

Министерство образования и науки Российской Федерации
Государственное образовательное учреждение
высшего профессионального образования
«Оренбургский государственный университет»

И.А. Никифоров

СТАТИСТИЧЕСКИЙ И АНАЛИЗ ГЕОЛОГИЧЕСКИХ ДАННЫХ

Рекомендовано Учёным советом Государственного образовательного учреждения высшего профессионального образования «Оренбургский государственный университет» в качестве учебного пособия для студентов, обучающихся по программам высшего профессионального образования направления «Прикладная геология»

Оренбург
ИПК ГОУ ОГУ
2010

УДК 55:004.4 (075.8)
ББК 26.3+32.973-018.2я73
Н62

Рецензент- профессор, доктор технических наук
В.И. Чепасов

Никифоров, И.А.

Н 62

Статистический анализ геологических данных: учебное пособие/И.А. Никифоров; Оренбургский гос. ун-т. – Оренбург: ОГУ, 2010. - 170 с.
ISBN

В учебном пособии излагаются основы применения статистических методов для решения широкого круга геологических задач. Рассмотрены методологические приёмы математической обработки геологической информации с помощью наиболее распространённых программных средств.

Учебное пособие предназначено для студентов специальностей направления 130300 – Прикладная геология

УДК 55:004.4 (075.8)

ББК 26.3+32.973-018.2я73

Н 1804010000

ISBN

© Никифоров И.А., 2010

©ГОУ ОГУ, 2010

Содержание

| | |
|---|----|
| Введение..... | 8 |
| 1 Введение в статистический анализ..... | 12 |
| 1.1 Элементы теории вероятности..... | 15 |
| 1.1.1 Предмет теории вероятностей..... | 15 |
| 1.1.2 Виды случайных событий..... | 16 |
| 1.1.2.1 Совместные события..... | 16 |
| 1.1.2.2 Несовместные события..... | 17 |
| 1.1.2.3 Полная группа событий..... | 17 |
| 1.1.2.4 Равновозможные события..... | 17 |
| 1.1.3 Классическое определение вероятности..... | 18 |
| 1.1.4 Операции над событиями..... | 21 |
| 1.1.4.1 Сложение событий..... | 21 |
| 1.1.4.2 Произведение событий..... | 22 |
| 1.1.5 Зависимые и независимые события..... | 23 |
| 1.1.5.1 Схема возвращённого шара..... | 23 |
| 1.1.5.2 Схема невозвращённого шара..... | 23 |
| 1.1.6 Основные формулы комбинаторики..... | 25 |
| 1.1.6.1 Перестановки..... | 26 |
| 1.1.6.2 Сочетания..... | 26 |
| 1.1.6.3 Размещения..... | 27 |
| 1.2 Случайные величины..... | 29 |

| | | |
|---------|--|----|
| 1.2.1 | Статистическое распределение случайной величины..... | 29 |
| 1.2.1.1 | Функции распределения..... | 29 |
| 1.2.1.2 | Описание распределения случайной величины..... | 32 |
| 1.2.1.3 | Моменты случайных величин | 36 |
| 1.2.2 | Примеры статистических распределений..... | 37 |
| 1.2.2.1 | Равномерное распределение | 37 |
| 1.2.2.2 | Нормальное распределение | 38 |
| 1.2.2.3 | Распределение Пирсона (хи - квадрат) | 41 |
| 1.2.2.4 | Распределение t Стьюдента | 42 |
| 1.2.2.5 | Распределение Фишера | 43 |
| 1.3 | Оценка параметров статистического распределения..... | 44 |
| 1.3.1 | Генеральная совокупность и выборка..... | 45 |
| 1.3.1.1 | Выборочный метод исследования..... | 45 |
| 1.3.2 | Оценка параметров генеральной совокупности..... | 50 |
| 1.3.2.1 | Статистики..... | 50 |
| 1.3.3 | Точечные оценки параметров распределения | 53 |
| 1.3.4 | Интервальные оценки параметров распределения | 55 |
| 1.3.4.1 | Доверительный интервал для среднего при известном σ | 57 |
| 1.3.4.2 | Доверительный интервал для среднего при неизвестном σ | 58 |
| 1.3.5 | Доверительный интервал для дисперсии..... | 60 |
| 1.4 | Статистическая проверка гипотез | 62 |
| 1.4.1 | Статистические критерии..... | 68 |

| | | |
|---------|--|-----|
| 1.4.1.1 | Односторонние и двусторонние критерии | 68 |
| 1.4.1.2 | Параметрические и непараметрические критерии | 70 |
| 1.5 | Гипотезы о параметрах распределения | 72 |
| 1.5.1 | Сравнение выборочного среднего с гипотетическим..... | 72 |
| 1.5.1.1 | Дисперсия генеральной совокупности известна | 72 |
| 1.5.1.2 | Дисперсия генеральной совокупности неизвестна | 75 |
| 1.5.2 | Сравнение двух выборочных дисперсий. Критерий Фишера | 77 |
| 1.5.3 | Сравнение двух выборочных средних. Критерий Стьюдента..... | 79 |
| 1.5.3.1 | Неизвестные, но равные генеральные дисперсии | 80 |
| 1.5.3.2 | Неизвестные и неравные дисперсии | 83 |
| 1.6 | Непараметрические методы проверки гипотез..... | 84 |
| 1.6.1 | Проверка распределения по χ^2 -критерию Пирсона..... | 84 |
| 1.6.2 | Критерий Вилкоксона | 89 |
| 1.7 | Корреляционный и регрессионный анализ при решении прогнозных задач | 93 |
| 1.7.1 | Корреляционный анализ..... | 95 |
| 1.7.1.1 | Корреляционное отношение | 98 |
| 1.7.1.2 | Регрессия..... | 99 |
| 1.7.1.3 | Множественная регрессия | 104 |
| 1.7.1.3 | Частная корреляция | 105 |
| 1.7.1.4 | Предположения и ограничения корреляционно-регрессионного анализа..... | 106 |
| 1.8 | Дисперсионный анализ | 108 |
| 1.8.1 | Теоретические предпосылки | 108 |

| | |
|--|-----|
| 1.8.2 Цели и методы дисперсионного анализа | 110 |
| 1.8.2.1 Структура дисперсии и разбиение суммы квадратов | 110 |
| 1.8.2.2 SS ошибок и эффекта..... | 111 |
| 1.8.2.3 Проверка значимости | 112 |
| 1.8.3 Геологические приложения дисперсионного анализа..... | 112 |
| 1.8.3.1 Однофакторный дисперсионный анализ | 112 |
| 1.8.3.2 Двухфакторный дисперсионный анализ | 117 |
| 1.9 Кластерный анализ | 123 |
| 1.9.1 Основная цель и терминология..... | 124 |
| 1.9.2 Область применения | 125 |
| 1.9.3 Процедура кластеризации | 126 |
| 1.9.4 Типы расстояний | 128 |
| 1.9.4.1 Евклидово расстояние | 128 |
| 1.9.4.2 Расстояние городских кварталов (Манхэттенское расстояние)..... | 129 |
| 1.9.5 Методы объединения в кластеры | 129 |
| 1.9.5.1 Иерархические методы..... | 130 |
| 1.9.5.2 Неиерархические методы..... | 134 |
| 1.10 Факторный анализ | 136 |
| 1.10.1 Подготовка исходных данных | 138 |
| 1.10.2 Вычисление матрицы взаимосвязей признаков | 139 |
| 1.10.3 Факторизация..... | 139 |
| 1.10.3.1 Методы факторизации..... | 142 |
| 1.10.3.2 Число выделяемых факторов..... | 145 |

| | | |
|----------|---|-----|
| 1.10.4 | Вращение факторов..... | 147 |
| 1.10.4.1 | Ортогональное вращение..... | 147 |
| 1.10.4.2 | Косоугольное вращение..... | 148 |
| 2 | Статистический анализ геологических данных..... | 155 |
| 2.1 | Восстановление геологического поля..... | 156 |
| 2.1.1 | Выделение региональной составляющей..... | 158 |
| 2.1.1.1 | Методы скользящего среднего..... | 158 |
| 2.1.1.2 | Аппроксимация алгебраическими полиномами..... | 161 |
| 2.1.1.3 | Аппроксимация гармониками..... | 163 |
| 2.1.1.4 | Аппроксимация сплайнами..... | 165 |
| 2.1.2 | Обособление локальной составляющей..... | 167 |
| | Список использованных источников..... | 169 |

Введение

Отечественная нефтяная и газовая промышленность близка к “пику” своего развития. Специалисты добывающих отраслей сталкиваются с непрерывным осложнением горно-геологических и природно-климатических условий работ. На практике оно выражается в необходимости освоения глубокопогруженных залежей, в поиске поднадвиговых месторождений и ловушек неантиклинального типа, в активизации поисково-разведочных работ на континентальном шельфе и за его пределами. Этот процесс сопровождается прогрессивным нарастанием информационных потоков, объединяющих сведения по большинству разделов современного естествознания. Их синтаксическая и смысловая фильтрация является главным делом геологического персонала, уровень подготовки которого должен обеспечить оперативность и качество выполнения этой непростой задачи. Успех её решения складывается из многих составляющих, но, вероятно, две из них доминируют.

Во первых- это ясное понимание научно-обоснованной методологии математической обработки крайне противоречивых эмпирических данных, характеризующих состояние недр.

Во вторых- это умение и практические навыки эксплуатации компьютерной техники и специализированного программного обеспечения, применяемого в отрасли.

Настоящее пособие предназначено для студентов начальных курсов и содержательно связано с проблематикой первой доминанты, т.е. с изучением основ аппарата математической статистики- главного средства первичной обработки геологических данных. В практическом плане, статистические алгоритмы реализованы в широком спектре пакетов прикладных программ, не имея явной геологической специализации. Умение свободно работать в их среде, отвлекаясь от конкретики предметной области, являются показателем профессиональной зрелости инженерного персонала, вообще, но для геологов - в особенности. Это объяс-

няется спецификой геологических исследований, которая заключается в невозможности непосредственного чувственного восприятия объектов недр. Оно заменяется их информационным описанием, служащим единственным источником сведений для производственно-управленческих решений и научных выводов.

Учитывая, что добыча геологических сведений- процесс чрезвычайно трудоёмкий и дорогостоящий, такое описание никогда не было и не будет достаточно подробным и обстоятельным. Именно это ставит на повестку сегодняшнего дня подготовку людей, способных действовать в особом междисциплинарном режиме, объединяющем геологию, математику и программирование. Выпускники университета должны обладать широким кругозором и глубокими знаниями во всех трех направлениях.

Сказанное определяет содержание курса и особенности его изложения в данном пособии, которое может быть полезно для преподавания дисциплин «Математические методы моделирования в геологии», «Применение ЭВМ в геологии» и «Основы компьютерных технологий решения геологических задач».

Оно рассчитано на студентов направления 130300- «Прикладная геология», обучающихся по специальностям:

130301- Геологическая съёмка, поиски и разведка месторождений полезных ископаемых.

130304- Геология нефти и газа.

Текст разбит на две неодинаковые по объёму части, связанные с основными, по мнению автора, вопросами обработки и интерпретации геологических данных:

- введение в статистический анализ;
- статистический анализ геологических данных.

Несмотря на то, что математическое содержание курса в пособии сведено к минимуму, для освоения предмета студент должен быть знаком с принципами дифференциального и интегрального исчисления, основами матричной алгебры и уметь работать с компьютером.

Структура изложения материала приведена в нижеследующем списке:

1) введение в статистический анализ:

- роль теории вероятностей и математической статистики в геологии;
- законы статистических распределений;
- оценки параметров статистического распределения;
- проверка статистических гипотез;
- корреляционный и регрессионный анализ;
- дисперсионный анализ;
- кластерный анализ;
- факторный анализ;

2) статистический анализ геологических данных:

- сглаживание скользящим средним;
- полиномиальная аппроксимация геологических данных;
- аппроксимация данных рядами Фурье;
- аппроксимация числовых полей сплайнами.

Изложение курса сопровождается практическими занятиями, в ходе которых студенты должны закрепить навыки компьютерной обработки данных в среде программных пакетов MS Excel, Statistica 6, MatLab 6.5, MathCAD 2001.

В ходе практических и лабораторных занятий студентам необходимо выполнить следующие задания:

- компьютерная реализация алгоритмов преобразований градусных мер;
- расчёт искривления ствола скважины по данным инклинометрических наблюдений;
- статистическая обработка тестового массива геологических данных;
- проверки гипотез согласия с помощью ранговых критериев;
- расчёт уравнения регрессии методом наименьших квадратов;
- кластерный анализ гидрохимической выборки;
- факторный анализ данных методом главных компонент;
- сглаживание учебного набора данных, распределённых по стволу скважины,

методом скользящего среднего.

Программой курса предусматривается выполнение курсового проекта, предназначенного для закрепления полученных знаний по статистическому анализу геологических данных и демонстрации студентами своих способностей.

1 Введение в статистический анализ

По большому счёту, главной целью естественных наук является описание, объяснение и прогноз явлений действительности на основе законов природы. Они в реальном времени воздействуют на все процессы и объекты мироздания, из которых только ничтожная часть может быть нами осознана и исследована.

Современный уровень науки позволяет рассчитать орбиту спутника, передавать информацию посредством радиосигналов, создавать довольно сложные механизмы и многое другое. Однако, мы не в состоянии излечивать онкологические заболевания, решить проблему эффективной генерации электроэнергии, антигравитации и т.п. В этой связи интересно, каким образом всё-таки добыты основные человеческие знания и может ли этот опыт быть использован для их расширения?

Чтобы ответить на этот вопрос необходимо вспомнить о тех далёких временах, когда людям не было известно не одного закона природы. Они жили в том же мире, что и мы, сталкиваясь с теми же явлениями и событиями, объяснить которые сейчас в силах любой грамотный человек.

Невероятная любознательность древних- вот ключ к действительно судьбоносным открытиям, которые до сих пор питают современную цивилизацию. Даже есть пословица- «Нельзя открыть ничего такого, о чём бы уже не писали древние греки!». Разумеется, это в равной степени относится к древним египтянам и китайцам, коренным народам Америки и другим мудрецам, сумевшим организовать научную систему наблюдений всего непонятого и непознанного. Способы и результаты этой организации поражают до сих пор. Грандиозные пирамиды Египта и Мексики, компас и порох, карта Птолемея и великие географические открытия 15 века тому доказательство!

Таким образом, можно утверждать, что в основе научных знаний лежит наблюдение. Везде, где результат опыта неочевиден, а это бывает в огромном большинстве случаев, необходимо обратиться к опыту предков. На самом деле, разни-

ца между нами не так уж и велика. Все мы равны перед непознанным, но к счастью за столетия цивилизации разработан математический аппарат, сокращающий путь от умозрительных рассуждений и сомнительных эмпирических закономерностей к более-менее совершенным вариантам законов природы. Этот аппарат называется «Математическая статистика», и в его основе лежит многократное наблюдение природного феномена в одинаковых условиях.

Смысл статистической обработки состоит в том, что мы сначала не вдаёмся в изучение конкретных причин (законов), приводящих к тому или иному событию. Вместо этого изучаются сами события многократно воспроизводимые физическим или иным экспериментом. Однако, после статистической обработки этих наблюдений выявляются количественные закономерности, влияющие на результаты проводимого опыта. Другими словами статистический прогноз в некотором смысле является прогнозом «от противного». Например, мы в реальном времени можем рассматривать график изменения курса доллара на валютной бирже ФОРЕКС. Понятно, что за каждым скачком или падением курса стоят конкретные события, но времени вдаваться в такие подробности у трейдеров просто нет. В своей работе они руководствуются формой графика, его трендом, т.е. статистикой. Судя потому, что среди биржевых игроков много миллионеров, такой подход часто бывает полезен.

