на вопрос о факторах локализации этих россыпей.

Обозначим событие, заключающееся в том, что некоторый объект обладает признаком A через A, а признаком B через B. Отсутствие признака A обозначим через α, а отсутствие признака B через β. Пусть общее число рассматриваемых объектов будет N, а число тех объектов, которые обладают признаком A, обозначим (A), признаком B — (B), α — (α), β — (β). Число объектов, обладающих признаками A и B одновременно, обозначим (AB) и т. д.

В случае отсутствия какой бы то ни было связи между двумя качественными признаками A и B имеет место равенство

$$\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)}. \qquad (1)$$

Пример. Пусть из 713 проб, взятых на четных горизонтах оловянного рудника, 11 проб оказались ураганными (пробы с выдающимся высоким содержанием олова), а из 1206 проб, взятых на нечетных горизонтах,

ураганных проб было 19. Таким образом, $(AB) = 11$, $(A\beta) = 19$, $(B) = 713$, $(\beta) = 1206$. Чтобы узнать, есть ли связь ураганных проб с четностью номеров горизонтов, подставим в эту формулу полученные значения групповых объемов. При этом оказывается, что

$$\frac{(AB)}{(B)} = \frac{11}{713} = 0{,}0154; \quad \frac{(A\beta)}{(\beta)} = \frac{19}{1206} = 0{,}0158.$$

Таким образом,

$$\frac{(AB)}{(B)} \simeq \frac{(A\beta)}{(\beta)},$$

а это доказывает отсутствие связи.

Запишем приведенную выше формулу с обратными знаками и прибавим к ее правой и левой части по единице:

$$1 - \frac{(AB)}{(B)} = 1 - \frac{(A\beta)}{(\beta)}.$$

Ввиду того что $(B) - (AB) = (\alpha B)$, а $(\beta) - (A\beta) = (\alpha\beta)$, получим

$$\frac{(\alpha B)}{(B)} = \frac{(\alpha\beta)}{(\beta)}.$$

Таким же путем получим

$$\frac{(AB)}{(A)} = \frac{(\alpha B)}{(\alpha)}$$

и

$$\frac{(A\beta)}{(A)} = \frac{(\alpha B)}{(\alpha)}.$$

Приведенные в последних формулах групповые объемы можно свести в таблицу (табл. 124).

Таблица 124

Признак	B	β	Итого
A α	(AB) (αB)	$(A\beta)$ $(\alpha\beta)$	(A) (α)
Всего	(B)	(β)	N

Во всех приведенных формулах фигурируют как позитивные, так и негативные признаки. На практике приходится иметь дело с разными сочетаниями признаков, в том числе и с одними только позитивными признаками, а также с исходной совокупностью. В связи с этим формулу (1) можно преобразовать так:

$$\frac{(AB)}{(B)} = \frac{(AB) + (A\beta)}{(B) + (\beta)} = \frac{(A)}{(N)}.$$

Таким образом, мы имеем

$$\frac{(AB)}{(B)} = \frac{(A)}{(N)}. \qquad (2)$$

Эта формула показывает, что в случае взаимной независимости признаков A и B доля объемов с A в числе тех объектов, для которых выполнено условие B, такова же, как и во всей совокупности.

Подобно этому получим:

$$\frac{(AB)}{(A)} = \frac{(B)}{(N)},$$

$$(AB) = \frac{(A)(B)}{(N)},$$

$$\frac{(AB)}{(N)} = \frac{(A)}{(N)} \cdot \frac{(B)}{(N)}.$$

Рассмотрим пример. При поисках месторождения серного колчедана в Карелии изучался состав галечников. Доля белых галек $\frac{(A)}{N}$ в одном из поисковых шурфов составила 0,85, а доля галек с признаками оруденения $\frac{(B)}{N}$ была равна 0,11. При какой доле $\frac{(AB)}{N}$ мы могли бы сделать вывод о независимости оруденения от цвета галек? Ответ на этот вопрос находим так:

$$\frac{(AB)}{(N)} = \frac{(A)}{(N)} \cdot \frac{(B)}{(N)} = 0{,}85 \cdot 0{,}11 = 0{,}09.$$

Здесь подразумевается доля во всей совокупности.

Из приведенных выше формул можно вывести ряд новых. Так, например,

$$\frac{(A\beta)}{(\beta)} = \frac{(AB) + (A\beta)}{(B) + (\beta)} = \frac{(A)}{(N)},$$

откуда

$$(A\beta) = \frac{(A)(B)}{(N)}.$$

Таким же путем получим

$$(\alpha B) = \frac{(\alpha)(B)}{(N)},$$

$$(\alpha\beta) = \frac{(\alpha)(\beta)}{(N)}.$$

В приведенных выше формулах фигурируют не все, а лишь некоторые групповые объемы (в каждой формуле свои). Критерий несвязанности * двух признаков будет более полным, если воспользоваться одним из следующих выражений:

$$(AB)(\alpha\beta) = \frac{(A)(B)(\alpha)(\beta)}{(N)^2},$$

$$(\alpha B)(A\beta) = \frac{(A)(B)(\alpha)(\beta)}{(N)^2}.$$

Оба эти критерия равноправны. Каждый из них позволяет выявить несвязанность признаков.

Рассмотрим пример. Пусть на месторождении пьезокварца было добыто и осмотрено некоторое количество кристаллов кварца. Одни из кристаллов были «скрученными» (признак A), другие — с включениями

* Иногда его называют критерием независимости, но тут надо отличать независимость испытаний от независимости признаков.

турмалина («стрелы Амура») или рутила («волосы Венеры»). Наличие включений будем считать признаком B. Количество кристаллов с различными комбинациями признаков таково: $(AB) = 312$, $(A\beta) = 254$, $(\alpha B) = 1912$, $(\alpha\beta) = 3491$.

Необходимо узнать, являются ли признаки A и B несвязанными друг с другом? Если признаки независимы, то

$$\frac{(AB)}{(A\beta)} = \frac{(\alpha B)}{(\alpha\beta)}.$$

В нашем случае равенство не выполняется, так как

$$\frac{312}{254} > \frac{1912}{3491}.$$

Отсюда следует, что признаки A и B нельзя считать несвязанными друг с другом.

Полное отсутствие взаимной связи между какими-либо двумя качественными признаками очень редко имеет место. Почти всегда признаки хоть слабо, но связаны друг с другом. В наличии этой связи можно убедиться по тем же критериям, которые выше были приведены для установления несвязанности, но в них вместо знака равенства нужно поставить знак неравенства. Например, можно записать

$$(AB) > \frac{(A)(B)}{(N)}$$

или

$$(AB) < \frac{(A)(B)}{(N)}.$$

Первое из этих неравенств указывает на наличие положительной связи между признаками A и B, а второе — на наличие отрицательной связи между теми же признаками.

Положительная связь состоит в увеличении тенденции к совместному появлению признаков A и B, а отрицательная — к совместному появлению признаков α и B или признаков A и β.

Если все единицы наблюдения, обладающие признаком A, одновременно обладают и признаком B, или если все единицы наблюдения с признаком B одновременно имеют и признак A, то мы получим полную положительную связь.

При этом первое условие может осуществиться при $(A) \leqslant (B)$, а второе — при $(A) \geqslant (B)$.

В случае полной положительной связи осуществляется одно из следующих двух равенств или оба равенства сразу:

$$(AB) = (A),$$
$$(AB) = (B),$$

В противоположность этому случай, когда ни одна единица наблюдения, обладающая признаком A, не имеет одновременно и признака B или когда ни одна единица наблюдения, обладающая признаком B, не имеет одновременно и признака A, даст нам полную отрицательную связь. Полная отрицательная связь будет и тогда, когда ни одна из единиц наблюдения, имеющая признак α, не обладает одновременно и признаком β, т. е. когда осуществляются следующие два равенства сразу или какое-либо одно из них:

$$(AB) = 0,$$
$$(\alpha\beta) = 0.$$

Исходя из приведенного выше, мы можем написать

$$(AB) = (A) + (B) - (N).$$

Полная связь (положительная или отрицательная), как и отсутствие связи, на практике встречается очень редко. Чаще всего мы имеем неполную связь (положительную или отрицательную). Реальное количество связанных признаков A и B, или α и β, а также признаков α и B или A и β, почти всегда находится где-то в промежутке между двумя границами, из которых одна означает отсутствие связи, а другая — полную положительную или полную отрицательную связь.

Степень близости реальной связи ко второй из этих двух границ характеризует силу связи. При этом связь бывает тесная, или сильная, и нетесная, или слабая.

Сила связи может очень близко подходить к той или другой границе, в частности она может быть очень близкой к тому, что мы называем отсутствием связи. Выводы о наличии существенной или несущественной связи нужно делать с большой осторожностью.

Некоторое количество единиц наблюдения может одновременно иметь оба признака, как следствие случайного совпадения. Делать на этом основании выводы о существенной связи было бы неосторожным. Поэтому необходим более полный анализ явления.

Поскольку в наличии связи мы можем убедиться путем сравнения (AB) с $\frac{(A)(B)}{(N)}$, то силу этой связи мы могли бы измерить, исходя из установленных соотношений. Но на практике измерение силы связи признаком A и B производят путем сравнения доли A в B с долей A в β. Это сравнение можно производить разными способами. Выбор способа зависит от того, какой вопрос нам нужно решить с помощью измерения силы связи.

Для измерения силы связи качественных признаков необходимо учитывать доли признаков в общей совокупности. Поэтому введем следующие обозначения для этих долей при условии, что зависимости нет:

$$(AB)_0 = \frac{(A)(B)}{(N)},$$

$$(\alpha B)_0 = \frac{(\alpha)(B)}{(N)},$$

$$(\alpha\beta)_0 = \frac{(\alpha)(\beta)}{(N)},$$

$$(A\beta)_0 = \frac{(A)(\beta)}{(N)}.$$

Поскольку равенство $(AB) = \frac{(A)(B)}{(N)}$ является критерием независимости, для нас представляет интерес разница между реальным (AB) и тем теоретическим значением (AB), которое имело бы место в случае независимости признаков, т. е.

$$(AB) - (AB)_0 = \delta.$$

Если нас интересуют другие групповые объемы, то и для них мы можем вычислить подобную разницу:

$$(\alpha B) - (\alpha B)_0 = \delta,$$
$$(\alpha\beta) - (\alpha\beta)_0 = \delta,$$
$$(A\beta) - (A\beta)_0 = \delta.$$

Из последних восьми равенств мы можем получить следующее общее выражение для δ:

$$\delta = \frac{(AB)(\alpha\beta) - (\alpha B)(A\beta)}{N},$$

откуда

$$N\delta = (AB)(\alpha\beta) - (\alpha B)(A\beta).$$

Силу связи можно измерить с помощью коэффициента связи, в определение которого входит δ. Этот коэффициент будет более удобным, если его значения уложатся в пределах от 0 до ±1. При этом нуль указывал бы на отсутствие связи, единица — на полную связь, а все прочие значения — на неполную связь. Знак плюс говорил бы о положительной связи, а знак минус — об отрицательной.

Д. Юл и М. Кендэл (1960) приводят два коэффициента, удовлетворяющие приведенному здесь условию. Первый из них определяется по формуле

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}.$$

Так как числитель в этой формуле представляет собой не что иное, как произведение $N\delta$, определяемое по предыдущей формуле, мы можем писать

$$Q = \frac{N\delta}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}.$$

Если связь отсутствует, то $\delta = 0$ и $Q = 0$. Если связь полная положительная, то $(A\beta)(\alpha B) = 0$, а $Q = +1$. Если, наконец, связь полная отрицательная, то $(AB)(\alpha\beta) = 0$, а $Q = -1$.

Коэффициент Q Д. Юл и М. Кендэл (1960) называют коэффициентом связи.

Второй коэффициент определяется по формуле

$$Y = \frac{1 - \sqrt{\frac{(AB)(\alpha B)}{(AB)(\alpha\beta)}}}{1 + \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}.$$

Этот коэффициент называется коэффициентом взаимосвязанности или коллигации.

Между величинами Q и Y существует такое взаимоотношение:

$$Q = \frac{2Y}{1 + Y^2}.$$

Ввиду того что каждый коэффициент имеет свои достоинства и свои недостатки, рекомендуется пользоваться тем и другим одновременно.

Пример. Пусть A — рабочие смены с повышенной шумностью массива угля в одной шахте (шумность измерялась звукометрическим методом), α — рабочие смены с обычной (неповышенной) шумностью, B — рабочие смены с внезапными выбросами угля, β — рабочие смены без выбросов угля. Статистика показала, что за полгода произошло 6 внезапных выбросов угля (по одному в смену), при этом $(AB) = 5$, $(\alpha B) = 1$, $(A\beta) = 3$, $(\alpha\beta) = 443$.

Требуется выяснить, какова сила связи выбросов с шумностью.

По приведенным выше формулам вычисляем:

$$Q = \frac{5 \cdot 443 - 3 \cdot 1}{5 \cdot 443 + 3 \cdot 1} = \frac{2212}{2218} = 0{,}99.$$

$$Y = \frac{1 - \frac{3 \cdot 1}{5 \cdot 443}}{1 + \frac{3 \cdot 1}{5 \cdot 443}} = \frac{0{,}9632}{1{,}0368} = 0{,}93.$$

По полученным значениям коэффициентов связи можно сделать вывод, что связь между шумностью и выбросами положительная и очень

сильная (почти полная), а это позволяет предсказывать наступление выброса.

Для измерения связи признаков, используется также коэффициент ассоциации Юла: Так, например,

$$K_A = \frac{ad - bc}{ad + bc}.$$

Здесь величины a, b, c и d — суть не что иное, как соответствующие частоты из таблицы сопряженности признаков (табл. 125).

Вычисление коэффициента контингенции Пирсона:

$$K_k = \frac{ad - bc}{\sqrt{(a+b)(b+d)(a+c)(c+d)}},$$

где a, b, c и d — имеют те же значения.

Таблица 125

a	b
c	d

Связь двух качественных признаков друг с другом, их зависимость друг от друга, можно измерить коэффициентом корреляции.

Коэффициент корреляции двух зависимых друг от друга событий или качественных признаков A и B при независимых испытаниях имеет следующую формулу:

$$r = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{\sqrt{(A)(\alpha)(B)(\beta)}}.$$

Этот коэффициент чутко устанавливает связь признаков.

Пример. Пусть A — одна конкретная причина производственного травматизма, а именно обвал породы в угольных шахтах; α — прочие причины производственного травматизма (нарушение правил ведения взрывных работ, неправильное использование подземного транспорта, взрывы газа и пыли и пр.); B — самый тяжелый вид травматизма — со смертельным исходом; β — прочие виды травматизма (общий и тяжелый, без смертельного исхода). За 1959 г. в угольной промышленности РСФСР (Черемшанцев, 1961) произошло некоторое количество случаев травматизма. Допустим, что из них 10% было особо тяжелых, повлекших за собой смерть пострадавшего. Сочетание признаков (по числу случаев) пусть будет таким: $(AB) = 3{,}87\%$, $(A\beta) = 30{,}65\%$, $(\alpha B) = 6{,}13\%$, $(\alpha\beta) = 59{,}35\%$. Требуется выяснить характер и силу связи летальности (смертельного исхода) травматизма и обвалов породы. Для этого необходимо вычислить коэффициент корреляции. В последней формуле приведены объемы групп, но коэффициент корреляции не изменится, если вместо объемов групп мы возьмем доли групп (в процентах к неизвестному объему всей совокупности). В нашем примере как раз и фигурируют не объемы групп, а их доли. Для вычисления коэффициента корреляции необходимо знать объемы групп первого порядка. Эти объемы легко подсчитываются (табл. 126).

Таблица 126

Признак	B	β	Итог
A	3,87	30,65	34,52
α	6,13	59,35	65,48
Всего. . . .	10,00	90,00	100,00

Теперь можно вычислить коэффициент корреляции

$$r = \frac{3{,}87 \cdot 59{,}35 - 30{,}65 \cdot 6{,}13}{\sqrt{34{,}52 \cdot 65{,}48 \cdot 10{,}00 \cdot 90{,}00}} = 0{,}028.$$

Ввиду того что коэффициент корреляции очень мал, делаем вывод: предположение о возможной связи смертельных исходов травматизма с обвалами породы не подтвердилось. Смертельные исходы могут зависеть не только от обвалов породы, но и от других причин. Обвалы же вызывают не только смерть людей, но и различные виды травматизма (легкого и тяжелого, но необязательно со смертельным исходом). Для более глубокого анализа этого вопроса в книге Я. Н. Черемшанцева (1961) нет достаточных данных.

Приведем еще один пример. На одном золоторудном месторождении в Забайкалье было замечено, что промышленное оруденение (в тонких кварцевых жилах), по-видимому, связано с зеленоватым оттенком вмещающей породы. Это предположение можно проверить по результатам анализа 1434 проб, сведенным в табл. 127.

Таблица 127

Характеристика проб по величине содержания в них золота	Вмещающие породы		Всего проб
	с зеленоватым оттенком B	без зеленоватого оттенка α	
Промышленные A	734	146	880
Непромышленные и пустые α . . .	99	455	554
Итого проб	833	601	1434

Коэффициент корреляции A и B таков:

$$r = \frac{734 \cdot 455 - 146 \cdot 99}{\sqrt{880 \cdot 554 \cdot 833 \cdot 601}} = 0{,}648.$$

Полученное число — положительное и не малое (более 0,5). Поэтому предположение о связи признаков A и B остается в силе. Оно, по крайней мере, не опровергнуто. Для окончательного принятия этого предположения необходимо еще проверить реальность полученного значения выборочного коэффициента корреляции.

Для решения вопроса о том, насколько реально значение коэффициента корреляции и не определено ли оно одними только случайностями, А. А. Чупров (1926) рекомендует вычислить среднюю квадратическую ошибку коэффициента корреляции σ_r по формуле

$$\sigma_r = \frac{1 - r^2}{\sqrt{N}}.$$

Здесь N — число наблюдений, а не 100%.

Для последнего примера имеем

$$\sigma_r = \frac{1 - 0{,}648^2}{\sqrt{1434}} = 0{,}015.$$

В. И. Романовский (1938) считает, что если абсолютное значение коэффициента корреляции не менее трех σ_r, то связь признаков реальная. В данном случае $3\sigma_r = 0{,}045$. Поэтому

$$|r| = 0{,}648 > 3\sigma_r = 0{,}045.$$

Теперь мы можем сделать вывод: между промышленным содержанием золота и зеленоватым оттенком вмещающей породы на одном из забайкальских рудников имеется прямая и довольно сильная связь.

Приведем еще пример связи признаков.

В. В. Эз (1956), М. С. Анциферов, А. Г. Константинова, Л. Б. Переверзев (1960) и другие изучали внезапные выбросы угля и газа в шахтах. В. В. Эз изучал связь внезапных выбросов угля с тектоникой угольных пластов и с крепостью угля. В табл. 103 его работы приведены данные о выбросах угля. По этим данным можно вычислить коэффициент корреляции между крепостью выбрасываемого угля и крутым залеганием пласта в местах выброса.

Обозначим крепкий (и средний по крепости) уголь буквой A, мягкий — α, крутое падение пластов — B, пологое — β. Групповые объемы таковы:

$$(AB) = 154,$$
$$(A\beta) = 4,$$
$$(\alpha B) = 37,$$
$$(\alpha\beta) = 16.$$

Исходя из этих данных, находим:

$$(A) = 158,$$
$$(B) = 191,$$
$$(\alpha) = 53,$$
$$(\beta) = 20,$$
$$(N) = 211.$$

Коэффициент корреляции таков:

$$r = \frac{154 \cdot 16 - 4 \cdot 37}{\sqrt{158 \cdot 191 \cdot 53 \cdot 20}} = \frac{2316}{5648} = 0{,}41.$$

Далее находим

$$\sigma_r = \frac{1 - 0{,}41^2}{\sqrt{211}} = 0{,}057.$$

Сравниваем с критерием Романовского:

$$r = 0{,}41 > \frac{3}{\sqrt{211 - 1}} = 0{,}207.$$

Вывод: связь признаков A и B — положительная, существенная. Следовательно, внезапные выбросы угля на шахтах с крутым падением пластов зависят от крепости угля, а на крепких углях — от крутизны падения.

Связь двух признаков друг с другом, обнаруженная с помощью коэффициента корреляции, не всегда является причинно-следственной связью. Возьмем, например, высокое содержание алмаза в россыпях по месторождениям Советского Союза как признак A и залесенность территории месторождений как признак B. Корреляционный анализ покажет нам положительную и существенную связь этих признаков, но причинно-следственной связи между ними нет. Возьмем также крупность алмазов и наличие вечной мерзлоты. Связь между этими признаками будет отрицательной, но один (любой) из этих признаков не является причиной другого, а другой — следствием первого.

Наличие корреляционной связи в этих и подобных им случаях вызвано не причинно-следственными отношениями, а наличием какого-либо третьего признака, с которым каждый из двух изучаемых признаков (A и B) самостоятельно связан причинно-следственной связью (не всегда непосредственно).

В первом из приведенных здесь примеров содержание алмазов связано с геологическими особенностями месторождений (в основном якутских и уральских), а эти особенности связаны с историей земной коры в соответствующих районах. В историю земной коры входит и история

поверхности земли, а с историей поверхности связана та или иная залесенность районов. Связь содержания алмазов с залесенностью осуществляется через посредство длинной цепи причин и следствий.

Во втором примере более значительная крупность уральских алмазов, по сравнению с якутскими алмазами, определена также особыми чертами геологической истории Урала, а вечная мерзлота в Якутии связана с историей земной коры, а может быть и солнечной системы через посредство многих промежуточных звеньев причинно-следственной цепи явлений. Получается так называемая ложная корреляция, или ложная связь признаков (крупности алмазов и мерзлоты).

Вместо термина «ложная корреляция» употребляют также термин «косвенная связь», «мнимая взаимосвязь», «частная ассоциация» и другие.

Отыскание связи признаков представляет собой сложную задачу. Если взять два каких-либо признака A и B и применить к ним какую-либо из приведенных выше формул, то в случае обнаружения связи мы еще не гарантированы от того, что эта связь мнимая, а в случае обнаружения несвязанности мы не можем быть уверенными в том, что эта несвязанность действительная, а не кажущаяся. Во всех этих случаях необходимо учитывать возможное влияние третьего или вообще третьих признаков и только после проверки этого влияния делать окончательный вывод о наличии или отсутствии связи между признаками A и B.

Косвенная или частная связь признаков A и B обнаруживается только там, где каждый из этих признаков связан с третьим признаком C. Прямая же или общая связь признаков A и B не зависит от признака C.

Ища косвенную связь, объемы (A), (α), (B), (β), (AB), $(A\beta)$, (αB) и $(\alpha\beta)$, необходимые при поисках прямой связи, заменяем соответственно объемами (AC), (αC), (BC), (βC), (ABC), $(A\beta C)$, (αBC) и $(\alpha\beta C)$.

Подставив эти объемы в обычную формулу коэффициента корреляции, получим коэффициент корреляции A и B в классе C

$$r_C(AB) = \frac{(ABC)(\alpha\beta C) - (A\beta C)(\alpha BC)}{(AC)(\alpha C)(BC)(\beta C)}.$$

Если нас интересует связь признаков A и B в классе γ, то таким же путем получим коэффициент корреляции A и B в классе γ:

$$r_\gamma(AB) = \frac{(AB\gamma)(\alpha\beta\gamma) - (A\beta\gamma)(\alpha B\gamma)}{(A\gamma)(\alpha\gamma)(B\gamma)(\beta\gamma)}.$$

Из сопоставления двух последних коэффициентов корреляции мы можем сделать заключение о том, какова связь A и B — частная или общая. Если эти два коэффициента равны между собой, то связь будем считать общей, независящей от C. В противном случае, т. е. при $r_C(AB) \neq r_\gamma(AB)$ связь A и B будет частной.

