

**КОМПЬЮТЕРНЫЕ
ТЕХНОЛОГИИ И
СТАТИСТИЧЕСКИЕ МЕТОДЫ В
ЭКОЛОГИИ И
ПРИРОДОПОЛЬЗОВАНИИ
(КОНСПЕКТ ЛЕКЦИЙ)**

Геолого-математическое моделирование

Понятийные модели являются мыслимым образом природных явлений. Основаны на наблюдениях, служат для выражения изучаемого явления в идеализированной форме, отвечают существующему уровню знаний. Основная часть процессов и явлений в науках о Земле описана на уровне понятийных моделей.

Математические модели – абстрактный аналог физических, геометрических, понятийных моделей, в которых силы, события, соотношения участков, площадей, понятия и другие элементы заменены математическими символами, связанными между собой определенными соотношениями.

Геолого-математическое моделирование

Из-за сложности геологических объектов ни одна математическая модель не может воспроизвести все их свойства. Поэтому для описания различных свойств одного и того же объекта часто приходится использовать различные математические модели.

Детерминированная модель – аналитическое представление закона, при котором для данной совокупности входных значений на выходе может быть получен единственный, всегда постоянный результат.

$y = f(x_1, x_2, \dots, x_k)$, где y – зависимая переменная (функция), $x_1 - x_k$ – независимы (аргументы).

Геолого-математическое моделирование

Стохастическая (статистическая, вероятностная) модель содержит случайный элемент ϵ , имеет вид $y = f(x_1, x_2, \dots, x_k) + \epsilon$. Если на входе задана некоторая совокупность значений, на выходе получаются близкие, но различающиеся между собой результаты. Различие их обуславливается влиянием случайных, неуправляемых воздействий неучтенных факторов. При характеристике результатов, полученных на основе использования таких моделей, говорят не о законе, а о *закономерности*.

Немного о теории вероятности

Теория вероятностей – математическая наука, позволяющая по вероятностям одних случайных событий находить вероятности других случайных событий, связанных каким-либо образом между собой.

Первичными понятиями в теории вероятности являются: событие, вероятность, случайная величина, статистическая устойчивость эксперимента.

Осуществление каждого отдельного наблюдения, опыта или измерения при изучении эксперимента называют *испытанием*. Результат испытания называют *событием*.

События принято обозначать A , B , C и т.д.

Немного о теории вероятности

Вероятность появления какого-то события прямо пропорционально числу m случаев, благоприятствующих появлению этого события, и обратно пропорционально числу n всех равновозможных случаев, могущих произойти при данном испытании.

$$p = m/n$$

Относительной частотой $P^*(A)$ события A в серии испытаний называется отношение m^* (числа испытаний, в которых появилось событие A) к n^* (общему числу проведенных испытаний), то есть

$$P^*(A) = m^*/n^*.$$

Из данного определения следует, что относительная частота случайного события всегда заключена между нулем и единицей:

$$0 \leq P^*(A) \leq 1.$$

Немного о теории вероятности

Статистической вероятностью $P(A)$ события A называется предел, к которому стремится относительная частота $P^*(A)$ при неограниченном увеличении числа испытаний, то есть

$$P(A) = \lim_{n \rightarrow \infty} P^*(A) = \lim_{n \rightarrow \infty} \frac{m^*}{n^*}.$$

Геологические данные, специфика геологических образований и процессов как объектов изучения.

В.И. Вернадский:

“Человечество закономерным движением, длившимся миллион - другой лет, со всеусиливающимся в своем проявлении темпом, охватывает всю планету, выделяется, отходит от других организмов как новая небывалая геологическая сила”

“...в геологической истории биосферы перед человеком открывается огромное будущее, если он поймет это и не будет употреблять свой разум и труд на самоистребление”

Геологические данные, специфика геологических образований и процессов как объектов изучения.

Геологические процессы (Г.П.) и образования обладают специфическими особенностями, в значительной мере определяющими методику их изучения:

- Г.П. представляют собой совокупность физических, химических и биологических природных явлений, между которыми существуют сложные причинно-следственные связи, поэтому свойства геологических образований зависят от множества факторов, характеризуются сильной изменчивостью, а сами объекты, как правило, имеют весьма сложное строение;
- Г.П. длительны, а геологические образования имеют значительные размеры и часто скрыты в недрах, что исключает возможность их полного и всестороннего изучения путем непосредственного наблюдения.

Геологические данные, специфика геологических образований и процессов как объектов изучения.

При решении своих задач геолог (геоэколог) располагает конечным числом наблюдений, характеризующих, как правило, незначительную часть изучаемого объекта.

Сложность, неоднородность объектов изучения наук о Земле заставляет рассматривать их как *природные системы*. Под *системой* понимается совокупность элементов, находящихся в отношениях и связях друг с другом, которая образует определенную целостность, единство.

Г.П. относятся к классу *динамических систем*, т.е. систем, изменяющих свое состояние во времени, а геологические образования, ввиду медленного протекания геологических процессов могут рассматриваться как *статические системы*.

Геологические данные, специфика геологических образований и процессов как объектов изучения.

Геологические данные делятся на количественные (признак характеризуется числом, например содержание U в пробе 3 г/т), полуколичественные (изучаемые объекты могут быть упорядочены по усилению какого-либо свойства, например, техногенных частиц нет – 0, мало – 1, много – 2), качественные (констатирует наличие или отсутствие признака, например, объектов ЯТЦ в радиусе 500 км нет – 0, есть – 1).

Математической обработке поддаются все указанные типы данных, и обоснованные геологические (геоэкологические) выводы могут быть получены даже по качественной информации при ограниченном числе наблюдений.

Некоторые типы геологических задач, решаемых математическими методами:

- оценка средних значений измеряемых признаков;
- характеристика изменчивости их;
- математическое описание распределения значений признаков на объектах изучения;
- установление характера и силы связи между признаками, отражающими специфичность неоднородности строения объектов и факторами, определяющими направленность протекания процессов, реализуемость явлений;
- математическое описание установленных корреляционных зависимостей;

Некоторые типы геологических задач, решаемых математическими методами:

- решение вопросов сходства – различия изучаемых объектов; процессов и явлений на основе сравнения средних значений, характеристик изменчивости, законов распределения измеряемых параметров, характера и тесноты корреляционных зависимостей между их значениями;
- установление закономерной и случайной составляющих изменчивости изучаемых параметров на линии, площади, в объеме;
- выбор наиболее информативных признаков и последующие классификация объектов изучения, выделение слабых сигналов на фоне случайных помех;

Некоторые типы геологических задач, решаемых математическими методами:

- построение карт комплексных показателей;
- оценка прогнозных ресурсов, техногенной нагрузки изучаемых площадей;
- выбор сети наблюдений;
- подсчет запасов на основе методов пространственно-статистического анализа;
- моделирование геологических (геоэкологических) явлений.

Геолого-математическое моделирование

Моделирование с целью познания процессов и явлений применяется при изучении систем, не поддающихся экспериментальным исследованиям и строгому описанию одновременно действующих многочисленных факторов.

Физические модели отражают подобие форм геометрических соотношений и происходящих в них физических процессов (пр-р: изучение процессов складкообразования наклоном плоскости, на которую нанесены слои песка).

Геометрические модели представляют собой объекты, геометрически подобные прототипу, дают внешнее представление (пр-р: слепки самородков, геологические, геохимические и геофизические карты).

Статистические характеристики используемые в геологии (геоэкологии)

Случайной величиной X называется такая величина, которая в результате эксперимента может принимать конкретное значение, заранее неизвестное.

Множество всех значений, которые может принимать случайная величина X в результате эксперимента, называется **генеральной совокупностью** случайной величины.

Множество значений случайной величины принимаемое в результате конкретного эксперимента называется **выборкой**.

Статистические характеристики используемые в геологии (геоэкологии)

Проиллюстрируем введенные понятия на простейшем примере случайной величины - бросании обычной игральной кости. В этом случае случайная величина X может принимать неизвестные до эксперимента значения $1, 2, \dots, 6$. Генеральной совокупностью такой случайной величины является множество $\{1, 2, 3, 4, 5, 6\}$. Если в результате эксперимента, заключающегося в трехкратном бросании игральной кости, выпали значения $2, 5, 5$, то множество $\{2, 5, 5\}$ называется выборкой случайной величины X .

Статистические характеристики используемые в геологии (геоэкологии)

Величины , которые могут принимать лишь отдельные (не обязательно целые значения), являются *дискретными*, а любые значения заданного интервала – *непрерывными*.

Число появления события в серии испытаний называется его *частотой*, а отношение числа появлений события к общему числу опытов в серии – *частотью*.

Функция распределения $F(x)$ выражает вероятность того, что выборочное значение случайной величины ξ окажется меньше некоторого предела, ограниченного x , где x – заданная переменная, т.е. вероятность события $\xi \leq x$.

Функция плотности распределения $f(x)$ характеризует вероятность попадания выборочного значения случайной величины ξ в заданный интервал от x до $x+\Delta x$, т.е. вероятность события $x < \xi < x+\Delta x$.

Статистические характеристики используемые в геологии (геоэкологии)

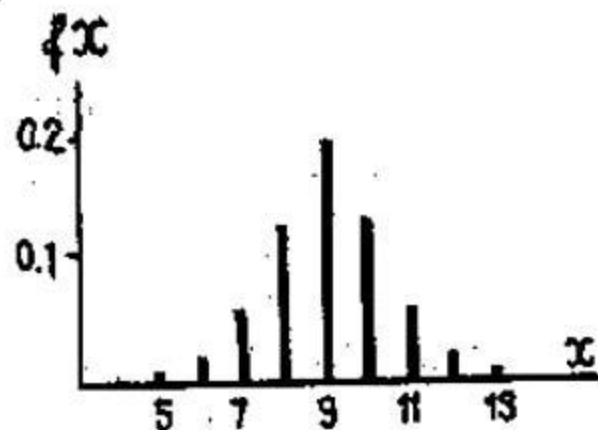
Величины , которые могут принимать лишь отдельные (не обязательно целые значения), являются *дискретными*, а любые значения заданного интервала – *непрерывными*.

Число появления события в серии испытаний называется его *частотой*, а отношение числа появлений события к общему числу опытов в серии – *частотью*.

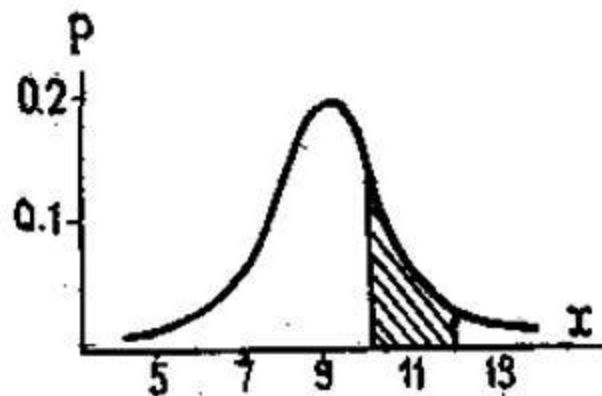
Функция распределения $F(x)$ выражает вероятность того, что выборочное значение случайной величины ξ окажется меньше некоторого предела, ограниченного x , где x – заданная переменная, т.е. вероятность события $\xi \leq x$.

Функция плотности распределения $f(x)$ характеризует вероятность попадания выборочного значения случайной величины ξ в заданный интервал от x до $x+\Delta x$, т.е. вероятность события $x < \xi < x+\Delta x$.

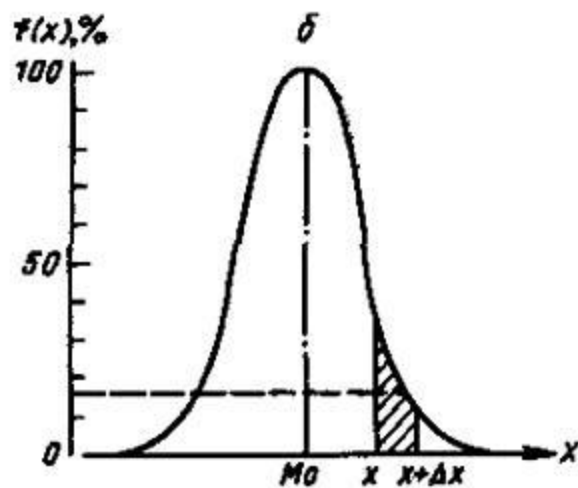
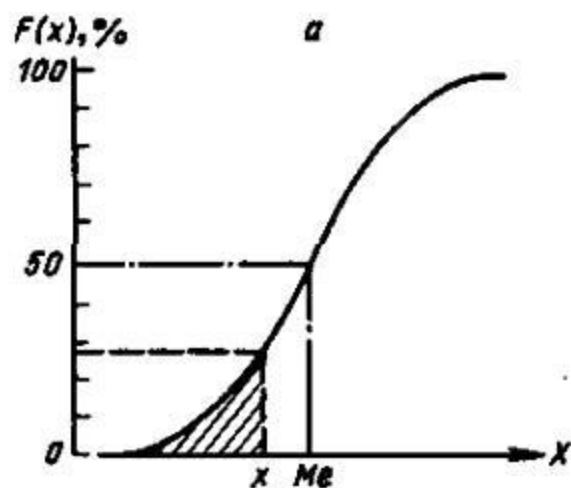
Статистические характеристики используемые в геологии (геоэкологии)



Дискретное распределение



Непрерывное распределение



Графики функций распределения:
 a — интегральная функция распределения; b — функция плотности распределения (дифференциальная функция распределения)

Статистические характеристики используемые в геологии (геоэкологии)

Для наглядного представления свойств случайной дискретной величины широко используются следующие виды графического представления, которые получаются по результатам наблюдения случайной величины.

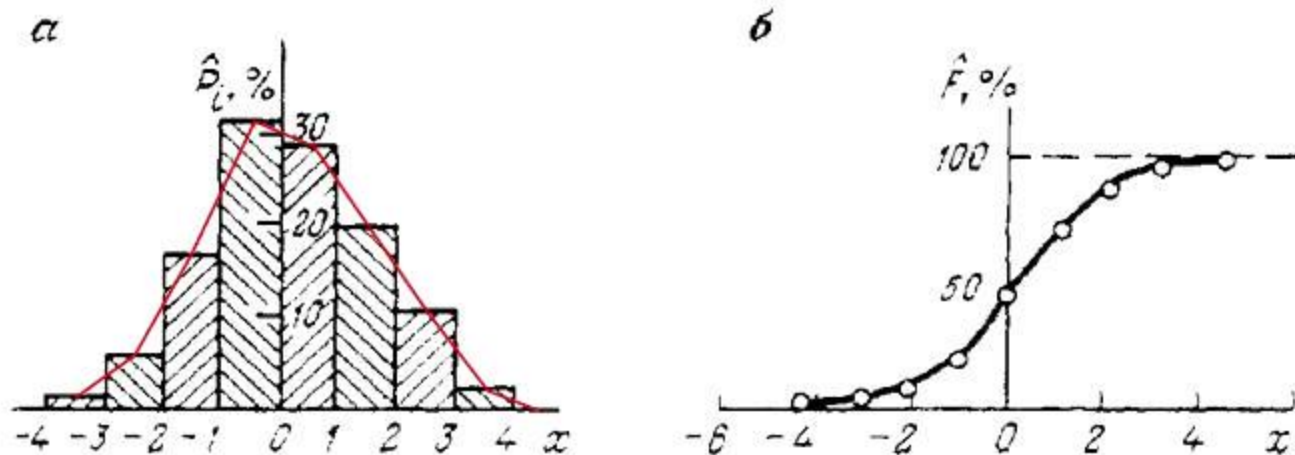
Гистограмма. Для ее построения в прямоугольной системе координат по оси абсцисс откладывают отрезки, изображающие интервалы значений случайной величины, и на этих отрезках, как на основании, строят прямоугольники с высотами, равными **частотам** $w_i = n_i/N$ соответствующего интервала (здесь – n_i число значений случайной величины попадающих в определенный интервал, N – общее число наблюдений случайной величины - **объем выборки**). В результате получают ступенчатую фигуру, состоящую из прямоугольников, которую и называют *гистограммой*.

Статистические характеристики используемые в геологии (геоэкологии)

Если соединить ломаной линией центры вершин прямоугольников составляющих гистограмму, то получится кривая, получившая название **полигон**.

Кумулятивная кривая (кривая накопленных частот) строится следующим образом. В прямоугольной системе координат строят точки с координатами $(x_i, \sum_{k \leq i} w_k)$.

Полученные точки соединяют отрезками.



Гистограмма и полигон (красная линия) (а) и кумулятивная кривая случайной величины (б)

Статистические характеристики используемые в геологии (геоэкологии)

Таблица . Частотное распределение содержания SiO_2 в неогеновых лавах

Содержание SiO_2 , %		Класс, от—до	Частота	Частость, %	Накопленная частость, %
целые числа	десятые доли				
56 57 58	6	56,0—58,9	1	3	3
59 60 61	5 5,7 2,2,1	59,0—61,9	6	20	23
62 63 64	9,4 7,1,6,8 6,6	62,0—64,9	8	27	50
65 66 67	8,3 8,3 8,7,5	65,0—67,9	7	23	73
68 69 70	2,3,9	68,0—70,9	3	10	83
71 72 73	6,4 5 2,2	71,0—73,9	5	17	100

Статистические характеристики используемые в геологии (геоэкологии)

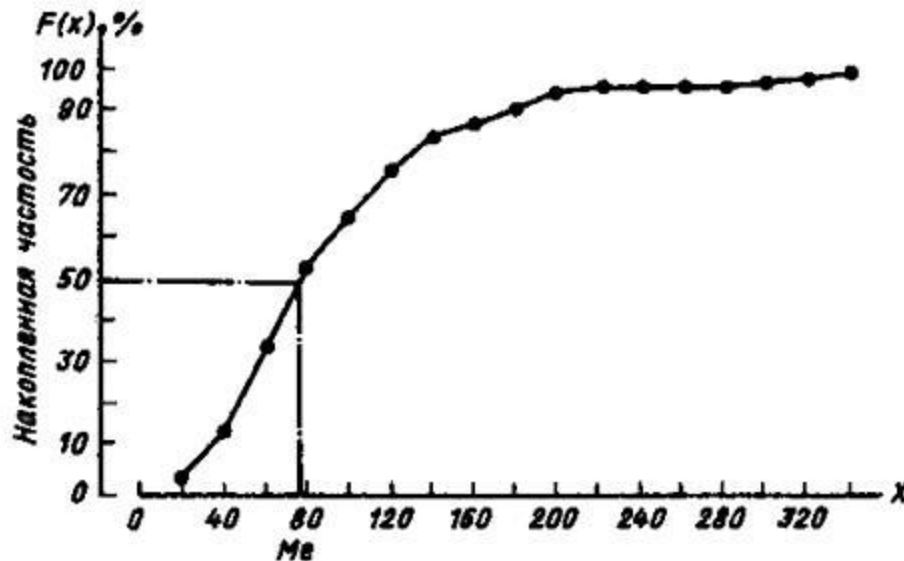
Необходимо отметить, что важнейшим параметром, определяющим вид гистограммы, кумулятивной кривой и полигона является величина интервала группирования результатов наблюдения (ширина основания прямоугольников, слагающих гистограмму). Для определения оптимального интервала, такого, при котором гистограмма не была бы слишком громоздкой и в то же время позволяла выявить характерные черты случайной величины, принято использовать **формулу Стэрджеса**:

$$\Delta x = \frac{x_{\max} - x_{\min}}{(1 + 4 \lg N)}$$

,где x_{\max} и x_{\min} — соответственно максимальное и минимальное значение в выборке случайной величины.