Приведём другой пример из области газовой индустрии. Был установлен факт незапланированного падения давления газа в подземном хранилище газа (ПХГ). Разумеется, специалисты ПХГ изначально понимали, что развитие неизвестного пока негативного фактора подчиняется главным физическим законам, например закону Дарси, Ньютона, Гука, Паскаля и многим другим. Их действие распространяется одновременно на все инженерные сооружения комплекса и недра как вместилище газовой полости. Однако, в практическом плане эти знания малополезны. Параметры этих законов, складываясь под влиянием непрерывно изменяющихся режимов эксплуатации и характеристик внешней среды, не поддаются непосредственным замерам, в связи с чем классический прогноз развития ситуации просто невозможен. Вместо этого имеет смысл создать систему посто-

янной регистрации состояний ключевых узлов всего комплекса. Это позволит путём статистической обработки результатов найти количественную зависимость падения давления с конкретным агрегатом или с состоянием пласта-коллектора.

Сказанное подводит нас к одному из множества определений математической статистики, которое звучит так:

Математическая статистика- это наука, изучающая методы обработки результатов наблюдений случайных массовых явлений, обладающих статистической закономерностью, с целью выявления этой закономерности.

С помощью статистических методов можно получать ответы на актуальные вопросы геологической и промысловой практики и проверять гипотезы. Например:

- оказывает ли влияние применение промывочных жидкостей на вскрытие продуктивных пластов?
- различаются ли средние проходки за рейс при разных способах бурения?
- как зависит коэффициент извлечения нефти от темпов отбора жидкости, плотности нефти и проницаемости коллекторов?

С практической точки зрения статистика как наука, прежде всего, связана с числовыми данными- с множеством фиксируемых измерений. Она включает в себя систему методов регулярного сбора, организации и обобщения данных. Они используются для получения содержательных выводов с помощью методологии теории вероятностей.

В отличие от математической статистики, имеющей дело с результатами выборочных наблюдений случайных явлений, теория вероятности изучает идеальные модели случайных явлений и величин. Она играет роль теоретической базы для статистических выводов, делая их объективно значимыми и полезными.

1.1 Элементы теории вероятности

1.1.1 Предмет теории вероятностей

Наблюдаемые нами события (явления) можно подразделить на следующие три вида: достоверные, невозможные и случайные.

Достоверным называют событие, которое обязательно произойдёт, если будет осуществлена определённая совокупность условий S .

Например, если в сосуде содержится вода при нормальном атмосферном давлении и температуре $20\text{ }^{\circ}\text{C}$, то событие «вода в сосуде жидкая» есть достоверное. В этом примере заданное атмосферное давление и температура составляют совокупность условий S .

Невозможным называют событие, которое заведомо не произойдёт, если будет осуществлена определённая совокупность условий S .

Например, событие «вода в сосуде находится в твёрдом состоянии» заведомо не произойдёт, если будет осуществлена совокупность условий предыдущего примера.

Случайным называют событие, которое при осуществлении совокупности условий S может либо произойти, либо нет.

В математике принято вместо фразы «совокупность условий S осуществлена» говорить кратко «произведено испытание». ***Таким образом, событие рассматривается как результат испытания.***

Например, если брошена монета, то может выпасть или «герб» или «надпись». Поэтому событие «при бросании монеты выпал герб» - случайное.

Мы должны понимать, что каждое случайное событие, в частности выпадение «герба» есть следствие воздействия очень многих причин. Это и сила броска, форма монеты, гладкость поверхности и т.п. Учесть их точное влияние на результат нельзя, поскольку число их слишком велико и законы действия оценке не поддаются. Поэтому теория вероятностей не ставит перед собой задачу предска-

зять исход единичного события- она просто не в состоянии этого сделать [1].

По иному обстоит дело, если рассматриваются случайные события, которые могут многократно наблюдаться при осуществлении одних и тех же условий S , т.е. если речь идёт о массовых однородных случайных событиях. Оказывается, что достаточно большое число однородных случайных событий независимо от их конкретной природы подчиняется определённым закономерностям, а именно вероятностным закономерностям. Их установлением и занимается теория вероятностей.

Предметом теории вероятностей является изучение вероятностных закономерностей массовых однородных случайных событий.

Методы теории вероятности широко применяются в различных отраслях естествознания и техники. Эти направления практически сливаются в геологических предметных областях, в связи с чем здесь знание основ теории вероятностей играет особую роль. Дело в том, что изучаемые геологами объекты в большинстве своём недоступны для непосредственных наблюдений. Они обычно производятся дистанционно с помощью горных выработок и буровых скважин или развитой приборной базы. При этом, естественно, нельзя говорить об однозначности полученных результатов, поскольку они осложняются множеством случайных факторов маскирующих цель геологических исследований.

1.1.2 Виды случайных событий

1.1.2.1 Совместные события

Несколько событий называют *совместными*, если в результате испытания наступление одного из них не исключает появления других. Например, при подбрасывании трёх монет выпадение цифры на одной не исключает появления цифр на других монетах.

1.1.2.2 Несовместные события

Несколько событий называют *несовместными* в данном испытании, если появление одного из них исключает появление других. Например, студент может провалиться на зачёте или сдать зачёт. Эти два события несовместны.

1.1.2.3 Полная группа событий

Несколько событий образуют *полную группу*, если в результате испытания проявится хотя бы одно из них. Другими словами, появление хотя бы одного из событий полной группы есть достоверное явление. В частности, если события полной группы попарно несовместны, то в результате испытания появится одно и только одно из этих событий.

Например, стрелок произвёл выстрел по цели. Обязательно произойдёт одно из следующих событий: попадание или промах. Эти два несовместных события образуют полную группу.

1.1.2.4 Равновозможные события

События называют *равновозможными*, если есть основания считать, что ни одно из них не является более возможным, чем другое.

Например:

- появление герба или цифры при бросании монеты- равновозможные события;
- появление того или иного числа очков на брошенной игральной кости- равновозможные события.

1.1.3 Классическое определение вероятности

В повседневной жизни мы часто используем слово «вероятный». Например, «к вечеру, вероятно, пойдёт дождь», «это невероятный случай», «вероятнее всего Сидоров пропустит лекцию».

При употреблении этого слова интуитивно оценивается возможность наступления того или иного события. Можно даже сказать, что одно событие наступит чаще, чем другое. Так студент Сидоров чаще пропускает лекции, чем идут дожди. В принципе, он может быть и не такой уж и лодырь, поскольку дожди в нашем засушливом крае случаются редко. В любом случае при оценке вероятности того или иного события нам помогает здравый смысл и жизненный опыт. Однако иногда случаются события, предсказать возможность появления которых крайне трудно. Например, событие «при бросании монеты герб выпадет два раза из пяти» может, конечно, произойти, но не тогда, когда вы на это рассчитываете.

Приведём выдержку из книги «Математика для любознательных» великого популяризатора науки Якова Исидоровича Перельмана, ставящую под сомнение рассудительность и здравый смысл.

У студента спрашивают:

«Верно ли утверждение, что среди 10 пешеходов, пересекающих улицу, всегда будет не менее 2 женщин?»

Без особых раздумий, тот отвечает: «Разумеется, поскольку число мужчин и женщин на Земле приблизительно одинаково!»

Оба поглядели в окно. Как раз, в этот момент, через перекрёсток проходила рота солдат, в которой, разумеется, не было ни одной женщины.

Из приведённых примеров видно, что каждое событие обладает определённой степенью возможности наступления, т.е. определённой оценкой. Такую оценку и называют *вероятностью* события.

Вероятность события- это числовая мера объективной возможности его появления.

Каким же образом определяется вероятность?

Предположим, что подбрасывают игральную кость и выигрывают, если выпадает 1 или 2 очка. Поскольку существует 6 равновозможных чисел и выигрыш наступает, при появлении любого из двух исходов (двух чисел), то вероятность выигрыша вычисляется как отношение двух выигрышных случаев к шести возможным и будет равна $2/6$ или 0.3333 .

Вероятностью появления события A называют отношение числа исходов, благоприятствующих наступлению этого события, к общему числу всех равновозможных, несовместных элементарных исходов, образующих полную группу.

Обозначим число благоприятствующих событию A исходов через M , а число всех исходов через N , тогда

$$P(A) = \frac{M}{N},$$

где M - целое неотрицательное число в интервале $0 \leq M \leq N$.

Это и есть формула классического определения вероятности.

С вероятностью события связано понятие *относительной частоты* (частоты) появления события. Так обозначается отношение числа испытаний m , при которых событие появилось к общему числу проведённых испытаний n :

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

где m -целое неотрицательное число в интервале $0 \leq m \leq N$

Несмотря на совершенно одинаковый вид выражений, приведённых для вычисления вероятности и относительной частоты, между ними есть принципиальное различие. Для определения вероятности выигрыша в кости (см. пример на предыдущей странице) нам надо знать «модель игры», в данном случае- это кость с шестью гранями. Мы можем определить наши шансы на успех *теоретически*, без подбрасывания кости, т.е. *априорно*.

Во втором случае мы определяем частоту только по результатам опыта, т.е. *апостериорно*. С увеличением числа испытаний относительная частота прояв-

ляет тенденцию к стабилизации, приближаясь с затухающими отклонениями к постоянному числу, называемому статистической вероятностью. Так, например, известный французский естествоиспытатель Бюффон по 4040 бросаниям монеты получил относительную частоту появления герба, равную 0.50693.

У английского статистика Пирсона по результатам 23000 бросаний монеты относительная частота оказалась равной 0.5005.

Приведём без доказательства три свойства вероятности, вытекающие из классического определения:

Вероятность достоверного события равна 1.

Действительно, если событие достоверно, то каждый исход испытания благоприятствует событию. В этом случае $M=N$, следовательно,

$$P(A)=M/N=1.$$

Вероятность невозможного события равна 0.

Действительно, если событие невозможно, то ни один из элементарных исходов не благоприятствует событию. В этом случае $M=0$, следовательно,

$$P(A)=M/N=0/N=0.$$

Вероятность случайного события есть положительное число, между нулём и единицей.

Действительно, случайному событию благоприятствует лишь часть из общего числа элементарных исходов испытания.

В этом случае $0 < M < N$, т.е. $0 < M/N < 1$, или, $0 < P(A) < 1$.

Итак, вероятность любого события удовлетворяет двойному неравенству:
 $0 \leq P(A) \leq 1$.

1.1.4 Операции над событиями

Над событиями можно производить некоторые логические операции, аналогичные *булевым*. Они позволяют упростить форму записи и построение логических рассуждений. Мы рассмотрим две из них- сложение и умножение событий.

1.1.4.1 Сложение событий

Суммой нескольких событий называется событие, состоящее в наступлении хотя бы одного из них в результате испытания.

Пусть имеются два совместных события A и B . Тогда $A+B$ означает, что наступит событие A или B или оба события вместе. Если же эти события несовместны, то событие $A+B$ означает, что наступит событие A или B .

Про событие $A+B+C$ можно сказать, что оно состоит в наступлении одного из событий A, B, C или в совместном наступлении пары событий A и B , A и C , B и C или в совместном наступлении этих трёх событий.

Например, в урне находятся красные, белые и чёрные шары. Допустим следующие события: A - «вынут красный шар», B - «вынут белый шар» и C -«вынут чёрный шар». Тогда событие $A+B$ означает, что произошло событие «не вынут чёрный шар», а событие $B+C$ означает, что произошло событие «не вынут красный шар».

Теорема. *Вероятность появления одного из двух несовместных событий, безразлично какого, равна сумме вероятностей этих событий.*

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

Пусть в урне для голосования находится 30 шаров: 10 красных, 5 синих и 15 белых. Найти вероятность появления цветного шара.

Чтобы решить эту задачу необходимо просто рассуждать логически. После-

довательность наших умозаключений может быть такой:

- а) появление цветного шара означает появление либо красного либо синего шара;
- б) вероятность появления красного шара (событие A) составляет:
 $P(A)=10/30=1/3$;
- в) вероятность появления синего шара (событие B) составляет:
 $P(B)=5/30=1/6$;
- г) события A и B несовместны (появление шара одного цвета исключает появление шара другого цвета), поэтому теорема сложения вероятностей применима: $P(A+B)=P(A)+P(B)=1/3+1/6=1/2$.

1.1.4.2 Произведение событий

Произведением нескольких событий называется событие, состоящее в совместном наступлении всех этих событий в результате испытания.

Произведение обозначается знаком « \times », который можно опускать.

Например, если произошло событие ABC это означает, что произошло событие « A и B и C ».

Например, бросается игральный кубик. Рассмотрим следующие события: A - «число выпавшее число очков меньше 5», B - «число выпавших очков больше 2» и C -« число выпавших очков чётное». Тогда событие ABC означает, что выпало 4 очка!

Перед тем как рассматривать более сложный вариант умножения вероятностей следует обратить внимание на то, что сами события могут зависеть одно от другого.

1.1.5 Зависимые и независимые события

Представим себе такую ситуацию: в урне для голосования находятся два белых и три чёрных шара.

Вопрос: чему равна вероятность выбора белого шара при извлечении из урны? Ответ на этот вопрос зависит от схемы голосования, которое может происходить по двум схемам:

- схеме возвращённого шара, т.е. шар возвращается в урну;
- схеме невозвращённого шара, т.е. шар в урну не возвращается.

1.1.5.1 Схема возвращённого шара

Шар после выбора возвращается в урну. Пусть событие A - появление белого шара при первом извлечении. Поскольку шаров всего 5, а белых- 2, то вероятность $P(A)=2/5$.

Пусть событие B - появление белого шара при втором извлечении. Поскольку белый шар был возвращён в урну, вероятность его выбора не изменилась: $P(B)=2/5$.

Приведённая схема гарантирует *независимость* события, заключающегося в выборе белого шара из урны.

События A и B называются независимыми, если вероятность каждого из них не зависит от того, произошло или нет другое событие.

1.1.5.2 Схема невозвращённого шара

Шар после испытания не возвращается в урну, следовательно, в ней остаются только один белый и три чёрных шара. Чему равна вероятность события B в этих условиях?

Её обозначают $P(B/A)$ и она равна уже $1/4$, поскольку в урне остался только

один белый шар. Такая вероятность называется условной вероятностью, а события A и B - зависимыми.

События A и B называются зависимыми, если вероятность каждого из них зависит от того, произошло или нет другое событие.

Понятие условной вероятности имеет важное методологическое значение. Естествоиспытатель, пытающийся раскрыть суть некоторого явления, стремится зафиксировать условия, при которых это явление проявляется. Например, в геологии часто осуществляется поиск золота по аксессуариям, алмазы связывают с кимберлитами.

Теорема. Вероятность произведения двух зависимых событий A и B равна произведению вероятности одного из них на условную вероятность другого.

$$P(AB)=P(B) \times P(B/A)$$

За сухой формулировкой теоремы кроется глубокий смысл, который может помочь в решении обыденных, казалось бы, вопросов. В качестве примера рассмотрим следующую ситуацию.

Студент пришёл на экзамен, выучив из 25 вопросов только 20. Известно, что экзаменатор задаёт только три вопроса, но они связаны друг с другом.

Какова вероятность того, что студент ответит на все три вопроса.

Решение.

Обозначим условия задачи:

A -студент знает все три вопроса.

$A1$ -студент знает первый вопрос.

$A2$ -студент знает второй вопрос.

$A3$ -студент знает третий вопрос.