Частная связь может быть как в классе C, так и в классе γ. Если $r_C(AB) \neq 0$, а $r_\gamma(AB) = 0$, то признаки A и B связаны друг с другом в классе C, но не связаны в классе γ. Если же $r_\gamma(AB) \neq 0$, а $r_C(AB) = 0$, то признаки A и B связаны между собой в классе γ, но не связаны в классе C.

Знак при коэффициенте корреляции (плюс или минус) укажет нам на характер связи — положительной или отрицательной (соответственно).

Поясним сказанное примерами.

Пример 1. Пусть на свинцовом (галенитовом) месторождении были взяты богатые (A) и бедные (α) пробы, причем применялось только два метода опробования: бороздовой (B) и горстевой (β). Одни из проб взяты

по ненарушенной разломами зоне (C), другие — по ненарушенной (γ). Количество проб по различным сочетаниям признаков таково:

$$(ABC) = 834,$$
$$(AB\gamma) = 107,$$
$$(A\beta C) = 144,$$
$$(A\beta\gamma) = 776,$$
$$(\alpha BC) = 693,$$
$$(\alpha B\gamma) = 206,$$
$$(\alpha\beta C) = 851,$$
$$(\alpha\beta\gamma) = 92.$$

По этим данным необходимо выяснить, связаны ли друг с другом признаки A и B в общей совокупности и в классах C и γ (отдельно в каждом классе).

Это можно сделать путем вычисления коэффициентов корреляции, но сначала необходимо найти объемы групп второго порядка, фигурирующих во всех формулах коэффициентов корреляции и групп первого порядка, фигурирующих в одной из них.

Найдем объем следующих групп:

$$(AB) = 834 + 107 = 941,$$
$$(AC) = 834 + 144 = 978,$$
$$(BC) = 834 + 693 = 1527,$$
$$(A\beta) = 144 + 776 = 920,$$
$$(A\gamma) = 107 + 776 = 883,$$
$$(B\gamma) = 107 + 206 = 313,$$
$$(\alpha B) = 693 + 206 = 899,$$
$$(\alpha C) = 693 + 851 = 1544,$$
$$(\beta C) = 144 + 851 = 995,$$
$$(\alpha\beta) = 851 + 92 = 943,$$
$$(\alpha\gamma) = 206 + 92 = 298,$$
$$(\beta\gamma) = 776 + 92 = 868,$$
$$(A) = 941 + 920 = 1861,$$
$$(B) = 941 + 899 = 1840,$$
$$(C) = 978 + 1544 = 2522,$$
$$(\alpha) = 899 + 943 = 1842,$$
$$(\beta) = 920 + 943 = 1863,$$
$$(\gamma) = 833 + 298 = 1181.$$

Теперь мы можем вычислить коэффициенты корреляции:

$$r(AB) = \frac{941 \cdot 943 - 920 \cdot 899}{\sqrt{1861 \cdot 1840 \cdot 1842 \cdot 1863}} = \frac{60\,283}{3\,422\,000} = 0{,}018;$$

$$r_C(AB) = \frac{834 \cdot 851 - 144 \cdot 693}{\sqrt{978 \cdot 1527 \cdot 1544 \cdot 995}} = \frac{610\,200}{1\,510\,000} = 0{,}404.$$

$$r_\gamma(AB) = \frac{107 \cdot 92 - 776 \cdot 206}{\sqrt{883 \cdot 298 \cdot 313 \cdot 868}} = -\frac{150\,012}{266\,500} = -0{,}564.$$

Эти коэффициенты корреляции показывают, что признаки A и B в общей совокупности не связаны друг с другом, но в частных совокупностях между ними имеется существенная связь. При наличии признака C эта связь положительная, а при наличии признака γ — связь отрицательная, наиболее сильная. Из этого можно сделать вывод, что высокое содер-

жание свинца в бороздовых пробах наблюдается только в ненарушенной зоне. В нарушенной зоне бороздовые пробы сильно преуменьшают содержание металла (по-видимому, из-за выкрашивания галенита).

Пример 2. В геологоразведочной партии было пробурено много скважин диаметром от 160 до 76 *мм*. По каждой из них подсчитывался выход керна отдельно для гранитов (с ними связано вкрапленное оруденение) и для туфогенных пород (других пород скважины не пересекали). Выход керна по многим скважинам был ниже нормы, установленной проектом разведки. Представляет интерес вопрос о возможной зависимости выхода керна от диаметра бурения.

Обозначим приемлемый выход керна через A, неприемлемый — через α, диаметры бурения 114 *мм* и выше — через B, менее 114 *мм* — β, граниты — C, туфогенные породы — γ.

Суммарный метраж бурения в зависимости от установленных здесь признаков разбивается на такие группы:

$$(ABC) = 419,$$
$$(AB\gamma) = 8,$$
$$(A\beta C) = 418,$$
$$(A\beta\gamma) = 11\,892,$$
$$(\alpha BC) = 9,$$
$$(\alpha B\gamma) = 1189,$$
$$(\alpha\beta C) = 59,$$
$$(\alpha\beta\gamma) = 6$$

По тому же плану, как и в первом примере, находим объемы следующих групп:

$$(AB) = 427,$$
$$(A\beta) = 12\,310,$$
$$(\alpha B) = 1198,$$
$$(\alpha\beta) = 65,$$
$$(AC) = 837,$$
$$(A\gamma) = 11\,900;$$
$$(\alpha C) = 68,$$
$$(\alpha\gamma) = 1195,$$
$$(BC) = 428,$$
$$(B\gamma) = 1197,$$
$$(\beta C) = 477,$$
$$(\beta\gamma) = 11\,898,$$
$$(A) = 12\,737,$$
$$(B) = 1625,$$
$$(C) = 946,$$
$$(\alpha) = 1263,$$
$$(\beta) = 12\,375,$$
$$(\gamma) = 13\,054.$$

Затем вычисляем коэффициенты корреляции:

$$r(AB) = \frac{427 \cdot 65 - 1198 \cdot 12\,310}{\sqrt{12\,737 \cdot 1263 \cdot 1625 \cdot 12\,375}} = -0{,}819;$$

$$r_C(AB) = \frac{419 \cdot 59 - 9 \cdot 418}{\sqrt{837 \cdot 68 \cdot 428 \cdot 477}} = +0{,}194;$$

$$r_\gamma(AB) = \frac{8 \cdot 6 - 11\,892 \cdot 1189}{\sqrt{11\,900 \cdot 1195 \cdot 1197 \cdot 11\,898}} = -0{,}993.$$

Первый из этих трех коэффициентов корреляции показывает, что при бурении скважин большого диаметра снижается выход керна, но это, несмотря на близость коэффициента корреляции к единице, противоречит здравому смыслу. Разбивка всей совокупности (суммарного метража) на две частные совокупности (в гранитах и туфогенных породах) и вычисление частных коэффициентов показывает, что все дело в крепости пород. При бурении скважин большого диаметра в гранитах выход керна повышается (хотя и в малой степени), а в туфогенных породах резко понижается. Граниты залегают под туфогенными породами, поэтому бурение по ним ведется после того, как перешли с бурения скважин большого диаметра на скважины малого диаметра. Туфогенные породы лежат сверху и пробуриваются коронками большого диаметра; выход керна по ним даже при больших диаметрах бурения очень низок, т. к. они очень рыхлые.

Таким образом, один фактор (крепость пород) наложился в этом примере на другой фактор (диаметр бурения), замаскировал и даже исказил его влияние на выход керна.

На основании этих примеров можно сделать выводы:

Если у нас имеется три признака и два из них (A и B) связаны с третьим признаком (C или γ), а также друг с другом в классе C и γ, то из этого нельзя сделать заключение о характере связи A и B в общей совокупности, т. е. в объединенном классе $C + \gamma$. Связь признаков A и B с C может как усилить, так и ослабить и изменить на противоположную связь их с γ.

Если признаки A и B не связаны между собой ни в классе C, ни в классе γ, то в общей совокупности (в объединенном классе $C + \gamma$) они могут быть связанными (при наличии их связи с C).

В случае, если в изучаемой совокупности ни один признак не связан ни с каким другим признаком, имеет место следующее равенство:

$$\frac{(ABC)}{(N)} = \frac{(A)}{(N)} \cdot \frac{(B)}{(N)} \cdot \frac{(C)}{(N)}.$$

При неограниченном числе признаков это равенство (критерий независимости) примет следующий вид:

$$\frac{(ABC\ldots)}{(N)} = \frac{(A)}{(N)} \cdot \frac{(B)}{(N)} \cdot \frac{(C)}{(N)} \cdot \ldots$$

XIV. СВЯЗЬ КАЧЕСТВЕННЫХ ПРИЗНАКОВ ПРИ МНОГОРАЗРЯДНОЙ ГРУППИРОВКЕ

В предыдущей главе мы рассматривали связь только двух признаков, но в явлениях природы и в деятельности людей нередко проявляется связь многих признаков.

Выявление и измерение связи многих качественных признаков представляет собой чрезвычайно трудную задачу, которая к настоящему времени разрешена только в общих чертах.

Связь двух признаков мы обнаруживали с помощью двухразрядной группировки, при которой признаки делились на позитивные и негативные. При этом последние отличались неопределенностью. О признаке α, например, мы могли сказать только одно — что он не является A.

Теперь мы можем уменьшить эту неопределенность негативных признаков и разделить их на несколько новых признаков, а это означает, что от двухразрядной группировки мы переходим к многоразрядной (в ней может быть любое целое число разрядов большее двух). Лишь один из новых признаков, получаемых нами путем уточнения и раздробления негативного признака, может еще оставаться неопределенным, да и то не всегда. Множественную группировку можно получить из простой (двухразрядной)

как путем выделения дополнительных групп из негативных, так и путем дробления позитивных групп.

Для пояснения первого пути приведем пример с якутскими алмазами.

Пусть A будет характеризовать форму алмазов, B — их цвет и C — место нахождения. При двухразрядной группировке мы могли бы буквой A обозначить октаэдрические алмазы, α — алмазы всех других форм; B — бесцветные, а β — окрашенные в любой цвет алмазы; C — алмазы из Алакитского района, а γ — из всех других алмазоносных районов. При многоразрядной группировке мы, оставив в силе признаки A, B и C, можем подразделить α, β и γ на ряд новых признаков. В конечном счете мы можем получить такие признаки: A_1 (ранее обозначенный как A) — октаэдрические алмазы, A_2 — алмазы с полицентрически растущими гранями, A_3 — алмазы, сложенные уменьшающимися тригональными слоями роста, A_4 — алмазы всех других форм; B_1 (ранее обозначавшиеся B) — бесцветные алмазы, B_2 — светло-серые алмазы, B_3 — алмазы всех других окрасок; C_1 — (ранее обозначавшиеся C) — алмазы из Алакитского района, C_2 — алмазы из Мунского района и C_3 — алмазы из Оленекского района.

В этой системе обозначений мы получаем по форме алмазов три определенных и один неопределенный признак (неопределенным является признак A_4), по цвету алмазов два определенных и один неопределенный (B_3) признак и по местоположению — три определенных и ни одного неопределенного признака (кроме трех учтенных, других алмазоносных районов в Якутии нет).

Второй путь получения множественной группировки из простой на практике встречается реже.

Ввиду сложности теории множественной группировки приведем (по Юлу и Кендэлу, 1960) некоторые ее положения без вывода.

Пусть число разрядов группировки по признаку A будет k, по признаку B — l, по C — m и т. д., тогда:

1. Общее число групп, включая исходную совокупность N, будет равно

$$(k + 1)\,(l + 1)\,(m + 1)\ldots$$

2. Число заключительных групп равно произведению чисел разрядов группировки, т. е. $klm\ldots$

3. Условием совместимости заданных групповых объемов является неотрицательный характер объема каждой заключительной группы.

4. Для полной характеристики изучаемой совокупности заданных признаков достаточно иметь $klm\ldots$ алгебраически независимых групповых объемов (в случае, когда некоторые из этих объемов неизвестны, имеющиеся данные позволяют вычислить пределы, в которых эти неизвестные лежат).

Изучение связи многих признаков в настоящее время возможно только в общих чертах. Все признаки мы делим на две группы и ищем связь между этими группами, а не между отдельными признаками. Так, мы можем найти связь между группой A_1, A_2, A_3, A_4 и группой B_1, B_2, B_3 или между группой B_1, B_2, B_3 и группой C_1, C_2, C_3, но не можем найти связь сразу между тремя группами.

При этом в каждую из двух групп, связь между которыми мы ищем, входят не какие угодно, а только родственные признаки (в одну группу, например, входят признаки формы, в другую — признаки цвета алмазов). Так что по существу и тут мы ищем связь не многих, а только двух признаков, правда, разделяющихся на несколько разновидностей.

В большинстве курсов математической статистики теория качественных признаков вообще не излагается, но практика исследовательской работы, особенно в геологии, нуждается в такой теории.

Ниже мы приведем основные положения упомянутой теории.

Для записи и систематизации исходных данных множественной группировки производят сочетание признаков по два, например A_iB_h или B_hC_m и т. д. и составляют по ним таблицы следующего вида (табл. 128).

Таблица 128

A	B				
	B_1	B_2	. . .	B_l	Итог
A_1	(A_1B_1)	(A_1B_2)	. . .	(A_1B_l)	(A_1)
A_2	(A_2B_1)	(A_2B_2)	. . .	(A_2B_l)	(A_2)
.	.	.	. . .	.	.
.	.	.	. . .	.	.
.	.	.	. . .	.	.
A_k	(A_kB_1)	(A_kB_2)	. . .	(A_kB_l)	(A_k)
Итог	(B_1)	B_2	. . .	(B_l)	(N)

Здесь (A_kB), (A_k) и другие символы в скобках означают объемы групп, т. е. число индивидов, обладающих данными признаками.

К. Пирсон и А. Чупров называют такую запись таблицей сопряженности признаков A и B.

Подобные таблицы составляются и для других пар признаков — для B_lC_m, C_mD_h, A_k C_m и т. д.

В изучении связи или ассоциации признаков необходимо различать три операции: выявление связи, измерение ее силы и определение ее характера (т. е. знака плюс или минус при коэффициенте связи).

Выявить наличие связи это значит установить, что признаки A и B не являются независимыми друг от друга.

Критерий независимости признаков A и B дается следующим равенством:

$$(A_iB_k) = \frac{(A_i)\,(B_h)}{N}.$$

В этой формуле индексы i и h могут быть любыми (в пределах таблицы сопряженности признаков). Иначе говоря, здесь берется любая из kl групп (i — номер строки от 1 до k, h — номер столбца от 1 до l).

Эту формулу можно переписать так:

$$N\,(A_iB_h) = (A_i)\,(B_h).$$

Если правую часть этого равенства разделить на (N), то получим

$$\frac{(A_i)\,(B_h)}{(N)} = (A_iB_h)_0.$$

Подставив правую часть последнего равенства в первое равенство, получим критерий независимости признаков A и B в другом выражении:

$$(A_iB_h) = (A_iB_h)_0,$$

но на практике это равенство осуществляется очень редко.

Подобно тому, как мы это делали при двухразрядной группировке, найдем разность $(A_iB_h)-(A_iB_h)_0$ и обозначим ее через δ_{ih}, т. е. получим

$$\delta_{ih} = (A_iB_h) - (A_iB_h)_0$$

или

$$\delta_{ih} = (A_iB_h) - \frac{(A_i)\,(B_h)}{N}.$$

Величина (A_iB_k) нам фактически дана, а величина $(A_iB_h)_0$ вычислена для того воображаемого случая, когда признаки A и B во всех их разновидностях независимы друг от друга. Разность этих величин определена взаимной связью изучаемых признаков. Поэтому ее можно положить в основу для вывода коэффициента связи.

Величина δ_{ih} равна нулю при отсутствии связи; она положительна, когда связь прямая, и отрицательна, когда связь обратная.

Индексы при δ играют очень важную роль. Величина δ_{ik} вообще не равна δ_{hi}, поскольку индекс первого признака i, пробегая ряд значений от 1 до k, может принять значение h, а индекс второго признака h, пробегая ряд значений от 1 до l, может принять значение i, но i как скользящий номер строки в таблице сопряженности — не одно и то же, что и h — скользящий номер столбца в той же таблице.

Ввиду того что при условии независимости

$$\left.\begin{aligned}\sum_i (A_iB_h) &= \sum_i (A_iB_h)_0 \\ \sum_h (A_iB_h) &= \sum_h (A_iB_h)_0\end{aligned}\right\},$$

сумма всех значений δ_{ih} по каждой строке, как и по каждому столбцу, равна нулю, т. е.

$$\left.\begin{aligned}\delta_{i\,1} + \delta_{i\,2} + \ldots + \delta_{ih} + \ldots + \delta_{il} &= 0 \\ \delta_{1h} + \delta_{2h} + \ldots + \delta_{ih} + \ldots + \delta_{kh} &= 0\end{aligned}\right\}.$$

Отрицательные значения δ_{ih} компенсируются положительными.

Величина δ_{ih}, разная для каждого сочетания индексов, является индикатором связи в каждом сочетании признаков A и B, но таких сочетаний много (число их равно произведению числа строк k на число столбцов l, т. е. равно kl), а нам нужен только один коэффициент, который бы характеризовал связь в целом, т. е. связь во всех kl группах.

Такой, общий, коэффициент нельзя получить путем сложения величин δ_{ih}, так как их сумма равна нулю. Чтобы избавиться от влияния знака плюс или минус, стоящего при величине δ_{ih}, берут квадрат этой величины.

Для выявления факта связи пользуются следующей формулой:

$$\chi^2 = \sum_{ih} \frac{\delta_{ih}}{(A_iB_h)_0}.$$

Знак $\sum_{ih}$ в этой формуле говорит о суммировании сначала по индексу i, затем по индексу h (иначе говоря, сначала подвести итог по строчкам, затем общий итог, или сначала итог по столбцам, т. е. по h, затем — общий, т. е. по i). Этот знак можно заменить знаком двойного суммирования $\sum_h \sum_i$, что мы и будем в дальнейшем делать.

Если проверяемая гипотеза об отсутствии сопряженности между признаками верна, то случайная величина χ^2 будет распределена по закону Пирсона χ^2 с $(k-1)(l-1)$ степенями свободы. Таким образом, сопряженность следует считать доказанной, если вычисленное значение χ^2 превысит допустимое $\chi^2_{q.\,(k-1)\,(l-1)}$ при уровне значимости q и $(k-1)(l-1)$ степенях свободы (приложение 14).

Величину χ^2 называют «квадратической сопряженностью». Далее χ^2 делят на число случаев, т. е. на N и получают «среднюю квадратическую сопряженность»

$$\varphi^2 = \frac{\chi^2}{N}.$$

Необходимо заметить, что

$$N = \sum_i (A_i) = \sum_h (B_h) = \sum_h \sum_i (A_iB_h).$$

Величина φ^2 следующим образом связана с коэффициентом корреляции, если последний вычислить для количественных признаков, т. е. для случая, когда каждому значению A_i и B_h соответствуют их количественные выражения, например числа x_i и y_i:

$$\varphi^2 = \frac{r^2}{1 - r^2}.$$

Величина φ^2 не имеет определенных границ. К. Пирсон предложил коэффициент средней квадратической сопряженности C (сам Пирсон называет его первым коэффициентом сопряженности)

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}} = \sqrt{\frac{\varphi^2}{1 + \varphi^2}}.$$

Подставив значение χ^2, получим

$$C = \sqrt{\frac{\sum_h \sum_i \frac{\delta_{ih}^2}{(A_i B_h)_0}}{N + \sum_h \sum_i \frac{\delta_{ih}^2}{(A_i B_h)_0}}}.$$

Как замечают Юд и Кендэл, в последних двух формулах «перед корнем не должно ставить никакого знака: этот коэффициент говорит лишь о том, зависимы или независимы друг от друга два данных признака».

Коэффициент C, по мнению Юла и Кендэла, имеет «одно серьезное неудобство», заключающееся в том, что он равен единице только при бесконечно большом числе групп. Если $k = l = t$, то зависимость максимального значения C от t выразится следующим образом:

t	2	3	4	5	6	7	8	9	10	∞
C_{max} . .	0,707	0,816	0,866	0,894	0,913	0,926	0,935	0,943	0,949	1

Таким образом, величина C зависит от того, на сколько разрядов мы разобьем тот и другой признак, т. е. от характера группировки, а это недопустимо. К. Пирсон предложил поправку к своему коэффициенту, но это делает его очень грамоздким.

Юл и Кендэл считают, что коэффициентом C можно пользоваться лишь при группировках типа 5×5 или более дробных (при этом условии ошибка будет не очень большой).

А. А. Чупров (1926) для определения силы связи признаков предложил коэффициент T, определяемый исходя из формулы

$$T^2 = \frac{\varphi^2}{\sqrt{(k-1)(l-1)}}.$$

Коэффициент T называют коэффициентом сопряженности признаков A и B, или коэффициентом Чупрова.

Коэффициент Чупрова в развернутом виде имеет такое выражение:

$$T^2 = \frac{1}{\sqrt{(k-1)(l-1)}} \sum_h \sum_i \left[\frac{\delta_{ih}^2}{(A_i B_h)_0} \right].$$

Если величину δ_{ih} заменить выражением $(A_i B_h) - (A_i B_h)_0$, а величину $(A_i B_h)_0$ величиной $\frac{(A_i)(B_h)}{N}$, то коэффициент Чупрова получит следующее выражение:

$$T^2 = \frac{1}{\sqrt{(k-1)(l-1)}} \sum_h \sum_i \frac{[N(A_i B_h) - (A_i)(B_h)]^2}{(A_i)(B_h)}.$$

Для вычислений может быть использована как та, так и другая формула.

Коэффициент Чупрова свободен от того недостатка, которым обладает коэффициент Пирсона. Коэффициент T всегда лежит в пределах от 0 до 1. Его нулевое значение является необходимым и достаточным условием полной независимости признаков A и B, а его значение, равное единице, наоборот, является необходимым и достаточным условием того, что между признаками A и признаками B имеется полная взаимосвязь. Коэффициент Чупрова пригоден для измерения связи не только между качественными, но и между количественными признаками (для этого вместо A_i и B_h достаточно взять количественное выражение признаков, например x_i и y_i.

Коэффициент Чупрова следующим образом связан с коэффициентом Пирсона:

$$T^2 = \frac{C^2}{(1 + C^2)\sqrt{(k-1)(l-1)}}.$$

Произведение $(k-1)(l-1)$, входящее в приведенные выше формулы, является числом степеней свободы δ_{ih}.

Средняя квадратическая ошибка коэффициента сопряженности, т. е. величины T, по исследованию А. А. Чупрова (1926, стр. 89), имеет очень сложное, хотя и приближенное выражение. По мнению В. И. Романовского, этим выражением почти невозможно воспользоваться на практике.

Среднюю квадратическую ошибку величины φ^2 в приближенном значении определил К. Пирсон. Он задался условием, что изучаемая совокупность индивидов с признаками A и B является случайной выборкой S' из общей совокупности S, построенной по нормальному закону. В этом случае средняя квадратическая ошибка определяется по формуле

$$\sigma_{\varphi^2} = \frac{2}{\sqrt{N}}\sqrt{\varphi^2(1 + \varphi^2)}.$$

Если же закон распределения общей совокупности нам неизвестен, то

$$\sigma_{\varphi^2} = \frac{2}{\sqrt{N}}\sqrt{\psi + \varphi^2 - \varphi^4},$$

где

$$\psi^3 = \sum_h \sum_i \frac{[N(A_iB_h) - (A_i)(B_h)]^3}{(A_i)(B_h)}.$$

Кроме сложности и громоздкости вычислений коэффициенты C и T имеют еще два серьезных недостатка.