Статистические характеристики используемые в геологии (геоэкологии)

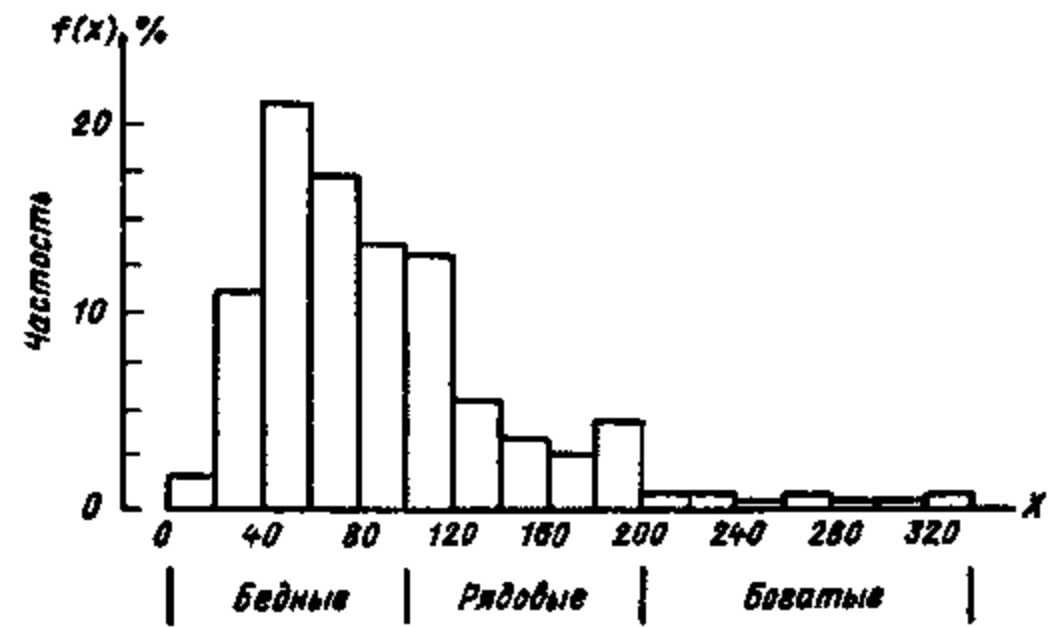
Изображение распределений выборочных данных в виде кумулянт используются при решении задач, связанных с поиском оптимальных границ геологических (геоэкологических) совокупностей. Пр-р: кумулянта содержаний полезного компонента в пробах позволяет оперативно оценить как будут меняться общие запасы месторождения при изменении требований к минимальной концентрации полезного компонента в руде.



Кумулянта частотного распределения содержания молибдена в руде.

Статистические характеристики используемые в геологии (геоэкологии)

Гистограммы можно применять для решения задач, связанных с разделением геологических (геоэкологических) совокупностей на несколько самостоятельных совокупностей по величине изучаемого свойства.



Гистограмма частотного распределения содержания молибдена в руде.

Статистические характеристики используемые в геологии (геоэкологии)

Наиболее существенные особенности распределения случайной величины могут быть выражены с помощью **числовых характеристик положения и разброса.**

К важнейшим характеристикам положения относятся **математическое ожидание, мода и медиана.**

Математическое ожидание M_x (средним по распределению) дискретной (прерывной) случайной величины X называют сумму произведений всех возможных значений случайной величины на соответствующие им вероятности.

$$M(X) = x_1 p_1 + x_2 p_2 + \dots + x_k p_k = \sum_{i=1}^k x_i p_i$$

Статистические характеристики используемые в геологии (геоэкологии)

Учитывая предыдущие записи, и что $\sum_{i=1}^k p_i = 1$, иногда пишут:

$$M(X) = \sum_{i=1}^k x_i p_i / \sum_{i=1}^k p_i.$$

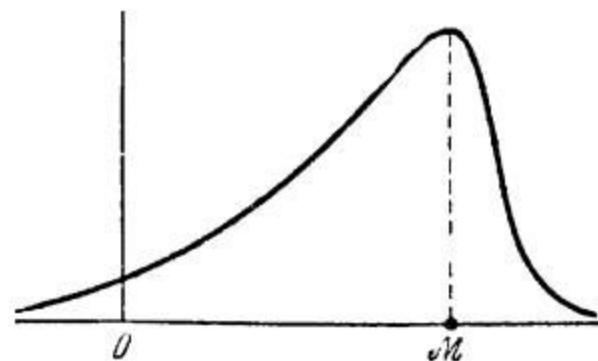
Математическим ожиданием непрерывной случайной величины X называется интеграл:

$$M(X) = \int_{-\infty}^{\infty} x f(x) dx$$

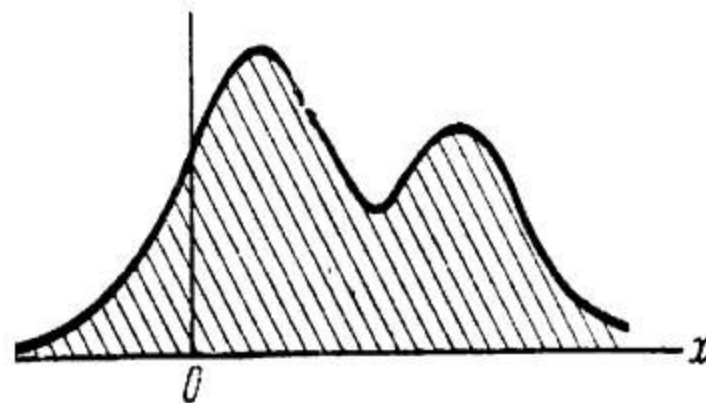
Математическое ожидание характеризует среднее значение случайной величины и поэтому его оценки широко используются для количественного описания весьма изменчивых свойств геологических объектов.

Статистические характеристики используемые в геологии (геоэкологии)

Модой случайной величины X называется такое ее значение x_{Mo} при котором, функция плотности распределения вероятностей максимальна $f(x_{Mo}) = \max$.

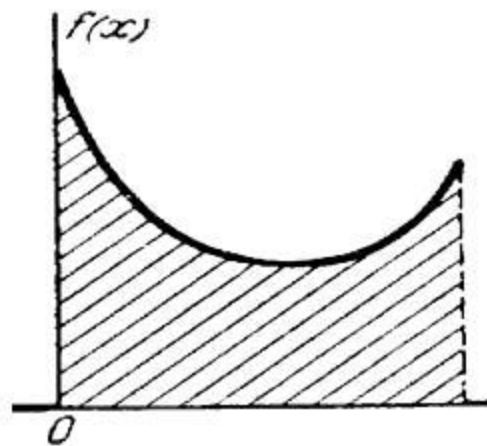


Если функция $f(x)$ имеет несколько выраженных максимумов, то такое распределение называется **полимодальным**.



Статистические характеристики используемые в геологии (геоэкологии)

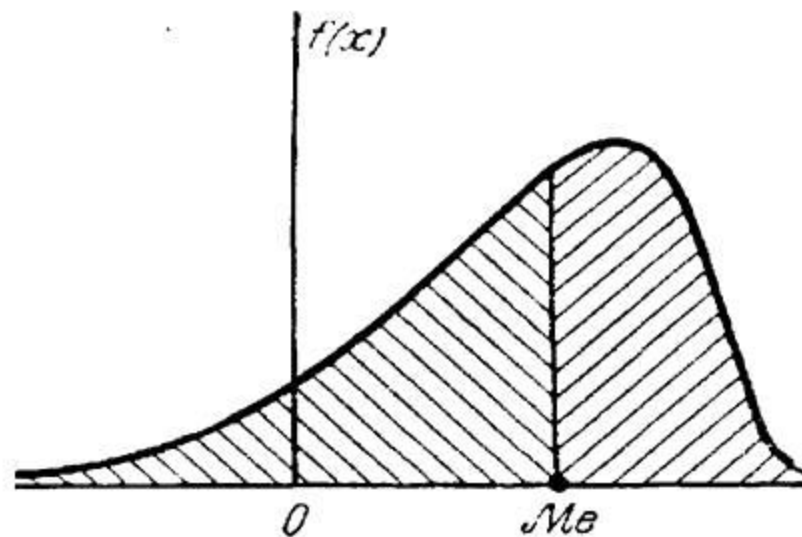
Функция, плотность распределения вероятностей которой не имеет максимумов, называется **антимодальным**.



Для случайной дискретной величины мода соответствует такому значению случайной величины, которое наблюдалось максимальное число раз (соответствующее максимуму гистограммы).

Статистические характеристики используемые в геологии (геоэкологии)

Медианой случайной величины X называется такое значение x_{Me} при котором вероятность принятия случайной величиной значений меньших x_{Me} равна вероятности принятия случайной величиной значений больших x_{Me} и равна 0.5, то есть $P(X < x_{Me}) = P(X > x_{Me}) = 0.5$.



Статистические характеристики используемые в геологии (геоэкологии)

*Характеристиками разброса, определяющими степень отклонения значений случайной величины от ее математического ожидания служат **размах варьирования** (т.е. интервал возможных значений случайной величины) и **центральные моменты** различных порядков.*

Размах варьирования R , равен разности между наибольшим и наименьшим значениями в выборке, то есть:

$$R = x_{\max} - x_{\min}$$

Размах - приближенный показатель изменчивости, так как часто почти не зависит от изменения выборки, а крайние значения случайной величины, которые используются для его вычисления, как правило, ненадежны.

Статистические характеристики используемые в геологии (геоэкологии)

Более содержательными являются меры рассеяния наблюдений вокруг средних величин. **Средним линейным отклонением d** называют среднюю арифметическую абсолютных величин отклонений результатов наблюдений от их средней арифметической, то есть:

$$d = \frac{\sum_{i=1}^n |x_i - \bar{X}|}{n}$$

Центральным моментом порядка s случайной величины X называется математическое ожидание s -ой степени соответствующей **центрированной** случайной величины.

$$M(\bar{X}^s) = M(X - M(X))^s$$

здесь $M(X)$ – математическое ожидание случайной величины X .

Статистические характеристики используемые в геологии (геоэкологии)

Для случайной дискретной величины s -й центральный момент выражается суммой

$$M(X^s) = \sum_{i=1}^n (x_i - M(X))^s p_i$$

здесь - $M(X)$ – математическое ожидание случайной величины X ;

$x_1, x_2, \dots, x_n, \dots$ - значения, принимаемые случайной величиной;

$p_1, p_2, \dots, p_n, \dots$ - соответствующие им вероятности;

а для непрерывной – интегралом:

$$M(\bar{X}^s) = \int_{-\infty}^{\infty} (x - M(X))^s f(x) dx$$

- здесь $f(x)$ – функция плотности распределения случайной величины.

Статистические характеристики используемые в геологии (геоэкологии)

Главной характеристикой разброса случайной величины служит центральный момент второго порядка – *дисперсия*.

$$D(x) = \sum_{i=1}^n (x_i - M(X))^2 p_i$$
$$D(X) = \int_{-\infty}^{\infty} (x - M(X))^2 f(x) dx$$

-соответственно для дискретной и случайной непрерывной величины.

Дисперсия случайной величины есть характеристика *рассеивания*, разбросанности значений случайной величины около ее математического ожидания. Само слово «дисперсия» означает «рассеивание».

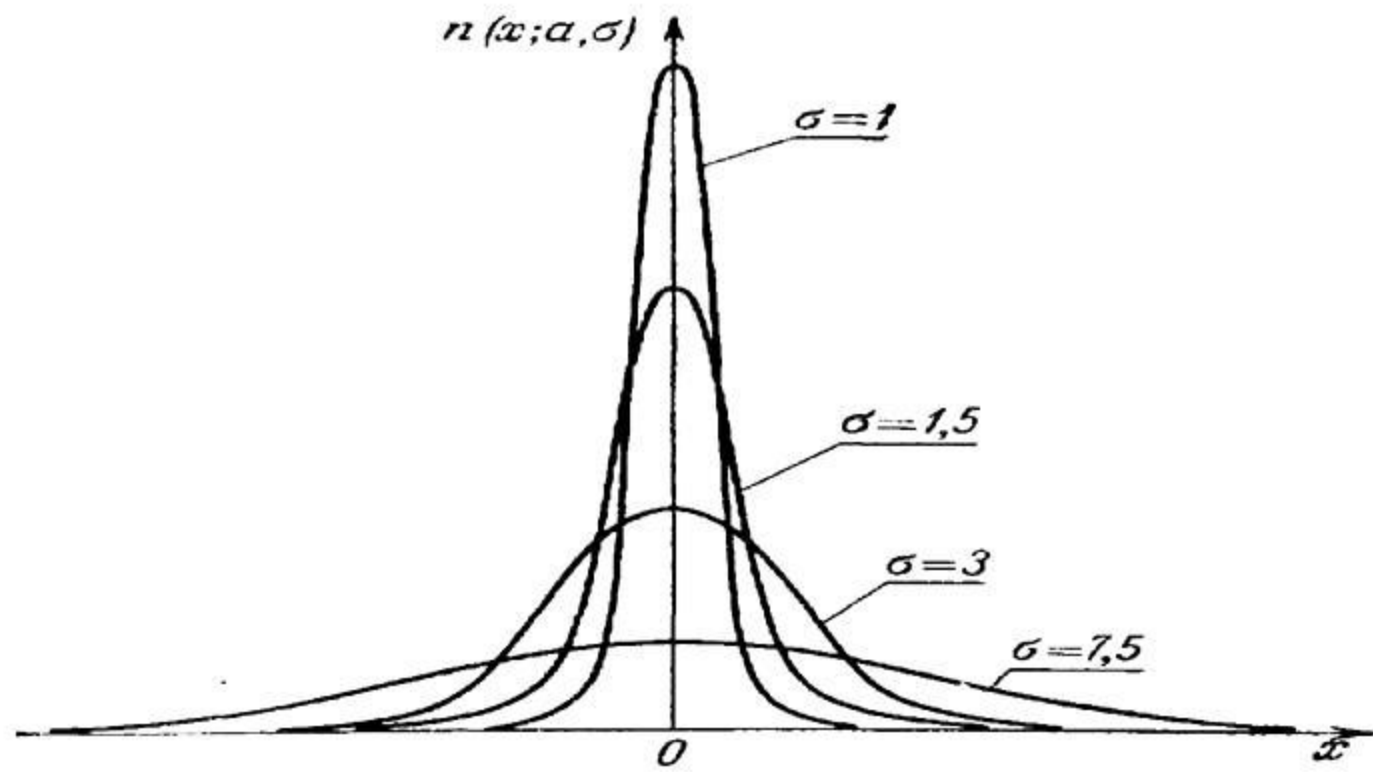
Статистические характеристики используемые в геологии (геоэкологии)

Дисперсия случайной величины имеет размерность квадрата случайной величины; для наглядной характеристики рассеивания удобнее пользоваться величиной, размерность которой совпадает с размерностью случайной величины. Для этого из дисперсии извлекают квадратный корень.

Среднеквадратическим (стандартным) отклонением

$\sigma_x = \sqrt{D(X)}$ случайной величины X называется арифметическое значение корня квадратного из ее дисперсии.

Статистические характеристики используемые в геологии (геоэкологии)



Функция плотности вероятности случайной величины для различных значений ее дисперсии

Статистические характеристики используемые в геологии (геоэкологии)

Коэффициент вариации $V = (\sigma/M_x) \times 100\%$ - величина безразмерная, поэтому его применяют в тех случаях, когда необходимо сравнить по степени изменчивости свойства, выраженные в разных единицах измерений, например, мощность рудного тела и содержание в нем полезного компонента.

Статистические характеристики используемые в геологии (геоэкологии)

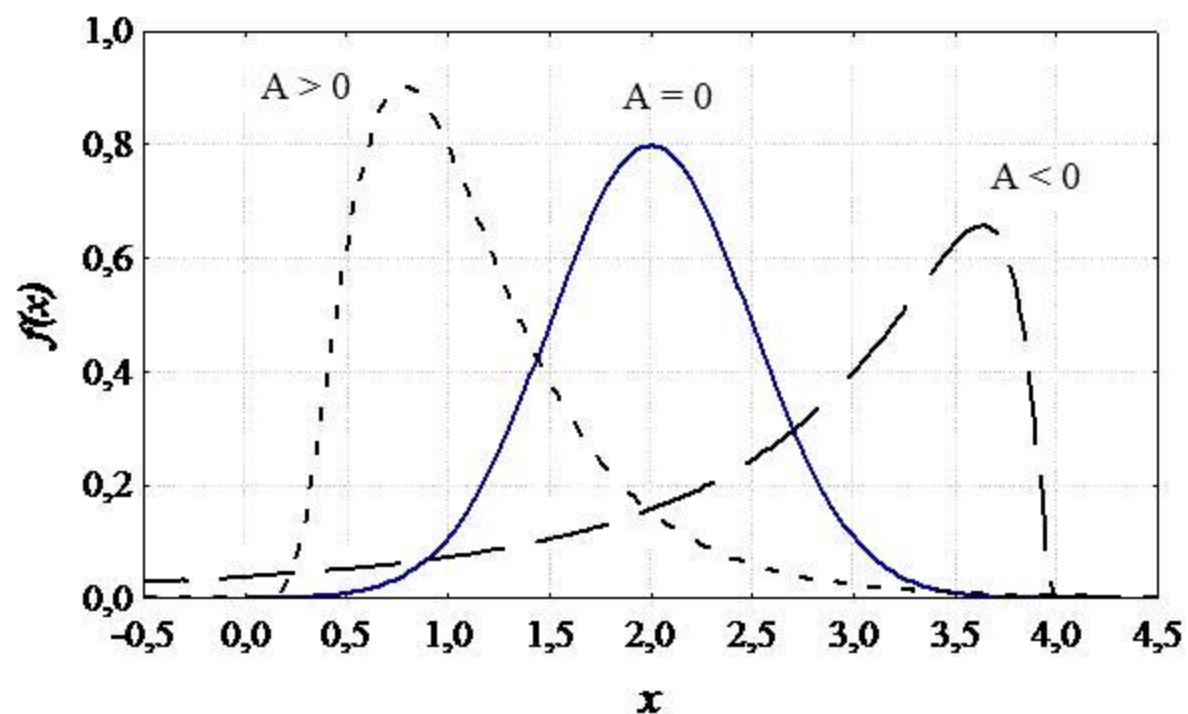
Третий центральный момент служит для характеристики **асимметрии** (или «скошенности») распределения. Если распределение симметрично относительно математического ожидания, то все моменты нечетного порядка (если они существуют) равны нулю:

$$A(x) = \sum_{i=1}^n (x_i - M(X))^3 p_i$$
$$A(X) = \int_{-\infty}^{\infty} (x - M(x))^3 f(x) dx$$

- соответственно для дискретной и случайной непрерывной величины.