События $A1$, $A2$, $A3$ - зависимые.

$$P(A)=P(A1) \times P(A2/A1) \times P(A3/A1 \cdot A1)$$

$$P(A)=20/25 \times 19/24 \times 18/23=0.496$$

Таким образом, студент недоучил всего 20 %, а вероятность успеха меньше половины!

Операции над событиями имеют полезную геометрическую интерпретацию (рисунок 1), которую называют диаграммами Венна, по имени английского логика Джона Венна (1834-1923 г.г.).

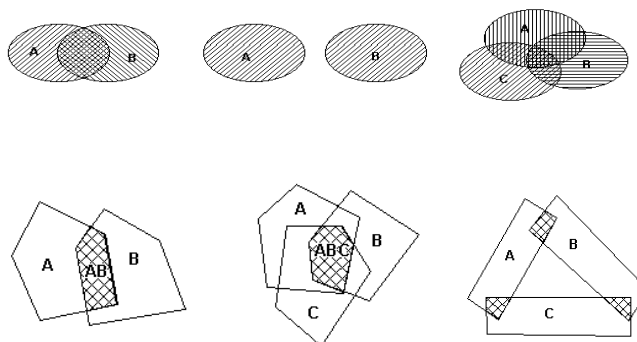


Рисунок 1- Диаграммы Эйлера-Венна

На рисунке представлены следующие случаи:

- а) сложение двух событий;
- б) случай двух несовместных событий;
- в) сложение трёх несовместных событий;
- г) произведение двух событий;
- д) произведение трёх событий;
- е) случай, когда произведение событий A , B и C является невозможным, но попарно они совместны.

1.1.6 Основные формулы комбинаторики

Очень часто при решении практических вопросов оценки вероятности тех или иных событий приходится вычислять число возможных исходов, т.е. объем полной группы событий. Эту задачу обслуживает специальный раздел дискретной математики- комбинаторика.

Комбинаторика изучает количества комбинаций, подчинённых определённым условиям, которые можно составить из элементов заданного множества. При

непосредственном вычислении вероятностей используются формулы комбинаторики, наиболее употребительные из которых приводятся здесь без доказательств.

1.1.6.1 Перестановки

Перестановками называются комбинации, состоящие из одних и тех же n различных элементов и отличающиеся только порядком их расположения.

Число всех возможных перестановок определяется выражением:

$$P_n = n!,$$

где $n! = 1 \times 2 \times 3 \dots n$.

Напомним, что восклицательный знак в формуле не означает, что её надо произносить громко или с особым выражением. Это просто обозначение факториала, т.е. произведения n натуральных чисел от 1 до n . Вычисление его довольно трудоёмко и обычно, для этой цели используется функция MS Excel, которая называется «ФАКТР».

Таким образом:

- число перестановок из одного элемента равно 1.
- число перестановок из двух элементов a, b равно двум: ab, ba ;
- число перестановок из трёх элементов a, b, c равно шести: $abc, acb, bac, bca, cab, cba$.

Вычислим теперь, сколько трёхзначных чисел можно составить из цифр 1, 2, 3, если каждая входит в число только один раз?

Решение:

Искомое число трёхзначных чисел $P_3 = 3! = 1 \times 2 \times 3 = 6$.

1.1.6.2 Сочетания

Сочетаниями из n элементов по m в каждом называются такие соединения, из которых каждое содержит m элементов, взятых из числа данных n элементов, и которые отличаются друг от друга, по крайней мере, одним элементом. Таким об-

разом:

- из одного элемента можно составить только одно сочетание;
- из двух элементов a и b можно составить два сочетания по одному элементу: a , b и лишь одно сочетание по два элемента: ab ;
- из трёх элементов a , b , c можно составить три сочетания по одному элементу: a , b , c , три сочетания по два элемента: ab , ac , bc и одно сочетание по три элемента: abc .

Число сочетаний из n элементов по m в каждом при условии, что m находится в интервале $0 \leq m \leq n$ вычисляется согласно следующим выражениям:

$$C_n^m = \frac{A_n^m}{P_m} \quad \text{или} \quad C_n^m = \frac{n!}{m!(n-m)!}$$

Разберём следующий пример. Пусть правление коммерческого банка выбирает из 10 кандидатов трёх человек на *одинаковые* должности (Все 10 кандидатов имеют равные шансы). Сколько всевозможных групп по три человека можно составить из 10 кандидатов?

Решение:

$$C_{10}^3 = 10! / (3! \times 7!) = 120$$

1.1.6.3 Размещения

Размещениями из n элементов по m в каждом называются такие соединения, из которых каждое содержит m элементов, взятых из числа данных n элементов, и которые отличаются друг от друга либо самими элементами (хотя бы одним), либо лишь порядком их расположения.

Итак:

- а) из одного элемента можно составить только одно размещение;
- б) из двух элементов a и b можно составить два размещения по два элемента: ab , ba ;
- в) из трёх элементов a , b , c можно составить:
 - 1) три размещения по одному;

- 2) шесть размещений по 2 элемента- ab, ac, ba, bc, ca, cb ;
- 3) шесть размещений по 3 элемента $abc, acb, bac, bca, cab, cba$.

Число размещений из n элементов по m обозначается как A_n^m . Оно вычисляется по формуле:

$$A_n^m = \frac{n!}{(n-m)!}$$

Разберём ещё один пример. Пусть правление коммерческого банка выбирает из 10 кандидатов трёх человек на *различные* должности (Все 10 кандидатов имеют равные шансы). Сколько всевозможных групп по три человека можно составить из 10 кандидатов?

Решение:

$$N = A_{10}^3 = 10 \times 9 \times 8 = 720$$

Вопросы для самопроверки:

- 1 Может ли быть невозможное событие достоверным?
- 2 Зависит ли количество случайных событий от уровня развития личности?
- 3 Сколькими способами можно усадить 5 гостей за круглым столом?
- 4 Слово «нефть» составлено из букв разрезной азбуки. Наудачу извлекаются четыре карточки и складываются в ряд друг за другом в порядке появления.
 - сколько возможных соединений можно получить из букв этого слова?
 - какова вероятность получения при этом слова «тень»?

1.2 Случайные величины

Случайной величиной или стохастической переменной, называют величину, значение которой зависит от случая.

Так, общее число очков, выпавшее при троекратном бросании игральной кости, является случайной величиной. В зависимости от природы явлений случайная величина может быть дискретной или непрерывной.

Непрерывной случайной величиной называется величина, которая при испытании может принять любое значение из заданного диапазона.

Примерами непрерывной случайной величины может служить содержание химического элемента в горной породе, пластовое давление или температура.

Дискретные случайные величины в отличие от непрерывных могут принимать лишь избранные значения на числовой оси.

Примерами дискретных случайных величин являются показания цифрового измерительного прибора или число бракованных изделий m при выборочном контроле партии объемом n изделий. Распределение дискретной величины представляет собой линейчатую функцию. Каждое значение этой функции является вероятностью того, что рассматриваемая случайная величина будет обладать конкретным значением.

1.2.1 Статистическое распределение случайной величины

1.2.1.1 Функции распределения

Если X -случайная величина, а её значение x изменяется от $-\infty$ до $+\infty$, то вероятность события $X < x$ представляет собой неубывающую функцию от x , которая называется *функцией распределения* $F(x) = P(X < x)$. Она однозначно определяет вероятность того, что случайная величина принимает заданное значение

или принадлежит к некоторому заданному интервалу.

В практике особенно важны два случая:

Первый- если случайная величина принимает конечное число значений $x_1, x_2, x_3, \dots, x_n$ с вероятностями $p_1, p_2, p_3, \dots, p_n$. Тогда $F(x)$ представляет собой *ступенчатую функцию*, график которой в точке x_i имеет скачок, равный по величине p_i .

Сумма произведений дискретной случайной величины на соответствующие им вероятности называется математическим ожиданием дискретной случайной величины.

Она же является её взвешенным средним значением, где роль весов играют вероятности.

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

Причиной для выбора такого названия состоит в том, что среднее значение случайной величина есть оценка, которую ожидают получить.

Второй случай связан с непрерывными случайными величинами, когда они могут принимать бесконечно много значений. Это возможно лишь тогда, когда вероятностное пространство, на котором определена случайная величина, состоит из бесконечного числа элементарных событий. Тогда распределение задается набором вероятностей $P(a < X < b)$ для всех пар чисел a, b таких, что $a < b$. При этом ясно, что $P(a < X < b) = F(b) - F(a) = \int_a^b f(x) dx$.

Если же перейти от a и b к $-\infty$ до $+\infty$, то вероятность случайного события в бесконечном интервале будет равна 1, т.е.

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

Функция $f(x)$ называется *плотностью вероятности величины x* . Выражаясь популярно можно утверждать, что $f(x)dx$ есть вероятность того, что значение случайной величины заключено в пределах между x и $x+dx$. Другими словами плотность вероятности $f(x)$ является производной функции распределения $F(x)$.

Итак, если $f(x)dx$ - вероятность попадания случайной величины x в интервал

длиной Δx , то по аналогии с дискретными случайными величинами можно определить математическое ожидание непрерывной случайной величины. Это- несобственный интеграл вида:

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

Используемые в прикладных исследованиях функции распределения бывают либо дискретными, либо непрерывными, либо их комбинациями.

Дискретные функции распределения соответствуют дискретным случайным величинам, принимающим конечное число значений или же значения из множества, элементы которого можно перенумеровать натуральными числами (такие множества в математике называют счетными). Их график имеет вид ступенчатой лестницы (рисунок 2).

Пусть число X дефектных упаковок цемента в партии принимает значение 0 с вероятностью 0.3, значение 1 с вероятностью 0.4, значение два с вероятностью 0.2 и значение три с вероятностью 0.1. График функции распределения случайной величины X изображен на рисунке 2.

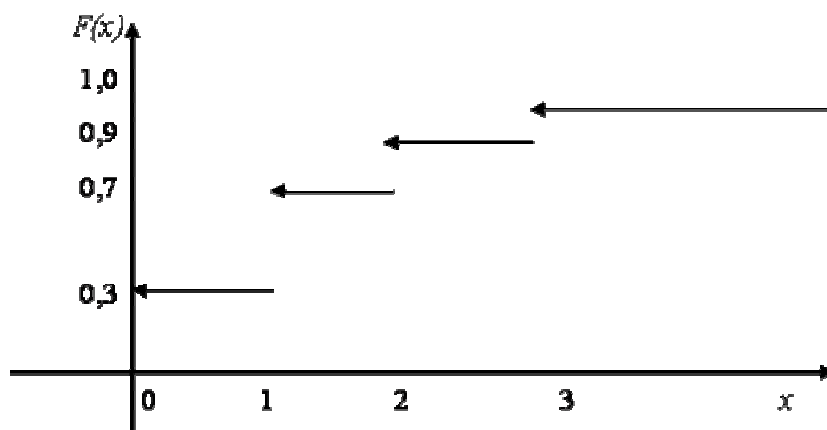


Рисунок 2- График функции распределения числа дефектных изделий.

Непрерывные функции распределения не имеют скачков. Они монотонно возрастают при увеличении аргумента – от 0 при $x \rightarrow -\infty$ до 1 при $x \rightarrow +\infty$. Так, например, выглядят графики функции распределения и плотности вероятности ошибок астрономических наблюдений (рисунок 3).

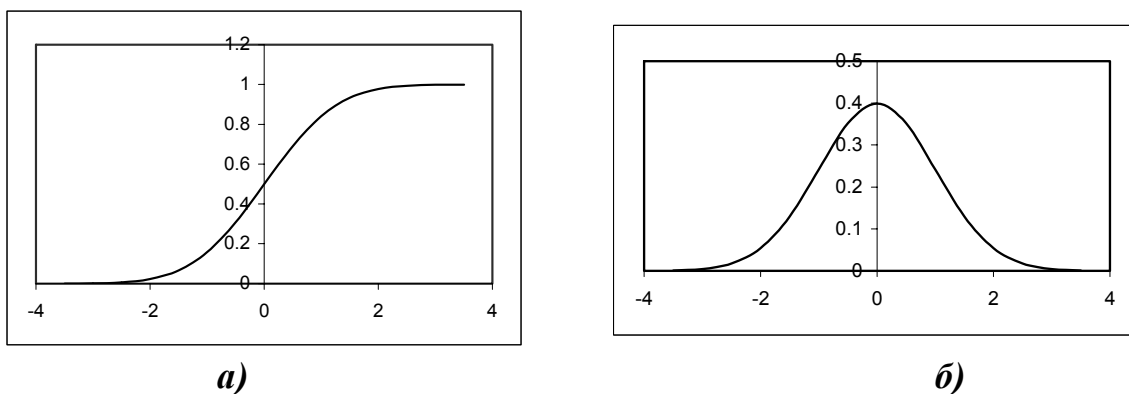
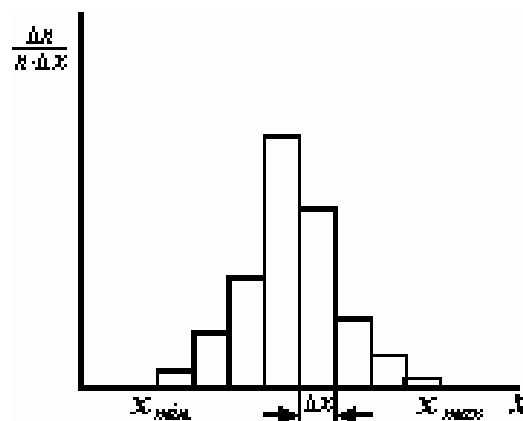


Рисунок 3- Графики функции распределения а) и плотности вероятностей б)

1.2.1.2 Описание распределения случайной величины

Рассмотрение вопроса начнем с предполагаемого эксперимента, в котором выполняются многократные прямые измерения какой-то случайной физической величины, проводимые без изменения условий эксперимента. Закономерности распределения величины отражаются на специальном графике, который называется статистической гистограммой.

Гистограмма представляет собой ступенчатую диаграмму, показывающую как часто при измерениях появляются результаты, попадающие в тот или иной интервал Δx между наименьшим x_{min} и наибольшим x_{max} из измеренных значений величины x . Гистограмму строят в следующих координатах: по оси абсцисс откладывают измеряемую величину x , по



оси ординат $-\Delta n/n\Delta x$. Здесь n – полное количество проведенных измерений, Δn – количество результатов, попавших в интервал $[x, x+\Delta x]$.

Отношение $\Delta n/n$ есть доля результатов, оказавшихся в указанном интервале. Оно имеет смысл вероятности попадания результата отдельного измерения в данный интервал. Выражение $\Delta n/(n \times \Delta x)$, получаемое после деления $\Delta n/n$ на ши-

рину интервала Δx , приобретает смысл плотности вероятности.

При очень большом количестве измерений ($n \rightarrow \infty$) весь диапазон изменения величины x можно разбить на бесконечно малые интервалы dx , как это делается в математике, и найти количество результатов dn в каждом из них. В этом случае гистограмма превратится в график плотности вероятности.

Распределение выступает в роли исчерпывающей характеристики случайной величины. Закон распределения можно задать в виде функционального выражения, графика, таблицы или каким-то другим способом. При любом варианте задания устанавливается связь между вероятностью того, что результат однократного измерения случайной величины попадет в заданный интервал возможных значений, и шириной этого интервала.

Распределение содержит наиболее полную информацию о случайной величине, однако пользоваться им не всегда удобно. Опираясь на результаты проведенного эксперимента, вместо функции распределения обычно пользуются числовыми характеристиками меры:

- среднего положения распределения (арифметическое среднее значение, медиана, мода и др.).
- рассеяния, характеризующие изменчивость распределения (дисперсия, стандартное отклонение, размах).