Первый недостаток, общий для всех среднеквадратичных величин, заключается в преувеличении влияния крайних отклонений. В результате этого преувеличения получается неравномерное изменение коэффициентов в связи с изменением измеряемой величины.

Второй недостаток заключается в отсутствии взвешивания величины отношения $\frac{\delta_{ih}^2}{A_iB_h}$ по групповым объемам, т. е. по величине (AB), а это приводит к искажению результата. В формуле Чупрова отношение $\frac{\delta_{ih}^2}{A_iB_h}$ присутствует в скрытом виде, вместо δ_{ih}^2 здесь взято равнозначное выражение

$$[N(A_iB_h) - (A_i)(B_h)]^2.$$

Рассмотрим пример, иллюстрирующий процедуру проверки гипотезы об отсутствии зависимости между двумя качественными признаками.

В табл. 129 приведено число случаев различных аварий на буровых скважинах, остановленных в различных породах.

Таблица 129

Породы, A	Аварии, B				
	Падение инструмента на забой B_1	Потеря промывочной жидкости B_2	Обрыв штанг B_3	Обвал в скважине B_4	(A_i)
Сланец A_1	6	10	11	17	44
Туфоген A_2	5	19	68	63	155
Кератофир A_3 . . .	54	76	40	29	199
Гранит A_4	12	17	5	2	36
(B_h) . . .	77	122	124	111	434

В основе всех трех коэффициентов, которые нам предстоит вычислить для этого примера, т. е. в основе C и T лежит сравнение фактического, заданного в этой таблице, распределения числа аварий с неким идеальным распределением, построенным в предположении, что признаки A и B независимы.

Вычислить теоретические значения групповых объемов $(A_iB_h)_0$ можно, исходя из следующего выражения:

$$(A_iB_h)_0 = \frac{(A_i)\,(B_h)}{N},$$

где N — общее число наблюдений.

Полученные в результате значения $(A_iB_h)_0$ будут вычислены в предположении о независимости признаков.

Вычисляем теоретические объемы групп, отвечающие этому условию, и располагаем полученные результаты в табл. 130.

Таблица 130

A	B				
	B_1	B_2	B_3	B_4	A_i
A_1	7,81	12,37	12,57	11,25	44
A_2	27,55	43,61	44,28	39,56	155
A_3	35,26	55,90	56,87	50,97	199
A_4	6,38	10,12	10,28	9,22	36
(B_h)	77	122	124	111	434

Сущность операции определения связи признаков заключается в сравнении таблиц 129 и 130. Таблица 129 — реальное распределение признаков, таблица 130 — условное (теоретическое) распределение, вычисленное исходя из предположения о полной независимости изучаемых признаков.

Определение и измерение связи делается по-разному.

По Пирсону, определяем сначала χ^2, для чего вычисляем δ_{ih}, возводим его в квадрат, делим на $(A_iB_h)_0$ и частное от деления дважды суммируем (сначала по строкам, затем эти суммы складываем по всей таблице). После этого величину χ^2 делим на N, получая при этом φ^2. Последней операцией будет вычисление C.

Проделаем все это с конкретными числами.

Вычисляем δ_{ih}. Для этого надо из каждого элемента табл. 129 вычесть соответствующий элемент табл. 130 в соответствующих клетках. Результат вычитания записываем следующим образом (табл. 131).

Таблица 131

A	B				
	B_1	B_2	B_3	B_4	$\sum_i \lvert \delta_{ik} \rvert$
A_1	—1,81	—2,37	—1,57	5,75	11,50
A_2	—22,55	—24,61	23,72	23,44	94,32
A_3	18,74	20,10	—16,87	—21,97	77,68
A_4	5,62	6,88	—5,28	—7,22	25,00

Каждое из чисел, проставленных в табл. 131, возведем в квадрат, а результаты, т. е. величину δ_{ih}, запишем в табл. 132.

Таблица 132

A	B			
	B_1	B_2	B_3	B_4
A_1	3,28	5,62	2,46	33,06
A_2	508,50	605,65	562,64	549,43
A_3	351,19	404,01	284,60	482,68
A_4	31,58	47,33	27,88	52,13

Данные табл. 132 делим на $(A_iB_k)_0$, т. е. на данные, проставленные в табл. 130, а частное от этого деления запишем в табл. 133 и подсчитаем в ней итоги, в результате чего получим величину χ^2.

Таблица 133

A	B				
	B_1	B_2	B_3	B_4	Сумма в строке
A_1	0,420	0,454	0,196	2,939	4,009
A_2	18,457	13,888	12,706	13,889	58,940
A_3	9,960	7,227	5,004	9,470	31,661
A_4	4,950	4,677	2,654	5,654	17,935
Сумма в столбце	33,787	26,246	20,560	31,952	$\chi^2 = 112{,}545$

Допустимое значение χ^2 при уровне значимости 0,001 и 25 степенях свободы равно 52,62. Так как вычисленное значение $\chi^2 = 112{,}54$ значительно превышает 52,62, то сопряженность признаков следует считать доказанной (приложение 14).

Далее вычисляем φ^2:

$$\varphi^2 = \frac{112{,}545}{434} = 0{,}259.$$

И, наконец, находим коэффициент Пирсона

$$C = \sqrt{\frac{0{,}259}{1 + 0{,}259}} = 0{,}453.$$

Вычислим коэффициент Чупрова

$$T^2 = \frac{0{,}259}{3} = 0{,}0863,$$

откуда

$$T = \sqrt{0{,}0863} = 0{,}294.$$

На основании полученных результатов можно сделать вывод, что тип аварии зависит от типа пород.

Выше мы видели, как велико значение группировки для статистического изучения качественных признаков. При этом мы всегда имели в виду такую группировку, когда на любой (отдельно взятой) ее ступени выделяются одни и те же признаки, например, как в A, так и в α выделяются B и β или в C_1, C_2, C_3 и т. д. выделяются D_1, D_2, D_3 и т. д.

Такая группировка называется гомогенной.

На практике гомогенная группировка встречается редко. Она есть в кристаллографии (группировка видов симметрии), бывает в подсчете запасов (разделение запасов по признаку категории запасов и по признаку категории промышленного назначения запасов), встречается в геохимии и т. д. Логически правильную гомогенную классификацию осадочных горных пород предложил Л. В. Пустовалов (1962).

Гораздо чаще на практике встречается другая группировка, при которой принципы классификации меняются для разных признаков на одной и той же ступени. Это — гетерогенная группировка. Хорошим примером ее может служить классификация геологических типов послемагматических рудных полей (Королев и Шехтман, 1959).

В основу последней положены следующие пять признаков.

1. Положение рудного поля по отношению к основным геотектоническим элементам (2 категории): A_1 — в подвижных поясах, A_2 — на платформах.

2. Положение рудного поля внутри основного геотектонического элемента (4 группы): B_1 — главным образом во внутренних частях геотектонического элемента, B_2 — главным образом в краевых частях геотектонического элемента, B_3 — во внутренних и краевых частях подвижных поясов, B_4 — на докембрийских щитах.

3. Характер толщи вмещающих пород (6 подгрупп): C_1 — мощные толщи пластичных дислоцированных осадочно-метаморфических пород, C_2 — перемежающиеся по составу осадочно-метаморфические и эффузивные толщи пород, C_3 — интрузивные и интрузивно-осадочные комплексы пород, C_4 — маломощные слабодислоцированные осадочные породы, C_5 — перемещающиеся по составу осадочно-эффузивно-интрузивные формации слабодислоцированных пород, C_6 — докембрийские интенсивно метаморфизованные и дислоцированные мощные толщи пород.

4. Формации вмещающих пород по их составу (10 классов): D_1 — флишоидная, D_2 — карбонатная, D_3 — молассовая, D_4 — карбонатно-терригенная, D_5 — эффузивно-осадочная, D_6 — эффузивная, D_7 — интрузивная (главным образом гранитоидная), D_8 — осадочно-эффузивно-интрузивная, главным образом в контактовых зонах, D_9 — осадочная, вмещающая интрузивные тела, D_{10} — кристаллические и метаморфические сланцы с различными изверженными породами.

5. Ведущий тип рудных тел (3 подкласса): E_1 — согласные, E_2 — секущие, E_3 — сложные.

Типы рудных полей, выделенные по сочетанию перечисленных пяти признаков, имеют подразделения по структурам и по преобладанию тех или иных металлов.

Всего выделено 26 типов рудных полей, а число признаков (с подразделениями) более 25. При этом в каждое сочетание входит по крайней мере 5 признаков.

Вся совокупность рудных полей делится на две категории: A_1 и A_2. Категория A_1 делится на три группы — B_1, B_2 и B_3. Категория A_2 делится тоже на другие три группы — B_1, B_2 и B_4. При этом каждому признаку B соответствует дополняющий его признак C по схеме: B_1—C_1, B_2—C_2, B_3—C_3, B_1—C_1, B_2—C_5, B_4—C_6.

Группы C делятся на классы D по схеме: группа C_1 — на D_1 и D_2; группа C_2 — на D_3, D_4, D_5 и D_6; группа C_3 — на D_7 и D_8. В группе C_4 имеется только один класс D_4; в группе C_5 — два класса: D_5 и D_9; в группе C_6 — один класс D_{10}.

Классы делятся на типы (а типы иногда на подтипы) рудных полей: D_1 на E_1 и E_2; в D_2 — только один тип E_3; D_3 делится на E_1, E_2 и E_3; в D_4 имеются E_1 и E_3; в D_5 — E_1, E_2 и E_3; в D_6 — E_1 и E_2 (тип E_2 делится на подтипы); в D_7 — E_1 и E_2 (есть подтипы); в D_8 — E_1 (с подтипами), E_2 (с подтипами), E_2 (без подтипов), E_3 (две разновидности); D_4 — один тип E_1; в D_5 — один тип E_2; в D_9 — E_1 и E_2; в D_{10} — E_1 (с подтипом) и E_2.

Гетерогенной является также палеонтологическая систематика, минералогическая классификация и другие группировки.

Гетерогенные группировки, как и гомогенные, нужны в геологической науке, но их статистическое изучение очень трудно.

Для того чтобы совокупность индивидов, разбитую на гетерогенные группы, можно было статистически изучить, необходимо эти группы превратить в гомогенные. Это делается путем уточнения и дробления одних признаков и объединения других.

С помощью теории качественных признаков можно объективно решать много различных задач в геологических исследованиях, например в области выявления связей:

1) между осветлением пород и скоплением нефти в нижележащих толщах. Это послужит поисковым признаком (связь данных признаков нашли Л. П. Задов и С. Я. Вайнбаум, 1952, но при этом они не пользовались методами статистики);

2) между вкусом воды в источнике, колодце или скважине и характером пород. Это позволит составить представление о путях миграции подземной воды;

3) между запахом породы и ее генезисом;

4) между формой алмазов и их генезисом (экспериментальное исследование связи этих признаков провел Вадило, 1961);

5) между характером тектонической складки и содержанием хлористого калия на месторождениях калийных солей;

6) между интенсивностью ртутного оруденения и глубиной складки, а также между интенсивностью того же оруденения и положением крыла складки;

7) между характером аварий на буровых скважинах и характером пород;

8) между успехом поисковых работ геолога и его квалификацией;

9) между характером дыма вулканов и силой последующего извержения;

10) между положением луны и солнца с одной стороны, и горными ударами в шахтах — с другой, по Хофер (Höfer, 1960) и Пилнэр (Pelnâr, 1960);

11) между характером пород и характером палеонтологических остатков и т. д.

ЛИТЕРАТУРА

Андреев С. Е., Зверевич В. В., Перов В. А. Дробление, измельчение, грохочение полезных ископаемых. Госгортехиздат, 1961.

Анциферов М. С., Константинова А. Г., Переверзев Л. Б. Сейсмоакустические исследования в шахтах. Изд. АН СССР, 1960.

Башаринов А. Е. и Флейшман Б. С. Методы статистического последовательного анализа и их радиотехнические приложения. Изд. «Сов. радио», 1962.

Беллман Р. Динамическое программирование. Изд. Ин. лит., 1960.

Билибин В. В. Методы математической статистики в подсчете подземных запасов нефти. Баку, 1930.

Блекуэлл Д. и Гиршик М. А. Теория игр и статистических решений. Изд. Ин. лит., 1958.

Бобриевич А. П., Бондаренко М. Н., Гневушев М. А., Красов А. М., Смирнов Г. И., Юркевич Р. К. Алмазные месторождения Якутии. Госгеолтехиздат, 1959.

Богацкий В. В. К вопросу о сокращении числа проб при разведке рудных месторождений. «Разведка недр», 1948, № 2.

Богацкий В. В. Оценка величины расхождения результатов разведочного опробования и экстраполяции. «Колыма», 1954, № 10.

Богацкий В. В. Влияние количества и размеров проб на точность результатов разведки полезных ископаемых. Сб. «Методика опробования рудных месторождений при разведке и эксплуатации». Свердловск, 1960.

Богацкий В. В. Оценка достоверности запасов полезного ископаемого и количество наблюдений, необходимых при разведочных работах. Тр. Сиб. научно-исследовательского ин-та геол. и геофиз. мин. сырья, 1962.

Богацкий В. В. Аналитические исследования результатов детальной разведки шахтных полей Назаровского буроугольного и Черногорского каменноугольного месторождений. Сб. «Материалы по методике разведки полезных ископаемых». Госгеолтехиздат, 1962.

Богацкий В. В. Влияние количества и размеров проб на точность результатов разведки полезных ископаемых. Сб. «Вопросы методики опробования рудных месторождений при разведке и эксплуатации». Госгеолтехиздат, 1962.

Богацкий В. В. Математический анализ разведочной сети. Госгеолтехиздат, 1963.

Боярский А. Я., Старовский В. Н., Хотимский В. И., Ястремский Б. С. Теория математической статистики. Планхозгиз, 1930.

Браунли К. А. Статистические исследования в производстве. Изд. Ин. лит., 1949.

Бухарцев В. П., Скороспелова Т. П., Строева Е. А. и Устинова З. С. К морфологии литофациального замещения в среднем девоне востока Русской платформы. ДАН СССР, т. 139, № 5, 1961.

Бухарцев В. П., Скороспелова Т. П., Строева Е. А. К методике количественного измерения несоответствия структурных планов. Новости нефт. и газ. техн. «Геология», 1961, № 11.

Бухарцев В. П., Мирчинк М. Ф. К методике геолого-статистического анализа локальных структур. Сб. «Опыт применения математической статистики при изучении локальных структур Волго-Уральской нефтегазоносной области». Изд. ЦНИИ ИТЭИ Нефтегаз, 1962.

Бухарцев В. П., Строева Е. А. Анализ морфологии и истории формирования Туймазинской структуры. Сб. «Опыт применения математической статистики при изучении локальных структур Волго-Уральской нефтегазоносной области». Изд. ЦНИИ ИТЭИ Нефтегаз, 1962.

Вадило П. С. Габитус кристаллов алмаза как отражение условий их образования. Зап. Всес. мин. о-ва, т. 90, № 2, 1961.

Вальд А. Последовательный анализ. Физматгиз, 1960.

Ван дер-Варден. Математическая статистика. Физматгиз, 1960.

Вентцель Е. С. Теория вероятностей. Изд. II. Физматгиз, 1962.

Вистелиус А. Б. Заметки по аналитической геологии. ДАН СССР, т. 44, № 1, 1944.

Вистелиус А. Б. Распределение частот коэффициентов пористости и эпигенетические процессы в спириферовых слоях Бугурусланского района. ДАН СССР, т. 49, № 1, 1945.

Вистелиус А. Б. О выражении результатов фоссилизации колебательных движений земной коры с помощью ряда. ДАН СССР, т. 49, № 7, 1945.

Вистелиус А. Б. Ритмы пористости и явление фазовой дифференциации осадочных толщ. ДАН СССР, т. 54, № 6, 1946.

Вистелиус А. Б. О корреляции мезоритмов в нижнепермских отложениях Закамской Татарии и их стратиграфическом значении. ДАН СССР, т. 55, № 3, 1947.

Вистелиус А. Б. О корреляционной связи между апатитом и нефелином в Кукисвумчорр-Юкспорском сфеновом месторождении (Хибинские тундры). ДАН СССР, т. 56, № 2, 1947.

Вистелиус А. Б. Новое подтверждение наблюдений Гольдшмидта о положении германия в каменных углях. ДАН СССР, т. 58, № 7, 1947.

Вистелиус А. Б. Мера связи между членами парагенезиса и методы ее изучения. Зап. Всес. мин. о-ва, ч. 77, вып. 2, 1948.

Вистелиус А. Б. Простейшие задачи математической обработки в литологии и пути их решения. Литологический сборник (ВНИГРИ), вып. 1, 1948.

Вистелиус А. Б. К геологии нижнеказанских отложений Бугурусланского района. «Советская геология», 1948, сб. 28.

Вистелиус А. Б. О некоторых аналитических методах исследования ритмичности. «Советская геология», 1948, сб. 28.

Вистелиус А. Б. О распространении магнезита в палеозое востока Русской платформы. Литологический сборник (ВНИГРИ), вып. 2, 1948.

Вистелиус А. Б. Сульфаты кальция в палеозойских отложениях востока Русской платформы. Тр. ВНИГРИ, вып. 28 — геохим., сб. 1, 1949.

Вистелиус А. Б. Пористость и химический состав карбонатных толщ палеозоя Поволжья и Заволжья. Тр. Лабор. гидрогеол. проблем АН СССР, т. 2, 1949.

Вистелиус А. Б. К вопросу о механизме слоеобразования. ДАН СССР, т. 65, № 2, 1949.

Вистелиус А. Б. К вопросу о механизме связи при слоеобразовании. ДАН СССР, т. 65, № 4, 1949.

Вистелиус А. Б. О минеральном составе тяжелой части песков нижнего отдела продуктивной толщи Апшеронского полуострова, Чокрака, южного Дагестана и аллювия Волги. ДАН СССР, т. 71, № 2, 1950.

Вистелиус А. Б. К вопросу о связи между содержанием меди в буровых водах Азербайджана и степенью их минерализации. ДАН Азерб. ССР, т. 6, № 1, 1950.

Вистелиус А. Б. О распространенности энантиоморфных типов кварца. Зап. Всес. мин. о-ва, ч. 79, вып. 3, 1950.

Вистелиус А. Б. К вопросу о палеогеографическом значении связи между мощностями слоев (на материале продуктивной толщи Апшеронского полуострова). Литологический сборник (ВНИГРИ), вып. 3, 1950.

Вистелиус А. Б. О состоянии обработки литологических наблюдений и мерах ее улучшения. «Изв. АН СССР, серия геол.», 1951, № 3.

Вистелиус А. Б. О необходимом числе зерен, подсчитываемых при иммерсии. Зап. Всес. мин. о-ва, ч. 80, вып. 3, 1951.

Вистелиус А. Б. Ритмы пористости в нижнеказанских отложениях Южной Татарии. Тр. Ленингр. о-ва естествоиспытат., т. 68, вып. 2, 1951.

Вистелиус А. Б. К минералогии песчано-алевритовых отложений миоцена юга Азербайджана. ДАН СССР, т. 85, № 5, 1952.

Вистелиус А. Б. Об обработке микроструктурных диаграмм. Зап. Всес. мин. о-ва, ч. 82, № 4, 1953.

Вистелиус А. Б. Проблема изучения связи в минералогии и петрографии. Зап. Всес. мин. о-ва, ч. 85, вып. 1, 1956.

Вистелиус А. Б. Расчленение немых толщ по количественным минералогическим, петрографическим или химическим признакам. Зап. Всес. мин. о-ва, № 1, 1957.

Вистелиус А. Б. К статистике микроструктурных диаграмм. Зап. Всес. мин. о-ва, ч. 86, № 6, 1957.

Вистелиус А. Б. Морфометрия обломочных частиц. Тр. лабор. аэрометодов АН СССР, 9, 1960.

Вистелиус А. Б. Фосфор в гранитоидах Центрального Тянь-Шаня. Геохим., № 2, 1962.

Вистелиус А. Б. О функциях распределения вероятностей концентрации фосфора в гранитоидах Швейцарии, Гвианы и Экваториальной Африки. ДАН СССР, т. 152, № 6, 1963.

Вистелиус А. Б. Проблемы математической геологии. «Геолог. и геофиз.» № 7, 1963.

Вистелиус А. Б., Белоусова В. Т. О применении коэффициента корреляции при исслед. парагенезисов минералов в терригенных отложениях. ДАН СССР, т. 55, № 4, 1947.

Вистелиус А. Б., Зульфугарлы Д. И. Естественные парагенезисы некоторых компонентов нефтей Азербайджана. «Изв. АН СССР», 1952, № 1.

Вистелиус А. Б., Сарманов О. В. Стохастическое обоснование одного геологически важного распределения вероятностей. ДАН СССР, т. 58, № 4, 1947.

Вистелиус А. Б., Сарманов О. В. Замечания по статье проф. П. А. Рыжова «Об оценке точности подсчета запасов месторождений полезных ископаемых». Исследования по вопросам маркшейдерного дела. Сб. 29, 1954.

Вистелиус А. Б., Яновская Т. Б. Программирование задач геологии и геохимии при использовании универсальных электронных вычислительных машин. Геология рудных месторождений, т. 5, № 3, 1963.

Владимирский В. И. К вопросу о нормальном ряде водоподъемного оборудования для опытных откачек. «Разведка и охрана недр», 1958, № 11.

Гихман И. И., Гнеденко Б. В., Смирнов Н. В. Непараметрические методы статистики. Тр. 3-го Всес. матем. съезда, 1956, т. 3. Изд. АН СССР, 1958.

Гнеденко Б. В. Последовательный анализ. Тр. 2-го Всесоюзного совещания по математической статистике. Ташкент, 1942.

Гнеденко Б. В. Курс теории вероятностей. Изд. II. Гостехиздат, 1954.

Гудков В. М. О применимости формул математической статистики при оценке результатов разведки. Сборник науч. тр. Моск. горн. ин-та, № 25, 1959.

Дементьев Л. Ф. О возможности использования горной геометрии при решении задач разработки. Тр. ВНИИ, вып. 14, Гостоптехиздат, 1958.

Дементьев Л. Ф. К вопросу о точности карт изопахит. «Татарская нефть», 1959, № 3—4.

Дементьев Л. Ф. Применение математической статистики к оценке результатов разведки. Тр. ВНИИ, в. 24. Гостоптехиздат, 1959.

Дементьев Л. Ф. Применение математической статистики при обобщении результатов изучения геологического строения продуктивных горизонтов нефтяных месторождений платформенного типа. Сб. «Вопросы разработки нефтяных месторождений и добычи нефти». Башкнигоиздат, 1960.

Дементьев Л. Ф. Применение математической статистики и теории вероятностей к оценке результатов разведки. Тр. ВНИИ, вып. 23. Гостоптехиздат, 1960.

Дементьев Л. Ф. Выбор рациональной величины сечения между изолиниями. «Татарская нефть», 1960, № 10.

Дементьев Л. Ф., Азаматов В. И. Об определении средней пористости и нефтенасыщенности при подсчете запасов и проектировании разработки. Новости нефт. и газ. техн. «Геология», 1961, № 1.

Доборжинский С. Ю. Некоторые общие правила ведения разведочных работ. Казань, 1908.

Доборжинский С. Ю. Некоторые общие правила разведочных работ. Изв. Томск. технолог. ин-та, т. 14, № 2, 1909.

Доборжинский С. Ю. Разработка разведочного материала. Горные и золотопромышленные известия, № 4—23, 1911.

Доборжинский С. Ю. Данные для оценки металлических руд и некоторых других ископаемых. Горные и золотопромышленные известия, № 9—15, 1912.