Статистические характеристики используемые в геологии (геоэкологии)

Коэффициент асимметрии характеризует степень асимметрии распределения случайной величины относительно ее математического ожидания. Для симметричных распределений $A = 0$. Если пик графика функции $f(x)$ смещен в сторону малых значений («хвост» на графике функции $f(x)$ справа), то $A > 0$. В противном случае $A < 0$ (см. рис.).



Статистические характеристики используемые в геологии (геоэкологии)

Мерой остроты графика функции плотности распределения (**эксцесса**) служит центральный момент четвертого порядка:

$$E(x) = \sum_{i=1}^n (x_i - M(X))^4 p_i - 3$$

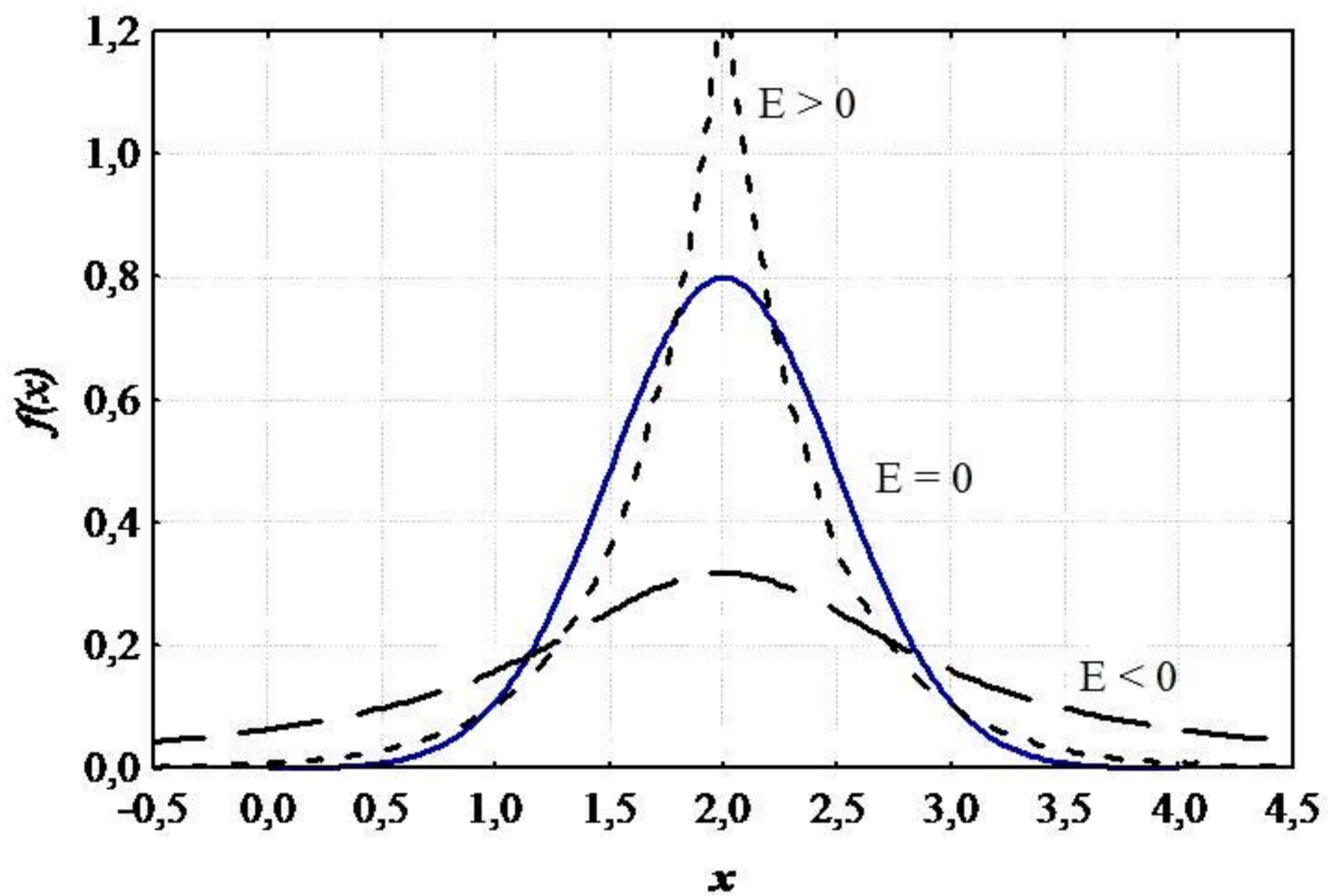
$$E(X) = \int_{-\infty}^{\infty} (x - M(x))^4 f(x) dx - 3$$

-соответственно для дискретной и случайной непрерывной величины.

Число 3 вычитается из отношения потому, что для весьма важного и широко распространенного в природе нормального закона распределения (которое рассматривается в дальнейшем) $E(X)=3$.

Статистические характеристики используемые в геологии (геоэкологии)

Коэффициент эксцесса является мерой остроты графика функции плотности распределения $f(x)$ (рис.).



Статистические характеристики используемые в геологии (геоэкологии)

Асимметрия и эксцесс имеют размерность куба и четвертой степени случайной величины, поэтому чтобы получить безразмерную характеристику третий момент делят на куб среднего квадратичного отклонения (показатель асимметрии), а эксцесс на среднее квадратичное отклонение возведенное в четвертую степень (показатель эксцесса):

$$A = A(x)/\sigma^3$$

$$E = E(x)/\sigma^4$$

Оценки моментов третьего и четвертого порядка используются для решения вопроса о соответствии выборочных данных определенному типу распределения.

Статистические характеристики используемые в геологии (геоэкологии)

При оценке характеристик положения и разброса по сгруппированным выборочным данным вероятности (p_i) в формулах заменяются частотами попадания значений в каждый интервал группирования (n_j), вместо значений величины x подставляются значения центров интервалов группирования (\bar{x}_j), вместо математического ожидания Mx – его оценка (\bar{x}), а операция интегрирования для непрерывных величин заменяется суммированием.

Таким образом, формулы для расчета оценка математического ожидания (\bar{x}), дисперсии (S^2), показателя асимметрии (A) и эксцесса (E) для сгруппированных и несгруппированных данных принимают следующий вид (см. следующий слайд):

Статистические характеристики используемые в геологии (геоэкологии)

$$\bar{x} = \frac{\sum_{j=1}^k n_j \hat{x}_j}{n} = \frac{\sum_{i=1}^n (x_i)}{n};$$

$$S^2 = \frac{\sum_{j=1}^k n_j (\hat{x}_j - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1};$$

$$A = \frac{\sum_{j=1}^k n_j (\hat{x}_j - \bar{x})^3}{nS^3} = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{nS^3};$$

$$E = \frac{\sum_{j=1}^k n_j (\hat{x}_j - \bar{x})^4}{nS^4} - 3 = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{nS^4} - 3,$$

где k – количество классов группировки.

Статистические характеристики используемые в геологии (геоэкологии)

Математическое ожидание, мода, медиана, начальные и центральные моменты и, в частности, дисперсия, среднее квадратичное отклонение, асимметрия и эксцесс представляют собой наиболее употребительные числовые характеристики случайных величин.

Основные статистические законы распределения, используемые в геологии (геоэкологии)

Для аппроксимации (приближенного описания) эмпирических свойств геологических объектов в большинстве случаев можно ограничиться *нормальным и логарифмическим нормальным* для непрерывных величин, *биномиальным распределением и распределением Пуассона* для дискретных величин.

Основные статистические законы распределения, используемые в геологии (геоэкологии)

1. Биномиальное распределение. Вероятность появления события A m раз при n независимых испытаниях определяют формулой:

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

, где p – вероятность появления события A при отдельном испытании, q – вероятность не появления этого события, равная $1-p$; $n!$ (n – факториал) есть произведение всех натуральных чисел от 1 до n включительно, причем $0! = 1$. Число появлений события будет случайной величиной, принимающей значения $m = 0, 1, 2, \dots, n$ с соответствующими вероятностями $P_{m,n}$.

Основные статистические законы распределения, используемые в геологии (геоэкологии)

$M(x) = a = np$, т.е. математическое ожидание числа проявлений события в независимых испытаниях равно произведению числа испытаний на вероятность появления события в отдельном испытании.

$D(m) = \sigma^2 = npq$, т.е. дисперсия числа проявлений события m в n независимых испытаниях равна произведению числа испытаний на вероятность появления и не появления события в отдельном испытании.

*Основные статистические законы распределения,
используемые в геологии (геоэкологии)*

График биномиального распределения симметричен только при $p = q = 0.5$

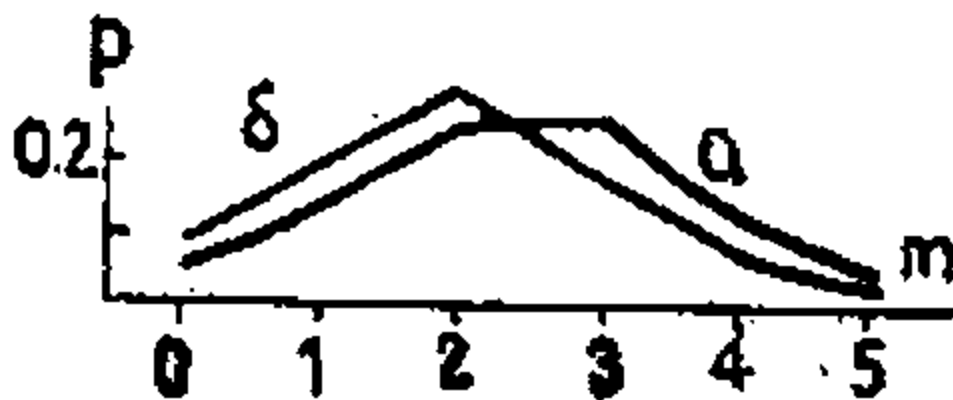


Рис. Биномиальное распределение (a) $p = q$; (б) $p < q$

Основные статистические законы распределения, используемые в геологии (геоэкологии)

2. Распределение Пуассона является предельным случаем биномиального, когда вероятности появления событий p и q очень малы, а число n испытаний достаточно большое. Малость p или q определяется тем, что произведение np при изменении n практически остается неизменным.

Основные статистические законы распределения, используемые в геологии (геоэкологии)

Функция распределения вероятностей Пуассона имеет вид:

$$P_m = \frac{\lambda^m \times e^{-\lambda}}{m!},$$

где m – число появлений события, принимающее значения 0, 1, 2, ..., n раз; λ – параметр распределения, равный произведению числа испытаний на вероятность появления события при отдельном испытании; т.е. $\lambda = n \times p$; P_m – вероятность того, что событие появится m раз. Для практического пользования существуют таблицы значений P_m при различных значениях λ .

*Основные статистические законы распределения,
используемые в геологии (геоэкологии)*

$M(x) = \lambda$, т.е. математическое ожидание случайной величины, распределенной по закону Пуассона, равно параметру распределения λ .

$D(x) = \lambda$, т.е. дисперсия числа появлений события равна параметру распределения λ . Чем больше λ , тем больше рассеяние случайной величины. Равенство $D(x) = P(x)$ служит критерием опознания такого распределения.

Основные статистические законы распределения, используемые в геологии (геоэкологии)

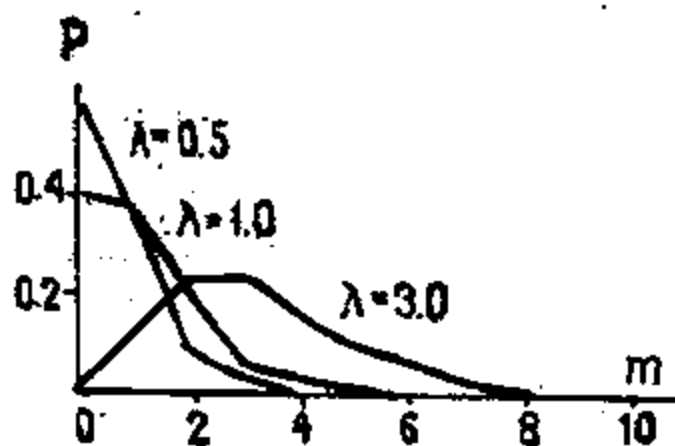


Рис. Кривые распределения Пуассона для различных значений λ .

График распределения Пуассона всегда ассиметричен с $\alpha_1 > 0$. При $\lambda \geq 9$ он приближается к симметричному, а распределение Пуассона может быть заменено нормальным.

Основные статистические законы распределения, используемые в геологии (геоэкологии)

3. Нормальное распределение (или Гауссово)
непрерывное, полностью определяемое двумя параметрами – математическим ожиданием Mx и дисперсией σ^2 :

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

Соответственно функция плотности распределения $f(x)$ будет иметь вид:

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

Нормальное распределение симметрично относительно математического ожидания, медиана и мода совпадают. Его асимметрия и эксцесс, равны нулю.

Основные статистические законы распределения, используемые в геологии (геоэкологии)

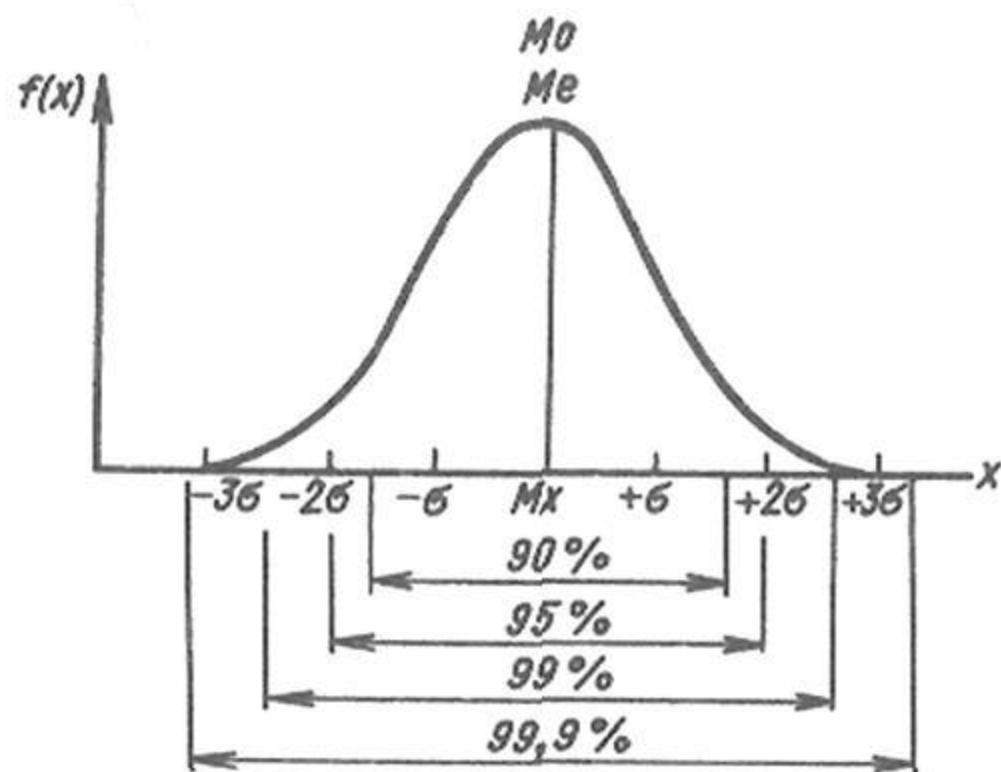


Рис. График функции плотности нормального распределения.

Основные статистические законы распределения, используемые в геологии (геоэкологии)

Нормальное распределенные случайные величины характеризуют такие свойства природных (геологических) объектов, которые зависят от очень большого количества независимых факторов, когда влияние каждого из них равномерное и незначительное.

Пр-р: близки к нормальному распределения содержания породообразующих минералов и входящих в их состав химических элементов, содержания некоторых полезных компонентов в рудах, когда они ***составляют целые проценты.***

Основные статистические законы распределения, используемые в геологии (геоэкологии)

4. Логарифмически нормальное распределение.

Значения случайной величины X распределены логарифмически нормально, если логарифмы этих значений распределены нормально. Функция распределения $F(X)$ такой случайной величины определяется выражением:

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

соответственно плотность вероятности определяется выражением:

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

Основные статистические законы распределения, используемые в геологии (геоэкологии)

Логарифмически нормальное распределение является положительно асимметричным и имеет положительный эксцесс. Математическое ожидание, мода и медиана логнормально распределенной случайной величины не совпадают, причем $M_0 < M_e < M(x)$.

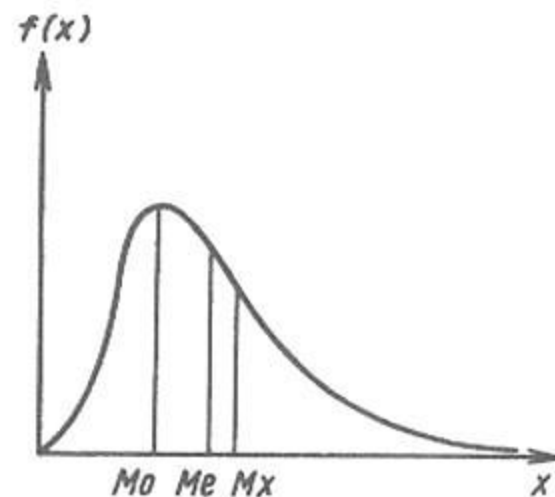


Рис. График функции плотности логнормального распределения

Основные статистические законы распределения, используемые в геологии (геоэкологии)

Логнормальное распределение связано с воздействием на исследуемое свойство множества факторов. ***Однако сила воздействия каждого фактора не одинакова, а пропорциональна изменению силы проявления данного фактора.***

Бликие к логарифмическому нормальному распределения характерны для свойств природных (геологических) объектов, изменчивость которых ограничена определенным жестко заданным интервалом, а их математическое ожидание смещено от середины этого интервала в сторону малых значений.

Основные статистические законы распределения, используемые в геологии (геоэкологии)

Хотя содержание любого элемента в горной породе или почве не может быть больше 0 и больше 100 %, фактический интервал варьирования конкретных химических элементов значительно уже. Например, среднее содержание (кларк) урана в почвах 2-3 г/т. Однако на участках воздействия предприятий по добычи и переработке руд этого элемента содержание U по отдельным пробам может достигать высоких значений.

В литосфере в целом размах варьирования содержаний микро- и редких элементов в сторону больших значений ($x_{min} - M(x)$) в несколько миллионов раз больше, чем в сторону малых значений ($M(x) - x_{min}$)).

Статистические гипотезы и критерии их проверки

К основным задачам математической статистики относится статистическая проверка гипотез о законах распределения и о параметрах распределения случайной величины.