Среднее значение \bar{x} измеряемой величины x указывает центр распределения, около которого группируются результаты отдельных измерений

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

Медиана \tilde{x} - является важной числовой характеристикой распределения, особенно в тех случаях, когда оно асимметрично.

Ассиметричные с одной вершиной распределения характеризуются тем, что большая часть значений расположена с одной стороны от среднего, в то время как меньшая часть значений расположена на большом удалении с другой стороны. Широко известным примером ярко выраженного асимметричного распределения является логарифмически нормальное распределение, когда нормально распреде-

лены не сами значения, а их логарифмы. По этому закону очень часто распределены проницаемости неоднородного пласта. Рассчитанная среднеарифметическая проницаемость оказывается слишком большой, иначе говоря среднее значение лежит слишком далеко вправо. Более достоверную картину даёт медиана, равная значению, которое делит распределение на две равные части, так что каждая содержит 50 % всего распределения.

Для вычисления медианы результаты измерений представляются в виде ранжированного ряда:

$$x^1 \leq x^2 \leq \dots \leq x^{(n)}$$

Если объём выборки n - нечётное число, то медиана находится точно в середине упорядоченной последовательности, т.е.

$$\bar{x} = x^{\left(\frac{n+1}{2}\right)}$$

При чётном n медиана равна среднему арифметическому двух расположенных в середине ряда значений:

$$\bar{x} = \frac{x^{\left(\frac{n}{2}\right)} + x^{\left(\frac{n+1}{2}\right)}}{2}$$

Допустим, что в лаборатории физики пласта по результатам отбора керна для каждой из двух скважин были получены значения проницаемости (мД), представленные в таблице 1.

Таблица 1- Результаты опробования проницаемости пласта

| Скважина | Проницаемость в мД | | | | | | |
|----------|--------------------|-----|-----|-----|-----|-----|------|
| | №1 | 5.2 | 5.9 | 6.0 | 6.5 | 7.7 | 8.2 |
| №2 | 7.5 | 7.6 | 7.8 | 8.1 | 8.5 | 8.6 | 10.3 |

Обратите внимание, данные уже отсортированы.

Тогда для первой скважины

$$\bar{x} = \frac{x^{(3)} + x^{(4)}}{2} = \frac{6.0 + 6.5}{2} = 6.25$$

Для второй $\bar{x} = x^{(4)} = 8.1$

Мода представляет собой наиболее вероятное или часто встречающееся значение в таблице частот.

Размах представляет собой простейшую меру рассеяния. Это разность между максимальным и минимальным значением в выборке: $R = x_{\max} - x_{\min}$

Дисперсию выводят как средний квадрат отклонения отдельных результатов от среднего значения случайной величины

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1} = \frac{1}{n - 1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Среднее квадратичное отклонение, называемое также стандартным, определяют как квадратный корень из дисперсии:

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

Как следует из способа вычисления, эта величина характеризует разброс результатов отдельных измерений вокруг среднего значения, получаемого после обработки всех данных многократного измерения. Конечно, точные значения σ и \bar{x} являются предельными величинами, так как могут быть получены лишь тогда, когда полное количество проведенных измерений достаточно велико, в пределе при $n \rightarrow \infty$.

При конечных n правильнее использовать термин *экспериментальная оценка*, который в равной мере относится и к среднему значению и к дисперсии.

Отметим, что среднее значение случайной величины нельзя расценивать как однозначный результат измерения. Иначе надо было бы полагать, что случайная величина всегда имеет только одно постоянное значение, чего не может быть в действительности из-за ее случайной природы.

Случайные факторы, характеризующие форму распределения случайной величины, не связаны только с возможной неточностью измерительных приборов. Среднее квадратичное отклонение, объективно отражает характер поведения ис-

следуемой случайной величины, поскольку её изменчивость объясняется множеством не поддающихся строгому учёту явлений совершенно разной природы.

1.2.1.3 Моменты случайных величин

Оценка распределения случайных величин может производиться с помощью так называемых моментов- начальных и центральных.

Начальным моментом k -го порядка (ν_k) случайной величины X называется математическое ожидание её *k -ой* степени:

Таблица 2- Начальные моменты случайных величин

| Дискретная случайная величина | Непрерывная случайная величина |
|----------------------------------|--|
| $\nu_k = \sum_{i=1}^n x_i^k p_i$ | $\nu_k = \int_{-\infty}^{+\infty} x^k f(x) dx$ |

Центральным моментом k -го порядка (μ_k) случайной величины X называется математическое ожидание *k -ой* степени отклонения случайной величины X от её математического ожидания:

Таблица 3- Центральные моменты случайных величин

| Дискретная случайная величина | Непрерывная случайная величина |
|---|---|
| $\mu_k = \sum_{i=1}^n [x_i - M(X)]^k p_i$ | $\mu_k = \int_{-\infty}^{+\infty} [x^k - M(X)]^k f(x) dx$ |

Обратите внимание на то, что:

- начальный момент первого порядка представляет собой математическое ожидание случайной величины;
- центральный момент второго порядка- дисперсию случайной величины;
- центральный момент третьего порядка применяется для характеристики скошенности или асимметрии распределения:

$$A_3 = \frac{\mu_3}{\sigma^3};$$

– центральный момент четвёртого порядка применяется для характеристики крутости или эксцесса распределения:

$$E_4 = \frac{\mu_4}{\sigma^4} - 3.$$

1.2.2 Примеры статистических распределений

1.2.2.1 Равномерное распределение

Непрерывная случайная величина X имеет равномерное распределение на отрезке $[a, b]$, если на этом отрезке плотность распределения случайной величины постоянна, а вне его- равна нулю.

$$f(x) = \begin{cases} 0, & x < a, \\ \frac{1}{b-a}, & a < x < b, \\ 0, & x > b \end{cases}$$

где $\frac{1}{b-a} = Const.$

На рисунке 4 показаны графики плотности $f(x)$ и функции $F(x)$ равномерного распределения.

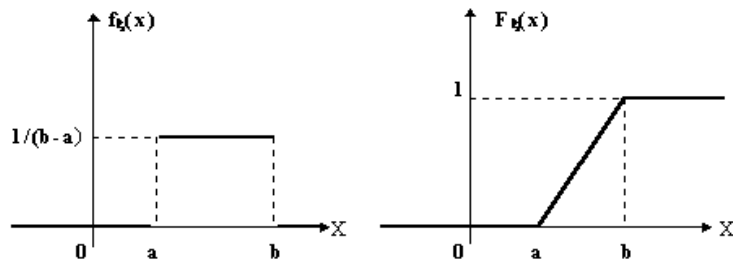


Рисунок 4- Графики плотности вероятностей и функции равномерного распределения.

Непрерывная случайная величина подчинена равномерному закону распределения, если ее возможные значения лежат в пределах некоторого определенного интервала, кроме того, в пределах этого интервала все значения случайной величины одинаково вероятны (обладают одной и той же плотностью вероятности).

С такими случайными величинами часто встречаются в измерительной практике при округлении от счетов измерительных приборов до целых делений шкал. Ошибка при округлении отсчета до ближайшего целого деления является случайной величиной, которая с постоянной плотностью вероятности принимает любое значение между соседними целыми делениями.

1.2.2.2 Нормальное распределение

Исключительно важную роль в теории вероятностей играет нормальное распределение (закон Гаусса). Если, помимо характерных для распределения значений величин \bar{X} и σ , известен функциональный вид распределения случайной величины, то можно получить полную информацию о вероятности реализации случайной величины в любом заданном интервале значений.

Рассмотрим это на примере *нормального*, или Гауссова, распределения, отображающего ситуацию, наиболее часто встречающуюся в природе.

Случайная величина, подчиняющаяся нормальному распределению, представляет собой сумму большого числа независимых случайных величин, каждая из которых играет в образовании всей суммы незначительную роль.

Например, нормально распределёнными являются следующие случайные величины:

- ошибки измерений;
- отклонения при стрельбе;
- рост человека.

Такое широкое распространение нормального закона связано с тем, что он является предельным законом, к которому приближаются многие другие (например, биномиальный). Доказано, что сумма очень большого числа случайных величин, влияние каждой из которых близко к 0, имеет распределение, близкое к нормальному. Этот факт является содержанием предельной теоремы Ляпунова.

Если случайная величина X представляет собой сумму очень большого числа взаимно независимых случайных величин, влияние каждой из которых на всю сумму ничтожно мало, то X имеет распределение, близкое к нормальному

Как следствие, нормальному закону распределения присуща особая роль, объясняемая тем, что при обработке данных измерений в науке и технике обычно предполагают нормальный закон распределения случайных погрешностей измерений.

Нормально распределенная случайная величина имеет следующие свойства:

- она может принимать непрерывный ряд значений от $-\infty$ до $+\infty$;
- центр распределения случайной величины одновременно является центром симметрии, т.е. одинаковые отклонения результатов измерения в меньшую и в большую стороны от центра встречаются одинаково часто;
- малые отклонения встречаются чаще больших, другими словами, реализуются с большей вероятностью.

Соответствующее функциональное выражение для распределения задает формула Гаусса

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

где \bar{X} и σ обозначают среднее значение и стандартное отклонение, соответственно;

e – основание натуральных логарифмов.

На рисунке 5 показано семейство графиков плотности нормального распределения при разных стандартных отклонениях, построенные в системе MathCad 2001.

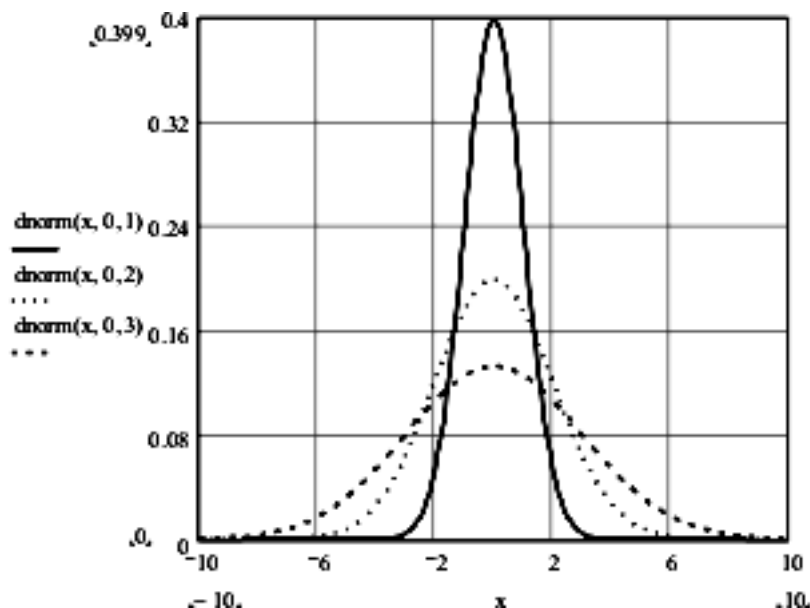


Рисунок 5- Графики нормального распределения для $x=0$; $s = 1$; 2 и 3.

Распределение, задаваемое функцией Гаусса, симметрично относительно максимума, находящегося при $x = \bar{x}$. Значение функции в максимуме

$$\rho_{\max} = \rho(x = \bar{x}) = \frac{1}{\sqrt{2\pi\sigma}}$$

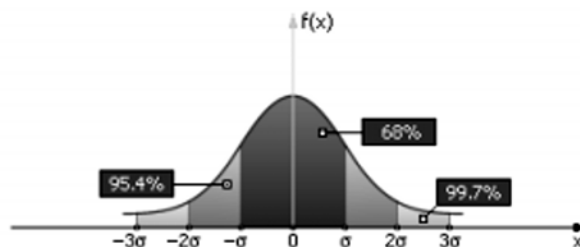
Значение аргумента x , при котором плотность вероятности максимальна называется *модой*.

Правило трёх сигм

Площадь, заключённая под кривой плотности вероятности нормального распределения можно табулировать, по числу значений стандартного отклонения.

Например:

- при $\sigma = 1$ она составляет 68 %;
- при $\sigma = 2$ она составляет 95 %;
- при $\sigma = 3$ она составляет 99.7 %.



Эти числа лежат в основе т.н. правила трёх сигм, которое гласит:

Вероятность того, что отклонение нормально распределённой случайной величины X от его математического ожидания $M(X)$ по абсолютной величине не превысит трёх стандартных отклонений близка к 1.

Правило означает, что случайная величина, распределенная нормально, практически не может отклониться от своего математического ожидания более чем на три средних квадратичных отклонения.

Таким образом, если наблюдаемые значения нормально распределённой случайной величины, выходят за границы интервала трёхкратного стандартного отклонения от среднего их можно отнести к *аномальным*, значит интересным в поисковом отношении. При этом выход за его левую границу будет означать отрицательные аномальные значения, а выход за правую будет фиксировать положительную аномалию.

1.2.2.3 Распределение Пирсона (хи - квадрат)

Распределение случайной величины:

$$X = X_1^2 + X_2^2 + \dots + X_n^2$$

где случайные величины X_1, X_2, \dots, X_n независимы и имеют одно и тоже распределение $N(0,1)$. При этом число слагаемых, т.е. n , называется «числом степеней свободы» распределения хи – квадрат. Напомним, что записью $N(0,1)$ обозначается нормально распределённая случайная величина, характеризующаяся нулевым математическим ожиданием и стандартным отклонением, равным единице. Такое нормальное распределение часто называют *стандартизованным*.

Распределение χ^2 (хи-квадрат) с n степенями свободы — это распределение суммы квадратов n независимых стандартных нормальных случайных величин. При этом, чем больше степеней свободы тем сильнее данное распределение сходится к нормальному закону распределения.

Распределение хи-квадрат используют при оценивании дисперсии (с помощью доверительного интервала), при проверке гипотез согласия, однородности,

независимости, прежде всего для качественных (категоризованных) переменных, принимающих конечное число значений, и во многих других задачах статистического анализа данных.

1.2.2.4 Распределение t Стьюдента

Распределение случайной величины:

$$T = \frac{U\sqrt{n}}{\sqrt{X}},$$

где случайные величины U и X независимы, называется распределением Стьюдента.

При этом U имеет стандартное нормальное распределение $N(0,1)$, а X – распределение хи – квадрат с n степенями свободы. При этом n называется «числом степеней свободы».

Распределение Стьюдента было введено в 1908 г. английским статистиком В. Госсетом, работавшем в то время на пивной фабрике. В те времена руководство этой компании, опасаясь конкуренции, запрещало своим сотрудникам любые публикации по технологии производства, в связи с чем В. Госсет был вынужден пользоваться псевдонимом «Студент».

История Госсета - Стьюдента показывает, что еще сто лет назад менеджерам Великобритании была очевидна большая экономическая эффективность вероятностно-статистических методов.

В настоящее время распределение Стьюдента – одно из наиболее известных распределений среди используемых при анализе реальных данных. Его применяют при оценивании математического ожидания, прогнозного значения и других характеристик с помощью доверительных интервалов, по проверке гипотез о значениях математических ожиданий, коэффициентов регрессионной зависимости, гипотез однородности выборок.

1.2.2.5 Распределение Фишера

Распределение случайной величины:

$$F = \frac{\frac{1}{k_1} X_1}{\frac{1}{k_2} X_2},$$

где случайные величины X_1 и X_2 независимы и имеют распределения хи – квадрат с числом степеней свободы k_1 и k_2 , соответственно, называется распределением Фишера.

Распределение случайной величины F названо в честь великого английского статистика Р.Фишера (1890-1962), активно использовавшего его в своих работах. Распределение Фишера используют при проверке гипотез об адекватности модели в регрессионном анализе, о равенстве дисперсий и в других задачах прикладной статистики.