Дунин-Барковский И. В. и Смирнов П. В. Теория вероятностей и математическая статистика в технике (общая часть). Гостехиздат, 1955.

Задов Л. П., Вайнбаум С. Я. Цветность пород как нефтепоисковый признак. «Нефтяное хозяйство», 1952, № 8.

Захариев Е. Изчисляване на средното съдържание на компонентите с оглед изчисляването на запасите. «Минно дело», 1952, № 10.

Захариев Е. Изчисляване на коефициентите свързани с опробоването и с точността на химическия анализ. «Минно дело», 1952, № 12.

Захариев Е. Определяне на някои физически свойства на полезното «Минно дело», 1952, № 11.

Каждан А. Б. и Соловьев Н. Н. К методике определения бортового содержания при подсчете запасов полезных ископаемых. «Разведка и охрана недр», 1958, № 12.

Казаковский Д. А. К вопросу о влиянии интервала между скважинами и способа их расположения на величину ошибки в подсчете объема тела полезных ископаемых. Тр. ВНИМИ, сб. 6, 1937.

Казаковский Д. А. Оценка ошибки аналогии и ошибки оконтуривания при подсчете запасов полезного ископаемого. Зап. ЛГИ, т. 14, 1941.

Казаковский Д. А. Оценка точности результатов в связи с геометризацией и подсчетом запасов месторождений. Углетехиздат, 1948.

Казаковский Д. А. К вопросу оценки ошибок аналогии при подсчете запасов месторождений. Исследования по вопросам маркшейдерского дела. Сб. 24. Тр. ВНИМИ, 1951.

Казаковский Д. А. Методика оценки точности подсчета запасов полезных ископаемых. Зап. ЛГИ, т. 26, вып. 1, 1952.

Казаковский Д. А. О приложении статистики к оценке точности подсчета запасов месторождений. Исследования по вопросам маркшейдерского дела. Сб. 29, Углетехиздат, 1954.

Казаковский Д. А. О применимости формул статистики к оценке точности подсчета запасов месторождений. Исследования по вопросам горного и маркшейдерского дела. Сб. 31, 1957.

Казаковский Д. А. О характеристике изменчивости залежей полезного ископаемого. «Колыма», 1959, № 5.

Казаковский Д. А. К вопросу определения среднего значения вторых разностей при оценке относительной изменчивости характеристик залежи. «Изв. вузов. Горный журнал», 1960, № 4.

Каллистов П. Л. Учет высоких проб и самородков при подсчете запасов месторождений золота. Изд. ОБТИ Главспеццветмет, 1952.

Колмогоров А. Н. Решение одной задачи из теории вероятности, связанной с вопросом о механизме слообразования. ДАН СССР, нов. серия, т. 65, № 6, 1949.

Конюс А. А. Исследование функциональных связей методом корреляционных уравнений. Тр. 2-го Всес. совещ. по математической статистике. Изд. АН УзССР, Ташкент, 1949.

Корнфельд М. К. К теории ошибок. ДАН СССР, т. 103, 1955.

Королев А. В., Шехтман П. А. Классификация послемагматических рудных полей. Закономерности размещения полезных ископаемых. Сб. 2, АН СССР, 1959.

Королева А. П., Шарапов И. П. О современном поднятии соляного купола Ходжа-Икан и о возможности аналитического определения скорости этого поднятия. «Проблемы советской геологии», 1936, № 12.

Королева А. П., Шарапов И. П. Находка оптического гипса в Ходжа-Икане. «Минеральное сырье», 1937, № 9.

Королева А. П., Шарапов И. П. Месторождения оптического гипса. «Разведка недр», 1939, № 10—11.

Королева А. П., Шарапов И. П. Открытие месторождений оптического гипса в Южном Узбекистане. «Соц. наука и техн.», 1939, № 11—12.

Королева А. П., Шарапов И. П. Месторождения оптического гипса Южного Узбекистана. «Советская геология», 1940, № 7.

Косыгин М. К. Методика разведки. Сб. «Ангаро-Илимские железные руды». Госгеолтехиздат, 1960.

Крамер Гаральд. Математические методы статистики. Изд. Ин. лит., 1948.

Кузьмин В. И. Методика эксплуатационной геометризации жильных золоторудных месторождений. Изд. Харьк. гос. Ун-та, 1952.

Кузьмин В. И. К вопросу о методике объединения проб. «Колыма», 1955, № 4.

Кузьмин В. И. Геометрия месторождений полезных ископаемых. Справочник маркшейдера. Металлургиздат, 1955.

Кузьмин В. И. К оценке ошибок аналогии средних значений показателей и запасов полезных ископаемых методами математической статистики. Науч. тр. Харьков. горн. о-ва, т. II, 1955.

Кузьмин В. И. Методика подсчета фактических запасов руды и металла в блоке. «Колыма», 1956, № 9.

Кузьмин В. И. О некоторых результатах экспериментальной оценки ошибок аналогии методами математической статистики. Сборник трудов ВНИМИ, вып. 31, 1957.

Кузьмин В. И. К подсчету оценки точности запасов, определяемых методом ближайшего района. Науч. Тр. Харьков. горн. ин-та, № 4, 1958.

Кузьмин В. И. К методике подсчета запасов жильных месторождений. «Колыма», 1960, № 4.

Кузьмин В. И. Вычисление среднего объемного веса по ограниченному числу определений. «Изв. вузов. Горный журнал», 1961, № 9.

Кузьмин В. И., Зарайский В. Н. К оценке ошибок аналогии запасов месторождений полезных ископаемых. Изв. вузов. «Горный журнал», 1961, № 2.

Купарадзе Г. К. Справочник экономиста. Изд. Груз. с.-х. ин-та, 1960.

Лукомский Я. И. Статистический анализ и контроль существенно положительных величин, характеризующих качество продукции. «Стандартизация», 1955, № 1 и 2.

Ляхович В. В., Родионов Д. А. К методике изучения акцессорных минералов в изверженных породах. Тр. ИМГРЭ, вып. 6, 1961.

Мещеряков Ю. П., Сетунская Л. Е. Приемы количественной характеристики взаимосвязей природных явлений по картам с помощью коэффициентов корреляции. «Изв. АН СССР, серия геогр.», 1960, № 1.

Мирчинк М. Ф., Бухарцев В. П. О возможности статистического исследования структурных соотношений. ДАН СССР, т. 126, № 5, 1959.

Митропольский А. К. Статистическое исчисление, т. 2. Изд. Всес. заочн. лесотех. ин-та, Л., 1952.

Наливкин Д. В. Группа Spirifer Anossofi Vern. и девон европейской части СССР. Зап. Российск. минералог. о-ва, ч. IV, вып. 2, 1925.

Налимов В. В. Применение математической статистики при анализе вещества. Физматгиз, 1960.

Обухов В. М. К вопросу о нахождении уравнения регрессии, удовлетворяющего эмпирическому статистическому ряду. Тр. ЦСУ, т. 16, вып. 2, 1923.

Огарков В. С. Новый способ решений основных задач методики разведки. Изд. Тульск. гор. ин-та, Тула, 1960.

О'Рурк А. Н. Таблицы умножения. Гостехиздат, 1953.

Перегудов Н. В. Теоретические вопросы индексного анализа. Госстатиздат, М., 1960.

Погребицкий Е. О. Месторождения ископаемых углей и их запасы. «Геология СССР», т. VII, Донбасс. Госгеолиздат, 1944.

Померанцев В. В. Промышленные условия для подсчета запасов месторождений цветных металлов. Изд. науч. техн. о-ва цветн. мет., 1957.

Прокофьев А. П. Определение минимального промышленного (бортового) содержания полезного компонента в пробах при подсчете запасов. «Разведка недр», 1950, № 2.

Псарев Н. Приложения теории вероятностей к вычислениям при разведках на золото. Вестн. золотопромышленности и горного дела вообще. № 15, 1 авг., 1899.

Пустовалов Л. В. Об основных принципах классификации осадочных горных пород. Уч. зап. ЛГУ, серия геол., вып. 12, 1962.

Раевский В. И., Шуруборю. В. Обработка данных внешнего контроля химических анализов геологических проб. «Изв. вузов, геолог. и разведка», 1958, № 11.

Разумовский Н. К. Механический состав россыпного золота и новые данные по методике подсчета запасов россыпей. «Сов. золотопромышленность», 1939, № 12.

Разумовский Н. К. Характер распределения содержаний металлов в рудных месторождениях. ДАН СССР, т. 28, № 9, 1940.

Разумовский Н. К. О значении логарифмически нормального закона распределения частот в петрологии и геохимии. ДАН СССР, т. 33, № 1, 1941.

Разумовский Н. К. Логарифмически нормальный закон распределения вещества и его свойства. Зап. ЛГИ, т. 20, 1948.

Разумовский Н. К. К вопросу о выделении аномалий на фоне обычных содержаний элементов в породе при поисковых работах. «Развед. геофизика», вып. 1, 1962.

Разумовский Н. К. Средняя арифметическая выборки и ее связи с логарифмическими моментами. Сб. «Вопр. развед. геофизики», вып. 1, 1962.

Родионов Д. А. Статистический закон распределения элементов на примере гранитоидов Западного Алтая. Сб. «Тезисы докладов на конференции молодых научных сотрудников ИМГРЭ», 1959.

Родионов Д. А. К вопросу о функциях распределения содержаний элементов в изверженных горных породах. ДАН СССР, т. 141, № 3, 1961.

Родионов Д. А. К вопросу о логарифмическом нормальном распределении содержаний элементов в изверженных горных породах, «Геохимия», 1961, № 4.

Родионов Д. А. О виде функций распределения содержаний акцессорных минералов в изверженных горных породах. Тр. ИМГРЭ АН СССР, вып. 6, 1961.

Родионов Д. А. Об оценивании среднего содержания и дисперсии логнормально распределенных компонентов в породах и рудах. «Геохимия», 1962, № 7.

Родионов Д. А. Задача сопоставления средних содержаний логнормально распределенных компонентов в породах. «Геохимия», 1962, № 8.

Родионов Д. А. Задача о корреляции содержаний компонентов в породах и минералах в условиях логнормального распределения. Тр. IV конфер. молод. научн. сотрудников ИМГРЭ, 1962.

Родионов Д. А. Трехпараметрические распределения содержаний элементов в породах. «Геохимия», 1963, № 2.

Родионов Д. А. Особенности распределения среднего арифметического в условиях асимметричных распределений содержаний. «Геохимия», 1963, № 7.

Родионов Д. А. Применение математической статистики для обоснования некоторых петрографических и геохимических выводов. «Советская геология», 1963, № 1.

Родионов Д. А., Ляхович В. В. О статистическом изучении пространственного распределения содержаний акцессорных минералов в гранитах Эльджуртинского массива. ДАН СССР, т. 134, № 5, 1960.

Родионов Д. А., Соболев С. Ф., Золотарев Б. П., Власова Е. В. О случайных погрешностях количественно-минералогического анализа шлихов и концентратов. Тр. ИМГРЭ АН СССР, вып. 4, 1960.

Романовский В. И. Математическая статистика. ГОНТИ НКТП СССР, М.—Л., 1938.

Романовский В. И. Применения математической статистики в опытном деле. Гостехиздат, М.—Л., 1947.

Рыжов П. А. Математические определения геологической ошибки («ошибки аналогий») при подсчете запасов полезных ископаемых. Тр. Каз. горнометаллург. ин-та, вып. 1. Алма-Ата, 1938.

Рыжов П. А. Математическая оценка точности подсчета запасов полезных ископаемых методом изолиний. Сборник статей по вопросам маркшейдерско-геологического обслуживания горнорудных предприятий цветных и редких металлов, кн. 1, 1940.

Рыжов П. А. Учет потерь полезного ископаемого на рудниках цветных металлов. Изд. Каз. фил. АН СССР, 1943.

Рыжов П. А. Геометрия недр. Углетехиздат, 1952.

Рыжов П. А. Об оценке точности подсчета запасов месторождений полезных ископаемых. Сборник ВНИМИ, № 28. Углетехиздат, 1954.

Рыжов П. А. О применении математической статистики к оценке точности подсчета запасов месторождений. Исследования по вопросам горного и маркшейдерского дела. Сб. 31, 1957.

Рыжов П. А., Гудков В. М. Некоторые особенности распределения металлов. «Изв. вузов, Горн. журнал», 1961, № 10.

Скрипиль В. И. Опыт применения метода корреляции для подсчета запасов элемента примеси. Информ. бюлл. № 1. Южно-Уральск. ГУ и НТГО. Уфа, 1957.

Слуцкий Е. Е. Теория корреляции и элементы учения о кривых распределения. Изв. Киевск. коммерч. ин-та. Киев, 1912.

Смирнов Н. В. Оценка расхождения между эмпирическими кривыми распределениями в двух независимых выборах. Бюлл. МГУ. Математика, т. II, вып. 2, 1933.

Смирнов Н. В., Дунин-Барковский И. В. Краткий курс математической статистики для технических приложений. Физматгиз, 1959.

Соловьев В. Г. К методике опробования и сортировки бокситов Тихвинского района. Тр. ВГРО, вып. 367. ОНТИ, 1934.

Соловьев В. Г. К методике определения степени разведанности месторождений полезных ископаемых. «Разведка недр», 1937, № 3.

Соловьев В. Г. Вариационная статистика в применении к разведке и подсчету запасов полезного ископаемого. «Разведка недр», 1938, № 1.

Соловьев В. Г. Методы вариационной статистики в приложении к разведке и подсчету запасов месторождений полезных ископаемых. Тр. ЦНИГРИ, вып. 115, 1939.

Соловьев В. Г. Об общих принципах методики разведки на примерах некоторых типов оловорудных месторождений. Мат-лы ВСЕГЕИ, сб. 3, 1946.

Соловьев В. Г. О положении с разработкой вопросов методики разведки. «Разведка недр», 1952, № 3.

Срезневский Б. И. Решение вопроса о корреляции двух переменных и метод равных повторяемостей. Метеорол. вести. № 3, 1914.

Срезневский В. И. Мои тезисы (единое решение корреляции и метод равных повторяемостей). Зап. Моск. метеорол. о-ва, вып. 3, 1928.

Фишер Ирвин. Построение индексов, 1922.

Хальд А. Математическая статистика с техническими приложениями. Изд-во ин. лит., 1956.

Черемшанцев Я. Н. Производственный травматизм и профессиональные заболевания, их предупреждение и профилактика. Госгортехиздат, 1961.

Четвериков Л. И. О принципах применения теории вероятностей для анализа развед. данных. «Изв. вузов, геология и разведка», 1962, № 9.

Чирвинский П. Н. Количественный минералогический и химический состав гранитов и грейзенов. Изд. Донского политехн. ин-та, 1911.

Чирвинский П. Н. Палласиты (краткое резюме большой работы). Приложение к Изв. Донского политехн. ин-та, т. 4, отд. 2, 1918.

Чирвинский П. Н. Геометро-химический анализ. ОНТИ, химтеорет. изд., 1937.

Чирвинский П. Н. Кларки комплекса магматических пород Восточно-Европейского щита и их космический смысл. Зап. Всес. мин. о-ва, ч. 70, № 1, 1941.

Чирвинский П. Н. Кларки магматических пород щитов и геосинклиналей. Сборник к 70-летию акад. Д. С. Белянкина. Изд. АН СССР, 1946.

Чирвинский П. Н. Средний количественно-минералогический состав главных пород габбрового массива южнее р. Баранчи на Урале. Зап. Всес. мин. о-ва, 2-я серия, ч. 80, 1952.

Чирвинский П. Н. Средний химический состав главных минералов изверженных, метаморфических и осадочных пород. Изд. Харьк. ун-та, 1953.

Чирвинский П. Н. Методика получения количественной характеристики агрегатов. Минер. сб. Львовск. геол. о-ва № 9, 1955.

Чупров А. А. Основные проблемы теории корреляции. Л., 1926.

Шаманский Л. И. К изучению структуры молибдено-медного месторождения Каялах-Узень. Вестн. Зап.-Сиб. геол. треста, вып. 5, 1935.

Шаманский Л. И. Математическая обработка разведочных материалов. ГОНТИ, 1936.

Шаманский Л. И. Оруденение 2-й Урской линзы в свете математического анализа. Вестн. Зап.-Сиб. геол. разв. треста, 1937.

Шаманский Л. И. К вопросу о принципах сокращения проб. «Разведка недр», 1938.

Шаманский Л. И. Точность подсчета запасов полезного ископаемого. Сборник статей по маркшейдерии, петрогр. и геодезии. Иркутск, 1938.

Шаманский Л. И. Отклонившиеся пробы. Сборник трудов треста Золоторазведка и НИГРИзолото, вып. VIII, 1938.

Шаманский Л. И. Опробование штокверковых месторождений. Тр. Иркутск. горнометаллург. ин-та, 1948.

Шарапов И. П. Значение самородков для разведки золотых россыпей. Сб. «Новая Лена», 1947, № 1—2.

Шарапов И. П. Об определении изменчивости и выдержанности месторождений полезных ископаемых. «Разведка недр», 1952, № 3.

Шарапов И. П. Преодолеть отставание методики разведки. «Разведка недр», 1952, № 6.

Шарапов И. П. О контрольных анализах геологических проб. «Разведка и охрана недр», 1954, № 1.

Шарапов И. П. Графическое определение содержания золота в россыпи по пробе, взятой из буровой скважины. Тр. Донецк. инд. ин-та, вып. 1, 1954.

Шарапов И. П. Об удельном весе каменного угля Донбасса. Сб. «Маркшейдерное дело», вып. 4, 1956.

Шарапов И. П. Об ошибках в определении мощности угольных пластов по буровым скважинам. «Разведка и охрана недр», 1956, № 8.

Шарапов И. П. К теории подсчета запасов элементов-примесей. «Разведка и охрана недр», 1957, № 1.

Шарапов И. П. К теории выдержанности месторождений полезных ископаемых. Тр. Пермск. НИУИ, сб. 4, 1962.

Шарапов И. П. Исследование мощности угольных пластов Кизеловского бассейна с помощью законов распределения. Науч. тр. Пермск. НИУИ, 1964.

Эз В. В. Микротектоника угольных пластов и внезапные выбросы. Сборник трудов геофиз. ин-та № 34 (161), 1956.

Юл Д., Кендэл М. Теория статистики. Госстатиздат, 1960.

Ястремский Б. С. Математическая статистика. Гостехиздат, 1936.

Ahrens L. H. A fundamental law of Geochemistry. Nature, v. 172, No. 4390, 1953.

Ahrens L. H. Quantitative Spectrochemical Analysis of Silicates. Pergamon Press, London, 1954.

Ahrens L. H. The lognormal distribution of the elements, Part I, Geochim. et Cosmochim. Acta, v. 5, No. 2, 1954.

Ahrens L. H. The lognormal distribution of the elements. Part II, Geochim. et Cosmochim. Acta, v. 6, No. 2—3, 1954.

Ahrens L. H. Lognormal-type distribution — III. Geochim. et Cosmochim. Acta, v. 11, No 4, 1957.

Ahrens L. H. Negatively skewed distributions of silica and potassium in igneous rocks. Nature (Engl.), v. 198, No. 4878, 1963.

Ahrens L. H. Lognormal-type distributions in igneous rocks. IV, Geochim. et Cosmochim. Acta, v. 27, No. 4, 1963.

Aswathanarayana U. Some statistical aspects of the distribution of radioactivity in the Salem gneises of Madras State. Publ. Bur. Centr. Seism. Internnat. No. 19. 1956.

Bintig Karl-Heinz. Zur Theorie der Vorratsberechnung von Begleit-oder Spurenelementen. Zeitschr. angew. Geol., v. 5, No 12, 1959.

Bintig Karl-Heinz. Vorratsberechnung von Begleitelementen mit Hilfe der Korrelationsrechnung. Zeitschr. angew. Geol., No 6, 1960.

Bintig Karl-Heinz. Fehlertheorie und Rundungsintervall von Vorratsberechnungen. Fehlerzusammensetzung und Fehlereinfluss. Zeitschr. angew. Geol., v. 7, No 2, 1961.

Bintig Karl-Heinz. Die Ermittlung und Behandlung von Mammutgehalten bei Vorratsberechnungen. Ber. Geol. Ges. v. 7, No 1, 1962.

Burma, Benjamin H. Studies in quantitative paleontology. J. Paleontol. v. 22, No. 6.

Burma Benjamin H. Multivariate analysis — a new analytical tool for paleontology and geology. Paleontol. v. 23, No. 1, 1949.

Burma Benjamin H. Studies in quantitative paleontology, II, Multivariate analysis, J. Paleontol, v. 23, No 1, 1949.

Burma Benjamin H. An application of sequential analysis to the comparison of growth stages and growth series. J. Ceol., v. 61, No. 6, 1953.

Burma Benjamin H. Studies in Quantitative Paleontology, III. An Application of Sequential Analysis to the Comparison of Growth Stages and Growth Series. J. Geol. v. 161, No. 6, 1953.

Chayes Felix. Petrographic analyses by fragment counting. Econ. Geol., v. 39, No. 7, 1944.

Chayes, Felix. A simple point counter for thin-section analysis. Amer. Miner., v. 34, 1949.

Chayes, Felix. On ratio correlation in petrography. J. Geol., v. 57, No. 3, 1949.

Chayes, Felix. The theory of thin-section analysis. J. Geol., v. 62, No. 1, 1954.

Chayes, Felix. The longnormal distribution of elements: A Discussion. Geochim. et Cosmochim. Acta, v. 6, No. 2—3, 1954.

Chayes, Felix. Petrographic modal analysis. An elementary statistical appraisal. Y. Wiley & Sons, Inc. New York, XII, 1956.

Chayes, Felix. On correlation between Variables of Constant Sum. J. Geophys. Res., v. 65, No. 12, 1960.

Chayes, Felix. Numerical correlation and petrographic variation. J. Geol., v. & 0, No. 4, 1962.

Davies O. L. (Editor) Statistical methods in research and production, with special reference to the chemical industry. Third Edition (revised), Imp. chem. Ind., Ltd., Oliver & Boyd, London, 1957.

de Wijs H. J. Statistische methodes toegepast op de Schatting van erstreserven. Lustrumjaarboek Mijn-bouwk. Ver. te Delft., 1948.

de Wijs H. J. Statistics of ore distribution, J. Geol. en Mijnbouw, November, 1951.

de Wijs H. J. Statistics of ore distribution. Journal of the Roy. Netherlands Geol. & Min. Society, II, January, 1953.

de Wijs H. J. Die statistische Auswertung der Probenahme. Erzmetall. Zeitschr. für Etzbergbau und Metall hüttenwesen, Bd. 6, 1953.

Doborzynski, St. Niekore prawidla ogolne prowadzenia robot gornieczych wywiadowszych. Przegladzie Gornicko-Hutniczym, 1908.

Doborzynski, St. Przyczynki do teorii okreslania skladu zloz mineralow i niejednelitych mas w ogole. Prze—gladzie Gorniczo-Hutniczym, 1910.

Duval K. Note concernant l'echantillonnage. Ann. Mines, Paris, v. 138, No. 2, 1949.

Duval K. Application des notions de statistique mathematique en matiere de laveries. Congres des laveries des mines metalliques francaises, 1953. Rev. Ind. Miner., april, 1954.

Duval K. Contribution a l'etude de l'echantillonnage des gisements. Ann. Mines, No. 1, 1955.

Duval K. Contribution a l'etude de l'echantillonnage des gisements. Ann. Mines, No. XII, 1955.

Duval K. A propos de l'echantillonnage des gisements. Ann. Mines, No. XII, 1955.

Duval K. Controle de quanite et analyses statistiques en matieres de lavaries metalliques. Ann. Mines, No. 10, 1957.