При исследовании различных случайных величин на определённом его этапе появляется возможность выдвинуть ту или иную гипотезу о свойствах изучаемой величины, например, сделать предположение о законе распределения её, или, если закон распределения известен, но неизвестны его параметры, то сделать предположение о их величине.

Статистические гипотезы и критерии их проверки

Наиболее правдоподобную по каким - то соображениям гипотезу называют нулевой (основной) и обозначают H_0 . Наряду с основной гипотезой рассматривают другую (альтернативную) гипотезу H_1 , противоречащую основной. Выдвинутая нулевая гипотеза нуждается в дальнейшей проверке. При этом могут быть допущены ошибки двух типов:

- ошибка первого рода - отвергнута правильная гипотеза;
- ошибка второго рода - принята неправильная гипотеза.

Статистические гипотезы и критерии их проверки

Вероятность совершить ошибку первого рода (вероятность отвергнуть правильную гипотезу) обычно обозначают α и называют **уровнем значимости**. Случайную величину Z , служащую для проверки гипотезы, называют **критерием**. Совокупность значений критерия, при которых нулевую гипотезу отвергают, называют **критической областью**. Граничные точки критической области $z_{кр}$ называют **критическими точками**.

Статистические гипотезы и критерии их проверки

Статистические критерии согласия разделяются на параметрические и непараметрические.

Параметрические критерии выводятся из свойств тех или иных статистических законов распределения и могут использоваться лишь в том случае, если распределение выборочных данных согласуется с этим законом.

Непараметрические критерии могут применяться даже в том случае, если закон распределения изучаемых величин не известен или их закон не соответствует ни какому из известных законов.

Непараметрические критерии обладают несколько меньшей мощностью по сравнению с параметрическими аналогами, но область их применения значительно шире.

Проверка гипотез о законе распределения параметров природных (геологических) объектов

Проверка гипотезы о соответствии эмпирических распределений нормальному или логнормальному закону с помощью **критерия Пирсона** (χ^2).

Этот способ заключается в разделении выборочных данных на k класс-интервалов и сравнения эмпирических частот по классам (n_j) с теоретическим (n_j^l) для нормального распределения.

Способ проверки гипотез о законе распределения с помощью критерия χ^2 обычно применяется в том случае, когда объем выборки превышает 60 значений.

Проверка гипотез о законе распределения параметров природных (геологических) объектов

Критерий χ^2 определяют по формуле:

$$\chi^2 = \sum (n_i - \tilde{n}_i)^2 / \tilde{n}_i,$$

где n_i - эмпирическая частота; \tilde{n}_i - теоретическая частота. Если $\chi^2_{\text{эмп}} < \chi^2_{\alpha}(f)$, гипотеза о согласии эмпирического и теоретического распределения не отвергается. Число степеней свободы определяется в зависимости от применяемого теоретического закона. Для нормального закона $f = k - 3$ (k - число классов группировки), для закона Пуассона $f = k - 2$.

Проверка гипотез о законе распределения параметров природных (геологических) объектов

Наряду с критерием Пирсона, основанным на сравнении эмпирических и теоретических частот, применяется также **критерий Колмогорова – Смирнова (критерий λ)**, основанный на сравнении накопленных частот.

Для сравнения эмпирического распределения с теоретическим критерий λ определяют по формуле:

$$\lambda = D/\sqrt{n},$$

где $D = |\tilde{N}_i - N_i|_{\max}$ – наибольшее значение абсолютной разности между накопленными значениями абсолютной разности между накопленными значениями частот эмпирического и теоретического распределения.

Проверка гипотез о законе распределения параметров природных (геологических) объектов

Теоретическое значение λ не зависит от объема выборки и числа степеней свободы, а определяется только выбранным уровнем значимости. Для $\alpha = 0.05$; $\lambda = 1,36$; $\alpha = 0.01$; $\lambda = 1,63$.

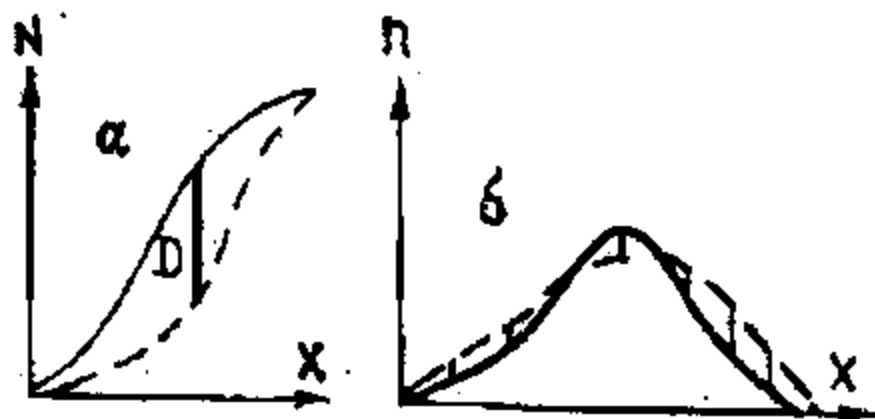


Рис. Графическая интерпретация существа критериев (сравниваемых характеристик распределения) λ (а) и χ^2 (б)

Проверка гипотез о законе распределения параметров природных (геологических) объектов

Для **выборок меньшего объема** в качестве критерия соответствия эмпирического распределения нормальному теоретическому используют также отношения коэффициентов асимметрии A и эксцесса E к их стандартным отклонениям σ_A и σ_E соответственно:

$$t_1 = \frac{A}{\sigma_A}, \quad t_2 = \frac{E}{\sigma_E}$$

Если эти отношения по абсолютной величине превышают 3, то гипотеза о нормальном распределении отвергается.

$\sigma_A \cong \sqrt{6/n}$, $\sigma_E \cong \sqrt{24/n}$ (где n – количество замеров в выборке).

Сравнение дисперсий двух выборочных совокупностей

Отметим одно важнейшее в математической статистике распределение случайной величины, которое непосредственно связано с оценками дисперсии, а именно определяемое как отношение двух несмещенных оценок дисперсий, полученных из независимых выборок, взятых из нормальных совокупностей:

$$F = \frac{D_1(X)}{D_2(X)}$$

Обычно, в качестве числителя, берут большую из двух несмещенных оценок дисперсии. Распределение такой случайной величины называется ***F-распределением (или распределением Р.Фишера)***. *F*-распределение зависит от двух параметров (степеней свободы) $f_1 = n_1 - 1$ и $f_2 = n_2 - 1$. Здесь n_1 и n_2 соответственно объемы выборок, на основании которых оценивались дисперсии D_1 и D_2 .

Сравнение дисперсий двух выборочных совокупностей

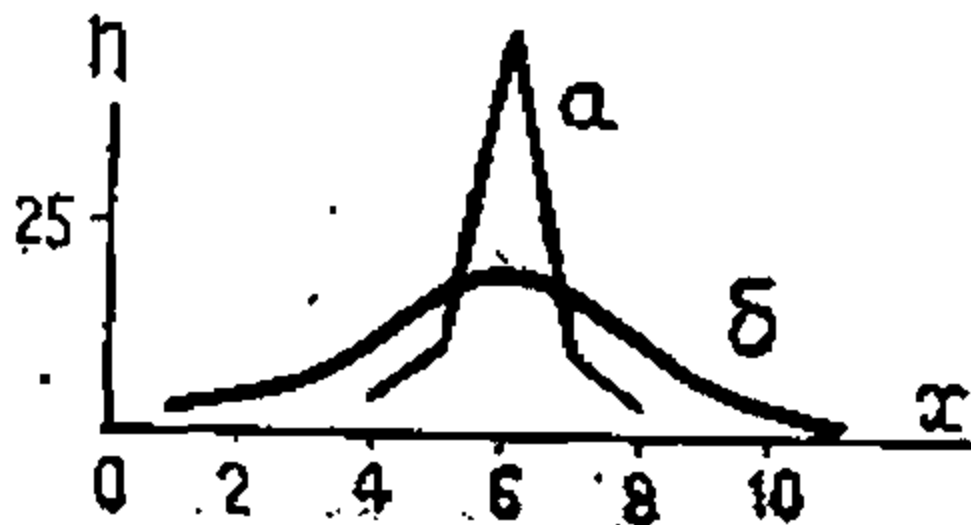


Рис. Ряды с одинаковыми средними и различными дисперсиями: $S^2(a) < S^2(б)$

Сравнение математических ожиданий двух выборочных совокупностей

Наиболее часто в геологической (геоэкологической) практике употребляют параметрический критерий Стьюдента t . Его применение основано на том, что если из **нормально** распределенной совокупности отобраны выборки X_1, X_2, \dots, X_k объемом в n_1 значений и выборки Y_1, Y_2, \dots, Y_k объемом n_2 значений, то величина

$$T_{\text{ст}} = \frac{\bar{x} - \bar{y}}{\sqrt{s_x^2/n_x + s_y^2/n_y}}$$

, где \bar{x}, \bar{y} – выборочные оценки среднего, а S_x^2, S_y^2 – выборочные оценки дисперсии, подчиняются закону распределения Стьюдента с $n_1+n_2 - 2$ степенями свободы.

Сравнение дисперсий двух выборочных совокупностей

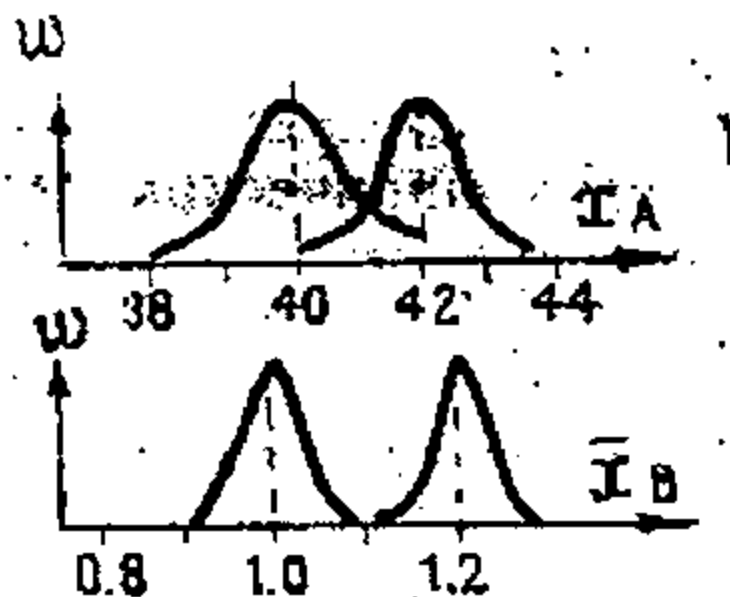


Рис. Графическое представление различий средних содержаний элементов А и Б

Сравнение математических ожиданий двух выборочных совокупностей

В случае если закон распределения не определен используют непарметрические критерии, которые особенно полезны для малых выборок.

Рассмотрим в качестве примера ***U*** - критерий Манна-Уитни для проверки гипотезы H_0 об однородности двух выборок, представляющий непарметрическую альтернативу t – критерию Стьюдента для независимых выборок.

U- критерий Манна-Уитни предполагает, что все значения по обеим выборкам случайных величин X и Y объемов n и m соответственно, ранжируются, то есть записываются в один ряд в порядке возрастания.

Сравнение математических ожиданий двух выборочных совокупностей

После этого каждый элемент выборок характеризуется рангом – порядковым номером каждого элемента выборок в общем ранжированном ряду из обеих выборок. Наблюдаемое значение критерия U вычисляется по формуле

$$U = W - \frac{1}{2}m(m+1) = \sum_{i=1}^n \sum_{j=1}^m \delta_{ij},$$

где W - значение критерия Уилкоксона, численно равное сумме рангов элементов второй выборки (объема m) в общем ранжированном ряду,

$$\delta_{ij} = \begin{cases} 1, & \text{если } X_i < Y_j \\ 0, & \text{если } X_i > Y_j \end{cases}$$

Сравнение математических ожиданий двух выборочных совокупностей

Распределение случайной величины U асимптотически нормально с параметрами $M[U] = nm/2$ и $D[U] = nm(n+t+1)/12$, чем и пользуются на практике, если $\min\{n,t\} > 25$, для определения критического значения $U_{кр}(\alpha, n, t)$, соответствующего заданному уровню значимости α . Для случаев, когда n и $t < 25$, пользуются специальными таблицами.

Однофакторный и двухфакторный дисперсионный анализ

Свойства сложных природных систем, часто зависят от ряда факторов, обуславливающих их изменчивость. Выявление этих факторов и оценка степени их влияния на изменчивость (неоднородность) свойств изучаемых объектов осуществляется с помощью **дисперсионного анализа**.

Этот статистический метод основан на следующем принципе: если на случайную величину действуют взаимонезависимые факторы A, B, \dots, D , то общую дисперсию этой случайной величины σ^2 можно рассматривать как сумму дисперсий $\sigma^2 = \sigma^2_A + \sigma^2_B + \dots + \delta^2_D$.

Однофакторный и двухфакторный дисперсионный анализ

С помощью дисперсионного анализа решается широкий круг геологических (геоэкологических) задач:

- установить воздействие предприятия металлургии на изменения содержания элемента Сг в почвах (один дискретный фактор – воздействие предприятия, может изменяться на уровнях: 1- СЗЗ, 2 – километровая зона, 3 – десяти километровая зона);
- определить влияние веса пробы и способа ее отбора на изменение содержания изучаемого компонента. Фактор А – способ отбора пробы дискретный, может варьировать на уровнях: 1 – точечный; 2 - методом конверта. Фактор В – вес пробы – интервальный (непрерывный). Уровнями его могут быть: 1 – вес до x_1 гр; 2 – вес от x_1 гр до x_2 гр; 3 – вес более x_2 гр и т.д.

Однофакторный и двухфакторный дисперсионный анализ

С помощью дисперсионного анализа анализа решается широкий круг геологических (геоэкологических) задач:

- установить влияние степени измельченности материала, методов сжигания, изменения силы тока, отрезка времени рабочего дня и исполнителя. Всего пять факторов, из них первый, третий, четвертый принимают непрерывные значения (интервальные), второй и пятый – дискретные, каждый может варьировать на нескольких уровнях.

И многие другие задачи.

Однофакторный и двухфакторный дисперсионный анализ

По количеству оцениваемых факторов дисперсионный анализ подразделяется на одно-, двух- и многофакторный.

Каждый фактор представляет собой переменную дискретную или непрерывную величину, которая разделяется на определенное количество постоянных интервалов (уровней). Если количество замеров изучаемой случайной величины на всех уровнях по всем факторам одинаково, дисперсионный анализ принято называть *равномерным*, а если разное – *неравномерным*.

Однофакторный и двухфакторный дисперсионный анализ

При равномерном однофакторном дисперсионном анализе случайной величины x относительно фактора A , имеющего k уровней (номер пробы) при количестве замеров на каждом уровне равном n , результаты наблюдений обозначаются как x_{ij} где i – номер наблюдения ($i = 1, 2, \dots, n$), а j – номер уровня фактора ($j = 1, 2, \dots, k$).

Номер изменения	Уровень фактора			
	A_1	A_2	...	A_k
1	x_{11}	x_{12}	...	x_{1k}
2	x_{21}	x_{22}	...	x_{2k}
·	·	·	...	·
·	·	·	...	·
·	·	·	...	·
n	x_{n1}	x_{n2}	...	x_{nk}
Групповые средние	\bar{x}_1	\bar{x}_2	...	\bar{x}_k

Однофакторный и двухфакторный дисперсионный анализ

Номер изменения	Уровень фактора			
	A_1	A_2	...	A_k
1	x_{11}	x_{12}	...	x_{1k}
2	x_{21}	x_{22}	...	x_{2k}
·	·	·	...	·
·	·	·	...	·
·	·	·	...	·
n	x_{n1}	x_{n2}	...	x_{nk}
Групповые средние	\bar{x}_1	\bar{x}_2	...	\bar{x}_k

Далее рассчитываются:

1) Общая сумма квадратов отклонений от общей средней \bar{x} :

$$C_{\text{общ}} = \sum_{j=1}^k \sum_{i=1}^n (x_{ij} - \bar{x})^2,$$

Однофакторный и двухфакторный дисперсионный анализ

Номер изменения	Уровень фактора			
	A_1	A_2	...	A_k
1	x_{11}	x_{12}	...	x_{1k}
2	x_{21}	x_{22}	...	x_{2k}
·	·	·	...	·
·	·	·	...	·
·	·	·	...	·
n	x_{n1}	x_{n2}	...	x_{nk}
Групповые средние	\bar{x}_1	\bar{x}_2	...	\bar{x}_k

2) факторная сумма квадратов отклонений групповых средних от общей средней, характеризующая рассеяние между группами:

$$C_{\text{факт}} = n \sum_{j=1}^k (x_j - \bar{x})^2,$$

Однофакторный и двухфакторный дисперсионный анализ

Номер изменения	Уровень фактора			
	A_1	A_2	...	A_k
1	x_{11}	x_{12}	...	x_{1k}
2	x_{21}	x_{22}	...	x_{2k}
·	·	·	...	·
·	·	·	...	·
·	·	·	...	·
n	x_{n1}	x_{n2}	...	x_{nk}
Групповые средние	\bar{x}_1	\bar{x}_2	...	\bar{x}_k

3) остаточная сумма квадратов отклонений наблюдаемых значений от своей групповой средней, характеризующая рассеяние внутри групп:

$$C_{ост} = \sum_{i=1}^n (x_{i1} - \bar{x}_1)^2 + \sum_{i=1}^n (x_{i2} - \bar{x}_2)^2 + \dots + \sum_{i=1}^n (x_{ik} - \bar{x}_k)^2,$$

Однофакторный и двухфакторный дисперсионный анализ

4) факторная и остаточная дисперсии:

$$S^2_{\text{факт}} = C_{\text{факт}} / (k-1); S^2_{\text{ост}} = C_{\text{ост}} / k(n-1),$$

5) Значение критерия Фишера:

$$F = S^2_{\text{факт}} / S^2_{\text{ост}}$$

Значения критерия Фишера сравнивается с критическим для заданного уровня значимости α и числа степеней свободы $k-1$ и $k(n-1)$.

Для упрощения вычислительных операций при однофакторном дисперсионном анализе используют равенство $C_{\text{ост}} = C_{\text{общ}} - C_{\text{факт}}$

Однофакторный и двухфакторный дисперсионный анализ

При **неравномерном однофакторном дисперсионном анализе**, когда количество наблюдений на уровне A_1 равно n_1 , на уровне A_2 – n_2 ..., на уровне A_k – n_k , а общее их число равно $N = \sum_{j=1}^k n_j$, факторная и остаточная дисперсия находятся по формулам:

$$S^2_{\text{факт}} = \frac{1}{k-1} \sum_{j=1}^k n_j (x_j - \bar{x})^2$$

$$S^2_{\text{ост}} = \frac{1}{N-1} \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2$$

Остальные операции выполняются так же, как при равномерном анализе.