Вопросы для самопроверки:

1 Укажите, какие величины являются случайными, а какие – нет:

- время отправления поезда с вокзала по расписанию;
- время прибытия поезда на конечную станцию;
- скорость света в вакууме;
- скорость автомобиля;
- число сидений в автобусе;
- число сидящих пассажиров в автобусе;
- номинальное напряжение в сети;
- реальное напряжение в сети;
- концентрация компонентов в буровом растворе;
- атмосферное давление;
- плотность нефти;
- концентрация посторонней примеси в добываемой нефти.

- 2 Что такое гистограмма случайной величины и как ее строят ?
- 3 При каких условиях гистограмма переходит в распределение плотности вероятности ?
- 4 Что характеризует среднее значение и среднее квадратичное отклонение ?
- 5 Почему нормальное распределение встречается чаще других ?
- 6 Что такое правило 3-х сигм и для чего оно применяется?

1.3 Оценка параметров статистического распределения

Задачи оценки неизвестных параметров статистического распределения обычны в геологической практике. Так при проектировании разработки нефтяных месторождений, определение её показателей для пористых коллекторов базируется на величинах мощности коллекторов, их пористости и проницаемости. В целом эффективность эксплуатации скважины определяется параметрами системы: пласт-скважина и точностью этой информации.

Например, пластовые нефти с высоким содержанием асфальтенов отличаются аномальными структурно-механическими свойствами. Это вязкость, динамическое напряжение сдвига и т.п. Они являются причиной многих осложнений, которые происходят при разработке таких залежей. В частности, на участках низких градиентов давлений нефть почти не движется, и большая часть залежи может быть представлена застойными зонами.

Приведённый пример показывает насколько важное значение имеет правильная и своевременная оценка, как свойств самой нефти, так и свойств пласта.

Объективная оценка некоторых показателей месторождений позволяет производить их *типизацию* по данным продукции эксплуатируемых скважин. В качестве одного из критериев при этом используют эксплуатационный газовый фактор. К нефтяным залежам относят флюиды, газовый фактор которых ниже 600 м³/м³, а у газоконденсатных флюидов он выше 900 м³/м³. В промежуточном интервале 600-900 м³/м³ считается, что пластовый флюид в зависимости от темпе-

ратуры пласта может находиться как полностью в жидкой фазе (нефтяная залежь), так и в газовой (газоконденсатная залежь).

Нередко в нефте- и газопромысловой практике требуется оценка вероятности явлений, влияющих на ход эксплуатации месторождений. Например, разработка залежей в слабосцементированных коллекторах (Западная Туркмения), сопровождается частым пробкообразованием и даже техногенными землетрясениями (Газли, 1984). Необходимо по частоте пробкообразования численно оценить эффективность мероприятий, направленных на её снижение. В простейшей постановке эта задача сводится к оцениванию параметра генеральной совокупности на основе выборочных данных.

1.3.1 Генеральная совокупность и выборка

В математической статистике *генеральная совокупность* - понятие абстрактное, представляющее собой множество результатов всех мыслимых наблюдений, которые могли быть получены при данном комплексе условий. Например, пласт-коллектор во всём его объёме, численно охарактеризованный в каждой из бесконечного числа точек будет представлять генеральную совокупность [2].

Понятно, что такими подробными сведениями мы никогда не будем располагать и на практике для изучения генеральной совокупности, т.е. реального природного объекта применяют т.н. выборочный метод.

1.3.1.1 Выборочный метод исследования

Разведка месторождений полезных ископаемых представляет собой с точки зрения математической статистики не что иное, как взятие выборки из месторождения, которое рассматривается как генеральная совокупность.

Главная цель выборочного метода- найти важнейшие характеристики выборки и перенести их на генеральную совокупность с наименьшими потерями.

Другими словами- это умение по малому судить о многом, что является главным в статистическом исследовании вообще.

Заметим, что под выборкой понимается как процесс отбора элементов генеральной совокупности, так и множество отобранных в ходе данного процесса её элементов.

Удачно сделанная выборка считается представительной или репрезентативной, если в ней распределение и среднее значение изучаемого признака настолько близки к распределению и среднему в генеральной совокупности, что их расхождением можно пренебречь.

Выборочная совокупность - это часть объектов генеральной совокупности, от которых исследователь получает необходимые сведения (проводя, например, интервью), а затем экстраполирует (распространяет) полученные результаты на всю генеральную совокупность. Однако для этого необходимо отбирать объекты, входящие в выборку, с соблюдением определённых процедур. Не вдаваясь в детали, отметим, что основными требованиями к выборке можно считать:

- репрезентативность (представительность, способность быть отражением генеральной совокупности);
- случайность формирования (каждый объект генеральной совокупности должен иметь равную вероятность быть отобранным);
- достаточность объема для получения статистически значимых результатов.

Многие неверно считают, что большой объем выборки гарантирует ее репрезентативность. Они задают вопрос: как обследование незначительного числа респондентов может дать точные сведения о большом числе людей, которые не были опрошены? Чем больше людей будет опрошено, тем более точные результаты будут получены, считают они.

Однако *репрезентативность* не зависит от объема выборки. Она достигается только тогда, когда в выборку отобраны объекты из разных групп, при условии, что их доли в генеральной и выборочной совокупности равны. Репрезентативность выборки зависит только от методики отбора единиц из генеральной совокупности в выборочную совокупность и не зависит от объема последней. Ко-

нечно, чем больше объем выборки, тем выше ее точность, однако, неверно распределенная выборка в 5000 единиц намного хуже, чем хорошо распределенная выборка в 500 единиц.

Чем более однородна генеральная совокупность, тем меньший объем выборочной совокупности потребуется для получения точных результатов. Например, чтобы определить вкус каши достаточно съесть одну ложку, а не всю тарелку, конечно, при условии, что каша хорошо перемешана.

Практически выборку считают представительной, если интересующие нас характеристики в ней и генеральной совокупности отличаются друг от друга не более чем на допустимую, заранее заданную величину.

Процесс отбора элементов выборки следует производить наудачу. При этом его техника может быть различной.

Одним из признаков, по которым производится отбор, может служить номер членов совокупности. Такое становится возможным, если на каждый элемент совокупности завести карточки и пронумеровать их нарастающим итогом, невзирая на их содержание. Например, в качестве совокупности может выступать многолетняя база данных с химическими анализами по месторождению. Если каждая запись такой БД пронумерована, то мы с помощью генератора случайных чисел можем выбирать их и составлять соответствующие случайные выборки.

Выборки отличаются друг от друга по:

а) схеме отбора:

- 1) повторная выборка;
- 2) бесповторная выборка;

б) Способу отбора:

- 1) типическая выборка;
- 2) серийная выборка;
- 3) механическая выборка.

Повторная выборка

Выборка этого типа производится по схеме «возвращённого шара», т.е. выбранный однажды член генеральной совокупности может быть повторно выбран. Вероятность его выбора постоянна во время всего выборочного процесса.

Бесповторная выборка

Бесповторная выборка формируется по схеме «невозвращённого шара», т.е. уже выбранный член генеральной совокупности повторно не выбирается. Строго говоря, при бесповторной выборке объём генеральной совокупности всё время сокращается.

Типическая выборка

Типическая выборка создается, когда известно, что генеральная совокупность неоднородна и состоит из нескольких типичных разновидностей. Это могут быть зоны минерализации на месторождении, зоны окисления, выветривания и т.п.

Чтобы учесть эту неоднородность генеральная совокупность разделяется на несколько типических для исследуемого признака групп, после чего из каждой группы делается выборка наудачу.

Разбивку на группы следует производить так, чтобы в выборке были представлены все типы изучаемых объектов.

При производстве типической выборки нужно иметь в виду, что выделяемые типические группы почти всегда будут иметь неодинаковый объём. Поэтому возникает вопрос- как отбирать объекты из частной совокупности? Тут возможны 3 метода отбора:

- пропорционально объёму группы;
- непропорционально объёму группы;
- приблизительно пропорционально степени изменчивости признака в группе, т.е. в зависимости от внутригрупповой дисперсии признака (чем она выше-

тем большим должен быть объём).

Серийная выборка

Серийная выборка желательна там, где добываемое минеральное сырьё смешивается в участковые потоки. Другими словами, при перемешивании материала происходит его случайное перераспределение на механическом уровне, а не на уровне отбора.

Серийная выборка формируется так же как и типическая, т.е. выделяются типические группы, однако способ их опробования иной. Вместо отбора отдельных членов генеральной совокупности в серийной выборке производится отбор целых групп для их сплошного обследования. Зато другие группы остаются совершенно не затронутыми процессом отбора. Пример серийной выборки из практики контрольного опробования.

Для того, чтобы проконтролировать метод взятия проб иногда подвергают вторичному (контрольному) опробованию целый штрек или горизонт или 3-4 участка из общего, довольно значительного числа штреков, горизонтов, участков. Если по всему месторождению взято, например, 15000 проб, то в серийную выборку попадёт 500-1000.

Механическая выборка

Механическая выборка осуществляется также по группам, но последние выделяются по признакам, не имеющим никакой связи с исследуемой совокупностью. Для осуществления такого отбора:

- генеральную совокупность разбивают на группы чисто механически, например, по порядковым номерам, по квадратной сетке или в шахматном порядке;
- из каждой группы берётся также механически каждый первый или каждый второй или n -ый член и включается в выборку.

Механическая выборка иногда используется при сдаче на анализ оптического сырья: например, из каждого 5-го ящика на анализ сдаётся каждый 4-ый кри-

сталл.

Несколько напоминает механическую выборку опробование при разведке месторождения, если она производится скважинами или шурфами расположенными по геометрической сети. Месторождение разбивается на некоторое количество участков или ячеек, одинаковых по размеру, форме и ориентации. В пределах такой ячейки опробование производится случайно, но опробуется лишь один объект из ячейки.

1.3.2 Оценка параметров генеральной совокупности

Предположим, что нам заранее известно с точностью до значения неизвестного параметра распределение генеральной совокупности. Пусть, кроме того, дана выборка из n наблюдений x_1, \dots, x_n .

Требуется определить, используя эти наблюдения число, которое можно было бы принять в качестве неизвестного параметра, или интервал, о котором можно было бы утверждать, что он это значение содержит.

Например, в некоторых ситуациях можно считать, что x_1, \dots, x_n - равномерно распределённые случайные величины на отрезке $[a, b]$, или x_1, \dots, x_n - подчиняются нормальному распределению $N(\mu, \sigma^2)$. Обратите внимание, мы знаем или предполагаем тип распределения, но не знаем его параметров. В приведённых примерах неизвестны величины a и b или μ, σ^2 . По результатам выборочных наблюдений мы хотим оценить значение неизвестного параметра распределения- θ .

1.3.2.1 Статистики

Для ясности дальнейшего изложения материала необходимо ввести понятие «статистика». В данном случае это не будет названием, изучаемой нами математической дисциплины. Назовём статистикой $T(X_1, \dots, X_n)$ любую функцию, зависящую только от наблюдений. Так как наблюдения являются случайными вели-

чинами, то и статистики тоже будут случайными величинами. Рассмотрим распределение некоторых наиболее известных статистик.

Среднее арифметическое

Выборочная функция $\bar{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_n) = \frac{1}{n} \sum_{i=1}^n X_i$,

где n - число наблюдений,

распределена нормально с параметрами $N\left(\mu, \frac{\sigma^2}{n}\right)$, если соответствующая генеральная совокупность распределена нормально с параметрами $N(\mu, \sigma^2)$.

Обычно это допускается априори.

Выборочная дисперсия и стандартное отклонение

Статистика $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ называется дисперсией выборки.

Математическое ожидание (т.е. среднее генеральной совокупности) случайной величины S^2 равно дисперсии генеральной совокупности.

Стандартным отклонением называется квадратный корень из дисперсии, т.е.

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$$

Статистика Хи- квадрат

Выборочная функция $\chi^2 = \frac{(n-1) \times S^2}{\sigma^2}$ имеет непрерывную функцию распре-

деления $f_{\chi^2}(x) = C_m e^{-\frac{x}{2}} \times x^{\frac{m}{2}-1}$ для всех $x > 0$ при $m = n - 1$.

С возрастанием m (число степеней свободы) распределение χ^2 стремится к нормальному распределению, что хорошо заметно по изменению характера кривой плотности распределения. С увеличением числа степеней свободы она становится всё более симметричной (рисунок 6).

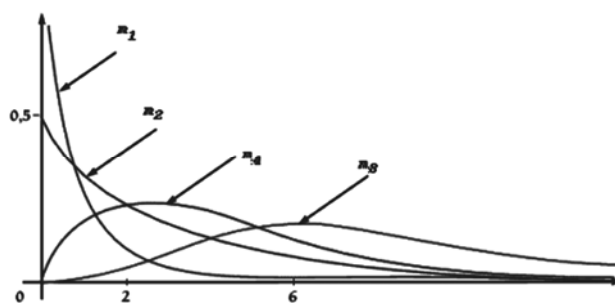


Рисунок 6- График плотности вероятности χ^2 при разных степенях свободы

Статистика Стьюдента (t)

Из выборочных средних (\bar{x}) и дисперсии (S) образуется "статистика Стьюдента":

$$t = \frac{\bar{X} - \mu}{S} \times \sqrt{n}$$

Её распределение (распределение Стьюдента) описывается функцией:

$$f_t(x) = D_m \left(1 + \frac{x^2}{m} \right)^{-\left(\frac{m+2}{2}\right)},$$

где $m=n-1$;

D_m - константа, зависящая только от m .

$f_t(x)$ симметрична относительно нулевой точки $t=0$ и тем более пологая, чем меньше m . При $m \rightarrow \infty$ плотность $f_t(x)$ переходит в плотность вероятности нормированного нормального распределения. Сказанное иллюстрируется рисунком 7, полученным с помощью пакета Statistica 6.

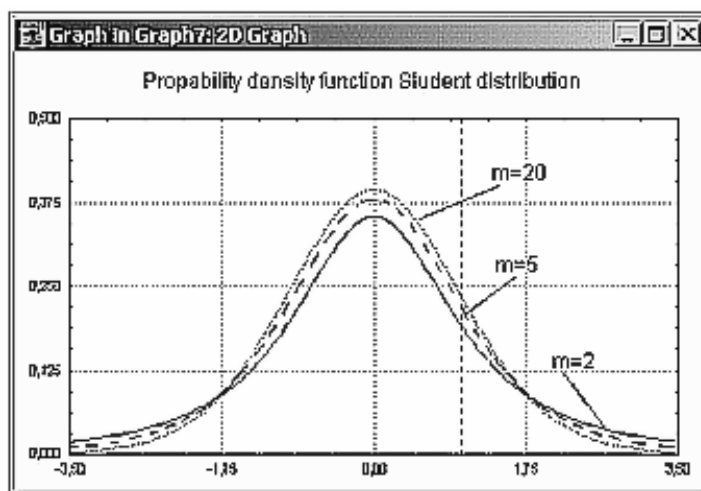


Рисунок 7- Графики $f_t(x)$ при разных степенях свободы.

Статистика Фишера (F)

Предположим, что мы имеем две выборки из двух генеральных совокупностей, которые распределены нормально с параметрами μ_1, σ_1^2 и μ_2, σ_2^2 . Кроме того, выполняется условие, что $\sigma_1^2 = \sigma_2^2$. В этом случае может быть рассчитана статистика Фишера:

$$F = \frac{S_1^2}{S_2^2}$$

Она имеет непрерывную функцию распределения, плотность вероятности которой равна:

$$f_F(x) = E_{m_1, m_2} \times x^{\frac{m_1}{2}-1} \times (m_2 + m_1 \times x)^{-\left(\frac{m_1+m_2}{2}\right)},$$

где $m_1 = n_1 - 1$; $m_2 = n_2 - 1$ и E_{m_1, m_2} - коэффициент, зависящий только от m_1 и m_2 .