Duval K. Etude graphique d'une distribution lognormale. Rev. Stat. Appl., v. VII, No. 1, 1959.

Duval K. Genese de certaines distributions dissymetriques. Rev. Stat. Appl., v. IX, No. 1, 1961.

Duval K. Nouvelle explication proposee de la dissymetrie constatee des distributions des frequences des tenures de gisements. Rev. Ind. miner., v. 6, No. 6, 1961.

Duval K. Deux questions souvent mal traitees dans les rapports sur les echantillonnages de gisement. Rev. Ind. miner., v. 44, No. 8, 1962.

Duval K., Levy, R. et Matheron G. Travaux de M. D. G. Krige sur l'evaluation des gisements dans les mines d'or sud-africaines. Ann. Mines, No. XII, 1955.

Eisenhart, Churchill. A test for the significance of lithological variations. J. Sed. Petrol., v. 5, 1935.

Gini, C. Une question importante pour la science de constitution et pour la medicine militaire; comment juger si les pronortitions d'un individu sont normales. Revue de l'Institut international de statistique., 1937, Juillet, Rev. antropol., No. 78, 1939.

Griffiths, J. C. Estimation of error in grain size analysis. J. Sed. Petrol., v. 23, 1953.

Griffiths, J. C. Statistics for the description of frequency distribution, generated by the measurement of petrographic properties of sediments. Compass of Sigma—Gamma—Epsilon, v. 31, No. 4, 1955.

Griffiths J. C. Statistical methods in sedimentary petrography. Milner H. B. Sedimentary petrography, 4th Edition, Chapter 16, New-York, 1957.

Griffiths J. C. Petrography and porosity K of the Cow Run Sand, St. Mary's West Virginia. J. Sediments. Petrogr., v. 28, No. 1, 1958.

Griffiths J. C. Geometrics in petroleum petrograph. Prod. Monthly, v. 22, No. 4, 1958.

Griffiths J. C. Size and Shape of Rock-fragments in Tuscarora Scree, Fishing creek, Lamar, Central Pennsylvania. J. Sediment. Petrogr., v. 29, No. 3, 1959.

Griffiths J. C. Modal analysys of sediments. Analyse quantitative statistique des sediments. Rev. geogr.-phys. et geol.-dynam., v. 3, No. 1, 1960.

Griffiths J. C. Aspects of measurement in the geosciences. Min. Ind. Bull., 29 (4).

Griffiths J. C. Measurement of properties of sediments. J. Geol., v. 69, No. 5, 1961.

Grubbs F. E. Simple criteria for testing outlying observations. Ann. Mathem. Stat., v. 21, No. 1, 1950

Gumbel E. J. Statistics of Extremes. Columbia University Press, New-York, London, Oxford University Press, 1958.

Höfer K. H. Gibt es eine Periodizität der Gebirgsschläge? Geologie und Bauwesen, 1960, Bd. 25, Nos. 2—3.

Holzinger K. J. & Harman H. H. Factor analysis. Synthesis of factorial methods. The University of Chicago Press, November, 1941. V—XII

Irwin I. O. On a criterion for the rejection of outlying observations. Biometrica, 17, Parts III & IV, December, 1925.

Jahns H. L'influence sur la precision de l'echantillonnage de la «tres grosse particule» dans la division en quarte des echantillons ainsi que celle des residus restant dans les appareils. Trad. No. AE 311, 1950, 56—45—129—7.

Jahns H. Die Genauigkeit der Probenahme von Kohlen und Erzsendungen. Glückauf, v. 88, H. 13—14, 1952.

Jahns H. Accuracy in Sampling Coal and Ore Shipments. Application of the Gauss System. Glückauf, 1952.

Jahns H. Principe de la classification des reserves du projet de normalisation. Glückauf, 1956, 92, 35—36, 1042—1047. Trad. 18—57.

Jahns H. Answahl der für die mathematische Behandlung geeigneten Verfahren der Pobenahme von Massengütern des Bergbaus. Bergbauarchiv, Bd. 18, No. 1, 1957.

Jahns H. Die Aussagesicherheit der Vorratsangaben von Lagerstätten. Teil II., Zeitschr. Erzbergbau und Metallhüttenwesen, 12, 7, 1959.

J a h n s H. Die Aussagesicherheit der Vorratsangaben von Lagerstätten, I Teil, Zeitschr. Erzbergbau und Metallhüttenwesen, 12, No. 5, 1959.

J a h n s H. Die statistische Auswetung von Porosi-Täts- und Permeabillitätsabmessungen. Erdöl und Kohle, Bd. 14, No. 2, 1961.

J o n e s H. E. Some Geometrical Considerations in the General Theory of Fitting Lines and Planes. Metrop., v. 13, No. 1, 1937.

K r i g e D. G. A statistical approach to some basic mine valuation problems on the Witwatersrand. J. Chem. Met. & Min. Soc. of South Africa, December, 1951. Discussions and replies, — March, 1952; May, 1952; July, 1952; August, 1952.

K r i g e D. G. Mine valuation and Statistics. The South African Min. and Engng J., v., LXIII, part 1, No. 3091, May, 1952.

K r i g e D. G. A statistical analysis of some of the borehole values in the Orange Free State goldfields. Chem. Met. and Min. Soc. of South Africa, September, 1952. Discussions and replies, November, 1952; February, 1953.

K r i g e D. G. Emploi statistique mathematique dans les problemes poses par l'evaluation des gisements. Ann. Mines, Dec., 1955.

K r i g e D. G. Analyse statistique des princiaux risques en relation avec l'investissement dans les nouvelles mine d'or d'Afrique du Sud. Tenicon, Oct., 1955.

K r i g e D. G. A study of the relationship between development values and recovery grades on the South African gold fields. J. of South Aftican Inst. Min. and Met., Janv., 1959. Discussions et reponses, Avr. et Oct. 1959.

K r i g e D. G. On the departure of ore value distributions from the lognormal model in South African Gold Mines. J. of South African Inst. Min. and Met., v. 61, No. 4, 1960.

K r i e g e D. G. On the departure of ore value distributions from the lognormal model in South African gold mines. J. of South African Inst. Min. and Met.,1961.

K r i e g e D. G. Developments in the valuation of gold mining properties from borehole results. Min. J. 1961, 256, No. 6563.

K r i e g e D. G. Statistical applications in mine valuation. J. Inst. Min. Surv. of S. A., v. XII, No. 2, 3, 1962.

K r i e g e D. G., N e c k e r m a n n H. J. Value contours and improved regression techniques for ore reserve valuations. J. of South African Inst. Min. and Met., May, 1963.

K r u m b e i n W. C. The probable Error of Sampling Sediments for mechanical analysis. Amer. J. Sci., v. 27, No. 159, 1934.

K r u m b e i n W. C. Size frequency distributions of sediments. J. Sed. Petrol. 4.

K r u m b e i n W. C. Thin-Section Mechanical Analysis of Indurated Sediments. J. Geol., v. XLIII, N. 5, 1935.

K r u m b e i n W. C. The use of quartile measures in describing and comparing sediments. Amer. J. Sci., 32: 98—111.

K r u m b e i n W. C. Sediments and exponential curves. J. Geol., v. 45, p. 577—601.

K r u m b e i n W. C. Size frequency distributions of Sediments and the Normal Phi Curve. J. Sed. Petrol., v. 8, No. 3, 1938.

K r u m b e i n W. C. Statistical problems of sample size and spacing on lake Michigan beaches. A book: «Coastal Engineering». Proceedings of the Fourth Conferenc, Chicago, 1953.

K r u m b e i n W. C. Latin square experiments in sedimentary Pet rology. J. Sed. Petrol., v. 25, No. 4.

K r u m b e i n W. C. Statistical designs for sampling beach sand. Am. Geophys. Union Transact., v. 34, 1953.

K r u m b e i n W. C. Application of Statistical Methods to Sedimentary Rocks. J. Amer. Statist. Assoc., v. 49, No. 265, 1954.

K r u m b e i n W. C. Statistical analysis of facies maps. J. Geol., v. 63, 1955.

K r u m b e i n W. C. Measurement and error in regional stratigraphic analysis. J. Sed. Petrol., v. 28, 1958.

K r u m b e i n W. C. The sorting out of geological variables illustrated by regression analysis etc. J. Sed. Petrol., v. 29, No. 4, 1959.

K r u m b e i n W. C. Trend surface analysis of contour type maps with irregular controlpoint spacing. J. Geophis. Res., v. 64, No. 7, 1959.

K r u m b e i n W. C. The sorting out of geological variables illustrated by regression analysis of factors controlling beach firmness. J. Sed. Petrol., v. 29, No. 4, 1959.

K r u m b e i n W. C. The geological population as a frame work for analysing numerical data in geology. Liverpool and Manchester Geol. J., v. 2, No. 3, 1960.

K r u m b e i n W. C. Some problems in applying statistics to geology. Appl. Statistics, v. 9, No. 2, 1960.

K r u m b e i n W. C. The computer in geology. Science, v. 136, No. 3522, 1962.

K r u m b e i n W. C. Open and closed number systems in stratigraphic mapping. Bull. Amer. Assoc. Petrol. Geol., v. 46, No. 12, 1962.

K r u m b e i n W. C. Some problems in applying statistics to Geology. Appl. Statist., v. 7, No. 2, 1962.

K r u m b e i n W. C. & Imbrie John. Stratigraphic factor maps. Bull. Amer. Assoc Petrol. Geol., v. 47, No. 4, 1963.

K r u m b e i n W. C. & Lieblein J. Geological application of extreme value methods to interpretation of cobbles and boulder in gravel deposits. Trans. Amer. Geophys. Union, v. 37, No. 313, 1956.

Krumbein W. C. & Miller R. L. Design of Experiments for Statistical Analysis of Geological data. J. Geol., v. 61, No. 6, 1953.

Krumbein W. C. & Tukey, J. W. Multivariate analysis of mineralogic, lithologic and chemical composition of rock bodies. J. Sed. Petrol., v. 26, No. 322, 1956.

Lombard Jean. L'actualité dans les grands problèmes de la géologie minière. (Coup d oeil sélectif sur notre «Chronique» de 1959). «Chronique mines outremer», v. 27, No. 282, 1959.

Lombard Jean. Die Beziehungen zwischen nachgewiesenen Vorratsmengen, Investitionsaufwand und ökonomischen Nureffect. Zeitschr. angew. Geol., Bd. 6, No. 1, 1960.

Matheron G. Application des méthodes statistiques à l'evaluation des gisements. Ann. Mines, No. XII, 1955.

Matheron G. Utilité des méthodes statistiques dans la recherche minière. Revue de l'Industrie Minérale, Janvier, No. Spec. 1—R, 1956.

Matheron G. Théorie lognormale de l'échantillonnage systématique des gisements. Ann. Mines, 1957.

Matheron G. Remarques sur la loi de Lasky. Chronique Mines Outre-mer, v. 27, No. 282, 1959.

Matheron G. Precision of exploring a stratified formation by boreholes with rigid spacing-application to a vauxite deposit. Intern. Symposium on Min. Res., v. 1, Pergamon Press, Oxford—London—New-York—Paris, 1962.

Matheron G. Traité de géostatistique appliquée. 1 (Mem. du Bureau de Récherches géologiques et min., 14) Paris, Eds. Technique, 1962.

Matheron G. & Formery Ph. Recherche d'optimum dans la reconnaissance et la mise en exploitation des gisements miniers. Ann. Mines, Mai, 1963.

Miller R. L. An application of the analysis of variance to paleontology. J. Paleontol., v. 23, No. 6, 1949.

Miller R. L. Introduction to special issues on statistics in geology. J. Geol., No. 16, 1954.

Miller R. L. An analysis of the interaction of quantitative variables in a modern environment of sedimentation. Bull. G. S. A., 65 : 12 part 2; 1285, 1954.

Miller R. L. Model for analysis of enviroments of sedimentation. J. Geol., v. 62, No. 1, 1954.

Miller R. L. Trend Surfaces; Their Application of Environments of Sediments. 1. The Relation of Sediment—Size Parameters to Current-Wave Systems and Physiography. J. Geol. v. 64, No. 5, 1956.

Miller R. L. & Goldberg E. D. The normal distribution in geochemistry. Geochim. et Cosmochim. Acta, v. 8, No. 1—2, 1955.

Miller R. L. & Kahn J. S. Statistical analysis in the geological sciences. Editors: John Wiley & Sons, Inc., New York—London, 1962. (Printed in USA).

Miller R. L. & Olsen E. C. A mathematical model applied to a study of the evolution of species. Evolution, v. 5, 1951.

Miller R. L. The statistical stability of quantitative properties as a fundamental criterion for the study of enviroments. J. Geol., v. 63, No. 4, 1955.

Murard R. Etude par sondages d'un gisement stratiforme. Comm. Cong. Cent. Ind. Min., 1955 juin; Rev. de l'ind. Min., Saint-Etienne, janv., 1956, No. Spec. 1—R.

Murard R. Probabilités et statistique. Rev. Industr. Minér., No. Spec. 3—122, Oct., 1960.

Neill A. A study of glare discomfort and disability from miners' cap lamps. Graham Sheffield University Min. Mag., 1959.

Oertel A. C. Frequency distribution of spectrographic error in d. c. arc exitation of soil samples. Australian J. Appl. Sci., v. 7, No. 2, 1956.

Pelnar Antonin. Statistica analysa pribromsych otresu. Sb. Ustavu pro vyzkum rud, 1958—1959 (1960), 3.

Scheidegger A. E. Statistical analysis of recent fault-plane solutions of earthquakes. Bull. Seismol. Soc. Amer. v. 49, No. 4, 1959.

Scheidegger A. E. Mathematical methods in geology. Amer. J. Sci., v. 258, No 3, 1960.

Schlecht W. G. & Stevens R. E. Results of chemical analysis of samples of granite and diabase. A comperative Investigation of Precision and Accuracy in Chemical, Spectrochemical and Modal Analysis of Silicate Rocks. Geol. Surv. Bull. No. 980, 1951.

Sichel H. S. An Experimental and Theoretical Investigation of Bias Error in Mine Sampling with special Reference to Narrow gold Reefs. Bull. Inst. Min. and Met. No. 483, 1947.

Sichel H. S. Mine valuation and maximum liklyhood of Master's thesis. University of the Witwatersrand, 1949.

Sichel H. S. New methods in the statistical evaluation of mine sampling data. Discussion. Bull. Inst. Min. and Met. No. 544, 548, 552, 1952.

Sichel H. S. Equilibrium theory of erosional slopes approached by frequency distribution analysis. Amer. J. Sci., v. 248, No. 11, 1950.

Sichel H. S. Quantitative analysis of watershed geomorphology. Trans. Amer. Geophys. Union, v. 38, No. 6, 1957.

Sichel H. S. Dimensional analysis applied to eluvially eroded landforms. Bull. Geol. Soc. Amer., v. 69, 1958.

Trembecki Adam. Wstep do ststystycznych metod obliczania zasobow. Przegl. Geol., t. 6, No. 1, 1958.

Trembecki Adam. Parametr zlozowy tractovany jako niezalezna zmienna losowa. Arch. Gorn., t. 6, No. 2, 1961.

Trembecki Adam. Parametr zlozowy tractovany jako zalezna zmienna losowa. Arch. Gorn., t. 6, No. 3, 1961.

Vistelius A. B. Paragenisis of sodium, potassium and uranium in volcanic rocks of Lassen Volcanic National Park, California. Geochim. et Cosmochim. Acta, v. 14, No. 1—2, 1958.

Vistelius A. B. The screw frequency distributions and the fundamental law of the geochemical processes. J. Geol., v. 68, No. 1, p. 1—22, 1960.

Vistelius A. B. Sedimentation Time Trend Functions and their Application for Correction of Sedimentary Deposits. J. Geol., v. 69, No. 6, 1961.

Vistelius A. B., Sarmanov O. V. On the correlation between percentage Values: Major Component Correlation in Ferromagnesian Micas. J. Geol., v. 69, No. 2, 1961.

Whitten E. H. Timothy. Quantitative areal model analysis of granitic complexes. Bull. Geol. Soc. Amer., v. 72, No. 9, 1961.

Whitten E. H. Timothy. Systematic quantitative areal variation in five granite massives from India, Canada and Great Britain. J. Geol., v. 69, No. 1, 1961.

Whitten E. H. T. Quantitative distribution of major and trace components in rock masses. Trans. Amer. Inst. Min and Met. Eng. (Mining), v. 223, 1961.

Whitten E. H. T. Sampling and trend-surface analysis of granites: A reply. Bull. J. Soc. Amer., v. 73, 1962.

Yamaguchi Takao. Tisutsugaku dzassi J. Geol. Soc. Japan, v. 64, No. 757, 1958.

Zubrzycki Stefan. Szacomanie ztoz jak zagadnienie statystyczne. Przegl. Geol., t. 7, No. 9, 1959.

Таблица слу

1	2	3	4	5	6	7	8	9	10
1534	7106	2836	7873	5574	7545	7590	5574	1202	7712
6128	8993	4102	2551	0330	2358	6427	7067	9325	2454
6047	8566	8644	9343	9297	6751	3500	8754	2913	1258
0806	5201	5705	7355	1448	9562	7514	9205	0402	2427
9915	8274	4525	5695	5752	9630	7172	6988	0227	4264
2882	7158	4341	3463	1178	5786	1173	0670	0820	5067
9213	1223	4388	9760	6691	6896	8214	8813	0611	3131
8410	9836	3899	3683	1253	1683	6988	9978	8026	6751
9974	2362	2103	4326	3825	9079	6187	2721	1489	4216
3402	8162	8226	0782	3364	7871	4500	5598	9421	3816
8188	6596	1492	2139	8823	6878	0613	7161	0241	3834
3825	7020	1124	7483	9155	4919	3209	5959	2364	2555
9801	8788	6338	5899	3309	0807	0968	0539	4205	8257
5603	1251	6352	6467	0231	3556	2569	9446	4174	9219
0714	3757	0378	8266	8864	1374	6687	1221	0678	3714
4617	5652	7627	0372	8151	3668	1994	4402	2124	0016
6789	6279	7306	1856	7028	9043	7161	7526	6913	6393
6705	4978	8621	1790	4433	6298	0854	9127	3445	1111
3840	1086	0774	9241	9297	4233	1739	7734	0119	2436
7662	3939	2965	3273	0551	1645	8477	1877	5327	8629
7639	2868	4391	2950	7122	7325	9727	0080	7464	7947
3237	7203	4246	7329	7936	0065	4146	0866	4916	8648
3917	6271	1721	5469	1914	8653	0387	2756	6073	8984
9138	9395	6005	6423	7977	1873	7103	4267	9316	7206
8358	5896	6286	9242	5040	8509	2941	3913	3028	1563
1030	5094	1745	2975	2018	7340	6547	0207	5587	0300
6606	6305	1564	6668	7822	7142	6564	1639	5369	1659
4533	8841	4922	9365	1361	6692	1633	6774	0747	3881
4258	2012	0992	0106	1542	4760	0392	4057	0092	5203
5224	5128	8949	7928	7267	0116	1476	2009	1772	3860
6872	7492	7962	1867	7437	1526	3516	9129	4153	8084
8638	8407	7198	0956	0950	7753	5144	3914	5596	6104
9958	7172	5822	4224	6701	7559	4985	4856	4461	6147
0265	3086	2996	0699	3584	9702	1665	0446	9107	6437
8987	5441	7878	9404	0487	2939	3805	9172	7887	5197
5552	3529	9627	9362	6298	6021	0024	9520	9154	0643
9383	6640	7394	9592	9903	7699	8939	9972	1257	0994
9903	4059	0332	9109	0182	6721	163	9008	2542	4461
6530	5070	7589	6928	6014	1832	9307	5107	1354	9257
8679	8953	8310	2060	6277	1773	7979	6741	6033	3588
5765	4987	1639	3512	9843	5286	3786	2384	4919	5661
7198	2447	6716	0291	5585	1106	5330	0504	6346	3679
2385	0605	2678	1399	2371	7968	1212	9569	8650	5841
0732	8732	8660	5836	9065	4603	0029	8042	0159	0345
1642	6094	3795	2600	4532	9740	0376	4384	9203	5387
4514	1956	7212	0687	7632	2106	0846	7055	4106	9157
8744	5580	8038	9087	7222	0424	0028	4511	3191	9846
3729	6225	5397	6790	2157	3414	6509	5204	4779	5641
8858	3147	8410	2783	1290	9796	8873	7585	7185	4726
3522	5601	6197	6051	3470	8283	5702	0103	8726	5282

11	12	13	14	15	16	17	18	19	20
5489	5583	3156	0835	1988	3912	0938	7460	0869	4420
3522	0935	7877	5665	7020	9255	7379	7124	7878	5544
7555	7579	2550	2487	9477	0864	2349	1012	8250	2633
5759	3554	5080	9074	7001	6249	3224	6368	9102	2672
6303	6895	3371	3196	7231	2918	7380	0438	7547	2644
7351	5634	5323	2623	7803	8374	2191	0464	0696	9529
7068	7803	8832	5119	6350	0120	5026	3684	5657	0304
3613	1428	1796	8447	0503	5654	3254	7336	9536	1944
5143	4534	2105	0368	7890	2473	4240	8652	9435	1422
9815	5144	7649	8638	6137	8070	5345	4865	2456	5708
5780	1277	6316	1013	2867	9938	3930	3203	5696	1769
1187	0951	5991	5245	5700	5564	7352	0891	6249	6568
4184	2179	4554	9083	2254	2435	2965	5154	1209	7069
2916	2972	9885	0275	0144	8034	8122	3213	7666	0230
5524	1341	9860	6565	6981	9842	0171	2284	2707	3008
0146	5291	2354	5694	0377	5336	6460	9585	3415	2358
4920	2826	5238	5402	7937	1993	4332	2327	6775	5230
7978	1947	6380	3425	7267	7285	1130	7722	0164	8573
7453	0653	3645	7497	5969	8682	4191	2976	0361	9334
1473	6938	4899	5348	1641	3652	0852	5296	4538	4456
8162	8797	8000	4707	1880	9660	8446	1883	9768	0881
5645	4219	0807	3301	4279	4168	4305	9937	3120	5547
2042	1192	1175	8851	6132	4635	5757	6656	1660	5389
5470	7702	6958	9080	5925	8519	0127	9233	2452	7341
4045	1730	6005	1704	0345	3275	4738	4862	2556	8333
5880	1257	6163	4439	7276	6353	6912	0731	9033	5294
9083	4260	5277	4998	4298	5204	3965	4028	8936	5148
1762	8713	1189	1090	8989	7273	3213	1935	9321	4820
2023	2589	1740	0424	8924	0005	1969	1636	7237	1227
7965	3855	4765	0703	1678	0841	7543	0308	9732	1289
7690	0480	8098	9629	4819	7219	7241	5128	3853	1921
9292	0426	9573	4903	5916	6576	8368	3270	6641	0033
0867	1656	7016	4220	2533	6345	8227	1904	5138	2537
0505	2127	8255	5276	2233	3956	4118	8199	6380	6340
6295	9795	1112	5761	2575	6837	3336	9322	7403	8345
6323	2615	3410	3365	1117	2417	3176	2434	5240	5455
8672	8536	2966	5773	5412	8114	0930	4697	6919	4569
1422	5507	7596	0670	3013	1351	3886	3268	9469	2584
2653	1472	5113	5735	1469	9545	9331	5303	9914	6394
0438	4376	3328	8649	8327	0110	4549	7955	5275	2890
2851	2157	0047	7085	1129	0460	6821	8373	2572	8962
7962	2753	3077	8718	7418	8004	3425	3706	8822	1494
3837	4098	0220	1217	4732	0150	1637	1097	1040	7372
8542	4126	9274	2251	0607	4301	8730	7690	6235	3477
0139	0765	8039	9484	2577	7859	1976	0623	1418	6685
6687	1943	4307	0579	8171	8224	8641	7034	3595	3875
6242	5582	5872	3197	4919	2792	5991	4058	9769	1918
6859	9606	0522	4993	0345	8958	1289	8825	6941	7685
6590	1932	6043	3623	1973	4112	1795	8465	2110	8045
3482	0478	0221	6738	7323	5643	4767	0106	2372	9862