Однофакторный и двухфакторный дисперсионный анализ

При двухфакторном дисперсионном анализе сумма квадратов отклонений от общего среднего разделяется на компоненты, отвечающие двум предполагаемым факторам изменчивости – A и B . Если по фактору выделяется A выделяется p уровней, а по фактору B выделяется q уровней, то общее количество групп будет равно $t = pq$.

A	Уровни фактора B						Среднее
	B_1	B_2	...	B_j	...	B_q	
A_1	x_{11}	x_{12}	...	x_{1j}	...	x_{1q}	\bar{x}_1
A_2	x_{21}	x_{22}	...	x_{2j}	...	x_{2q}	\bar{x}_2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
A_i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{iq}	\bar{x}_i
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
A_p	x_{p1}	x_{p2}	...	x_{pj}	...	x_{pq}	\bar{x}_p
Среднее	$\bar{x}_{.1}$	$\bar{x}_{.2}$...	$\bar{x}_{.j}$...	$\bar{x}_{.q}$	\bar{x}

Однофакторный и двухфакторный дисперсионный анализ

A	Уровни фактора B						Среднее
	B_1	B_2	...	B_j	...	B_q	
A_1	x_{11}	x_{12}	...	x_{1j}	...	x_{1q}	\bar{x}_1
A_2	x_{21}	x_{22}	...	x_{2j}	...	x_{2q}	\bar{x}_2
...
A_i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{iq}	\bar{x}_i
...
A_p	x_{p1}	x_{p2}	...	x_{pj}	...	x_{pq}	\bar{x}_p
Среднее	$\bar{x}_{.1}$	$\bar{x}_{.2}$...	$\bar{x}_{.j}$...	$\bar{x}_{.q}$	\bar{x}

Если для каждого сочетания факторов $A_i B_j$ осуществлено по n наблюдений, то в каждую клетку помещается n значений, а единичное наблюдение обозначается как x_{ijk} , где $k = 1, 2, \dots, n$.

Однофакторный и двухфакторный дисперсионный анализ

A	Уровни фактора B						Среднее
	B_1	B_2	...	B_j	...	B_q	
A_1	x_{11}	x_{12}	...	x_{1j}	...	x_{1q}	\bar{x}_1
A_2	x_{21}	x_{22}	...	x_{2j}	...	x_{2q}	\bar{x}_2
...
A_i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{iq}	\bar{x}_i
...
A_p	x_{p1}	x_{p2}	...	x_{pj}	...	x_{pq}	\bar{x}_p
Среднее	$\bar{x}_{.1}$	$\bar{x}_{.2}$...	$\bar{x}_{.j}$...	$\bar{x}_{.q}$	\bar{x}

Оценки средних значений по группам (\bar{x}_{ij}) по факторам ($x_{i..}$ и $x_{.j.}$) и общее среднее (\bar{x}) в этом случае рассчитываются по формулам

Для оценки степени зависимости двух случайных величин X и Y используется **корреляционный момент** K_{xy} . Для дискретных случайных величин корреляционный момент определяется формулой:

$$K_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})$$

-здесь n - число наблюдений случайных величин X и Y ; \bar{X} , \bar{Y} - средние значения случайных величин X и Y соответственно. Для независимых случайных величин корреляционный момент равен нулю, в противном случае существует зависимость между случайными величинами.

Двумерные статистические модели

Корреляционным отношением называется отношение дисперсий (стандартов) центров условных распределений к общей дисперсии (стандарту) величины

$\eta = \sigma(\bar{y}_i) / \sigma(y)$, где y – значения принимаемые зависимой переменной; \bar{y} – условные средние, соответствующие значениями x_i .

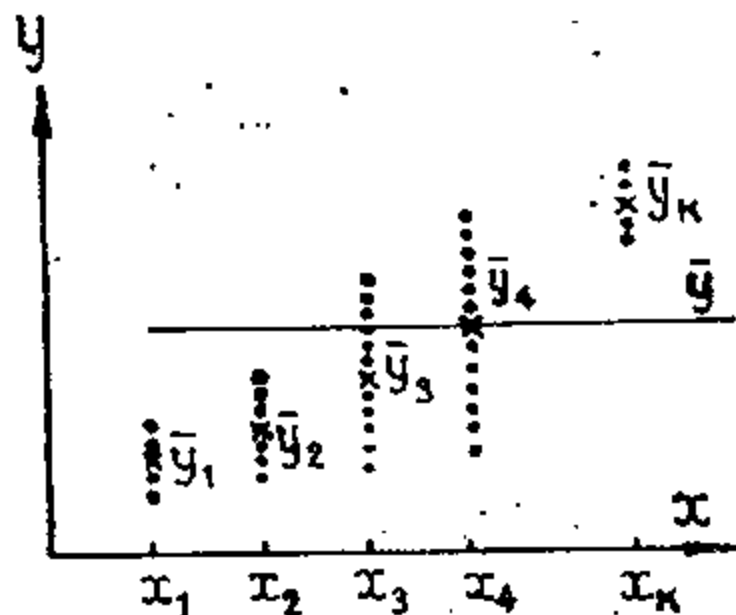


Рис. Разброс значений зависимой переменной y и ее условных средних \bar{y}

Равенство $\eta = 0$ – необходимое и достаточное условие отсутствия корреляционной зависимости. При $\eta = 1$ корреляционная связь переходит в функциональную, когда все значения переменной, соответствующие определенному x_i соответствует одно единственное y_j .

Для характеристики связи между величинами (X, Y) в чистом виде переходят от момента K_{xy} к безразмерной характеристике

$$r_{xy} = \frac{K_{xy}}{\sigma_x \sigma_y}$$

где - σ_x, σ_y среднеквадратические отклонения величин X, Y . Эта характеристика называется **коэффициентом корреляции** величин X и Y . Если коэффициент корреляции двух случайных величин равен нулю, то такие случайные величины называются **некоррелируемыми** (независимыми). При значениях коэффициента корреляции близких к единице – **коррелируемыми**. В случае, когда значение коэффициента корреляции, между двумя случайными величинами, близко к минус единице, считается, что между случайными величинами существует сильная **обратная корреляционная** связь.

Матрица, составленная из коэффициентов корреляции ,
представляет корреляционную матрицу:

$$r_{ij} = \begin{bmatrix} 1 & r_{21} & \dots & r_{n1} \\ r_{12} & 1 & \dots & r_{n2} \\ \dots & \dots & \dots & \dots \\ r_{1n} & r_{2n} & \dots & 1 \end{bmatrix}$$

Если не удастся проверить гипотезу о соответствии эмпирического распределения определенному закону из-за малого количества данных (или распределения существенно отличаются от нормального закона), то для проверки гипотезы о наличии корреляционной связи можно использовать ранговый коэффициент корреляции Спирмена.

Его расчет основан на замене выборочных значений исследуемых случайных величин их рангами в порядке возрастания. При этом предполагается, что если между значениями случайных величин нет корреляционной зависимости, то ранги этих величин тоже будут независимыми.

$$r = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)},$$

где d_i - разность рангов сопряженных значений изучаемых величин x_i и y_i ; n - количество пар в выборке

Результаты вычисления рангового коэффициента корреляции для значений m_r и ρ_k

№ скв.	m_r		ρ_k		d_i	d_i^2
	значение, %	ранг	значение, Ом·м	ранг		
1	67	9	253	10	-1	1
2	80	12	115	7	5	25
3	40	5	126	8	-3	9
4	24	2	82	6	-4	16
5	25	3	66	5	-2	4
6	38	4	25	1	3	9
7	18	1	44	3	-2	4
8	72	10	180	9	1	1
9	44	6	32	2	4	16
10	51	8	319	11	-3	9
11	76	11	421	12	-1	1
12	50	7	51	4	3	9

$$\sum d_i^2 = 104$$

Ранговый коэффициент корреляции $r = 1 - \frac{6 \times 104}{12(144 - 1)} = 1 - 0.39 = 0.61$

Множественная корреляция

Корреляция двух случайных величин X и Y – частный случай более распространенной в окружающей нас действительности множественной корреляции, когда изменение одной из переменных зависит от изменения множества других. При исследовании таких связей возникает две существенно отличных друг от друга задачи:

- определение тесноты связи между парами факторов, когда влияние других исключено;
- Тесноты линейной зависимости между одним из факторов (функцией) и остальными (аргументами).

Показатель, характеризующий тесноту линейной связи между двумя признаками X и Y , когда влияние других факторов исключено, называется **частным коэффициентом корреляции**.

Множественная корреляция

Для трех признаков – X, Y, Z вычисляются следующие частные коэффициенты корреляции:

- взаимодействие между X и Y при фиксированном Z

$$r_{xy.z} = \frac{r_{xz} - r_{xy} \times r_{zy}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}$$

- взаимодействие между X и Z при фиксированном Y

$$r_{xz.y} = \frac{r_{xz} - r_{xy} \times r_{zy}}{\sqrt{(1 - r_{xy}^2)(1 - r_{yz}^2)}}$$

- взаимодействие между Y и Z при фиксированном X

$$r_{yz.x} = \frac{r_{yz} - r_{yx} \times r_{zx}}{\sqrt{(1 - r_{yx}^2)(1 - r_{zx}^2)}}$$

Множественная корреляция

Показателем тесноты связи между одним из факторов и остальными является коэффициент множественной корреляции. Для зависимости X от Y и Z и его вычисляют по формуле

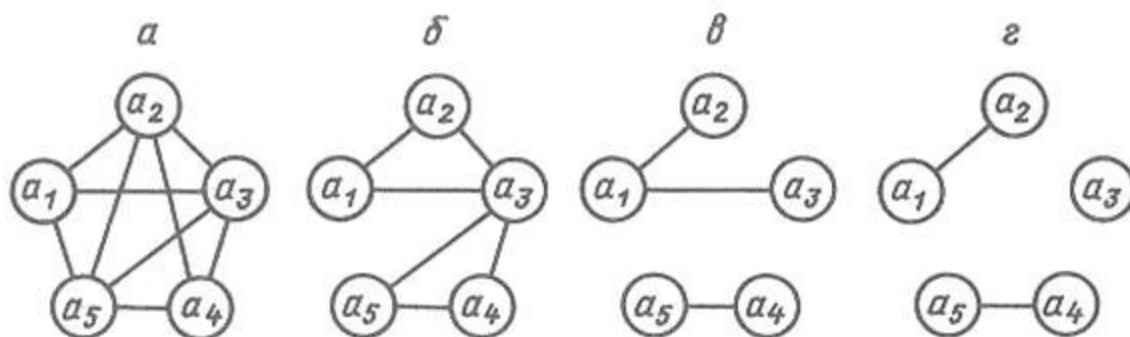
$$R_x = \sqrt{\frac{r_{xy}^2 + r_{xz}^2 - 2r_{xy}r_{xz}r_{yz}}{1 - r_{yz}^2}}$$

Статистические методы выделения ассоциаций химических элементов

Анализ корреляционной матрицы с позиции теории графов

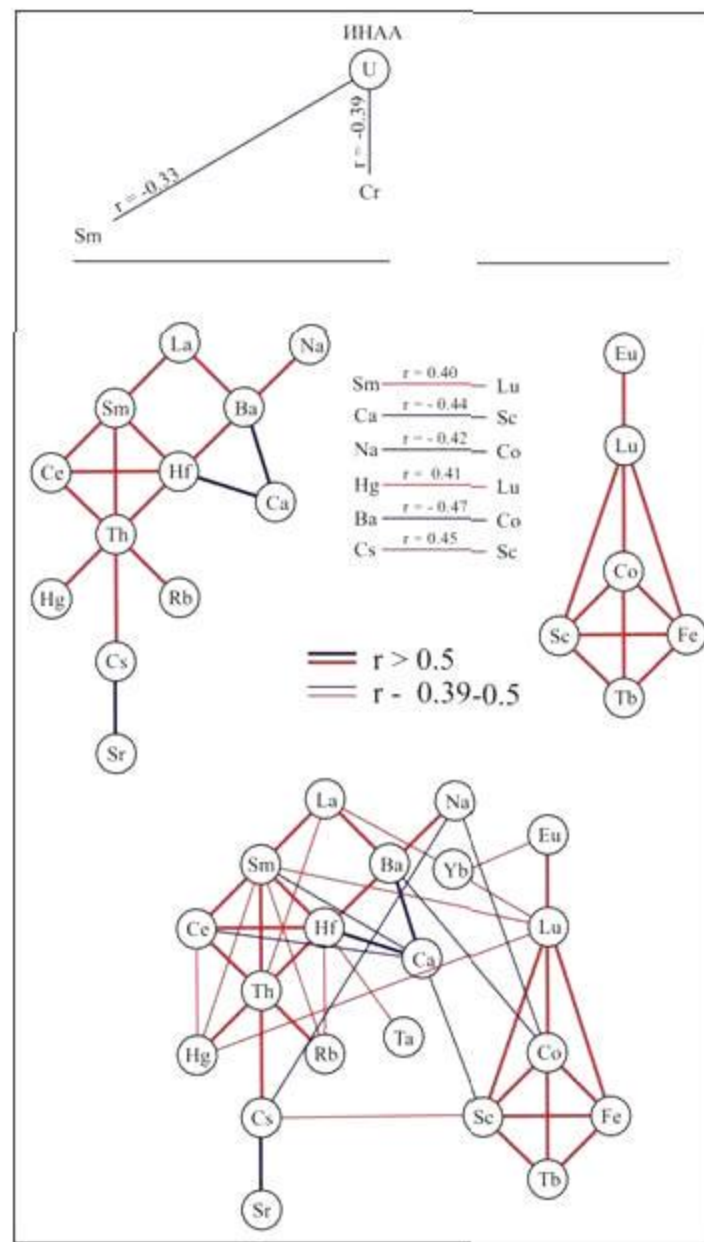
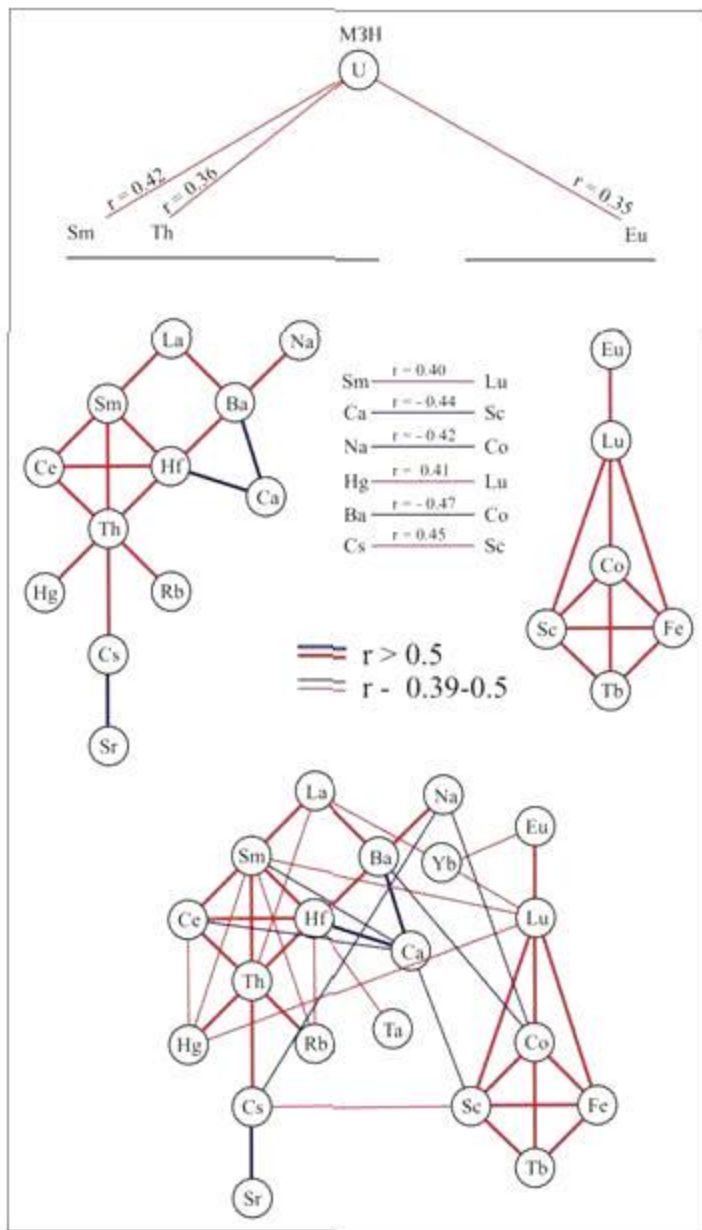
Графом $G(A)$ называется геометрическая схема, включающая две и более точки конечного множества