С увеличением объёма сравниваемых выборок распределение статистики F стремится к нормальному распределению.

1.3.3 Точечные оценки параметров распределения

Перечисленные свойства статистик, распределение которых стремится к нормальному с увеличением степеней свободы, позволяют использовать их для оценки неизвестного параметра генеральной совокупности θ . Это может быть математическое ожидание, генеральная дисперсия и др. Однако, тот факт, что статистики являются случайными величинами их значения в различных выборках могут сильно отличаться друг от друга. Отсюда ясно, что нельзя найти оценку, которая принимала бы близкие к θ значения во всех случаях. Мы должны выбрать такую процедуру «оценивания», которая давала бы хорошие результаты «в среднем» при многократном её воспроизведении. Не следует отвергать оценку, если она даёт плохие результаты для отдельных выборок, но от неё надо отказаться, если плохие результаты возникают слишком часто.

Основная проблема теории оценок состоит в выявлении свойств, которыми

должны обладать оценки.

Чтобы выбранная статистика или оценка была наилучшей она должна обладать рядом свойств. К их числу относятся: *несмещённость, состоятельность и эффективность.*

Статистическая оценка является несмещённой, если её математическое ожидание равно оцениваемому параметру генеральной совокупности.

Например, для случайной величины, подчиняющейся нормальному закону распределения $N(\mu, \sigma^2)$ выборочное среднее $\bar{X} = \left(\sum_{i=1}^n X_i \right) / n$ даёт несмещённую оценку математического ожидания μ .

В качестве несмещённой оценки дисперсии σ^2 нормально распределённой генеральной совокупности может быть использована выборочная дисперсия, определяемая выражением:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Статистическая оценка является состоятельной, если при $n \rightarrow \infty$ она сходится по вероятности к оцениваемому параметру θ .

Не приводя строгих математических доказательств, отметим, что рассмотренные выше выборочное среднее и выборочная дисперсия являются состоятельными оценками.

Эффективной оценкой распределения называется та, которая обладает минимальной дисперсией.

Поскольку существует множество оценок статистического распределения генеральной совокупности, возникает необходимость выбора из них наиболее эффективных. Например, в этом качестве могут использоваться дисперсии среднего арифметического и медианы:

$$D\bar{X} = \frac{\sigma^2}{n}, \quad D\tilde{X} = \sigma^2 \times \frac{\pi}{2n}$$

Из приведённых выражений ясно, что дисперсия среднего в $\pi/2=1.57$ раза меньше дисперсии медианы, что свидетельствует о её более высокой эффективности.

1.3.4 Интервальные оценки параметров распределения

К сожалению, точечные оценки не дают информации о точности конкретной оценки характеристик генерального распределения. Для устранения этого недостатка применяют доверительную или интервальную оценку, позволяющую по выборочным данным найти интервал, в котором с заданной вероятностью лежит истинное, но неизвестное значение параметра распределения генеральной совокупности.

Ширина интервала и доверительные пределы (граничные точки), определяются:

- степенью достоверности или доверительной вероятностью того, что искомый параметр лежит в этом интервале;
- размером выборки n ;
- выбранной для оценки статистикой.

Обозначим неизвестный параметр генеральной совокупности как θ . Пусть θ^* -его статистическая оценка.

Тогда вероятность того, что интервал $((\theta^* - \delta) < \theta \leq (\theta^* + \delta))$ включает в себе неизвестный параметр генеральной совокупности и есть доверительная вероятность γ .

Доверительная вероятность задаётся априорно. Обратную ей величину: $\alpha = 1 - \gamma$ называют уровнем значимости. Другими словами 95 % доверительной вероятности соответствует уровень значимости (т.н. *p-уровень*). Он вычисляется вычитанием доверительной вероятности из единицы: $\alpha = 1 - 95 = 0.05$.

На практике в качестве доверительной вероятности используют обычно следующие значения: 95, 99, 99.9 %, т.е. $\alpha = 5, 1$ и 0.1 %.

Если установите меньшее значение *p-уровня*, то интервал будет шире, и увеличится "уверенность" в оценке. И напротив, как мы знаем из прогнозов погоды, чем они "неопределеннее" (т.е. шире доверительный интервал), тем скорее они сбудутся.

Доверительным интервалом называется область вокруг выборочной статистики, в которой с заданным уровнем значимости содержится "истинное" значение её генерального аналога.

Следует различать двусторонние и односторонние доверительные интервалы.

Для двусторонних действует условие симметричности доверительных вероятностей и уровней значимости, т.е $P_1=\alpha/2$; $P_2=1-\alpha/2$,

где α - уровень значимости, соответствующий доверительной вероятности γ ;

P_1 - уровень значимости левой границы;

P_2 - уровень значимости правой границы.

Односторонние доверительные интервалы используются при решении вопросов о вероятности того, что выборочная статистика меньше, или напротив, больше её генерального аналога.

В этом случае говорят об одностороннем уровне значимости α или односторонней доверительной вероятности $1-\alpha$.

Заметим, что ширина доверительного интервала зависит от размера выборки и дисперсии наблюдений. Обратите внимание, что сама идея вычисления их границ основывается на предположении, что переменная в совокупности *нормально* распределена. Таким образом, если это предположение не выполнено и размер выборки мал, интервальная оценка будет неверна!

Рассмотрим наиболее часто встречающиеся случаи вычисления доверительных интервалов для среднего значения и дисперсии выборки.

При расчёте доверительного интервала для среднего значения следует различать два варианта:

- стандартное отклонение генеральной совокупности (σ) известно;
- стандартное отклонение генеральной совокупности (σ) неизвестно.

1.3.4.1 Доверительный интервал для среднего при известном σ

Пусть количественный признак X генеральной совокупности распределён нормально, причём нам известно генеральное стандартное отклонение σ , а математическое ожидание генеральной совокупности μ – нет. При этом за стандартное отклонение генеральной совокупности допустимо принять стандартное отклонение признака X , определённого по очень большим выборкам ($n > 1000$).

Требуется найти двусторонний интервал по выборочной средней \bar{x} , который содержит μ при уровне значимости α .

В силу требования симметричности границ двустороннего доверительного интервала последний определяется выражением:

$$\bar{x} - g \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + g \frac{\sigma}{\sqrt{n}}$$

Соответственно односторонний доверительный интервал:

$$\text{верхняя граница } \mu_{\text{в}} = \bar{x} + g_{\text{одн}} \frac{\sigma}{\sqrt{n}}.$$

$$\text{нижняя граница } \mu_{\text{н}} = \bar{x} - g_{\text{одн}} \frac{\sigma}{\sqrt{n}}.$$

Значения g для каждого уровня значимости табулировано и находится по таблице функции Лапласа. Оно также легко определяется встроенной Excel-функцией НОРМСТОБР(ВЕРОЯТНОСТЬ), где «вероятность» означает заданную доверительную вероятность (в долях единицы) построения интервала.

Следует учесть, что данная функция автоматически меняет знак g на минус для значений вероятности меньших 0.5. Это означает, что для вычисления доверительного интервала с использованием этой функции следует пользоваться выражением:

$$\bar{x} + g \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + g \frac{\sigma}{\sqrt{n}}.$$

Техника использования Excel-инструментария для разнообразных задач статистического анализа подробно описана в [3]

ПРИМЕР

В результате 45 замеров температуры раздела фракции бензин-авиакеросин на установке переработки нефти получено среднее значение 140.3° . Генеральное стандартное отклонение примем равным 42.0.

Найдём двусторонний доверительный интервал для математического ожидания при 99 % доверительной вероятности.

$$140.3 + \left(g_{0.01} \times \frac{42}{\sqrt{45}} \right) \leq \mu \leq 140.3 + \left(g_{0.99} \times \frac{42}{\sqrt{45}} \right),$$

$$\text{или } 140.3 - 2.326 \times 6.261 \leq \mu \leq 140.3 + 2.326 \times 6.261 = 125.73 \leq \mu \leq 154.86$$

ПРИМЕР

Необходимо определить тип пластового флюида месторождения Русский хутор [4]. Тип определяется по нижней границе газового фактора. Если он ниже или равен $900 \text{ м}^3/\text{м}^3$ - флюид относится к газоконденсатным, $900-1200 \text{ м}^3/\text{м}^3$ - к газоконденсатным, выше - к газовым.

Исходные данные:

$$n=1, \bar{X} = 1200, \sigma = 50 \text{ м}^3/\text{м}^3$$

Определим нижнюю 95 %-ную доверительную границу для газового фактора, т.е. здесь речь идёт об определении одностороннего доверительного интервала.

$$\bar{x} - g_{одн} \frac{\sigma}{\sqrt{n}} = 1200 - 1.645 \frac{50}{1} = 1117.75$$

Таким образом, это месторождение с надёжностью 95 % должно быть отнесено к газоконденсатным.

1.3.4.2 Доверительный интервал для среднего при неизвестном σ

Пусть количественный признак X генеральной совокупности распределён нормально, но нам неизвестны значения его математического ожидания (μ) и ге-

нерального стандартного отклонения (σ).

Требуется найти интервальную оценку μ .

Эта задача довольно просто решается с помощью вышерассмотренной статистики t :

$$t = \frac{\bar{x} - \mu}{S / \sqrt{n}}$$

Эта случайная величина распределена по закону Стьюдента с $m=n-1$ степенями свободы. Для расчёта доверительного интервала необходимо задание доверительной вероятности равной $1-\alpha$ и значения m :

$$\bar{x} - t_{\alpha, m} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha, m} \frac{S}{\sqrt{n}}$$

Как видим, значение неизвестной нам величины стандартного отклонения генеральной совокупности в данном выражении заменяется выборочным стандартным отклонением.

Значение $t_{\alpha, m}$ наиболее просто вычислить с помощью стандартной Excel-функции СТЬЮДРАСПОБР(ВЕРОЯТНОСТЬ;СТЕПЕНИ_СВОБОДЫ).

ПРИМЕР

При исследовании влияния водорастворимых полимеров на коэффициент конечной нефтеотдачи были проведены эксперименты на пористой среде- кварцевом песке с проницаемостью 4Д.

Необходимо определить 95 % доверительный интервал для коэффициента конечной нефтеотдачи, если результаты замеров дали следующие значения: 0.83; 0.78; 0.72; 0.75; 0.74.

Вычисляем выборочные:

- среднее: $\bar{x} = 0.764$
- стандартное отклонение: $S=0.042778$
- статистику $t_{0.05, 4}=2.776$

При этих значениях доверительный интервал вычисляется как:

$$0.764 - 2.776 \times \frac{0.043}{\sqrt{5}} \leq \mu \leq 0.764 + 2.776 \times \frac{0.043}{\sqrt{5}} = 0.71 \leq \mu \leq 0.817$$

ПРИМЕР

При разработке газовых месторождений, приуроченных к водонапорным системам, происходит интенсивное перемещение контурных и подошвенных вод в газонасыщенную часть пласта [4]. Часто этот процесс сопровождается «защемлением» значительных объёмов газа, которые в свою очередь влияют на фазовую проницаемость воды.

Для оценки этого процесса был исследован 41 образец цементированного песчаника. Эксперимент показал, что в сравнении с абсолютной проницаемостью, при защемлении газа коэффициент водопроницаемости уменьшился от 4.2 до 11.6 раз!

Статистическая обработка материала показала, что:

- среднее уменьшения коэффициента водопроницаемости: $\bar{x} = 7.0$;
- стандартное отклонение $S=6.0$;
- статистика $t_{0.05, 40}=2.021$.

Найдём доверительные 95 %-ные границы для средней степени снижения коэффициента водопроницаемости:

$$7.0 - 2.021 \frac{6}{\sqrt{41}} \leq \mu \leq 7.0 + 2.021 \times \frac{6}{\sqrt{41}} = 5.106 \leq \mu \leq 8.89$$

1.3.5 Доверительный интервал для дисперсии

Пусть дана случайная выборка x_1, x_2, \dots, x_n объёма n из нормально распределённой генеральной совокупности, причём среднее её значение (т.е μ) неизвестно. Задача состоит в том, чтобы в этих непростых условиях построить доверительный интервал, который бы с заданной вероятностью покрывал значение дисперсии генеральной совокупности.

Для этого вспомним статистику хи-квадрат, рассмотренную нами ранее:

$$\chi^2 = \frac{(n-1) \times S^2}{\sigma^2}$$

Она подчиняется распределению χ^2 с $m=n-1$ степенями свободы. Мы знаем, во-первых, что с возрастанием m распределение χ^2 стремится к нормальному, и во-вторых в состав статистики входит легко вычисляемая выборочная дисперсия.

Путём элементарных преобразований переписываем вышеприведённое выражение следующим образом:

$$\sigma^2 = \frac{(n-1) \times S^2}{\chi^2}$$

Общее выражение интервала для генеральной дисперсии при уровне значимости α и числе степеней свободы m записывается так:

$$\frac{(n-1) \times S^2}{\chi_{\frac{\alpha}{2}, m}^2} \leq \sigma^2 \leq \frac{(n-1) \times S^2}{\chi_{\left(1-\frac{\alpha}{2}\right), m}^2}$$

Значения χ^2 легко вычисляются с помощью Excel-функции ХИ2ОБР(ВЕРОЯТНОСТЬ;СТЕПЕНИ_СВОБОДЫ).

ПРИМЕР

Пусть при измерении эталонного образца на приборе были получены следующие величины отклонений от истинного значения: 0.24; 0.03; -0.12; -0.15; -0.31; -0.08; -0.26; 0.19.

Найти 95 %-ные доверительные границы для оценки генеральной дисперсии измерений на данном приборе.

Исходные данные:

- число наблюдений $n=8$;
- число степеней свободы $m=n-1=7$;
- дисперсия выборки $S^2=0.039$;
- уровень значимости $\alpha=1-0.95=0.05$.

Решение:

$$\frac{\alpha}{2} = \frac{0.05}{2} = 0.025;$$

$$1 - \frac{\alpha}{2} = 0.975;$$

$$\chi_{0.025, 7}^2 = 16.01;$$

$$\chi_{0.975,7}^2 = 1.689.$$

Тогда расчёт искомого интервала выполняется следующим образом:

$$\frac{(8-1) \times 0.039}{16.01} \leq \sigma^2 \leq \frac{(8-1) \times 0.039}{1.689} = 0.017 \leq \sigma^2 \leq 0.163$$

Вопросы для самопроверки:

- 1 В чём состоит различие между несмещённостью и состоятельностью точечных оценок статистического распределения?
- 2 Какая точечная оценка более эффективна: медиана или среднее арифметическое?
- 3 Что такое доверительный интервал и в чём его преимущества?
- 4 Чем отличается доверительная вероятность от уровня значимости?
- 5 В чём состоит различие между генеральной и выборочной совокупностью?
- 6 Чем отличается двусторонний доверительный интервал от одностороннего?
- 7 Как связан размах доверительного интервала с доверительной вероятностью?
- 8 Предположим, что генеральная совокупность имеет нормальное распределение. Какова вероятность того, что выборочное среднее будет меньше генерального среднего?
- 9 Какая Excel-функция используется для вычисления t-статистики распределения Стьюдента?
- 10 От чего зависит ширина доверительных интервалов?