Приложение 2

Треугольник Паскаля

n	Коэффициенты биномиальной строки
1	1 1
2	1 2 1
3	1 3 3 1
4	1 4 6 4 1
5	1 5 10 10 5 1
6	1 6 15 10 15 6 1
7	1 7 21 35 35 21 7 1
8	1 8 28 56 70 56 28 8 1
9	1 9 36 84 126 126 84 36 9 1
10	1 10 45 120 210 252 210 120 45 10 1
11	1 11 55 165 330 462 462 330 165 55 11 1
12	1 12 66 220 495 792 924 792 495 220 65 12 1
	C_n^0, C_n^1, C_n^2, ... C_n^{n-2}, C_n^{n-1}, C_n^n

Приложение 3

Значение функций $Z(t) = \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}}$ **для** $0{,}00 \leqslant t \leqslant 4{,}29$, **при этом** $Z(-t) = Z(t)$

t	,00	,01	,02	,03	,04	,05	,06	,07	,08	,09
0,0	0,3989	0,3989	0,3989	0,3988	0,3986	0,3984	0,3982	0,3980	0,3977	0,3973
1	3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
2	3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
3	3814	3802	3790	3778	3765	3752	3739	3725	3712	3697
4	3683	3668	3653	3637	3621	3605	3589	3572	3555	3538
5	3521	3503	3485	3467	3448	3429	3410	3391	3372	3352
6	3332	3312	3282	3271	3251	3230	3209	3187	3166	3144
7	3123	3101	3079	3056	3034	3011	2989	2966	2943	2920
8	2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
9	2661	2637	2613	2589	2565	2541	2516	2492	2468	2444
1,0	2420	2396	2371	2347	2323	2299	2275	2251	2227	2203
1,1	2179	2155	2131	2107	2082	2059	2036	2012	1989	1965
1,2	1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1,3	1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1,4	1497	1476	1456	1435	1415	1394	1374	1354	1339	1315
1,5	1295	1276	1257	1238	1219	1200	1182	1163	1145	1129
1,6	1109	1092	1074	1057	1040	1023	1006	0989	0973	0957
1,7	0940	0925	0909	0893	0878	0863	0848	0833	0818	0804
1,8	0789	0775	0761	0748	0734	0721	0707	0694	0681	0669
1,9	0656	0644	0632	0619	0608	0596	0584	0573	0562	0551
2,0	0540	0529	0519	0508	0498	0488	0478	0468	0459	0449
2,1	0440	0431	0422	0413	0404	0395	0387	0379	0371	0363
2,2	0355	0347	0339	0332	0325	0317	0310	0303	0296	0290
2,3	0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2,4	0224	0219	0213	0208	0203	0198	0194	0189	0184	0180
2,5	0175	0171	0167	0162	0158	0154	0151	0147	0143	0139
2,6	0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2,7	0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2,8	0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2,9	0059	0058	0056	0054	0053	0051	0050	0048	0047	0046
3,0	0044	0043	0042	0040	0039	0038	0037	0036	0035	0034
3,1	0033	0032	0031	0030	0029	0028	0027	0026	0025	0025
3,2	0024	0023	0022	0022	0021	0020	0020	0019	0018	0018
3,3	0017	0017	0016	0016	0015	0015	0014	0014	0013	0013
3,4	0012	0012	0011	0011	0011	0010	0010	0010	0009	0009
3,5	0009	0008	0008	0008	0008	0007	0007	0007	0007	0006
3,6	0006	0006	0006	0005	0005	0005	0005	0005	0005	0004
3,7	0004	0004	0004	0004	0004	0004	0003	0003	0003	0003
3,8	0003	0003	0003	0003	0002	0002	0002	0002	0002	0002
3,9	0002	0002	0002	0002	0002	0002	0002	0002	0001	0001
4,0	0001	0001	0001	0001	0001	0001	0001	0001	0001	0001
4,1	0001	0001	0001	0001	0001	0001	0001	0001	0001	0001
4,2	0001	0001	0001	0001	0000	0000	0000	0000	0000	0000

Например, для $t = 1{,}63$ величина $Z(t) = 0{,}1057$.

Таблица значений функций $\Phi_M(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\, dt$

z	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0113	0,0226	0,0338	0,0451	0,0564	0,0676	0,0789	0,0901	0,1013
0,1	1125	1236	1348	1459	1570	1680	1790	1900	2009	2118
0,2	2227	2335	2443	2550	2657	2763	2869	2974	3079	3183
0,3	3286	3389	3491	3593	3694	3794	3893	3992	4090	4187
0,4	4284	4380	4475	4569	4662	4755	4847	4937	5028	5117
0,5	5205	5292	5379	5465	5549	5633	5716	5798	5879	5959
0,6	6039	6117	6194	6271	6346	6420	6494	6566	6638	6708
0,7	6778	6847	6914	6981	7047	7112	7175	7238	7300	7361
0,8	7421	7480	7538	7595	7651	7707	7761	7814	7867	7918
0,9	7969	8019	8068	8116	8163	8209	8254	8299	8342	8385
1,0	8427	8468	8508	8548	8587	8624	8661	8698	8733	8768
1,1	8802	8835	8868	8900	8931	8961	8991	9020	9048	9076
1,2	9103	9130	9155	9181	9205	9229	9252	9275	9297	9319
1,3	9340	9361	9381	9400	9419	9438	9456	9473	9490	9507
1,4	9523	9539	9554	9569	9583	9597	9611	9624	9637	9639
1,5	9661	9673	9684	9695	9706	9716	9726	9736	9746	9755
1,6	9764	9772	9780	9788	9796	9804	9811	9818	9825	9832
1,7	9838	9844	9850	9856	9861	9867	9872	9877	9882	9886
1,8	9881	9895	9899	9904	9907	9911	9915	9918	9922	9925
1,9	9928	9931	9934	9937	9939	9942	9944	9947	9949	9951
2,0	9953	9955	9957	9959	9961	9963	9964	9966	9967	9969
2,1	9970	9972	9973	9974	9975	9976	9978	9979	9980	9981
2,2	9981	9982	9983	9984	9985	9985	9986	9987	9987	9988
2,3	9989	9989	9990	9990	9991	9991	9992	9992	9992	9993
2,4	9993	9994	9994	9994	9994	9995	9995	9995	9996	9996
2,5	9996	9996	9996	9997	9997	9997	9997	9997	9997	9998
2,6	9998	9998	9998	9998	9998	9998	9998	9998	9999	9999
2,7	9999	9999	9999	9999	9999	9999	9999	9999	9999	9999
2,8	9999	9999	9999	9999	9999	9999	1,0000	1,0000	1,0000	1,0000

Например, для $z = 1{,}52$ $\Phi_M(z) = 0{,}9684$.

Функция нормального распределения с параметрами 0, 1 (для отрицательных значений z)

$$F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}}\, dt \quad \text{для} \quad -3{,}89 \leqslant z \leqslant 0{,}00$$

z	0,00	,01	,02	,03	,04	,05	,06	,07	,08	,09
—0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
—,1	4602	4562	4522	4483	4443	4404	4364	4325	4286	4247
—,2	4207	4168	4129	4090	4052	4013	3974	3936	3897	3859
—,3	3821	3783	3745	3707	3669	3632	3594	3557	3520	3483
—,4	3446	3409	3372	3336	3300	3264	3228	3192	3156	3121
—,5	3085	3050	3015	2981	2946	2912	2877	2843	2810	2776
—,6	2743	2709	2676	2643	2611	2578	2546	2514	2483	2451
—,7	2420	2389	2358	2327	2297	2266	2236	2206	2177	2148
—,8	2119	2090	2061	2033	2005	1977	1949	1922	1894	1867
—,9	1841	1814	1788	1762	1736	1711	1685	1660	1635	1611
—1,0	1587	1562	1539	1515	1492	1469	1446	1423	1401	1379
—1,1	1357	1335	1314	1292	1271	1251	1230	1210	1190	1170
—1,2	1151	1131	1112	1093	1075	1056	1038	1020	1003	0985
—1,3	0968	0951	0934	0918	0901	0885	0869	0853	0838	0823
—1,4	0808	0793	0778	0764	0749	0735	0721	0708	0694	0681
—1,5	0668	0655	0643	0630	0618	0606	0594	0582	0570	0559
—1,6	0548	0537	0526	0515	0505	0495	0485	0475	0465	0455
—1,7	0446	0436	0427	0418	0409	0401	0392	0384	0375	0367
—1,8	0359	0351	0344	0336	0329	0332	0314	0307	0300	0294
—1,9	0287	0281	0274	0268	0262	0256	0250	0244	0238	0233
—2,0	0227	0222	0217	0212	0207	0202	0197	0192	0188	0183
—2,1	0179	0174	0170	0166	0162	0158	0154	0150	0146	0143
—2,2	0139	0135	0132	0129	0125	0122	0119	0116	0113	0110
—2,3	0107	0104	0102	0099	0096	0094	0091	0089	0087	0084
—2,4	0082	0080	0078	0075	0073	0071	0069	0068	0066	0064
—2,5	0062	0060	0059	0057	0055	0054	0052	0051	0049	0047
—2,6	0047	0045	0044	0043	0041	0040	0039	0038	0037	0036
—2,7	0035	0034	0033	0032	0030	0030	0029	0028	0027	0026
—2,8	0026	0025	0024	0023	0026	0022	0021	0020	0020	0019
—2,9	0019	0018	0017	0017	0016	0016	0015	0015	0014	0014
—3,0	0013	0013	0012	0012	0012	0011	0011	0011	0010	0010
—3,1	0010	0009	0009	0009	0008	0008	0008	0008	0007	0007
—3,2	0007	0007	0006	0006	0006	0006	0006	0005	0005	0005
—3,3	0005	0005	0004	0004	0004	0004	0004	0004	0004	0003
—3,4	0003	0003	0003	0003	0003	0003	0003	0003	0002	0002
—3,5	0002	0002	0002	0002	0002	0002	0002	0002	0002	0002
—3,6	0002	0002	0001	0001	0001	0001	0001	0001	0001	0001
—3,7	0001	0001	0001	0001	0001	0001	0001	0001	0001	0001
—3,8	0001	0001	0001	0001	0001	0001	0001	0001	0000	0000

Например, для $z = 2{,}97$ получаем $F(z) = 0{,}0015$.

Функция нормального распределения с параметром 0,1 (для положительных значений z)

$$F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt \quad \text{для } 0{,}00 \leqslant z \leqslant 3{,}69$$

z	,00	,01	,02	,03	,04	,05	,06	,07	,08	,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5136	0,5199	0,5239	0,5279	0,5319	0,5359
1	5398	5438	5478	5517	5557	5596	5636	5675	5714	5753
2	5793	5832	5871	5910	5948	5987	6026	6064	6103	6141
3	6179	6217	6255	6293	6331	6368	6406	6443	6480	6517
4	6554	6591	6628	6664	6700	6736	6772	6808	6844	6879
5	6915	6950	6985	7019	7054	7088	7123	7157	7190	7224
6	7257	7291	7324	7357	7389	7422	7454	7486	7517	7549
7	7580	7611	7642	7673	7703	7734	7764	7794	7823	7852
8	7881	7910	7939	7967	7995	8023	8051	8078	8106	8133
9	8159	8186	8212	8238	8264	8289	8315	8340	8365	8389
1,0	8413	8438	8461	8485	8508	8531	8554	8577	8599	8621
1,1	8643	8665	8686	8708	8729	8749	8770	8790	8810	8830
1,2	8849	8869	8888	8907	8925	8925	8944	8962	8980	9015
1,3	9032	9049	9066	9082	9100	9115	9131	9147	9162	9177
1,4	9192	9207	9222	9236	9251	9265	9278	9292	9306	9319
1,5	9331	9345	9357	9370	9382	9394	9406	9418	9429	9441
1,6	9452	9463	9474	9484	9495	9505	9515	9525	9535	9545
1,7	9554	9564	9573	9582	9591	9599	9608	9616	9625	9633
1,8	9641	9648	9656	9664	9671	9678	9686	9693	9699	9706
1,9	9713	9719	9726	9732	9738	9744	9750	9756	9761	9767
2,0	9772	9778	9783	9788	9793	9798	9803	9808	9812	9817
2,1	9821	9826	9830	9834	9838	9842	9846	9850	9854	9857
2,2	9861	9864	9868	9871	9874	9878	9881	9884	9887	9890
2,3	9893	9896	9898	9901	9904	9906	9909	9911	9913	9916
2,4	9918	9920	9922	9924	9927	9929	9931	9932	9934	9936
2,5	9938	9940	9941	9943	9945	9946	9948	9949	9951	9952
2,6	9953	9955	9956	9957	9959	9960	9961	9962	9963	9964
2,7	9965	9966	9967	9968	9969	9970	9971	9972	9973	9974
2,8	9974	9975	9976	9977	9977	9978	9979	9979	9980	9981
2,9	9981	9982	9982	9983	9984	9984	9985	9985	9986	9986
3,0	9986	9987	9987	9988	9988	9989	9989	9989	9990	9990
3,1	9990	9991	9991	9991	9991	9992	9992	9992	9993	9993
3,2	9993	9993	9994	9994	9994	9994	9994	9995	9995	9995
3,3	9995	9995	9995	9996	9996	9996	9996	9996	9996	9996
3,4	9997	9997	9997	9997	9997	9997	9997	9997	9997	9998
3,5	9998	9998	9998	9998	9998	9998	9998	9998	9998	9998
3,6	9998	9998	9998	9998	9999	9999	9999	9999	9999	9999

Например, для $z = 1{,}34$ $F(z) = 0{,}9100$.

Значения функций $\Phi^*(z) = \frac{1}{\sqrt{2\pi}} \int_0^z e^{-\frac{t^2}{2}} dt$

z	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	398	438	478	517	557	596	636	675	714	754
0,2	793	832	871	910	948	987	1026	1064	1103	1141
0,3	1179	1217	1255	1293	1331	1368	1406	1443	1480	1517
0,4	1554	1591	1628	1664	1700	1736	1772	1808	1850	1879
0,5	1915	1950	1985	2019	2054	2088	2123	2157	2190	2224
0,6	2258	2291	2324	2357	2389	2422	2454	2486	2518	2549
0,7	2580	2612	2642	2673	2704	2734	2764	2794	2823	2852
0,8	2881	2910	2939	2967	2996	3023	3051	3079	3106	3133
0,9	3159	3186	3212	3238	3264	3289	3315	3340	3365	3389
1,0	3413	3438	3461	3485	3508	3531	3554	3577	3599	3621
1,1	3643	3665	3686	3708	3789	3749	3770	3790	3810	3830
1,2	3849	3869	3888	3907	3925	3944	3962	3980	3997	4015
1,3	4032	4049	4066	4082	4099	4115	4131	4147	4162	4177
1,4	4192	4207	4222	4236	4251	4265	4279	4292	4306	4320
1,5	4332	4345	4357	4370	4382	4394	4406	4418	4430	4441
1,6	4452	4463	4474	4485	4495	4505	4515	4525	4535	4545
1,7	4554	4564	4573	4582	4591	4599	4608	4616	4625	4633
1,8	4641	4649	4656	4664	4671	4678	4686	4693	4700	4706
1,9	4713	4719	4726	4732	4738	4744	4750	4756	4762	4767
2,0	4773	4778	4783	4788	4793	4798	4803	4808	4812	4820
2,1	4821	4826	4830	4834	4838	4842	4846	4850	4854	4857
2,2	4861	4865	4870	4871	4875	4878	4881	4884	4887	4890
2,3	4893	4896	4898	4901	4904	4906	4909	4911	4913	4916
2,4	4918	4920	4922	4925	4927	4929	4931	4932	4934	4936
2,5	4938	4940	4941	4943	4945	4946	4948	4949	4951	4952
2,6	4953	4955	4956	4957	4959	4960	4961	4962	4963	4964
2,7	4965	4966	4967	4968	4969	4970	4971	4972	4973	4974
2,8	4974	4975	4976	4977	4977	4978	4979	4980	4980	4981
2,9	4981	4982	4983	4983	4984	4984	4985	4985	4986	4986
3,0	4987	4987	4987	4988	4988	4989	4989	4989	4990	4990
3,1	4990	4990	4991	4991	4992	4992	4992	4992	4993	4993
3,2	4993	4993	4994	4994	4994	4994	4994	4995	4995	4995
3,3	4995	4995	4996	4996	4996	4996	4996	4996	4996	4997
3,4	4997	4997	4997	4997	4997	4997	4997	4997	4998	4998
3,5	4998	4998	4998	4998	4998	4998	4998	4998	4998	4998
3,6	4998	4999	4999	4999	4999	4999	4999	4999	4999	4999
3,7	4999	4999	4999	4999	4999	4999	4999	4999	4999	4999
3,8	4999	4999	4999	4999	4999	4999	4999	5000	5000	5000

Например, для $z = 0{,}13$ функция $\Phi^*(z)$ равна 0,0557.

Значения функции $\Phi(z) = \frac{2}{\sqrt{2\pi}} \int_0^z e^{-\frac{t^2}{2}} dt$

z	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0080	0,0160	0,0239	0,0319	0,0399	0,0478	0,0558	0,0638	0,0717
0,1	797	876	955	1034	1113	1192	1271	1350	1429	1507
0,2	1585	1663	1741	1819	1897	1974	2051	2128	2205	2282
0,3	2358	2434	2510	2586	2661	2737	2812	2886	2961	3035
0,4	3108	3182	3255	3328	3401	3473	3545	3616	3688	3759
0,5	3829	3900	3969	4039	4108	4177	4245	4313	4381	4448
0,6	4515	4581	4647	4713	4778	4843	4908	4971	5035	5098
0,7	5161	5223	5285	5346	5407	5468	5528	5587	5646	5705
0,8	5763	5821	5878	5935	5991	6047	6102	6157	6211	6265
0,9	6319	6372	6424	6476	6528	6579	6629	6680	6729	6778
1,0	6827	6875	6923	6970	7017	7063	7109	7154	7199	7293
1,1	7287	7330	7373	7415	7457	7499	7540	7580	7620	7660
1,2	7699	7737	7775	7813	7850	7887	7923	7959	7995	8030
1,3	8064	8098	8132	8165	8198	8230	8262	8293	8324	8355
1,4	8385	8415	8444	8473	8501	8529	8557	8584	8611	8638
1,5	8664	8690	8715	8740	8764	8787	8812	8836	8859	8882
1,6	8904	8926	8948	8969	8990	9011	9031	9051	9070	9090
1,7	9109	9127	9146	9164	9181	9199	9216	9233	9249	9266
1,8	9281	9297	9312	9328	9342	9357	9371	9385	9399	9412
1,9	9426	9439	9451	9464	9476	9488	9500	9512	9523	9534
2,0	9545	9556	9566	9576	9587	9596	9606	9616	9625	9634
2,1	9643	9651	9660	9668	9677	9684	9692	9700	9707	9715
2,2	9722	9729	9736	9743	9749	9756	9762	9768	9774	9780
2,3	9786	9791	9797	9802	9807	9812	9817	9822	9827	9832
2,4	9836	9841	9845	9849	9853	9857	9861	9865	9869	9872
2,5	9876	9879	9883	9886	9889	9892	9895	9898	9901	9904
2,6	9907	9910	9912	9915	9917	9920	9922	9924	9926	9929
2,7	9931	9933	9935	9937	9939	9940	9942	9944	9946	9947
2,8	9949	9951	9952	9954	9955	9956	9958	9959	9960	9962
2,9	9963	9964	9965	9966	9967	9968	9969	9970	9971	9972
3,0	9973	9974	9975	9976	9976	9977	9978	9979	9979	9980
3,1	9981	9981	9982	9983	9983	9984	9984	9985	9985	9986
3,2	9986	9987	9987	9988	9988	9989	9989	9989	9990	9990
3,3	9990	9991	9991	9991	9992	9992	9992	9993	9993	9993
3,4	9993	9994	9994	9994	9994	9994	9995	9995	9995	9995
3,5	9995	9996	9996	9996	9996	9996	9996	9996	9997	9997
3,6	9997	9997	9997	9997	9997	9997	9998	9998	9998	9998
3,7	9998	9998	9998	9998	9998	9998	9999	9999	9999	9999
3,8	9999	9999	9999	9999	9999	9999	9999	9999	9999	9999
3,9	9999	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Например, для $z = 3{,}18$ функция $\Phi(z)$ равна 0,9985.

Приложение 9

Значения функции $1 - \Phi(z) = \frac{2}{\sqrt{2\pi}} \int_z^\infty e^{-\frac{t^2}{2}} dt$

z	0	1	2	3	4	5	6	7	8	9
0,0	1,0000	0,9920	0,9840	0,9761	0,9681	0,9601	0,9522	0,9442	0,9362	0,9283
0,1	9203	9124	9045	8966	8887	8808	8729	8650	8572	8493
0,2	8115	8337	8259	8181	8103	8026	7949	7872	7795	7718
0,3	7642	7566	7490	7414	7339	7263	7189	7114	7040	6965
0,4	6892	6818	6745	6672	6599	6527	6455	6384	6312	6241
0,5	6171	6101	6031	5961	5892	5823	5755	5687	5619	5552
0,6	5485	5419	5353	5287	5222	5157	5093	5027	4965	4902
0,7	4839	4777	4715	4654	4593	4533	4473	4413	4354	4295
0,8	4237	4179	4122	4065	4009	3953	3898	3843	3789	3735
0,9	3681	3628	3576	3524	3472	3421	3371	3321	3271	3222
1,0	3173	3125	3077	3030	2983	2937	2891	2846	2801	2757
1,1	2713	2670	2627	2585	2543	2501	2461	2420	2380	2341
1,2	2301	2263	2225	2187	2150	2113	2077	2041	2006	1971
1,3	1936	1902	1868	1835	1803	1770	1738	1707	1676	1645
1,4	1615	1585	1556	1527	1499	1471	1443	1416	1389	1362
1,5	1336	1310	1285	1260	1236	1211	1188	1164	1141	1118
1,6	1096	1074	1052	1031	1010	989	969	949	930	910
1,7	891	873	854	836	819	801	784	767	751	735
1,8	719	703	688	673	658	643	629	615	601	588
1,9	574	561	549	536	524	512	500	488	477	466
2,0	455	444	434	424	414	404	394	386	375	366
2,1	357	349	340	332	324	316	308	300	293	285
2,2	278	271	264	258	251	245	238	232	226	220
2,3	215	209	203	198	193	188	183	178	173	169
2,4	164	160	155	151	147	143	139	135	131	128
2,5	124	121	117	114	111	108	105	102	99	96
2,6	93	91	88	85	83	81	78	76	74	72
2,7	69	67	65	63	61	60	58	56	54	53
2,8	51	50	48	47	45	44	42	41	40	39
2,9	37	36	35	34	33	32	31	30	29	28
3,0	27	26	25	25	24	23	22	21	21	20
3,1	19	19	18	18	17	16	16	15	15	14
3,2	14	13	13	12	12	12	11	11	10	10
3,3	10	9	9	9	8	8	8	8	7	7
3,4	7	7	6	6	6	6	5	5	5	5
3,5	5	5	4	4	4	4	4	4	3	3
3,6	3	3	3	3	3	3	3	2	2	2
3,7	2	2	2	2	2	2	2	2	2	2
3,8	1	1	1	1	1	1	1	1	1	1
3,9	1	1	1	1	1	1	1	1	1	1
4,0	1	0	0	0	0	0	0	0	0	0

Например, для $z = 3{,}35$ функция равна 0,0008.