$$A = \{a_1, \dots, a_k, \dots, a_l, \dots, a_p\}$$

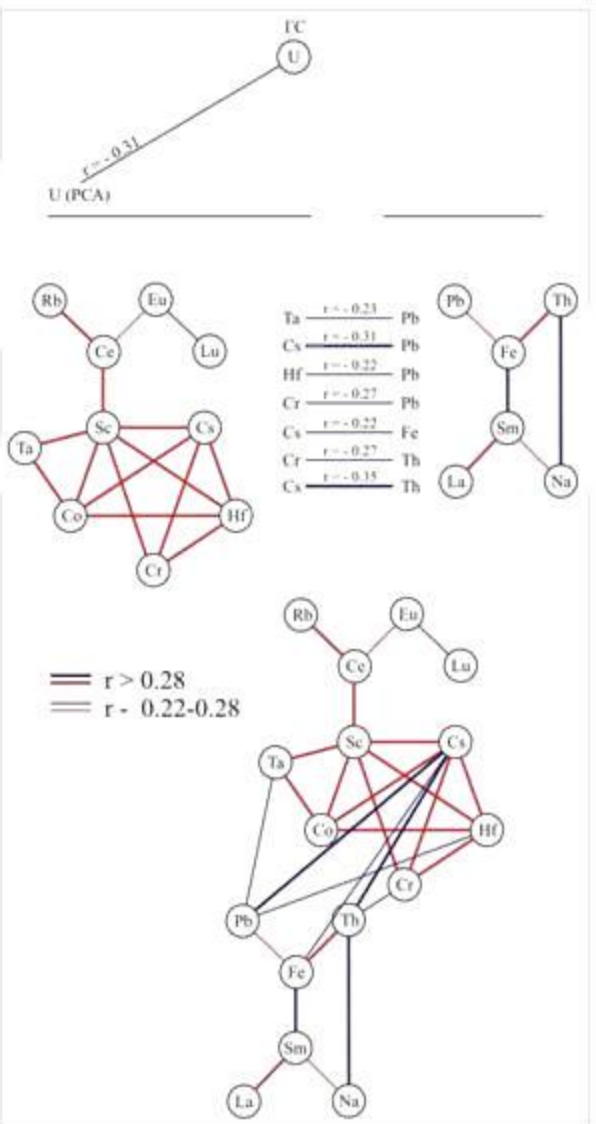
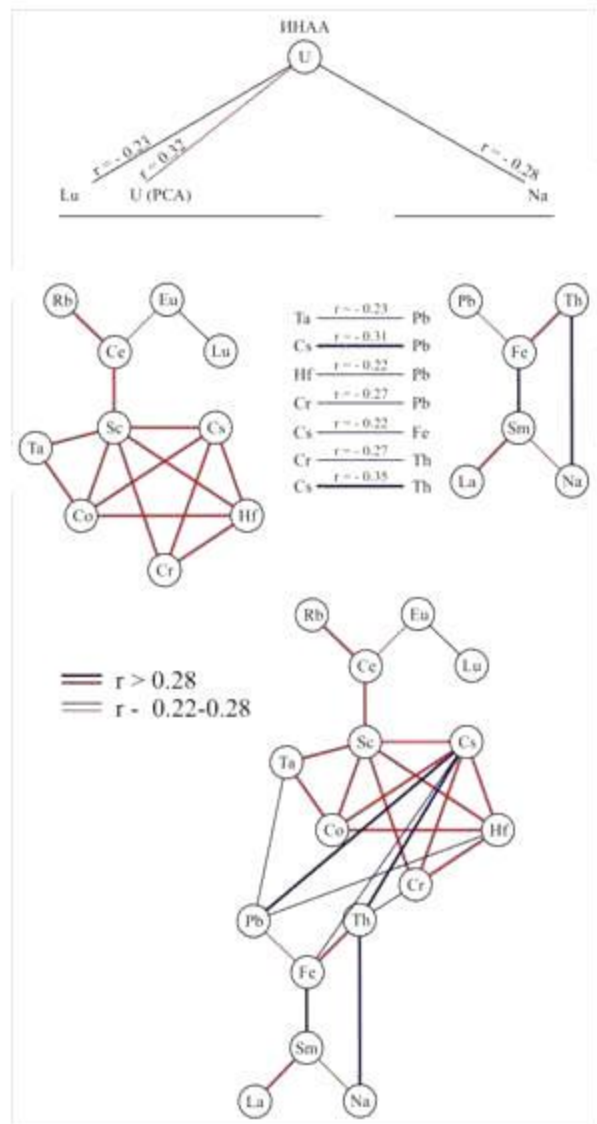
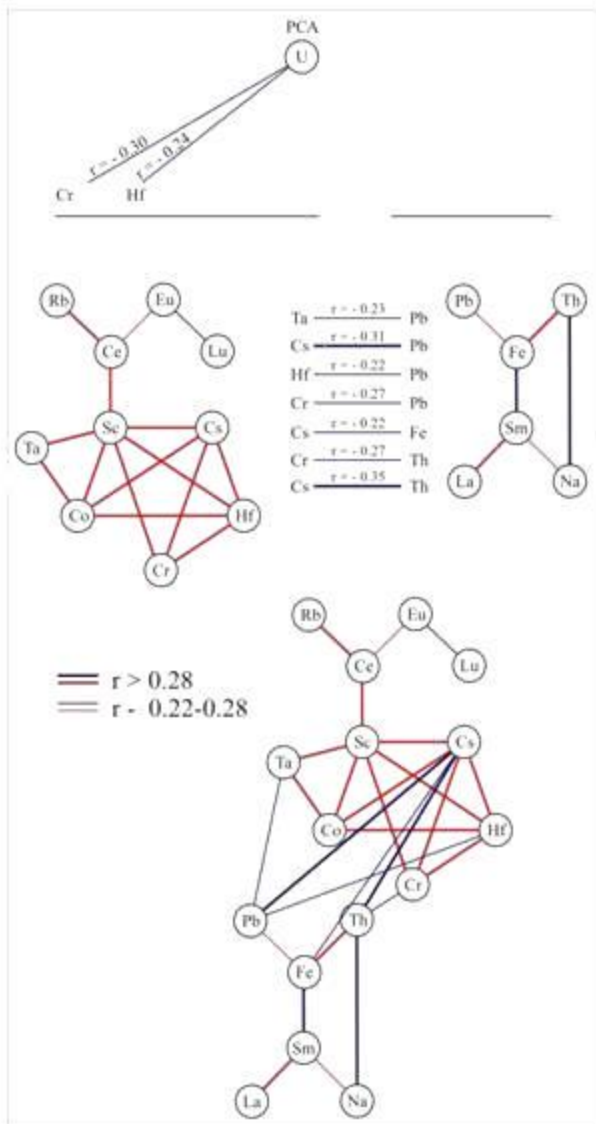


Пороговые значения: а – $r = 0.2$; б – $r = 0.3$; в – $r = 0.4$; г – $r = 0.5$

ГРАФЫ АССОЦИАЦИЙ ХИМ. ЭЛЕМЕНТОВ В ГРУНТАХ ИМБИНСКОЙ ГАЗОНОСНОЙ ПЛОЩАДИ



ГРАФЫ АССОЦИАЦИЙ ХИМ. ЭЛЕМЕНТОВ В ГРУНТАХ ЗАПАДНО-ПОЛУДЕННОЙ НЕФТЕНОСНОЙ ПЛОЩАДИ ГАЗОНОСНОЙ ПЛОЩАДИ



Кластерный анализ (дендрограммы)

Результаты кластер-анализа изображаются в виде древовидного графа – дендрограммы, в которой по оси абсцисс располагаются символические обозначения объектов исследования (векторов матрицы), а по оси ординат – минимальные значения дистанционных коэффициентов, соответствующих каждому шагу классифицированной процедуры.

Первый шаг состоит в выявлении высших коэффициентов корреляций между отдельными парами, которые объединяются и принимаются за центры групп. Число таких центров может изменяться от одного до трех (редко более).

Кластерный анализ (дендрограммы)

Второй шаг – матрица вычисляется снова, причем сгруппированные элементы считаются за один элемент, а коэффициенты их корреляции с другими группами вычисляются заново с помощью различных методов осреднения. По результатам вычисления составляется новая матрица, меньшей размерности, в которой изменяются лишь значения коэффициентов, связанные с членами объединенных групп. Сокращенная и пересчитанная матрица вновь подвергается сокращению, путем выявления и объединения пар с максимальными признаками сходства, и последующим осреднением новых групповых коэффициентов.

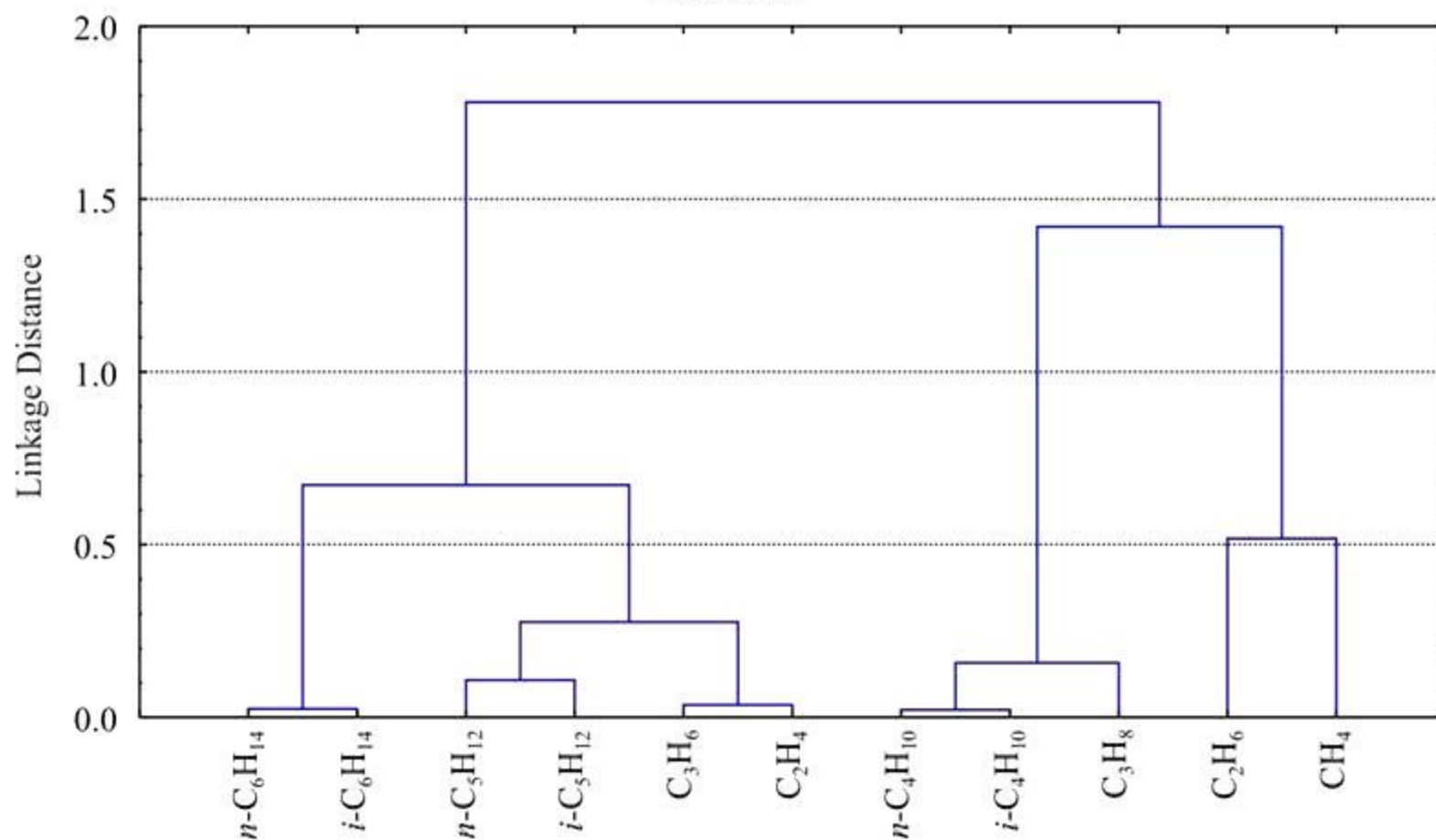
Кластерный анализ (дендрограммы)

Операция последовательного сокращения и пересчета коэффициентов матрицы повторяется до тех пор, пока значения групповых коэффициентов сходства не достигнут порогового значения или размерность матрицы не станет минимальной.

В качестве меры сходства используются непосредственно парные коэффициенты корреляции ($1 - r$), m -мерное эвклидово расстояние или другие дистанционные коэффициенты.

Дендрограмма корреляционной матрицы геохимического спектра углеводородов ($C_1 - C_6$) в пробах снега
($1 - r_{0.05} = 0.93, n = 801$)

Tree Diagram
Ward's method
1-Pearson r



Метод главных компонент

С увеличением размерности признакового пространства возрастают трудности изучения природных объектов и возникает проблема замены многочисленных наблюдаемых признаков меньшим их числом, без существенной потери полезной информации. Одним из наиболее распространенных методов решения этой задачи является метод главных компонент.

Основой метода является линейное преобразование m исходных переменных (признаков) в m новых переменных, где каждая новая переменная представляет собой линейное сочетание исходных. В процессе преобразования векторы наблюдаемых переменных заменяются новыми векторами (главными компонентами), которые вносят резко различные вклады в суммарную дисперсию многомерных признаков.

Метод главных компонент

Сокращение пространства признаков достигается путем отбора нескольких наиболее информативных компонент, обеспечивающих основную долю суммарной дисперсии, что приводит к заметному уменьшению их общего числа за счет наименее информативных компонент, отражающих малые доли суммарной дисперсии

Главные компоненты – это собственные векторы ковариационных матриц исходных признаков.

Координаты собственных векторов рассматриваются как **нагрузки** соответствующих переменных на тот или иной фактор.

Как правило, дисперсия одной из главных компонент достигает половины и более от суммарной дисперсии а в совокупности с дисперсиями еще одной-двух последующих компонент, их общий вклад в суммарную дисперсию превышает 90%.

Метод главных компонент

Таким образом, без существенной потери информации об изменчивости наблюдаемых признаков можно заметно сократить размерность пространства наблюдаемых признаков (до $p \leq m$), ограничившись данными по двум-трем наиболее главным компонентам.

Метод главных компонент используется в качестве основы **факторного анализа** многомерных совокупностей.

В практике используют два метода факторного анализа *R*- и *Q*-метод. *R*-метод основан на использовании не ковариационной, а корреляционной матрицы.

Для более содержательной интерпретации результаты факторного анализа подвергаются процедуре «**вращения**» вокруг центра координат, для выявления наиболее контрастных сочетаний факторных нагрузок.

Факторные нагрузки углеводородных компонентов в пробах снега эталонных скважин (R-метод главных компонент)

УВ компо- нит	до вращения						нагрузка варимакс (varimax)						нагрузка промакс (promax)					
	pF ₁	uF ₁	pF ₂	uF ₂	pF ₃	uF ₃	pVF ₁	uVF ₁	pVF ₂	uVF ₂	pVF ₃	uVF ₃	pPF ₁	uPF ₁	pPF ₂	uPF ₂	pPF ₃	uPF ₃
CH ₄	-0.44	-0.51	-0.15	-0.37	-0.25	-0.25	0.37	0.62	0.18	-0.13	0.32	0.23	0.04	-0.20	0.05	-0.63	-0.13	-0.22
C ₂ H ₆	-0.37	-0.76	-0.38	-0.25	0.01	-0.02	0.53	0.75	-0.03	0.21	0.09	0.18	-0.12	0.16	-0.11	-0.75	0.07	-0.09
C ₂ H ₄	0.11	-0.53	0.16	0.51	0.91	-0.53	-0.03	0.09	0.06	0.29	-0.93	0.85	0.06	0.16	0.05	-0.08	0.94	-0.77
C ₃ H ₈	-0.75	-0.75	-0.61	-0.61	0.14	0.00	0.97	0.96	0.07	-0.01	0.03	0.00	-0.01	-0.04	-0.92	-0.96	0.00	0.00
C ₃ H ₆	0.25	-0.35	0.01	0.36	0.91	-0.76	-0.02	0.04	-0.14	-0.03	-0.93	0.91	-0.11	-0.16	-0.06	-0.03	0.90	-0.89
i-C ₄ H ₁₀	-0.80	-0.78	-0.55	-0.57	0.14	0.06	0.96	0.97	0.15	0.06	0.04	-0.02	0.07	0.04	-0.91	-0.97	0.00	0.06
n-C ₄ H ₁₀	-0.75	-0.80	-0.60	-0.55	0.18	0.06	0.97	0.97	0.08	0.08	0.00	0.00	-0.01	0.05	-0.92	-0.96	0.00	0.01
i-C ₅ H ₁₂	-0.68	-0.53	0.60	0.75	-0.04	0.05	0.07	-0.05	0.91	0.78	0.08	0.49	0.91	0.73	0.00	0.04	-0.09	-0.34
n-C ₅ H ₁₂	-0.51	-0.43	0.50	0.75	0.23	-0.08	0.06	-0.13	0.72	0.66	-0.20	0.56	0.71	0.61	0.08	0.15	0.23	-0.44
i-C ₆ H ₁₄	-0.75	-0.60	0.62	0.58	-0.01	0.48	0.11	0.11	0.96	0.96	0.06	0.08	0.93	0.92	-0.05	-0.11	-0.04	0.14
n-C ₆ H ₁₄	-0.74	-0.55	0.59	0.46	0.00	0.58	0.12	0.15	0.93	0.90	0.05	-0.07	0.89	0.87	-0.07	-0.16	-0.05	0.26

Примечание: p – продуктивные скважины, u – непродуктивные скважины

Интерпретация полей статистических характеристик геополей

Среднее, мода, медиана, центральные моменты и другие характеристики случайной величины позволяют подчеркнуть различные свойства функции распределения конкретной случайной величины X : островершинность, асимметричность, положение среднего значения, степень разбросанности и т.д. При расчете статистических характеристик в скользящих окнах, получается распределение этих характеристик случайной величины X вдоль профиля, по площади или в пространстве.

Интерпретация статистических характеристик геополей

Пример (А.В. Петров, 2004): Модельные профильные данные, представленные 501-ой точкой наблюдения (**красный** цвет). Размер скользящего «окна», в котором рассчитывались все статистические характеристики, принимался равным 101 точке. Модельные данные представляют собой сумму нормальной помехи и аномалии:

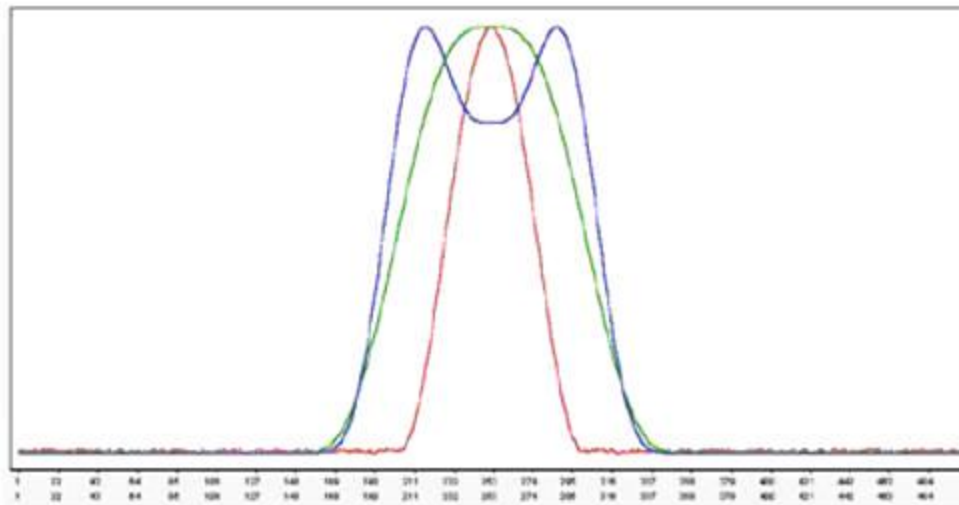


Рис. Статистические характеристики в скользящем окне: красный цвет - модельное поле с положительной аномалией, зеленый - асимметрия; синий - эксцесс.

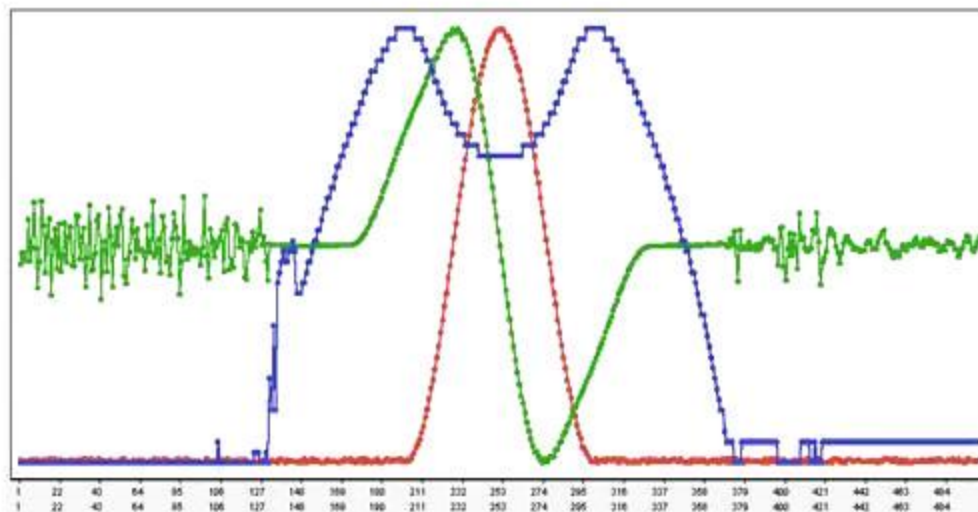


Рис. Статистические характеристики в скользящем окне: красный цвет - модельное поле с положительной аномалией, зеленый - коэффициент регрессии; синий - радиус корреляции.