1.4 Статистическая проверка гипотез

Информация по статистическим выборкам может быть использована для оценки правомерности некоторых предположений (гипотез) о генеральной совокупности.

Статистической гипотезой называется любое суждение относительно функции распределения наблюдаемых случайных величин или её параметров.

Примерами статистических гипотез могут служить следующие предположения:

- нормальное распределение имеет заданные среднее и дисперсию;
- нормальное распределение имеет заданное среднее (о дисперсии ничего не говорится);
- распределение нормально (с любыми возможными параметрами);
- два неизвестных непрерывных распределения одинаковы.

Статистические тесты и критерии позволяют проверить соответствия выборочных данных выдвинутой гипотезе.

В геологии, особенно нефтегазовой, существует огромное число прикладных задач, формулируемых в терминах проверки статистических гипотез. Перечислим наиболее типичные из них.

Например, проектирование разработки нефтяных месторождений предполагает использование надёжных методик расчёта извлекаемой из пласта жидкости (нефти и воды). В их основе лежит моделирование проницаемости неоднородного пласта, которая в сочетании со свойствами вытесняющих (вода) и вытесняемых (нефть) жидкостей определяет большинство показателей заводнения. При этом сама модель пласта представляет собой набор конгруэнтных цифровых сеток (гридов), однородных по простиранию, но отличных друг от друга по основным физическим константам. Минимальный их набор – мощность, пористость и проницаемость.

Для построения такой модели используют функции распределения проницаемости продуктивного пласта конкретного месторождения. Если в расчётах используется тот или иной вид закона распределения, то необходимо доказать, что статистическое распределение проницаемостей кернового материала, соответствует выбранному для моделирования распределению.

Очень важной проблемой, во многом решаемой в терминах проверки статистических гипотез является управление разработкой газовых залежей. Если объём

ективно существует значимая неоднородность проницаемости коллекторов, это приводит к значительному разбросу продуктивности эксплуатационных скважин. Определение показателей разработки таких объектов сводится к последовательному решению двух задач:

- определению вероятностного закона распределения числа проектных скважин по продуктивности;
- оптимизация показателей разработки с учётом существенной разнодебитности скважин.

Другая задача, связанная с подтверждением выбранного закона статистического распределения состоит в анализе надёжности систем нефтегазодобычи.

В качестве его основы используются накопленные сведения об отказах сходного оборудования в отрасли. Таким образом, сам анализ состоит в сравнении выборочного распределения отказов на конкретном промысле с отраслевой генеральной совокупностью отказов.

Ещё один круг задач приходится решать при оценке эффективности того или иного метода интенсификации добычи нефти и газа. Например, на основании статистического анализа данных необходимо выяснить (для конкретных условий, конечно) зависит ли продуктивность скважины от конструкции забоя, и если да, то какую предпочесть?

Аналогичные выводы требуются для подтверждения пользы от применения поверхностно-активных веществ (ПАВ), кислотных обработок, гидроразрыва пласта и т.п.

В приведённых примерах отчётливо выделяются два вопроса:

- соответствует ли данное эмпирическое (выборочное) распределение тому или иному теоретическому распределению?
- являются ли эмпирические совокупности выборками из одной и той же генеральной совокупности?

Утвердительный ответ на первый вопрос означает, что наблюдаемые различия между выборочным и теоретическим распределением носят случайный характер и могут быть обусловлены обычным расхождением между выборкой и гене-

ральной совокупностью.

Сущность второго вопроса похожа, но более глубока. Здесь оценивается степень расхождения двух эмпирических распределений. Она может быть незначительна и тогда принадлежность их к одной генеральной совокупности не снимается с повестки дня. Иначе, напротив, различия между выборками можно рассматривать как свидетельство их разной природы, т.е. что они отобраны из разных генеральных совокупностей.

Другими словами, в обоих случаях приходится решать, является ли наблюдаемое различие между объектами объективно реальным или же оно есть результат случайности.

Предположим, что наша задача сводится к проверке гипотезы об отсутствии реального различия. Эту гипотезу называют *нулевой* гипотезой и обозначают H_0 .

Так нулевая гипотеза для первого вопроса формулируется следующим образом: различие между эмпирическим и теоретическим распределением является случайным.

H_0 для второго вопроса состоит в утверждении, что обе сравниваемые выборки принадлежат к одной и той же генеральной совокупности.

Статистическая гипотеза противоположная нулевой, называется *альтернативной* и обозначается как H_1 .

Для рассмотренных выше двух случаев альтернативная гипотеза предполагает для первого - закономерное, а не случайное отличие сравниваемых выборочного и теоретического распределений. Соответственно, вопрос о принадлежности двух выборок к одной генеральной совокупности, согласно альтернативной гипотезе решается отрицательно.

Чрезвычайно важно понимать то, что статистические критерии могут устанавливать только *отличие*, но не *одинаковость* совокупностей. В этой связи нуль-гипотезу чаще всего выдвигают для проверки того- нет ли оснований для отказа от неё и принятия альтернативной.

В самом деле, при сравнении выборок мы обычно ничего не знаем о генеральной совокупности, из которой они отобраны. Вариации же в параметрах вы-

борок даже из одной совокупности наблюдаются практически во всех случаях. Чтобы решить, является ли это различие случайным или значимым, мы должны установить границы, где «господство» случайности заканчивается.

Итак, мы выдвигаем нуль-гипотезу и отвергаем её тогда, когда по выборке получаем результат, который при истинности нуль-гипотезы невозможен (маловероятен).

Выражение, которое для каждой выборки определяет, удовлетворяет ли она гипотезе или нет, называется статистическим критерием.

В результате обработки экспериментальных данных любым статистическим критерием можно получить одно из двух по-разному интерпретируемых утверждений:

Таблица 4- Интерпретация выдвинутых утверждений о нулевой гипотезе

| Утверждение | Интерпретация |
|---|-------------------------------|
| Нулевая гипотеза неверна | Верна альтернативная гипотеза |
| Нет достаточных оснований считать нулевую гипотезу неверной | Верна нулевая гипотеза |

Рассмотрим сначала первое утверждение и его интерпретацию. Хотя они и эквивалентны, это не свидетельствует об их абсолютной достоверности. Здесь может быть совершена т.н. *ошибка первого* рода, когда критерий отвергает верную нулевую гипотезу.

Ошибочное отрицание нулевой гипотезы называется ошибкой первого рода.

Любой статистический критерий строится таким образом, чтобы вероятность отвергнуть правильную нулевую гипотезу не превышала наперёд заданного числа α (уровня значимости).

Допустим, что мы имеем дело с тремя уровнями значимости: 0.1 (10 %), 0.05 (5 %) и 0.01 (1 %). Тогда если применение критерия приводит к первому утверждению, вероятность его правильности равна $1-\alpha$. Таким образом, мы считаем

практически достоверной альтернативную гипотезу, если вероятность того, что она правильна, превышает соответственно 0.90, 0.95 или 0.99.

Рассмотрим второе утверждение и его интерпретацию. Совершенно очевидно, что они не эквивалентны, поскольку из утверждения не вытекает его интерпретация как следствие. Из-за этого мы рискуем принять нулевую гипотезу как верную, тогда как в действительности верна альтернативная гипотеза.

Принятие ложной нулевой гипотезы называется ошибкой второго рода.

Обозначим вероятность ошибки второго рода через β и сведём разобранные ситуации в таблицу 5.

Таблица 5- Классификация ошибок оценки статистических гипотез

| Решение | Истинная гипотеза | |
|-------------------|------------------------|------------------------|
| | H_0 | H_1 |
| Принимается H_0 | Удовлетворительно | Ошибка II рода β |
| Отвергается H_0 | Ошибка I рода α | Удовлетворительно |

При заданном α и фиксированном объёме выборки, значение β будет тем больше, чем меньше принятое α .

Допустим, что альтернативная гипотеза состоит в том, что генеральная совокупность подчиняется распределению $N(2;1)$. Тогда величина β при уровнях значимости 0.1 и 0.001 будет равна соответственно 0.236 и 0.862. Это означает, что если нулевая гипотеза на самом деле неверна, то при $\alpha = 0.1$ в 100 случаях её проверки мы, тем не менее, будем примерно 24 раза принимать нулевую гипотезу, а при $\alpha = 0.001$ уже 86 раз!

Исходя из этого и выбираются конкретные значения α при решении практических задач. Сам выбор должен зависеть от того, какие ошибки являются наиболее дорогостоящими. Если это ошибки I рода, то может принимать минимальные значения (0.05, 0.01), если более дорогостоящи ошибки II рода, то $\alpha = 0.1$.

ПРИМЕР.

При цементировании обсадных колонн используется тампонажный раствор с определёнными характеристиками. Это плотность, температура затвердевания, срок схватывания и другие показатели, определяющие качество крепления обсадных колонн. Необоснованное принятие нуль-гипотезы «тампонажный раствор в норме» означает крайне опасную ошибку, связанную с трудноустраняемыми осложнениями при креплении скважин. И напротив, отказ от верной нуль-гипотезы (т.е. перестраховка) приведёт лишь к незначительным затратам на повторные испытания раствора.

В приведённом случае ошибка второго рода явно обойдётся дороже, в связи с чем надо минимизировать вероятность её события, выбрав $\alpha = 0.1$.

1.4.1 Статистические критерии

1.4.1.1 Односторонние и двусторонние критерии

Иногда цель исследования состоит в том, чтобы выявить различие параметров двух генеральных совокупностей и неизвестно, какой из этих параметров будет больше, а какой меньше. Допустим, при сравнении эффективности двух способов укрепления призабойной зоны принимается нуль-гипотеза «эффект от применения этих способов одинаков». В этом случае она состоит в том, что генеральные средние равны между собой ($H_0: \mu_1 = \mu_2$) а цель исследования — доказать обратное ($H_0: \mu_1 \neq \mu_2$), т.е. наличие различия между генеральными средними. Такие гипотезы называются *двусторонними*, поскольку никаких предположений о знаке ожидаемого отличия не делается.

Вместе с тем часто проведённые исследования дают основания для более конкретных формулировок гипотез. Например, гипотеза альтернативная приведённой выше может иметь вид: «эффект от применения нового способа ниже, чем

от старого», т.е. $H_0: \mu_2 < \mu_1$. Такие гипотезы называются *односторонними*. Соответственно этому критерию значимости, служащие для проверки двусторонних гипотез, называются двусторонними, а для односторонних — односторонними.

Возникает вопрос о том, какой из критериев следует выбирать в каждом конкретном случае. Ответ на него находится за пределами формальных статистических методов и полностью зависит от целей исследования. Если до проведения эксперимента допускается, что различие сравниваемых параметров может быть как положительным, так и отрицательным, то следует использовать двусторонний критерий. Если же есть дополнительная информация, например, из предшествующих экспериментов, на основании которой можно сделать предположение, что один из параметров больше или меньше другого, то используется односторонний критерий. Когда имеются основания для применения одностороннего критерия, его следует предпочесть двустороннему, потому что односторонний критерий полнее использует информацию об изучаемом явлении и поэтому чаще дает правильные результаты.

Другими словами мощность односторонних критериев выше двусторонних.

Мощностью статистического критерия называется вероятность отвергнуть нулевую гипотезу, когда верна альтернативная.

Мощность критерия зависит от конкретных законов распределения исследуемых случайных величин, в связи с чем, в разных обстоятельствах необходимо выбирать разные критерии. В общем случае, мощность критериев снижается с уменьшением уровня значимости α .

Критерий называется более мощным, когда по сравнению с другими возможными критериями при заданном α он показывает более высокую дискриминирующую способность (способность к разделению гипотез).

Как правило, мощность критерия тем выше, чем больше исходных предположений сделано о статистическом распределении исследуемых генеральных совокупностей.

Несмотря на большое число различных статистических критериев, обработ-

ку экспериментальных данных с помощью любого из них проводят по одной общей для всех случаев схеме.

Сначала выбирается ряд, либо два ряда независимых наблюдений и над их членами проделывают определённые логические или математические операции в соответствии с чётко сформулированными правилами. Само собой разумеется, что для различных критериев эти правила различаются. В результате проведённых выкладок получается некоторое число.

Для каждого критерия имеются таблицы, в которых в зависимости от заданного уровня значимости α и от объёма сравниваемых рядов приводятся граничные значения для получаемых в предыдущем пункте чисел. Если наше число не выходит за пределы табулированных значений, то принимается утверждение: «нет достаточных оснований считать нулевую гипотезу неверной». При развитии этого утверждения и переходе к утверждению «верна нулевая гипотеза», необходимо помнить о возможной ошибке II рода.

1.4.1.2 Параметрические и непараметрические критерии

Статистические критерии делятся на параметрические и непараметрические. Если вид функции распределения $F(x)$ задан отдельными параметрами и сама гипотеза строится именно по ним, то говорят о параметрических критериях. Например, высказывание о значении неизвестного параметра μ (математического ожидания) нормального распределения является такой параметрической гипотезой. В принятых обозначениях её можно записать так: $H_0: \mu = \mu_0$, где μ_0 – принятое значение.

Другим примером является гипотеза о равенстве двух средних выборок, сформулированная следующим образом: $H_0: \mu_1 - \mu_2 = 0$, где μ_1 – среднее для первой выборки, а μ_2 – среднее для второй выборки в нормально распределённых генеральных совокупностях.

Критерии, которые служат для проверки гипотез о параметре, называются параметрическими критериями.

В противоположность этому статистические гипотезы общего характера, например $H_0: F(x) = F_0(x)$, называются непараметрическими, а критерии их проверки - непараметрическими тестами. Непараметрические гипотезы не требуют дополнительных предположений о виде функции $F(x)$. Примером такой непараметрической гипотезы может служить предположение H_0 : принадлежат ли две сравниваемые выборки к одной и той же генеральной совокупности.

Особую роль в непараметрических критериях играют критерии согласия, которые проверяют, согласуется или нет наблюдаемое эмпирическое распределение с гипотетическим.

Непараметрические критерии по сравнению с параметрическими имеют меньшую мощность. Однако, при малых объёмах выборок непараметрические критерии часто эффективнее некоторых оптимальных параметрических критериев. В любом случае, если для анализа доступны несколько критериев, то обычно выбирают те, которые наиболее полно используют информацию, содержащуюся в статистических данных.

Вопросы для самопроверки:

- 1 Чем необходимо руководствоваться при выборе нулевой гипотезы?
- 2 Чем отличается нулевая гипотеза от альтернативной? Какая точечная оценка более эффективна: медиана или среднее арифметическое?
- 3 Что можно установить с помощью статистических критериев - отличие или одинаковость сравниваемых совокупностей?
- 4 В чём отличие ошибок первого и второго рода?
- 5 На чём основан выбор значения критерия значимости при оценке статистических гипотез?
- 6 Влияет ли формулировка статистических гипотез на выбор типа статистического критерия (одностороннего или двустороннего)?
- 7 Какие факторы определяют мощность статистического критерия?
- 8 Что такое дискриминирующая способность статистического критерия?

9 Опишите процедуру обработки экспериментальных данных с помощью статистических критериев.

10 Чем отличаются параметрические критерии от непараметрических?

1.5 Гипотезы о параметрах распределения

1.5.1 Сравнение выборочного среднего с гипотетическим

Многие практические задачи могут быть сформулированы в терминах проверки гипотезы о среднем значении. Так, при открытии нового месторождения многие параметры его эксплуатации ещё неизвестны, но декларируются. Например, коэффициент извлечения нефти (КИН) для таких случаев может быть установлен равным 0.5, но это утверждение нуждается в строгой доказательной базе. Она строится на сравнении этого гипотетического числа с эмпирическими значениями КИН продолжительно разрабатываемых месторождений, обладающими идентичными характеристиками геологического строения и свойствами нефти.