Приложение 10

Значения пределов, вне которых находится только один член нормальной совокупности, имеющей объем n

Объем n	Предел, $\pm k\sigma$	Объем n	Предел, $\pm k\sigma$	Объем n	Предел, $\pm k\sigma$	Объем n	Предел, $\pm k\sigma$
10	1,65	300	2,94	60	2,39	800	3,2
20	1,96	370	3,0	70	2,45	900	3,2
22	2,00	400	3,0	80	2,50	1 000	3,3
30	2,13	500	3,0	90	2,54	10 000	3,9
40	2,24	600	3,1	100	2,58	15 800	4,0
50	2,33	700	3,2	200	2,81	100 000	4,4
						1 740 000	5,0

Значения функции Пуассона $P_m = \frac{\lambda^m e^{-\lambda}}{m!}$

m	λ																	
	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1	2	3	4	5	6	7	8	9
0	0,9048	0,8187	0,7408	0,6703	0,6065	0,5488	0,4966	0,4493	0,4066	0,3679	0,1353	0,0498	0,0188	0,0067	0,0025	0,0009	0,0003	0,0001
1	905	1637	2223	2682	3033	3293	3476	3595	3659	3679	2707	1494	733	337	149	64	27	11
2	45	164	333	536	758	987	1216	1438	1647	1839	2707	2240	1415	842	446	223	107	50
3	2	11	33	71	126	198	284	383	494	613	1804	2240	1954	1404	893	521	286	150
4		1	3	7	16	30	50	77	111	153	902	1681	1954	1755	1335	912	573	337
5				1	2	4	7	12	20	31	361	1008	1563	1755	1606	1277	916	607
6							1	2	3	5	120	504	1042	1462	1606	1490	1221	911
7										1	35	216	595	1044	1377	1490	1396	1171
8											9	81	298	643	1033	1304	1396	1318
9											2	27	133	363	688	1014	1241	1318
10												8	53	181	413	710	993	1186
11												2	19	81	225	452	722	970
12												1	6	34	113	264	481	728
13													2	13	52	142	296	504
14													1	5	22	71	170	324
15														2	9	33	90	194
16														1	3	15	45	109
17															1	6	21	58
18																2	9	29
19																1	4	14
20																	2	6
21																	1	3
22																		1

Например: 1) вероятность того, что событие A появится 5 раз, если $\lambda = 4$, равна 0,1954; 2) для $\lambda = 4$ и $m = 14$ функция $P_m = 0{,}0001$.

Значения функции $P_n(\xi \geqslant m) = \sum_{k=m}^{m} \frac{\lambda^k e^{-\lambda}}{k!}$

m	λ																	
	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1	2	3	4	5	6	7	8	9	10
0	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
1	952	1813	2592	3297	3935	4512	5034	5507	5934	6321	8647	9502	9817	9933	9975	9991	9997	9999	1,0000
2	47	175	369	616	902	1219	1558	1912	2275	2642	5940	8009	9084	9596	9826	9227	9970	9988	9995
3	2	11	36	79	144	232	341	474	629	803	3233	5768	7619	8753	9380	9704	9862	9938	9972
4		1	3	8	18	34	58	91	135	190	1429	3528	5665	7350	8488	9182	9776	9788	9897
5					2	4	8	14	23	37	527	1847	3712	5595	7149	8270	9004	9450	9707
6								2	3	6	166	839	2149	3840	5543	6993	8088	8843	9329
7											45	335	1107	2378	3937	5503	6866	7932	8699
8											11	119	511	1334	2560	4013	5470	6761	7798
9											2	39	214	681	1528	2709	4074	5443	6672
10												11	81	318	839	1695	2834	4126	5421
11												3	28	137	427	985	1841	2940	4170
12													9	55	201	534	1119	1967	3032
13														20	88	270	638	1242	2084
14														7	36	128	342	739	1355
15															14	57	173	415	835
16															5	24	82	220	487
17																10	37	111	270
18																	16	53	143
19																	7	24	72
20																		11	35
21																		4	16
22																			7

Например: 1) для $\lambda = 4$ и $m = 5$ значение функции равно 0,3712; 2) для $\lambda = 0{,}8$ и $m = 6$ значение функции равно 0,0002.

Значения функции $P_n(\xi \leqslant m) = \sum_{k=0}^{m} \frac{\lambda^k e^{-\lambda}}{k!}$

m	λ																		
	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1	2	3	4	5	6	7	8	9	10
0	0,9048	0,8187	0,7408	0,6703	0,6065	0,5488	0,4966	0,4493	0,4066	0,3679	0,1353	0,0498	0,0183	0,0067	0,0025	0,0009	0,0003	0,0001	0,0000
1	9953	9825	9631	9384	9098	8781	8442	8088	7725	7358	4060	1991	916	404	174	73	30	12	5
2	9998	9989	9964	9921	9856	9779	9659	9526	9371	9197	6767	4332	2381	1247	620	296	138	62	28
3	1,0000	9999	9997	9992	9982	9976	9942	9909	9885	9810	8571	6472	4335	2650	1512	818	424	212	103
4		1,0000	1,0000	9999	9998	9996	9992	9986	9977	9963	9473	8153	6288	4405	2851	1730	996	550	293
5				1,0000	1,0000	1,0000	9999	9998	9997	9994	9834	9161	7851	6160	4457	3007	1912	1157	671
6							1,0000	1,0000	1,0000	9999	9955	9665	8893	7622	6063	4497	3134	2068	1301
7										1,0000	9989	9881	9487	8666	7440	5987	4530	3239	2202
8											9998	9962	9786	9318	8472	7291	5925	4557	3328
9											1,0000	9989	9919	9682	9161	8305	7166	5874	4579
10												9997	9972	9862	9574	9015	8159	7060	5830
11												9999	9991	9945	9799	9467	8881	8030	6968
12												1,0000	9997	9980	9912	9730	9362	8558	7916
13													9999	9992	9964	9872	9558	9261	8645
14													1,0000	9998	9986	9943	9827	9585	9166
15														9999	9995	9976	9918	9780	9513
16														1,0000	9998	9990	9963	9889	9730
17															9999	9996	9984	9947	9857
18															1,0000	9999	9994	9976	9928
19																1,0000	9997	9989	9966
20																	9999	9996	9984
21																	1,0000	9998	9993
22																		9999	9997
23																		1,0000	9999
24																			1,0000

Например: 1) для $\lambda = 4$ и $m = 5$ функция равна 0,7851; 2) для $\lambda = 10$ и $m = 1$ функция равна 0,0005.

Значения χ^2, определяемые из соотношения: $P(\chi^2) = \dfrac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} \int_{\chi^2}^{\infty} (\chi^2)^{\frac{k}{2}-1} e^{-\frac{\chi^2}{2}} d\chi^2$

Уровень значимости	Число степеней свободы k									
	1	2	3	4	5	6	7	8	9	10
0,99	0,000157	0,0201	0,115	0,297	0,554	0,872	1,239	1,646	2,088	2,558
0,98	0,000628	0,0404	0,185	0,429	0,752	1,134	1,564	2,032	2,532	3,059
0,95	0,00393	0,103	0,352	0,711	1,145	1,635	2,167	2,733	3,325	3,940
0,90	0,0158	0,211	0,584	1,064	1,610	2,204	2,833	3,490	4,168	4,865
0,80	0,0642	0,446	1,005	1,649	2,343	3,070	3,822	4,594	5,380	6,179
0,70	0,148	0,713	1,424	2,195	3,000	3,828	4,671	5,527	6,393	7,267
0,50	0,455	1,386	2,366	3,357	4,351	5,348	6,346	7,344	8,343	9,342
0,30	1,074	2,408	3,665	4,878	6,064	7,231	8,383	9,524	10,656	11,781
0,20	1,642	3,219	4,642	5,989	7,289	8,558	9,803	11,030	12,242	13,442
0,10	2,706	4,605	6,251	7,779	9,236	10,645	12,017	13,362	14,684	15,987
0,05	3,841	5,991	7,815	9,488	11,070	12,592	14,067	15,507	16,919	18,307
0,02	5,412	7,824	9,837	11,688	13,388	15,033	16,662	18,168	19,679	21,161
0,01	6,635	9,210	11,341	13,277	15,086	16,812	18,475	20,090	21,666	23,209
0,001	10,827	13,815	16,266	18,467	20,515	22,457	24,322	26,125	27,877	29,588

Уровень значимости	Число степеней свободы k									
	11	12	13	14	15	16	17	18	19	20
0,99	3,053	3,571	4,107	4,660	5,229	5,822	6,408	7,015	7,633	8,260
0,98	3,609	4,178	4,765	5,368	5,985	6,614	7,255	7,906	8,567	9,237
0,95	4,575	5,226	5,892	6,571	7,261	7,962	8,672	9,390	10,117	10,851
0,90	5,578	6,304	7,042	7,790	8,547	9,312	10,085	10,865	11,651	12,443
0,80	6,989	7,807	8,634	9,467	10,307	11,152	12,002	12,857	13,716	14,578
0,70	8,148	9,034	9,926	10,821	11,721	12,624	13,531	14,440	15,352	16,266
0,50	10,341	11,340	12,340	13,339	14,339	15,338	16,338	17,338	18,338	19,337

Уровень значимости	Число степеней свободы k									
	11	12	13	14	15	16	17	18	19	20
0,30	12,899	14,011	15,119	16,222	17,322	18,418	19,511	20,601	21,689	22,775
0,20	14,631	15,812	16,985	18,151	19,311	20,465	21,615	22,760	23,900	25,038
0,10	17,275	18,549	19,812	21,064	22,307	23,542	24,769	25,989	27,204	28,412
0,05	19,675	21,026	22,362	23,685	24,996	26,296	27,587	28,869	30,144	31,410
0,02	22,618	24,054	25,472	26,873	28,259	29,633	30,995	32,346	33,687	35,020
0,01	24,725	26,217	27,688	29,141	30,578	32,000	33,409	34,805	36,191	37,566
0,001	31,264	32,909	34,528	36,123	37,697	39,252	40,790	42,312	43,820	45,315

Уровень значимости	Число степеней свободы k									
	21	22	23	24	25	26	27	28	29	30
0,99	8,897	9,542	10,196	10,856	11,524	12,198	12,879	13,565	14,256	14,953
0,98	9,915	10,600	11,293	11,992	12,697	13,409	14,125	14,847	15,574	16,306
0,95	11,591	12,338	13,091	13,848	14,611	15,379	16,151	16,928	17,708	18,493
0,90	13,240	14,041	14,848	15,659	16,473	17,292	18,114	18,939	19,768	20,599
0,80	15,445	16,314	17,187	18,062	18,940	19,820	20,703	21,588	22,475	23,364
0,70	17,182	18,101	19,021	19,943	20,867	21,792	22,719	23,647	24,577	25,508
0,50	20,337	21,337	22,337	23,337	24,337	25,336	26,336	27,336	28,336	29,336
0,30	23,858	24,939	26,018	27,096	28,172	29,246	30,319	31,391	32,461	33,530
0,20	26,171	27,301	28,429	29,553	30,675	31,795	32,912	34,027	35,139	36,250
0,10	29,615	30,813	32,007	33,196	34,382	35,563	36,741	37,916	39,087	40,256
0,05	32,671	33,924	35,172	36,415	37,652	38,885	40,113	41,337	42,557	43,773
0,02	36,343	37,659	38,968	40,270	41,566	42,856	44,140	45,419	46,693	47,962
0,01	38,932	40,289	41,638	42,980	44,314	45,642	46,963	48,278	49,588	50,892
0,001	46,797	48,268	49,728	51,179	52,620	54,052	55,476	56,893	58,302	59,703

Например, для $P\ (\chi^2) = 0{,}02$ и $4k = 2$ величина $\chi^2 = 40{,}270$.

Приложение 15

Допустимые значения критерия Стьюдента при данном числе степеней свободы k и уровне значимости P

k	P												
	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,05	0,02	0,01	0,001
1	0,16	0,33	0,51	0,73	1,00	1,38	1,96	3,08	6,31	12,71	31,82	63,66	636,62
2	14	29	45	62	0,82	1,06	1,39	1,89	2,92	4,30	6,97	9,93	31,60
3	14	28	42	58	77	0,98	1,25	1,64	2,35	3,18	4,54	5,84	12,94
4	13	27	41	57	74	94	1,19	1,53	2,13	2,78	3,75	4,60	8,61
5	13	27	41	56	73	92	1,16	1,48	2,02	2,57	3,37	4,03	6,86
6	13	27	40	55	72	91	1,13	1,44	1,94	2,45	3,14	3,71	5,96
7	13	26	40	55	71	90	1,12	1,42	1,90	2,37	3,00	3,50	5,41
8	13	26	40	55	71	89	1,11	1,40	1,86	2,31	2,90	3,36	5,04
9	13	26	40	54	70	88	1,10	1,38	1,83	2,26	2,82	3,25	4,78
10	13	26	40	54	70	88	1,09	1,37	1,81	2,23	2,76	3,17	4,59
11	13	26	40	54	70	88	1,09	1,36	1,80	2,20	2,72	3,11	4,44
12	13	26	40	54	70	87	1,08	1,36	1,78	2,18	2,68	3,06	4,32
13	13	26	39	54	69	87	1,08	1,35	1,77	2,16	2,65	3,01	4,22
14	13	26	39	54	69	87	1,08	1,35	1,76	2,15	2,62	2,98	4,14
15	13	26	39	54	69	87	1,07	1,34	1,75	2,13	2,60	2,95	4,07
16	13	26	39	54	69	87	1,07	1,34	1,75	2,12	2,58	2,92	4,02
17	13	26	39	53	69	86	1,07	1,33	1,74	2,11	2,57	2,90	3,97
18	13	26	39	53	69	86	1,07	1,33	1,73	2,10	2,55	2,88	3,92
19	13	26	39	53	69	86	1,07	1,33	1,73	2,09	2,54	2,86	3,88
20	13	26	39	53	69	86	1,06	1,33	1,73	2,09	2,53	2,85	3,85
21	13	26	39	53	69	86	1,06	1,32	1,72	2,08	2,52	2,83	3,82
22	13	26	39	53	69	86	1,06	1,32	1,72	2,07	2,51	2,82	3,79
23	13	26	39	53	69	86	1,06	1,32	1,71	2,07	2,50	2,81	3,77
24	13	26	39	53	69	86	1,06	1,32	1,71	2,06	2,49	2,80	3,75
25	13	26	39	53	68	86	1,06	1,32	1,71	2,06	2,49	2,79	3,73
26	13	26	39	53	68	86	1,06	1,32	1,71	2,06	2,48	2,78	3,71
27	13	26	39	53	68	86	1,06	1,31	1,70	2,05	2,47	2,77	3,69
28	13	26	39	53	68	86	1,06	1,31	1,70	2,05	2,47	2,76	3,67
29	13	26	39	53	68	85	1,06	1,31	1,70	2,05	2,46	2,76	3,66
30	13	26	39	53	68	85	1,06	1,31	1,70	2,04	2,46	2,75	3,65
40	13	26	39	53	68	85	1,05	1,30	1,68	2,02	2,42	2,70	3,55
60	13	25	39	53	68	85	1,05	1,30	1,67	2,00	2,39	2,66	3,46
120	13	25	39	53	68	85	1,04	1,29	1,66	1,93	2,36	2,62	3,37
∞	13	25	39	52	67	84	1,04	1,28	1,65	1,96	2,33	2,58	3,29

Квантили (p-процентные нормы) для отношения выборочного размаха w_n к параметру σ исходного распределения (математическое ожидание α и среднее квадратическое отклонение β этого же отношения в долях параметра σ — исходного распределения) (И. В. Дунин-Барковский, Н. В. Смирнов, 1955)

n	α	β	γ	Вероятность в процентах																			
				0,05	0,1	0,5	1,0	2,5	5,0	10,0	20,0	30,0	40,0	50,0	60,0	70,0	80,0	90,0	95,0	97,5	99,0	99,5	99,9	99,95
2	1,128	0,853	0,756	0,00	0,00	0,01	0,02	0,04	0,09	0,18	0,36	0,55	0,74	0,95	1,20	1,47	1,81	2,33	2,77	3,17	3,64	3,97	4,66	4,92
3	1,693	0,888	0,525	0,04	0,06	0,13	0,19	0,30	0,43	0,62	0,90	1,14	1,36	1,59	1,83	2,09	2,42	2,90	3,31	3,68	4,12	4,42	5,06	5,31
4	2,059	0,880	0,427	0,16	0,20	0,34	0,43	0,59	0,76	0,98	1,29	1,53	1,76	1,98	2,21	2,47	2,78	3,24	3,63	3,98	4,40	4,69	5,31	5,56
5	2,326	0,864	0,371	0,31	0,37	0,55	0,66	0,85	1,03	1,26	1,57	1,82	2,04	2,26	2,48	2,73	3,04	3,48	3,86	4,20	4,60	4,89	5,48	5,72
6	2,534	0,848	0,335	0,47	0,54	0,75	0,87	1,06	1,25	1,49	1,80	2,04	2,26	2,47	2,69	2,94	3,23	3,66	4,03	4,36	4,76	5,03	5,62	5,86
7	2,704	0,833	0,308	0,61	0,69	0,92	1,05	1,25	1,44	1,68	1,99	2,22	2,44	2,65	2,86	3,10	3,39	3,81	4,17	4,49	4,88	5,15	5,73	5,96
8	2,847	0,820	0,288	0,75	0,83	1,08	1,20	1,41	1,60	1,83	2,14	2,38	2,59	2,79	3,00	3,24	3,52	3,93	4,29	4,61	4,99	5,26	5,82	6,04
9	2,970	0,808	0,272	0,88	0,96	1,21	1,34	1,55	1,74	1,97	2,28	2,51	2,71	2,92	3,12	3,35	3,63	4,04	4,39	4,70	5,08	5,34	5,90	6,12
10	3,078	0,797	0,259	1,00	1,08	1,33	1,47	1,67	1,86	2,09	2,39	2,62	2,83	3,02	3,23	3,46	3,73	4,13	4,47	4,79	5,16	5,42	5,97	6,19
11	3,173	0,787	0,248	1,10	1,20	1,45	1,58	1,78	1,97	2,20	2,50	2,72	2,93	3,12	3,32	3,55	3,82	4,21	4,55	4,86	5,23	5,49	6,04	6,25
12	3,258	0,778	0,239	1,21	1,30	1,55	1,68	1,88	2,07	2,30	2,59	2,82	3,01	3,21	3,41	3,63	3,90	4,29	4,62	4,92	5,29	5,54	6,09	6,31
13	3,336	0,770	0,231	1,30	1,39	1,64	1,77	1,97	2,16	2,39	2,68	2,90	3,09	3,29	3,48	3,70	3,97	4,35	4,69	4,99	5,36	5,60	6,14	6,36
14	3,407	0,762	0,224	1,38	1,48	1,72	1,86	2,06	2,24	2,47	2,75	2,97	3,17	3,36	3,55	3,77	4,03	4,41	4,74	5,04	5,40	5,65	6,19	6,40
15	3,472	0,755	0,217	1,46	1,56	1,80	1,93	2,14	2,32	2,54	2,83	3,04	3,23	3,42	3,62	3,83	4,09	4,47	4,80	5,09	5,45	5,70	6,23	6,45
16	3,532	0,749	0,212	1,53	1,63	1,88	2,01	2,21	2,39	2,61	2,89	3,11	3,30	3,48	3,67	3,89	4,14	4,52	4,85	5,14	5,49	5,74	6,28	6,49
17	3,588	0,743	0,207	1,60	1,69	1,94	2,07	2,27	2,45	2,67	2,95	3,17	3,35	3,54	3,73	3,94	4,19	4,57	4,89	5,18	5,54	5,79	6,32	6,52
18	3,640	0,738	0,203	1,66	1,75	2,01	2,14	2,34	2,51	2,73	3,01	3,22	3,41	3,59	3,78	3,99	4,24	4,61	4,93	5,22	5,57	5,82	6,35	6,56
19	3,689	0,733	0,199	1,72	1,82	2,07	2,20	2,39	2,57	2,79	3,06	3,27	3,46	3,64	3,83	4,03	4,29	4,65	4,97	5,26	5,61	5,86	6,38	6,59
20	3,735	0,729	0,195	1,78	1,88	2,12	2,25	2,45	2,63	2,84	3,11	3,32	3,51	3,69	3,87	4,08	4,33	4,69	5,01	5,30	5,65	5,89	6,41	6,62

Приложение 17

Значения u p-процентных норм для отклонения максимального члена x_{max} выборки из нормальной совокупности от центра распределения

n	p			
	90	95	99	99,9
1	1,28	1,65	2,33	3,09
2	1,63	1,96	2,58	3,29
3	1,81	2,12	2,71	3,39
4	1,94	2,23	2,81	3,49
5	2,04	2,32	2,88	3,54
6	2,11	2,39	2,94	3,59
7	2,17	2,44	2,98	3,65
8	2,22	2,49	3,02	3,68
9	2,27	2,53	3,05	3,70
10	2,31	2,57	3,09	3,72
12	2,38	2,63	3,14	3,79
14	2,43	2,68	3,19	3,82
16	2,48	2,73	3,23	3,86
18	2,52	2,77	3,26	3,86
20	2,56	2,80	3,29	3,89
25	2,64	2,87	3,95	3,98
30	2,70	2,93	3,40	4,08
35	2,75	2,98	3,44	4,08
40	2,79	3,01	3,49	4,17
45	2,83	3,05	3,51	4,17
50	2,86	3,08	3,54	4,17
60	2,92	3,14	3,59	4,17
70	2,97	3,18	3,65	4,27
80	3,01	3,22	3,67	4,27
90	3,04	3,26	3,70	4,27
100	3,08	3,28	3,72	4,27
150	3,19	3,40	3,82	4,27
200	3,28	3,48	3,89	4,42
300	3,39	3,59	4,08	4,67
500	3,53	3,72	4,17	4,68
1000	3,70	3,89	4,27	4,75

Таблица заимствована у И. В. Дунина-Барковского и Н. В. Смирнова (1955).