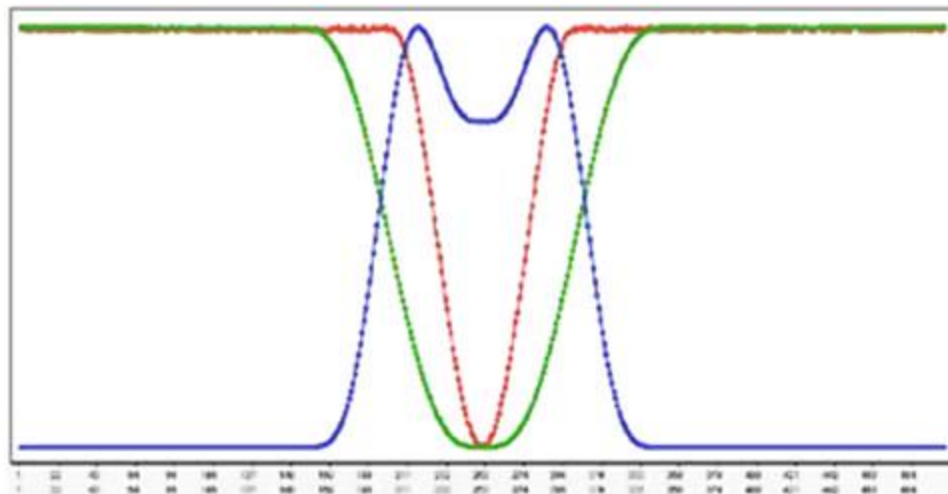


Рис. Статистические характеристики в скользящем окне: красный цвет - модельное поле с отрицательной аномалией, зеленый - среднее; синий - дисперсия.

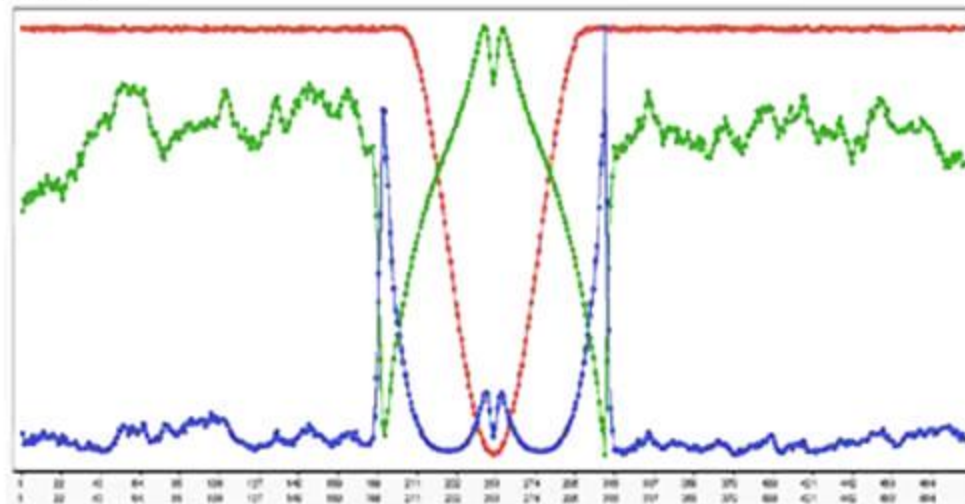


Рис. Статистические характеристики в скользящем окне: красный цвет - модельное поле с отрицательной аномалией, зеленый - асимметрия; синий - эксцесс.

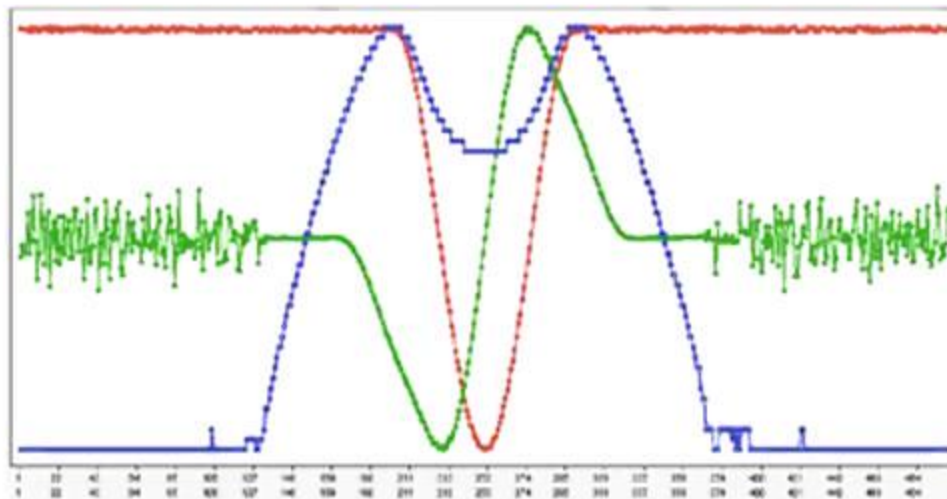


Рис. Статистические характеристики в скользящем окне: красный цвет - модельное поле с отрицательной аномалией, зеленый - коэффициент регрессии; синий - радиус корреляции.

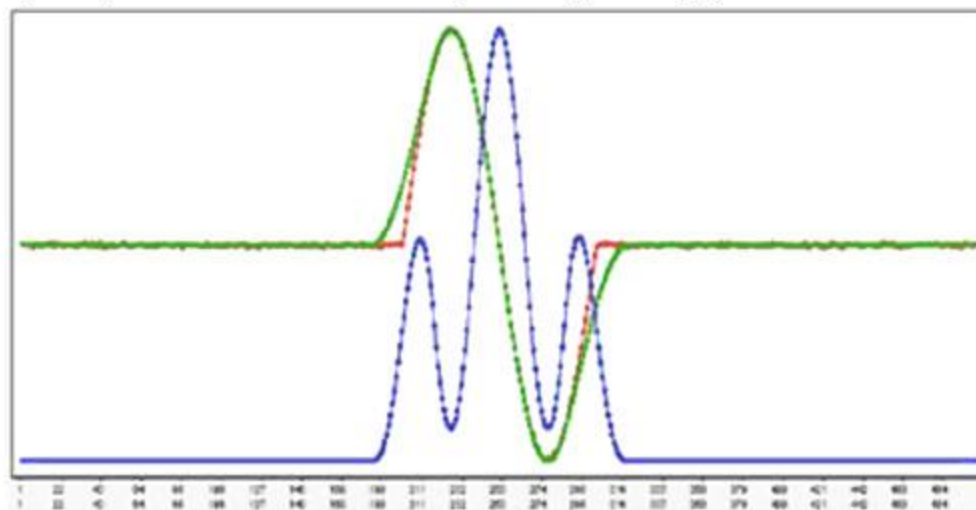


Рис. Статистические характеристики в скользящем окне: красный цвет - модельное поле со знакопеременной аномалией, зеленый - среднее; синий - дисперсия.

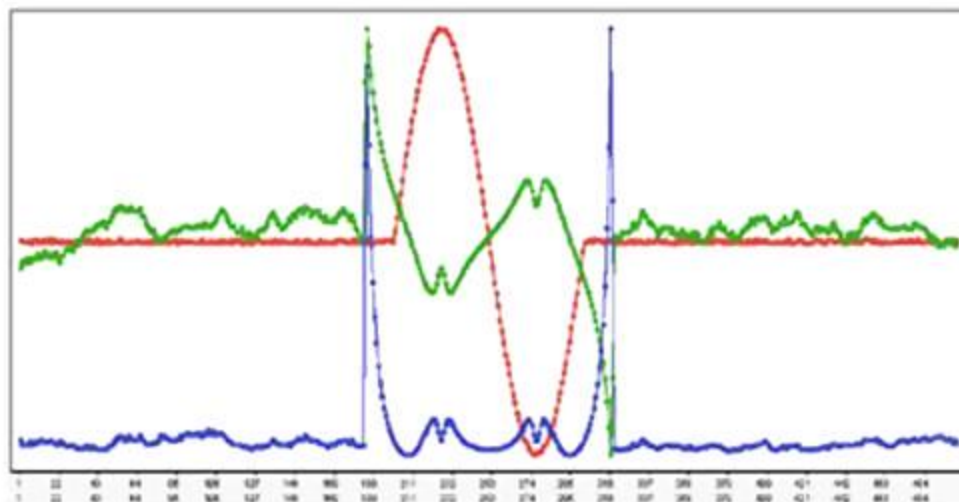


Рис. Статистические характеристики в скользящем окне: красный цвет - модельное поле со знакопеременной аномалией, зеленый - асимметрия; синий - эксцесс.

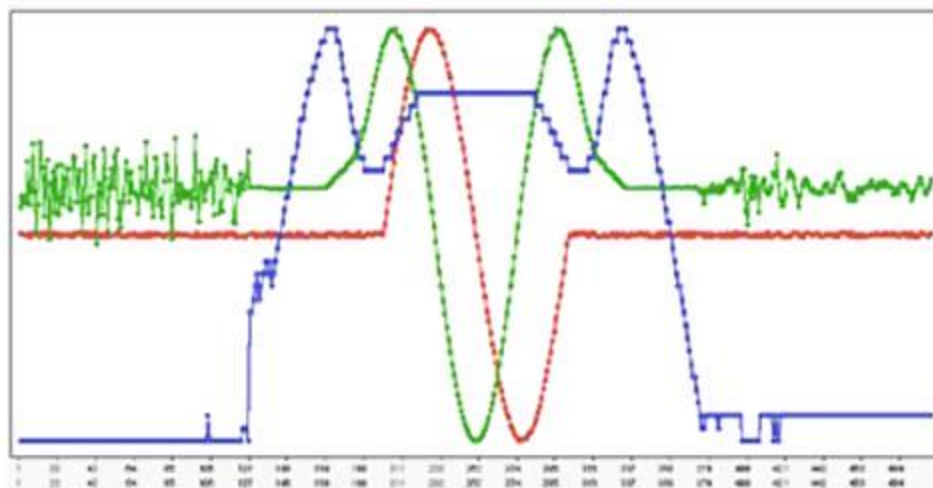


Рис. Статистические характеристики в скользящем окне: красный цвет - модельное поле со знакопеременной аномалией, зеленый - коэффициент регрессии; синий - радиус корреляции.

Интерпретация статистических характеристик геополей

В приведенных примерах, практически для всех статистических характеристик, за исключением среднего значения, очевидно следующее:

-границы аномальных объектов отмечаются экстремальными значениями распределения статистических характеристик вдоль профиля;

-аномалии простой формы в исходных данных, в полях статистических характеристик, представлены более дифференцированными по форме аномалиями;

-анализ аномальных значений в полях статистических характеристик, позволяет говорить о том что, отношение сигнал/помеха для них в несколько раз превышает аналогичное отношение для исходных данных.

Геообъекты, как поля пространственных переменных

Поле пространственной переменной называется область пространства, каждой точке которого поставлено в соответствие некоторое значение изучаемой переменной. В качестве геополя может рассматриваться область пространства, при этом каждому элементу последнего соответствует определенное значения изучаемого гео- признака.

По характеру распространения (областям существования) в земной коре гео-пространственные переменные разделяются на непрерывные и дискретные.

Геообъекты, как поля пространственных переменных

Непрерывные пространственные переменные выражают свойства горных пород, минеральных ассоциаций или полезных ископаемых, проявленные в любой точке поля, т.е. на всей площади (во всем объеме) исследуемого блока земной коры.

К числу дискретных пространственных переменных относятся пространственно ограниченные природные образования, области существования (размеры) которых пренебрежимо малы по сравнению с исследуемыми площадями или объемами недр.

Геообъекты, как поля пространственных переменных

Большинство обычно изучаемых гео-переменных относится к *скалярным* величинам, для задания которых достаточно знать их модуль и знак.

Реже в практике геоэкологических работ используются *векторные* пространственные переменные, для задания которых в каждой точке пространства необходимо знать не только модуль, но и направление переменной.

Фон, аномалии и поверхность тренда

В геоэкологических исследованиях проблема фона и аномалий менее актуальна, чем при поисках и разведке месторождений полезных ископаемых. В соответствующих нормативных документах для большинства компонентов приведены крайние значения (например, ПДК) выше которых их содержание представляет опасность здоровью человека.

Тем не менее, навыки выделения аномалий могут быть полезны при оценке динамики и прогнозе изменения состояния окружающей среды на перспективу, с целью принятия опережающих мер по снижению негативного воздействия сравнительно «молодых» объектов хозяйственной деятельности.

Фон, аномалии и поверхность тренда

Главной задачей изучения пространственных закономерностей является описание неслучайной (закономерной) компоненты поля, отражающей уровень его значений, характерных для отдельных частей изучаемой территории.

Неслучайная компонента, характеризующая основную часть моделируемого геологического поля, называется его фоном.

Методы выделения фоновой части геологического поля с разделением неслучайной и случайной составляющих изучаемых признаков по эмпирическим данным получили название *анализа поверхностей тренда*.

Фон, аномалии и поверхность тренда

Для целей тренд-анализа в основном используют два разных методических подхода:

1) Сглаживание исходных данных скользящими статистическими окнами.

В результате такого сглаживания по профилю закономерная изменчивость выявляется в виде плавной кривой, которая может быть описана функцией синусоидального типа, а для характеристики случайной изменчивости используется коэффициент вариации, вычисленный через отклонения каждого частного значения от скользящей средней, т.е. уровня неслучайной изменчивости.

Фон, аномалии и поверхность тренда

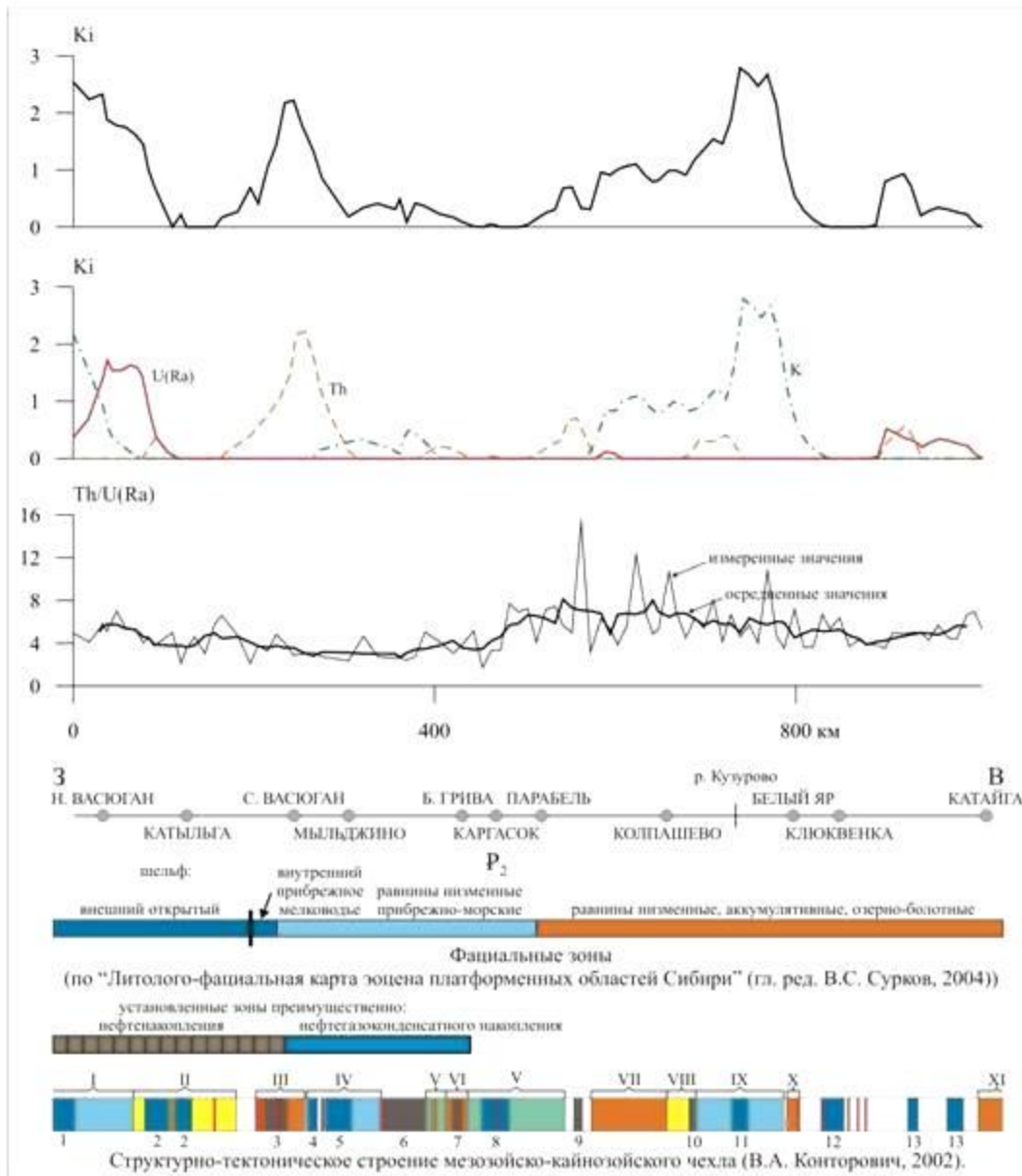


Рис. Графики изменения торий-уранового отношения и коэффициента перераспределения ЕРЭ (зоны перераспределения радионуклидов по Н.Г. Лященко) по региональному профилю

Фон, аномалии и поверхность тренда

Для решения аналогичной задачи не по профилям, а по площади, по точкам среднеарифметических значений признака, рассчитанных для центров разведочных ячеек путем двух- или трехкратного сглаживания, проводят изолинии, а дисперсия случайной составляющей рассчитывается через отклонения фактических значений от соответствующих изолиний.

Способы, основанные на сглаживании или преобразовании исходных данных, просты и наглядны, но обладают рядом недостатков:

Фон, аномалии и поверхность тренда

- они не дают объективных количественных критериев для оценки значимости выявленных закономерностей. Вопрос о наличии закономерностей решается по виду сглаженных поверхностей;
- результаты сглаживания существенно изменяются в зависимости от вида преобразования и размера площади трансформации. Выбор оптимального способа преобразования часто – эмпирический;
- любое преобразование обладает селективными свойствами только по отношению к закономерностям, близким по размеру с площадкой трансформации. Для выделения в наблюдаемой изменчивости закономерностей разного масштаба необходимо использовать различные варианты преобразований.