Приложение 18

Значения функции $\varphi(y) = e^{-e^{-y}}$ в зависимости от величины нормированного переменного y

y	$\varphi(y)$	y	$\varphi(y)$
—2,00	0,0006	1,75	0,8405
—1,75	32	2,00	8734
—1,50	113	2,25	9000
—1,25	305	2,50	9212
—1,00	660	2,75	9381
—0,75	1204	3,00	9514
—0,50	1923	3,25	9620
—0,25	2769	3,50	9703
0,00	3679	3,75	9768
0,25	4590	4,00	9819
0,50	5452	4,25	9858
0,75	6235	4,50	9890
1,00	6922	4,75	9914
1,25	7509	5,00	9933
1,50	8000	5,25	9948
		5,50	9959
		5,75	9968
		6,00	9975

Приложение 19

Нормированное отклонение y для закона распределения крайних членов вариационного ряда $P(x) = e^{-e^{-y}}$

$P(x)$	y	$P(x)$	y
0,0000	— ∞	0,55	0,51
0,0001	—2,22	0,60	0,67
0,0002	—2,14	0,65	0,84
0,0003	—2,09	0,70	1,03
0,0004	—2,06	0,75	1,25
0,0005	—2,03	0,80	1,50
0,001	—1,93	0,85	1,82
0,002	—1,83	0,90	2,25
0,003	—1,75	0,91	2,36
0,004	—1,75	0,92	2,48
0,005	—1,67	0,93	2,62
0,01	—1,53	0,94	2,78
0,02	—1,36	0,95	2,97
0,03	—1,25	0,96	3,20
0,04	—1,17	0,97	3,49
0,05	—1,10	0,98	3,90
0,06	—1,03	0,99	4,60
0,07	—0,98	0,995	5,30
0,08	—0,93	0,996	5,52
0,09	—0,88	0,997	5,81
0,10	—0,83	0,998	6,21
0,15	—0,64	0,999	6,91
0,20	—0,47	0,9995	7,60
0,25	—0,32	0,9996	7,82
0,30	—0,19	0,9997	8,11
0,35	—0,05	0,9998	8,52
0,37	0,00	0,9999	9,21
0,40	0,09	0,99995	9,90
0,45	0,23	0,99999	11,51
0,50	0,37	1,00000	∞

Значение функции $P_1(\lambda)$, т. е. вероятности того, что первый и второй члены упорядоченного ряда случайной выборки объема n отличаются друг от друга более чем в λ-кратное стандартное отклонение генеральной совокупности

n	λ																							
	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2,0	2,1	2,2	2,3	2,4	2,5
2	0,944	0,888	0,832	0,777	0,724	0,671	0,621	0,572	0,524	0,480	0,437	0,396	0,358	0,322	0,289	0,258	0,229	0,203	0,179	0,157	0,138	0,120	0,104	0,090	0,077
3	917	836	760	687	618	553	493	437	386	339	296	257	222	191	163	139	118	99	83	69	57	47	39	31	25
10	855	727	613	513	427	353	289	235	190	152	121	96	75	59	45	35	26	20	15	11	8	7	4	3	2
20	827	679	553	443	359	285	226	177	138	107	82	62	47	35	26	19	14	10	7	5	4	3	2	1	1
30	813	656	525	417	328	257	199	153	117	89	68	50	37	27	20	14	10	7	5	4	3	2	1	1	1
40	803	639	505	396	308	238	182	138	104	78	58	42	31	22	16	11	8	6	4	3	2	1	1	1	
50	796	628	491	382	294	225	170	128	96	71	52	38	27	19	14	10	7	5	3	2	1	1	1		
60	790	619	481	371	283	215	161	120	89	65	48	34	25	17	12	9	6	4	3	2	1	1	1		
70	785	611	471	361	274	206	154	114	84	61	44	32	22	16	11	8	5	4	2	2	1	1			
80	780	604	464	353	267	200	148	109	80	58	41	30	21	15	10	7	5	3	2	1	1	1			
90	788	600	459	348	262	195	144	106	77	55	40	28	20	14	10	7	4	3	2	1	1	1			
100	775	596	454	343	257	191	141	103	75	54	38	27	19	13	9	6	4	3	2	1	1	1			
200	757	567	422	311	227	165	118	84	60	42	29	20	14	9	6	4	3	2	1	1					
300	746	552	405	294	212	152	107	75	52	36	25	17	11	7	5	3	2	1	1	1					
400	740	543	394	284	203	144	101	70	48	33	22	15	10	7	4	3	2	1	1						
500	735	536	387	277	197	138	96	67	45	31	21	14	9	6	4	2	2	1	1						
600	731	530	381	271	191	134	93	64	43	29	20	13	9	5	4	2	1	1	1						
700	727	525	375	266	187	130	90	61	41	28	19	12	8	5	3	2	1	1	1						
800	725	521	371	262	183	127	87	59	40	27	18	12	8	5	3	2	1	1							
900	722	517	367	258	180	124	85	58	39	26	17	11	7	5	3	2	1	1							
1000	720	514	364	255	177	122	83	56	38	25	16	11	7	5	3	2	1	1							

n	λ																							
	2,6	2,7	2,8	2,9	3,0	3,1	3,2	3,3	3,4	3,5	3,6	3,7	3,8	3,9	4,0	4,1	4,2	4,3	4,4	4,5	4,6	4,7	4,8	4,9	5,0
2	0,066	0,056	0,048	0,040	0,034	0,028	0,024	0,020	0,016	0,013	0,011	0,009	0,007	0,006	0,005	0,004	0,003	0,002	0,002	0,001	0,001	0,001	0,001	0,001	—
3	21	16	13	10	8	6	5	4	3	2	2	1	1	1	1										
10	2	1	1	1																					
20	1																								

Например, для $n = 40$ и $\lambda = 1{,}7$ функция $P_1(\lambda) = 0{,}008$.

Приложение 21

Значение функции $P_2(\lambda)$, т. е. вероятности того, что второй и третий члены упорядоченного ряда случайной выборки объема n отличаются друг от друга более чем в λ-кратное стандартное отклонение генеральной совокупности

n	λ																			
	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2,0
3	0,917	0,836	0,760	0,687	0,618	0,553	0,493	0,437	0,386	0,339	0,296	0,257	0,222	0,191	0,163	0,139	0,118	0,099	0,083	0,069
10	772	589	444	331	244	177	128	91	64	44	30	21	14	9	6	4	2	2	1	
20	718	509	356	246	167	112	74	49	31	20	12	8	5	3	2	1	1			
30	693	473	319	211	138	89	57	35	22	13	8	4	3	1	1					
40	674	447	293	189	120	75	46	28	17	10	6	3	2	1	1					
50	659	428	274	172	107	65	39	23	14	8	4	2	1	1						
60	650	416	262	163	99	60	35	21	12	7	4	2	1	1						
70	642	406	252	155	93	56	32	19	11	6	3	2	1							
80	635	397	244	148	88	52	29	17	10	5	3	2	1							
90	628	388	237	142	84	49	27	16	9	5	3	1	1							
100	623	382	230	137	80	46	26	15	8	4	2	1	1							
200	596	349	201	114	64	35	19	10	5	3	1	1								
300	579	329	184	101	55	29	15	8	4	2	1									
400	565	314	172	92	49	25	13	6	3	2	1									
500	557	305	164	86	45	23	12	6	3	1	1									
600	550	297	158	82	42	21	11	5	2	1	1									
700	545	292	154	80	41	20	10	5	2	1										
800	541	288	150	77	39	19	9	5	2	1										
900	537	284	147	75	38	18	9	4	2	1										
1000	533	280	144	73	36	17	8	4	2	1										

n	λ																			
	2,1	2,2	2,3	2,4	2,5	2,6	2,7	2,8	2,9	3,0	3,1	3,2	3,3	3,4	3,5	3,6	3,7	3,8	3,9	4,0
3	0,057	0,047	0,039	0,031	0,025	0,020	0,016	0,013	0,010	0,008	0,006	0,005	0,004	0,003	0,002	0,002	0,001	0,001	0,001	0,001

Например, для $n = 50$ и $\lambda = 1{,}1$ функция $P_2(\lambda) = 0{,}004$.

Приложение 22

Значения верхних и нижних пределов общего числа серий R для различного числа наблюдений n

n	$R_{0,025}$	$R_{0,975}$	n	$R_{0,025}$	$R_{0,975}$
10	2	9	40	14	27
12	3	10	50	18	33
14	3	12	60	22	39
16	4	13	80	31	51
18	5	14	100	40	61
20	6	15	120	49	72
22	7	16	140	58	83
24	7	18	160	68	93
26	8	19	180	77	104
28	9	20	200	86	115
30	10	21	220	96	125
32	11	22			
34	11	24			
36	12	25			
38	13	26			

Приложение 23

Числа $H = \dfrac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \sqrt{\dfrac{2}{n}}$

n	H	n	H
4	0,7979	40	0,9811
5	0,8407	45	0,9832
6	0,8686	50	0,9849
7	0,8882	55	0,9863
8	0,9027	60	0,9874
9	0,9139	65	0,9884
10	0,9227	70	0,9892
11	0,9300	75	0,9900
12	0,9359	80	0,9906
13	0,9410	85	0,9911
14	0,9453	90	0,9916
15	0,9490	95	0,9921
16	0,9523	100	0,9925
17	0,9551		
18	0,9576		
19	0,9599		
20	0,9619		
25	0,9696		
30	0,9748		
35	0,9784		

Числа H позволяют по математическому ожиданию выборочного стандарта легко находить стандарт генеральной совокупности и решать практические задачи о вероятности нахождения величины $\dfrac{\bar{\sigma}}{\sigma H}$ в определенных пределах (здесь $\bar{\sigma}$ — выборочный стандарт, σ — стандарт генеральной совокупности, n — объем выборки).

Значение нижнего предела z, определяемое из соотношения

$$D(z) = \frac{2k_1^{\frac{k_1}{2}} k_2^{\frac{k_2}{2}} \Gamma\left(\frac{k_1+k_2}{2}\right)}{\Gamma\left(\frac{k_1}{2}\right)\Gamma\left(\frac{k_2}{2}\right)} \frac{t^{k_1 z}}{(k_1 e^{2z} + k_2)^{\frac{k_1+k_2}{2}}}$$

при вероятности $p = 0{,}05$ (по Фишеру)

k_2	k_1											
	1	2	3	4	5	6	8	12	16	24	50	∞
1	161,4	199,5	215,7	224,6	230,2	234,0	238,9	243,9	246,5	249,0	251,8	254,3
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,43	19,45	19,47	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,69	8,64	8,58	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,84	5,77	5,70	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,60	4,53	4,44	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,92	3,84	3,75	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,49	3,41	3,32	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,20	3,12	3,03	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,98	2,90	2,80	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,82	2,74	2,64	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,70	2,61	2,50	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,60	2,50	2,40	2,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,51	2,42	2,32	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,44	2,35	2,24	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,39	2,29	2,18	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,33	2,24	2,13	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,29	2,19	2,08	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,25	2,15	2,04	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,21	2,11	2,00	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,18	2,08	1,96	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,15	2,05	1,93	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,23	2,13	2,03	1,91	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,38	2,20	2,11	2,00	1,88	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,18	2,09	1,98	1,86	1,73
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,16	2,07	1,96	1,84	1,71
26	4,22	3,37	2,98	2,74	2,59	2,47	2,32	2,15	2,05	1,95	1,82	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,30	2,13	2,03	1,93	1,80	1,67
28	4,20	3,34	2,95	2,71	2,56	2,44	2,29	2,12	2,02	1,91	1,78	1,65
29	4,18	3,33	2,93	2,70	2,54	2,43	2,28	2,10	2,00	1,90	1,77	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,99	1,89	1,76	1,62
35	4,12	3,26	2,87	2,64	2,48	2,37	2,22	2,04	1,94	1,83	1,70	1,57
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,90	1,79	1,66	1,51
45	4,06	3,21	2,81	2,58	2,42	2,31	2,15	1,97	1,87	1,76	1,63	1,48
50	4,03	3,18	2,79	2,56	2,40	2,29	2,13	1,95	1,85	1,74	1,60	1,44
60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	1,92	1,81	1,70	1,56	1,39
70	3,98	3,13	2,74	2,50	2,35	2,23	2,07	1,89	1,79	1,67	1,53	1,35
80	3,96	3,11	2,72	2,49	2,33	2,21	2,06	1,88	1,77	1,65	1,51	1,32
90	3,95	3,10	2,71	2,47	2,32	2,20	2,04	1,86	1,76	1,64	1,49	1,30
100	3,94	3,09	2,70	2,46	2,30	2,19	2,03	1,85	1,75	1,63	1,48	1,28
125	3,92	3,07	2,68	2,44	2,29	2,17	2,01	1,83	1,72	1,60	1,45	1,25
150	3,90	3,06	2,66	2,43	2,27	2,16	2,00	1,82	1,71	1,59	1,44	1,22
200	3,89	3,04	2,65	2,42	2,26	2,14	1,98	1,80	1,69	1,57	1,42	1,19
300	3,87	3,03	2,64	2,41	2,25	2,13	1,97	1,79	1,68	1,55	1,39	1,15
400	3,86	3,02	2,63	2,40	2,24	2,12	1,96	1,78	1,67	1,54	1,38	1,13
500	3,86	3,01	2,62	2,39	2,23	2,11	1,96	1,77	1,66	1,54	1,38	1,11
1000	3,85	3,00	2,61	2,38	2,22	2,10	1,95	1,76	1,65	1,53	1,36	1,08
	3,84	2,99	2,60	2,37	2,21	2,09	1,94	1,75	1,64	1,52	1,35	1,00

Значение нижнего предела z, определяемое из соотношения

$$D(z) = \frac{2k_1^{\frac{k_1}{2}} k_2^{\frac{k_2}{2}} \Gamma\left(\frac{k_1+k_2}{2}\right)}{\Gamma\left(\frac{k_1}{2}\right)\Gamma\left(\frac{k_2}{2}\right)} \frac{e^{k_1 z}}{(k_1 e^{2z} + k_2)^{\frac{k_1+k_2}{2}}}$$

при вероятности $p = 0{,}01$ (по Фишеру)

k_2 \ k_1	1	2	3	4	5	6	8	12	16	24	50	∞
1	4052	4999	5403	5625	5764	5859	5981	6106	6169	6234	6302	6366
2	98,49	99,00	99,17	99,25	99,30	99,33	99,36	99,42	99,44	99,46	99,48	99,50
3	34,12	30,81	29,46	28,71	28,24	27,91	27,49	27,05	26,83	26,60	26,35	26,12
4	21,20	18,00	16,69	15,98	15,52	15,21	14,80	14,37	14,15	13,93	13,69	13,46
5	16,26	13,27	12,06	11,39	10,97	10,67	10,29	9,89	9,68	9,47	9,24	9,02
6	13,74	10,92	9,78	9,15	8,75	8,47	8,10	7,72	7,52	7,31	7,09	6,88
7	12,25	9,55	8,45	7,85	7,46	7,19	6,84	6,47	6,27	6,07	5,85	5,65
8	11,26	8,65	7,59	7,01	6,63	6,37	6,03	5,67	5,48	5,28	5,06	4,86
9	10,56	8,02	6,99	6,42	6,06	5,80	5,47	5,11	4,92	4,73	4,51	4,31
10	10,04	7,56	6,55	5,99	5,64	5,39	5,06	4,71	4,52	4,33	4,12	3,91
11	9,65	7,20	9,22	5,67	5,32	5,07	4,74	4,40	4,21	4,02	3,80	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,50	4,16	3,98	3,78	3,56	3,36
13	9,07	6,70	5,74	5,20	4,86	4,62	4,30	3,96	3,78	3,59	3,37	3,16
14	8,86	6,51	5,56	5,03	4,69	4,46	4,14	3,80	3,62	3,43	3,21	3,00
15	8,68	6,36	5,42	4,89	4,56	4,32	4,00	3,67	3,48	3,29	3,07	2,87
16	8,53	6,23	5,29	4,77	4,44	4,20	3,89	3,55	3,37	3,18	2,96	2,75
17	8,40	6,11	5,18	4,67	4,34	4,10	3,79	3,45	3,27	3,08	2,86	2,65
18	8,28	6,01	5,09	4,58	4,25	4,01	3,71	3,37	3,20	3,00	2,79	2,57
19	8,18	5,93	5,01	4,50	4,17	3,94	3,63	3,30	3,12	2,92	2,70	2,49
20	8,10	5,85	4,94	4,43	4,10	3,87	3,56	3,23	3,05	2,86	2,63	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,51	3,17	2,99	2,80	2,58	2,36
22	7,94	5,72	4,82	4,31	3,99	3,76	3,45	3,12	2,94	2,75	2,53	2,31
23	7,88	5,66	4,76	4,26	3,94	3,71	3,41	3,07	2,89	2,70	2,48	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,36	3,03	2,85	2,66	2,44	2,21
25	7,77	5,57	4,68	4,18	3,86	3,63	3,32	2,09	2,81	2,62	2,40	2,17
26	7,72	5,53	4,64	4,14	3,82	3,59	3,29	2,96	2,78	2,58	2,36	2,13
27	7,68	5,49	4,60	4,11	3,78	3,56	3,26	2,93	2,74	2,55	2,33	2,10
28	7,64	5,45	4,57	4,07	3,75	3,53	3,23	2,90	2,71	2,52	2,30	2,06
29	7,60	5,42	4,54	4,04	3,73	3,50	3,20	2,87	2,68	2,49	2,27	2,03
30	7,56	5,39	4,51	4,02	3,70	3,47	3,17	2,84	2,66	2,47	2,24	2,01
35	7,42	5,27	4,40	3,91	3,59	3,37	3,07	2,74	2,56	2,37	2,13	1,90
40	7,31	5,18	4,31	3,83	3,51	3,29	2,99	2,66	2,48	2,29	2,05	1,80
45	7,23	5,11	4,25	3,77	3,45	3,23	2,94	2,61	2,43	2,23	1,99	1,75
50	7,17	5,06	4,20	3,72	3,41	3,19	2,89	2,56	2,38	2,18	1,94	1,68
60	7,08	4,98	4,13	3,65	3,34	3,12	2,82	2,50	2,32	2,12	1,87	1,60
70	7,01	4,92	4,07	3,60	3,29	3,07	2,78	2,45	2,28	2,07	1,82	1,53
80	6,96	4,88	4,04	3,56	3,26	3,04	2,74	2,42	2,24	2,03	1,78	1,49
90	6,92	4,85	4,01	3,53	3,23	3,01	2,72	2,39	2,21	2,00	1,75	1,45
100	6,90	4,82	3,98	3,51	3,21	2,99	2,69	2,37	2,19	1,98	1,73	1,43
125	6,84	4,78	3,94	3,47	3,17	2,95	2,66	2,33	2,15	1,94	1,69	1,37
150	6,81	4,75	3,91	3,45	3,14	2,92	2,63	2,31	2,13	1,92	1,66	1,33
200	6,76	4,71	3,88	3,41	3,11	2,89	2,60	2,28	2,09	1,88	1,62	1,28
300	6,72	4,68	3,85	3,38	3,08	2,86	2,57	2,24	2,06	1,85	1,59	1,22
400	6,70	4,66	3,83	3,37	3,06	2,85	2,56	2,23	2,04	1,84	1,57	1,19
500	6,69	4,65	3,82	3,36	3,05	2,84	2,55	2,22	2,03	1,83	1,56	1,16
1000	6,66	4,63	3,80	3,34	3,04	2,82	2,53	2,20	2,01	1,81	1,54	1,11
	6,64	4,60	3,78	3,32	3,02	2,80	2,51	2,18	1,99	1,79	1,52	1,00

Пример: при $k_1 = 3$ и $k_2 = 35$ величина $z = 4{,}40$.

СОДЕРЖАНИЕ

Шарапов Иван Прокофьевич

ПРИМЕНЕНИЕ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ В ГЕОЛОГИИ

Редакторы издательства *В. В. Кузовкин, А. М. Антокольская*
Технический редактор *Л. Н. Ломилина*
Корректор *Л. Н. Ильина*

Сдано в производство 9/VII 1965 г. Подписано к печати 18/X 1965 г.
Формат 70×108¹/₁₆ Печ. л. 16,25 Усл. л. 22,75 Уч.-изд. л. 21,74
Т-13819 Тираж 3850 экз. Заказ № 538/2038—14 Цена 1 р. 67 к.
Объявлено в Сводном темплане изд-ва «Недра» 1965 г. № 417 Индекс 1—4—1

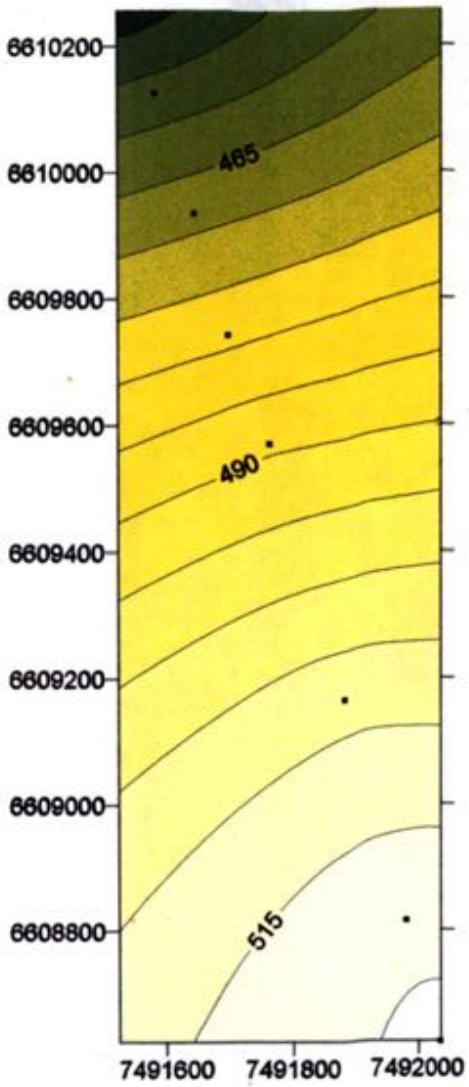
Издательство «Недра». Москва К-12, Третьяковский проезд, 1/19.

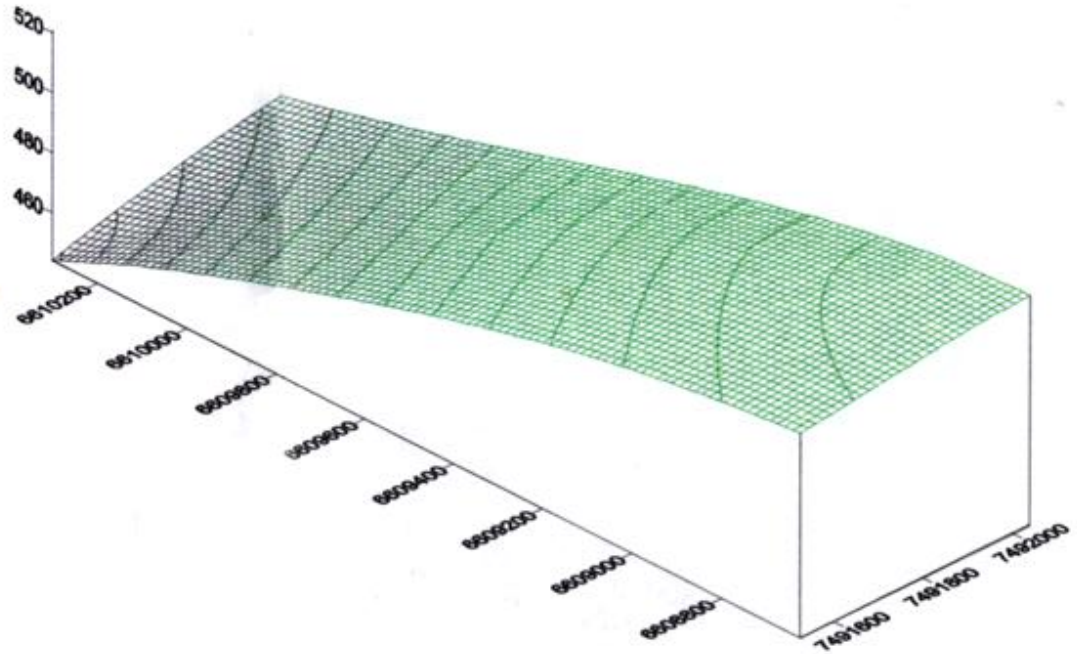
Ленинградская типография № 6 Главполиграфпрома Государственного комитета Совета Министров СССР по печати. Ленинград, ул. Моисеенко, 10

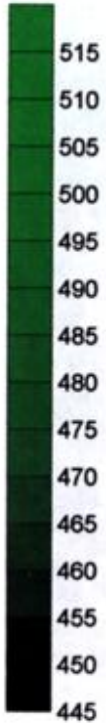
ОПЕЧАТКА

Стр.	Строка	Напечатано	Следует читать
251	1-я снизу	$4k = 2$	$k = 24$

Зак. № 538/2038—14. Шарапов И. П.


6610200
6610000
6609800
6609600
6609400
6609200
6609000
6608800
465
490
515
7491600
7491800
7492000


520
500
480
460
7491600
7491800
7492000


515
510
505
500
495
490
485
480
475
470
465
460
455
450
445