Фон, аномалии и поверхность тренда

Для целей тренд-анализа в основном используют два разных методических подхода:

2) Аппроксимация поверхностей тренда полиномами.

В качестве аппроксимирующей функции используются ортогональные полиномы различных степеней, тригонометрические полиномы и др.

Ортогональные полиномы обычно применяются в случае равномерной прямоугольной сети наблюдений. При этом тренд определяется как линейная функция пространственных координат, построенная по совокупности наблюдений таким образом, что сумма квадратов отклонений значений признака от плоскости тренда минимальна.

Фон, аномалии и поверхность тренда

Такая модель – вариант статистического метода множественной регрессии, в котором функция $\Phi(x, y) = \beta_0 + \beta_1 x + \beta_2 y$ (где x и y – координаты пространства; β_0 , β_1 и β_2 – полиномиальные коэффициенты).

Для оценки трех указанных коэффициентов используются уравнения

$$\sum u = \beta_0 n + \beta_1 \sum x + \beta_2 \sum y;$$

$$\sum xu = \beta_0 \sum x + \beta_1 \sum x^2 + \beta_2 \sum xy;$$

$$\sum yu = \beta_0 \sum y + \beta_1 \sum xy + \beta_2 \sum y^2,$$

где n – число точек наблюдения; u – значения признака в точках наблюдения; x и y – координаты точек наблюдений.

Фон, аномалии и поверхность тренда

Для решения уравнений они записываются в матричном
форме:

$$\begin{bmatrix} n & \sum x & \sum y \\ \sum x & \sum x^2 & \sum xy \\ \sum y & \sum xy & \sum y^2 \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \sum u \\ \sum xu \\ \sum yu \end{bmatrix}$$

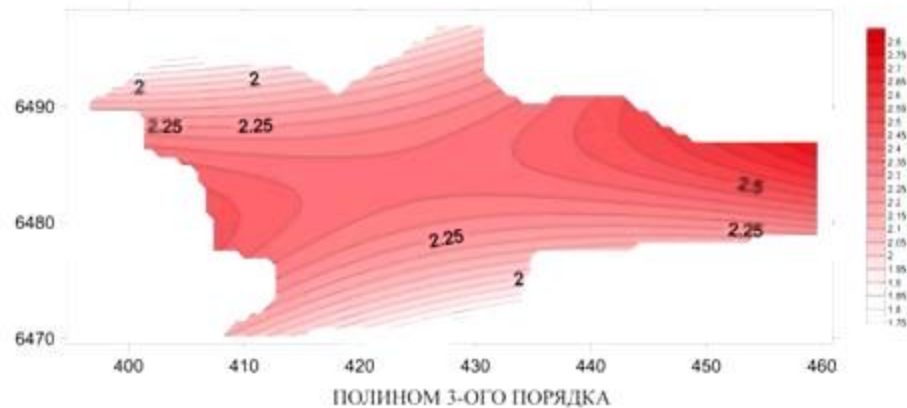
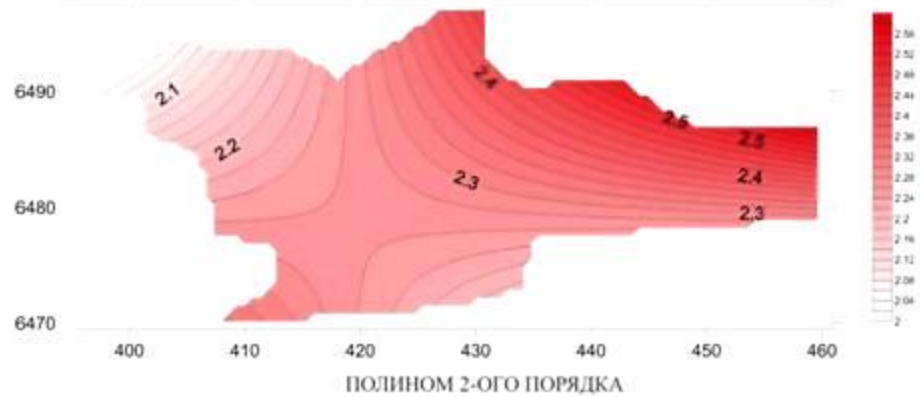
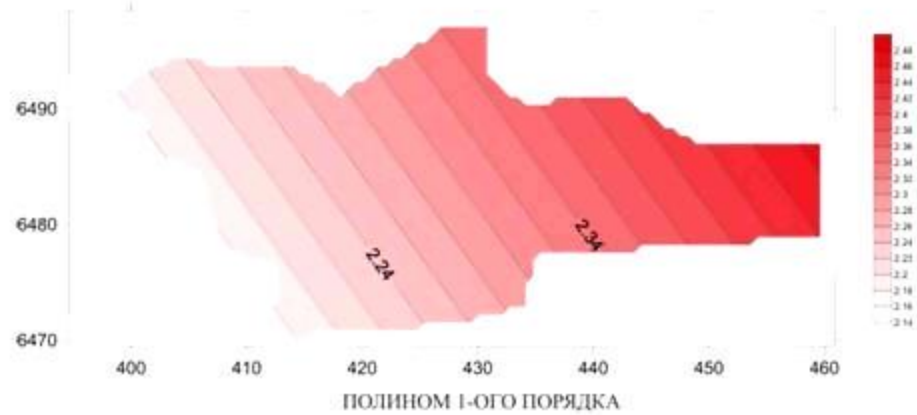
и решаются относительно $\beta_0, \beta_1, \beta_2$. Такой метод нахождения биномодальных коэффициентов носит название – метода наименьших квадратов.

Фон, аномалии и поверхность тренда

Выделение региональных закономерностей часто довольно хорошо решается путем аппроксимации эмпирических данных ортогональными полиномами первой степени.

В случаях, когда доля случайной изменчивости остается все же достаточно большей после аппроксимации линейными функциями, для выявления закономерной изменчивости более высокого порядка применяются полиномы второй, третьей и реже – более высоких степеней.

Фон, аномалии и поверхность тренда



Фон, аномалии и поверхность тренда

Задача отделения аномальных значений от фоновых не имеет строго математического решения и для ее решения используют различные подходы исходя из особенностей изучаемого объекта.

Однако, существенную помощь при выявлении аномалий и установления их природы могут оказать карты «остатков» («остаточных» аномалий) от тренда, которые строятся путем вычитания значений тренда из наблюдаемых значений поля в каждой точке.

Однофакторный и двухфакторный дисперсионный анализ

A	Уровни фактора B						Среднее
	B ₁	B ₂	...	B _j	...	B _q	
A ₁	x ₁₁	x ₁₂	...	x _{1j}	...	x _{1q}	\bar{x}_1
A ₂	x ₂₁	x ₂₂	...	x _{2j}	...	x _{2q}	\bar{x}_2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
A _i	x _{i1}	x _{i2}	...	x _{ij}	...	x _{iq}	\bar{x}_i
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
A _p	x _{p1}	x _{p2}	...	x _{pj}	...	x _{pq}	\bar{x}_p
Среднее	$\bar{x}_{.1}$	$\bar{x}_{.2}$...	$\bar{x}_{.j}$...	$\bar{x}_{.q}$	\bar{x}

$$\bar{x}_{ij.} = \frac{1}{n} \sum_{k=1}^n x_{ijk};$$

$$\bar{x}_{i..} = \frac{1}{qn} \sum_{j=1}^q \sum_{k=1}^n x_{ijk} = \frac{1}{q} \sum_{j=1}^q \bar{x}_{ij.};$$

$$\bar{x}_{.j.} = \frac{1}{pn} \sum_{i=1}^p \sum_{k=1}^n x_{ijk} = \frac{1}{p} \sum_{i=1}^p \bar{x}_{ij.};$$

Однофакторный и двухфакторный дисперсионный анализ

Схема вычисления дисперсий при двухфакторном дисперсионном анализе

Вид дисперсии	Сумма квадратов отклонений	Число степеней свободы	Дисперсия
Факторная по фактору A	$C_1 = nq \sum_{i=1}^p (\bar{x}_{i..} - \bar{x})^2$	$p - 1$	$S_1^2 = \frac{C_1}{p - 1}$
Факторная по фактору B	$C_2 = np \sum_{j=1}^q (\bar{x}_{.j.} - \bar{x})^2$	$q - 1$	$S_2^2 = \frac{C_2}{q - 1}$
Смешанная по факторам AB	$C_3 = n \sum_{i=1}^p \sum_{j=1}^q (\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x})^2$	$(p - 1)(q - 1)$	$S_3^2 = \frac{C_3}{(p - 1)(q - 1)}$
Остаточная	$C_4 = \sum_{i=1}^p \sum_{j=1}^q \sum_{k=1}^n (x_{ijk} - \bar{x}_{ij.})^2$	$pq(n - 1)$	$S_4^2 = \frac{C_4}{pq(n - 1)}$
Общая	$C = \sum_{i=1}^p \sum_{j=1}^q \sum_{k=1}^n (x_{ijk} - \bar{x})^2$	$npq - 1$	$S^2 = \frac{C}{npq - 1}$

Однофакторный и двухфакторный дисперсионный анализ

Проверка гипотезы о влиянии на изменчивость изучаемого свойства каждого фактора в отдельности и их совместного влияния производится по критерию Фишера:

$$F_A = S^2_1/S^2_4; \quad F_B = S^2_2/S^2_4; \quad F_{AB} = S^2_3/S^2_4.$$

Вид дисперсии	Сумма квадратов отклонений	Число степеней свободы	Дисперсия
Факторная по фактору A	$C_1 = nq \sum_{i=1}^p (\bar{x}_{i..} - \bar{x})^2$	$p-1$	$S^2_1 = \frac{C_1}{p-1}$
Факторная по фактору B	$C_2 = np \sum_{j=1}^q (\bar{x}_{.j.} - \bar{x})^2$	$q-1$	$S^2_2 = \frac{C_2}{q-1}$
Смешанная по факторам AB	$C_3 = n \sum_{i=1}^p \sum_{j=1}^q (\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x})^2$	$(p-1)(q-1)$	$S^2_3 = \frac{C_3}{(p-1)(q-1)}$
Остаточная	$C_4 = \sum_{i=1}^p \sum_{j=1}^q \sum_{k=1}^n (x_{ijk} - \bar{x}_{ij.})^2$	$pq(n-1)$	$S^2_4 = \frac{C_4}{pq(n-1)}$
Общая	$C = \sum_{i=1}^p \sum_{j=1}^q \sum_{k=1}^n (x_{ijk} - \bar{x})^2$	$npq-1$	$S^2 = \frac{C}{npq-1}$

Однофакторный и двухфакторный дисперсионный анализ

Данные схемы дисперсионного анализа основаны на свойствах нормального закона распределения и предположения о равенстве дисперсий на разных уровнях одного и того же фактора. *F-критерий* в случае выборок достаточно большого объема устойчив и для законов распределения, умеренно отклоняющихся от нормальных при небольшом различии в дисперсиях.

Для небольших выборок целесообразнее применять непараметрические критерии (Краскала – Уоллиса, Фридмана), использующих при расчете ранги (ранжированные ряды).

Двумерные статистические модели

Моделирование природных образований и процессов часто вызывают необходимость совместного рассмотрения нескольких их свойств с целью выяснения общей структуры изучаемого объекта.

В двумерной статистической модели объект исследований рассматривается как двумерная статистическая совокупность, а ее основной характеристикой является двумерная функция распределения случайных величин.

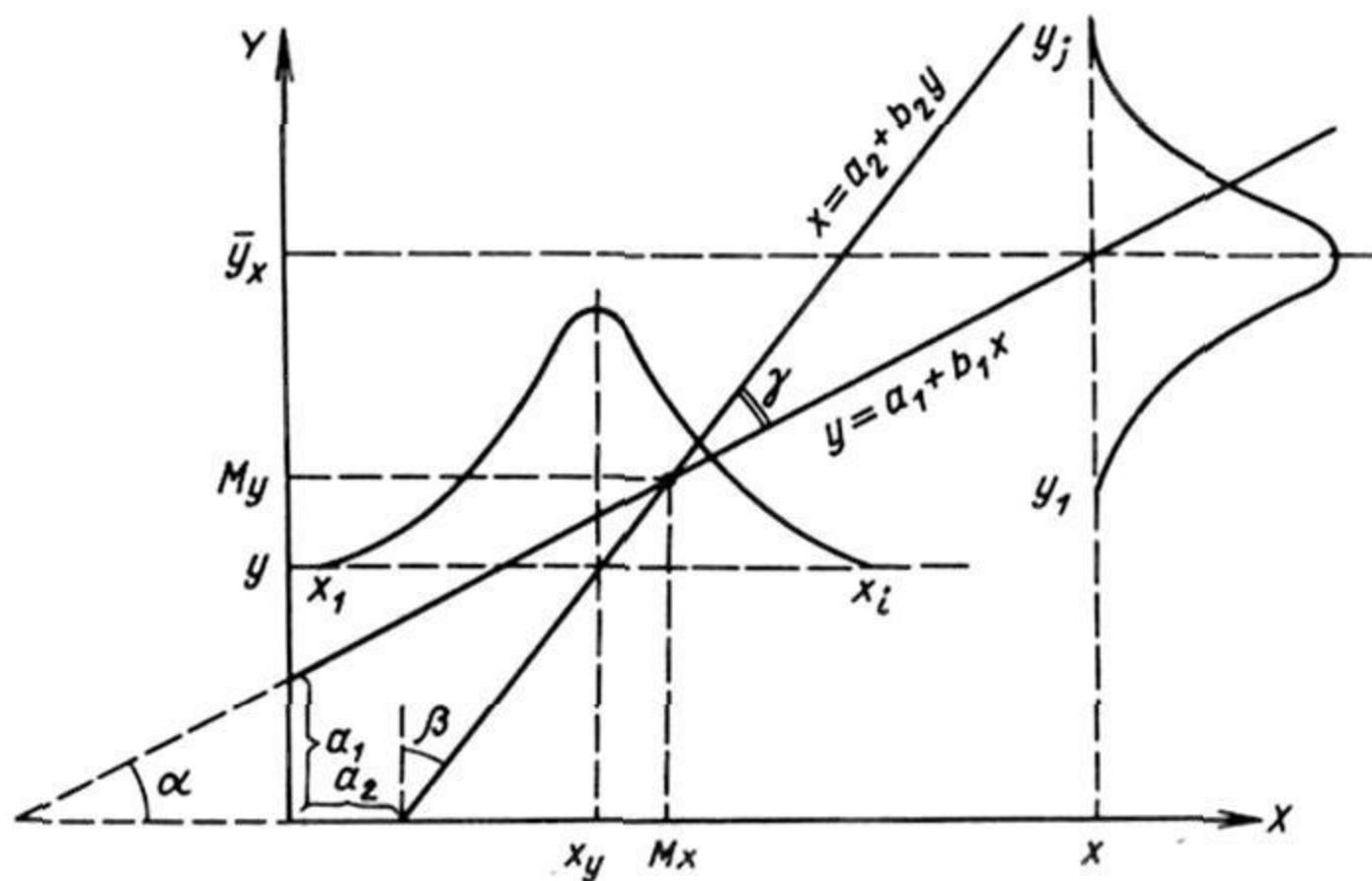
Между двумя случайными величинами проявляются вероятностные связи, когда заданному значению величины $X = x$ соответствует не какое-либо значение величины Y , а набор ее значений y_1, y_2, \dots, y_n , каждому из которых свойственна определенная вероятность p_1, p_2, \dots, p_n .

Функция распределения величины Y , соответствующая значению $X = x$, характеризуется математическим ожиданием μ_{y_x} и дисперсией $\sigma_{y_x}^2$

Распределение величины Y , соответствующие выбранным значениям величины X , называются *условными распределениями*, а дисперсии $\sigma_{y_x}^2$ - *условными дисперсиями*.

Геометрическое место точек, соответствующих центрам условных распределений y_x , называется *линией регрессии*, а уравнение этой линии – *уравнением регрессии*.

Двумерные статистические модели



Параметры двумерной случайной величины XY . M_x и M_y — математические ожидания величин X и Y ; \bar{Y} — центр условного распределения Y для $X=x$; \bar{X} — центр условного распределения X для $Y=y$; $b_1 = \operatorname{tg} \alpha$; $b_2 = \operatorname{tg} \beta$

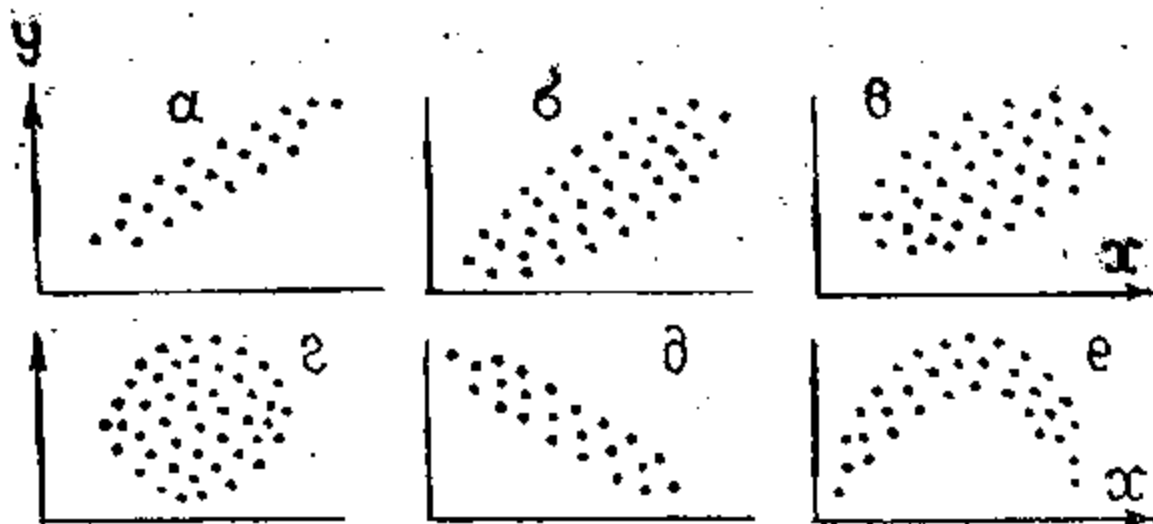
Двумерные статистические модели

Некоторые типы регрессионных уравнений:

- линейный $Y = A + B \times X$,
- гиперболический $Y = A + B/X$
- степенной $Y = A \times X^B$
- показательный $Y = A \times B^X$
- логарифмический $Y = A + B \times \lg X$
- парабалический $Y = A + B \times X + C \times X^2$

Двумерные статистические модели

Основными числовыми характеристиками двумерного распределения случайных величин являются показатели их связи: ковариация или корреляционный момент (момент связи), коэффициент корреляции и корреляционное отношение.



Графическое представление корреляционных зависимостей
а – сильная; б – средняя; в – слабая; г – отсутствует; а-в – прямая; д – обратная; а-в, д – линейная; е – нелинейная