Пусть мы имеем выборку x_1, x_2, \dots, x_n объёма n из нормально распределённой совокупности, причём среднее значение μ неизвестно. Мы хотим проверить параметрическую гипотезу о том, что среднее значение генеральной совокупности μ равно фиксированному значению μ_0 . Таким образом, нулевая гипотеза может быть оформлена в виде $H_0: \mu = \mu_0$. Её проверка имеет две модификации в зависимости от того известна или нет генеральная дисперсия σ^2 .

1.5.1.1 Дисперсия генеральной совокупности известна

Правильность гипотезы проверяется сравнением выборочного среднего с предполагаемым средним μ_0 . В качестве критерия значимости этого отличия рассматривают нормированную случайную величину:

$Z = \frac{\bar{X} - \mu_0}{\sigma} \sqrt{n}$, которая имеет стандартное $N(0,1)$ распределение, чьи критические значения g_α табулированы для разных уровней значимости α .

Формулировка статистического критерия для проверки гипотезы о среднем значении следующая:

- по результатам наблюдений вычисляют величину \widehat{Z} . Здесь крышечка сверху означает, что это значение получено эмпирически;
- при $|\widehat{Z}| \geq g_\alpha$ гипотеза H_0 отвергается, т.е. расхождение между выборочным средним \bar{x} и гипотетическим средним генеральной совокупности μ_0 признаётся значимым;
- при $|\widehat{Z}| < g_\alpha$ гипотеза H_0 принимается, т.е. расхождение между выборочным средним \bar{x} и гипотетическим средним генеральной совокупности μ_0 признаётся незначимым.

Рассмотренный вариант относится к т.н. двустороннему ограничению. В этом случае отклонение выборочного среднего \bar{x} от гипотетического среднего генеральной совокупности μ_0 оценивается по абсолютной величине, без учёта знака.

Односторонний тест соответствует случаю, когда представляет интерес отклонение только в одну сторону. При этом гипотеза H_0 отклоняется, когда выборочное среднее попадает в область, определяемую для слишком больших выборочных средних неравенством:

$$\widehat{Z} \geq g_\alpha \text{ или } \bar{x} \geq \mu_0 + \frac{\sigma}{\sqrt{n}} g_\alpha,$$

а для слишком малых выборочных средних неравенством:

$$\widehat{Z} \leq -g_\alpha \text{ или } \bar{x} \leq \mu_0 - \frac{\sigma}{\sqrt{n}} g_\alpha.$$

Таким образом, алгоритм проверки гипотезы о средних при известной дисперсии генеральной совокупности состоит из следующих шагов:

- выдвигается статистическая гипотеза $H_0: \mu = \mu_0$;
- по имеющимся выборочным данным x_1, x_2, \dots, x_n вычисляется выборочное среднее:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \text{ и величина критерия } \hat{Z} = \frac{\bar{X} - \mu_0}{\sigma} \sqrt{n};$$

- выбирается один из принятых в отрасли уровней значимости α (0.05, 0.01, 0.1) и табличным поиском или с помощью программного инструментария определяется соответствующее ему значение g_α ;
- производится формулировка критерия для двустороннего или одностороннего случая. Например, гипотеза отвергается при $|\hat{Z}| \geq g_\alpha$ (двустороннее ограничение).

ПРИМЕР

Коэффициенты газоотдачи Челбасского, Каневского и Ленинградского газоконденсатных месторождений Краснодарского края равны соответственно: 0.78, 0.73, 0.69. Считая, что $\sigma = 0.1$ проверить гипотезу о том, что среднее значение нефтеотдачи может быть выбрано равным 1.

Итак, проверяется нулевая гипотеза $H_0: \mu=1$

Для приведённой выборки:

$$\bar{x} = \frac{0.78 + 0.73 + 0.69}{3} = 0.733$$

$$\hat{Z} = \frac{0.733 - 1}{0.1} \sqrt{3} = -4.619$$

Для уровня значимости $\alpha=0.01$ с помощью MS Excel-функции НОРМСТОБР(0.01) определяем критическое значение g_α .

$$g_{0.01} = -2.32635$$

Поскольку $|-4.619| > -2.326$, то гипотеза H_0 отвергается.

Это означает, что для указанных месторождений коэффициент газоотдачи не может быть принят равным единице.

1.5.1.2 Дисперсия генеральной совокупности неизвестна

На практике применение критерия Z ограничено, поскольку обычной является ситуация, когда дисперсия генеральной совокупности неизвестна. В тех случаях, когда объём n выборки невелик, выборочной функцией Z модифицируется путём замены генерального среднеквадратичного отклонения на стандартное отклонение выборки.

$$t = \frac{\bar{X} - \mu_0}{S} \sqrt{n}$$

Если гипотеза H_0 верна, то статистика t имеет распределение Стьюдента с $m=n-1$ степенями свободы. Соответственно этому критерий, построенный на статистике t , называется критерием Стьюдента или t -критерием.

Согласно этому критерию, гипотеза H_0 отвергается и различие между выборочным средним \bar{x} и генеральным средним μ_0 считается значимым, если вычисленное по выборке значение \hat{t} окажется по модулю больше значения $t_{\alpha, m}$, соответствующего двусторонним границам. Здесь, как и в предыдущем случае α означает уровень значимости. При $|\hat{t}| < t_{\alpha, m}$ гипотеза H_0 принимается.

Для одностороннего критерия область непринятия гипотезы определяется неравенствами:

$$\hat{t} \geq t_{\alpha, m} \text{ либо } \hat{t} \leq -t_{\alpha, m}$$

Выбор того или иного неравенства определяется тем, проверяем ли мы гипотезу о больших положительных или отрицательных расхождениях между \bar{x} и μ_0 .

Односторонние *тодн* и двусторонние *тдвуст* границы связаны выражением:

$$t_{\alpha, m} \text{ тодн} = t_{2\alpha, m} \text{ тдвуст}$$

В целом, последовательность вычислений для проверки гипотезы о равенстве средних по критерию Стьюдента состоит из следующих шагов:

- выдвигается статистическая гипотеза $H_0: \mu = \mu_0$;
- по имеющимся выборочным данным x_1, x_2, \dots, x_n вычисляются выборочные статистики числовых характеристик:

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

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (\bar{x} - x_i)^2} ;$$

- вычисляется значение t - критерия и и величина критерия $\hat{t} = \frac{\bar{X} - \mu_0}{S} \sqrt{n}$;
- для выбранного уровня значимости α (0.05, 0.01, 0.1) и $m=n-1$ степеней свободы находятся граничные значения $t_{\alpha, m}$;
- производится сравнение вычисленного \hat{t} и табулированного $t_{\alpha, m}$. Если $|\hat{t}| < t_{\alpha, m}$, то гипотеза H_0 принимается.

ПРИМЕР

При настройке расходомера по эталонной ёмкости получены следующие ошибки измерения:

0.24, 0.03, -0.12, 0.15, -0.31, -0.08, -0.26, 0.19

Предполагая, что ошибки измерений распределены по нормальному закону, следует проверить гипотезу о том, что настройка произведена качественно, т.е. математическое ожидание ошибки уклонения показаний расходомера равно 0 для уровня значимости 5 %.

Решение:

$$\bar{x} = \frac{0.24 + 0.03 + (-0.12) + 0.15 + (-0.31) + (-0.08) + (-0.26) + 0.19}{8} = -0.02$$

$$S = \sqrt{\frac{\sum_{i=1}^8 (x_i - (-0.02))^2}{8-1}} = 0.206$$

$$\hat{t} = \frac{\bar{X} - \mu_0}{S} \sqrt{n} = \frac{-0.02 - 0}{0.206} * \sqrt{8} = -0.274$$

Для $\alpha=0.05$ и числе степеней свободы $m=8-1=7$ находим $t_{0.05, 7}=2.365$.

С этой целью удобно воспользоваться встроенной MS Excel-функцией СТЬЮД-РАСПОБР(0.05, 7).

Поскольку $|\hat{t}| = |-0.274| < 2.365$ гипотеза H_0 принимается.

1.5.2 Сравнение двух выборочных дисперсий. Критерий Фишера

Представим себе ситуацию, когда надо сравнить два разных способа измерений. С этой целью одни и те же изделия (эталонные) измеряются двумя методами или приборами, и по результатам этих опытов делается вывод о том, имеют ли эти методы одинаковую или различную точность. Другими словами проверяется предположение, что $\sigma_1^2 = \sigma_2^2$, где σ_1^2 - дисперсия наблюдений первым методом, а σ_2^2 - вторым.

Пусть в распоряжении исследователя имеются две независимые выборки данных: x_1, x_2, \dots, x_{n_1} и y_1, y_2, \dots, y_{n_2} объемом n_1 и n_2 соответственно. При этом есть основания предполагать, что обе они взяты из нормально распределенных генеральных совокупностей. Для этих выборок рассчитаны статистические характеристики \bar{x} и S_1^2 , а также \bar{y} и S_2^2 .

Мы уже знаем, что в нормально распределённых генеральных совокупностях выборочные дисперсии S_1^2 и S_2^2 являются несмещёнными и эффективными оценками соответствующих генеральных дисперсий σ_1^2 и σ_2^2 . Поэтому нулевая гипотеза, подлежащая проверке, будет выглядеть как $H_0: \sigma_1^2 = \sigma_2^2$.

Если S_1^2 и S_2^2 сильно различаются, то от нашей нулевой гипотезы придётся отказаться.

Сформулированная выше задача о равенстве генеральных дисперсий σ_1^2 и σ_2^2 была решена Р.А. Фишером, который вывёл статистику $F = S_1^2/S_2^2$ и нашёл её распределение при условии, что $\sigma_1^2 = \sigma_2^2$. Распределение этой статистики называют распределением Фишера с $m_1 = n_1 - 1$ и $m_2 = n_2 - 1$ степенями свободы.

При вычислении F обычно в числитель дроби подставляется большее значение дисперсии. Тогда, при использовании одностороннего критерия, если значение $F \gg 1$, сразу будет ясно, что гипотезу $H_0: \sigma_1^2 = \sigma_2^2$ надо отвергать.

Критическое значение статистики F , при котором нулевая гипотеза должна быть отвергнута, определяется из условия:

$$P\{F \geq F_{\alpha, m_1, m_2}\} = \alpha,$$

где α - уровень значимости;

m_1 - число степеней свободы числителя;

m_2 - число степеней свободы знаменателя выборочной статистики

$$\hat{F} = S_1^2 / S_2^2.$$

Если эта статистика больше найденного критического F_{α, m_1, m_2} , то гипотеза H_0 отвергается, т.к. расхождение между выборочными S_1^2 и S_2^2 является значимым.

В целом проверка гипотезы о равенстве дисперсий двух нормально распределённых совокупностей сводится к следующим шагам:

- выдвигается статистическая гипотеза $H_0: \sigma_1^2 = \sigma_2^2$;
- для двух выборок объёма n_1 и n_2 рассчитываются выборочные дисперсии:

$$S_1^2 = \sqrt{\frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (\bar{x} - x_i)^2}$$
 и $S_2^2 = \sqrt{\frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (\bar{y} - y_i)^2}$;
- вычисляется выборочное значение F -статистики: $\hat{F} = S_1^2 / S_2^2$;
- находится критическое значение F -статистики: F_{α, m_1, m_2} ;
- если выборочное \hat{F} удовлетворяет неравенству: $\hat{F} < F_{\alpha, m_1, m_2}$, то гипотезу H_0 принимают, т.е. считают, что различие между выборочными дисперсиями незначимо, а следовательно генеральные дисперсии равны между собой, т.е. $\sigma_1^2 = \sigma_2^2$.

ПРИМЕР

При анализе нефтей 289 длительно разрабатываемых месторождений Азейбарджана выяснено, что 160 залежей относятся к ньютоновской группе нефтей, а 129 залежей к неньютоновской. Для этих двух групп были рассчитаны дисперсии коэффициента извлечения нефти(КИН). Для ньютоновской группы $S_1^2 = 0.0049$, а для альтернативной $S_2^2 = 0.01$.

Необходимо проверить гипотезу о равенстве двух генеральных дисперсий КИН по выборочным данным для 5 %-го уровня значимости.

Решение:

– Находим выборочную статистику \hat{F} :

$$\hat{F} = S_1^2 / S_2^2 = \frac{0.01}{0.0049} = 2.04 ;$$

- с помощью встроенной Excel-функции: $F_{\text{РАСПОБР}}(0.05, 129, 160)$ определяем критическое значение статистики $F_{0.05, 129, 160} = 1.31$;
- поскольку $\hat{F} > F_{0.05, 129, 160}$, нуль-гипотеза о равенстве дисперсий КИН в нефтях двух типов вязкости отвергается.

1.5.3 Сравнение двух выборочных средних. Критерий Стьюдента

Одним из типичных вопросов, с которым всегда сталкивается исследователь- это вопрос о причинах различия двух рядов независимых случайных величин. Обусловлена ли эта разница влиянием каких-либо неслучайных, искусственных факторов (способом эксплуатации скважин, например) или она лежит в пределах обычного статистического разброса?

Так для выбора оптимального в конкретных геологических условиях способа бурения часто необходимо сравнить турбинный и роторно-турбинный способы. Критериями оценки могут служить: объём выбуренной породы одним долотом, объёмная скорость бурения и т.п. При этом обычно имеется два ряда наблюдений для каждого из сравниваемых способов.

Исследование эффективности работы долот часто основано на информации о промысловой их отработке и сводится к сравнительному анализу проходки долота в породах различных категорий твёрдости.

Математическая постановка задачи о сравнении средних состоит, как и в предыдущем случае в исследовании пары выборок: x_1, x_2, \dots, x_{n1} и y_1, y_2, \dots, y_{n2} объемом n_1 и n_2 соответственно.

Обозначим через \bar{x} и \bar{y} выборочные средние значения сравниваемых ря-

дов. Наверняка, между ними будет некоторая разница величиной $d = \bar{x} - \bar{y}$. Нам нужно выяснить, насколько эта разность должна быть велика, чтобы мы могли утверждать, что генеральные средние неравны между собой, т.е. $\mu_1 \neq \mu_2$.

Сделаем важное предположение о том, что оба ряда измерений - это выборки из нормально распределённых генеральных совокупностей со средними значениями μ_1 и μ_2 и дисперсиями σ_1^2 и σ_2^2 . Для решения задачи сформулируем нуль-гипотезу, что различие между \bar{x} и \bar{y} случайное, а значит и $\mu_1 = \mu_2$.

Проверка $H_0: \mu_1 = \mu_2$ сводится к определению значимости расхождения средних арифметических \bar{x} и \bar{y} обеих выборок и если это не так, то гипотеза принимается. Техника проверки связана с вычислением критерия Стьюдента или t -критерия с модификациями, зависящими от соглашений о дисперсиях сравниваемых независимых выборок. При этом обычно рассматриваются два случая:

- когда у нас есть основания предполагать равенство генеральных дисперсий, значения которых, увы, неизвестны;
- когда мы вынуждены признать, что у нас нет оснований судить о равенстве и значениях генеральных дисперсий.

1.5.3.1 Неизвестные, но равные генеральные дисперсии

Иногда, опытные данные и статистические исследования свидетельствуют о равной генеральной дисперсии сравниваемых выборок.

Поскольку $\sigma_1^2 = \sigma_2^2$ для оценки единой дисперсии измерений σ^2 объединяют обе выборки и в качестве оценки берут величину [5]

$$S^2 = \frac{\sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{i=1}^{n_2} (y_i - \bar{y})^2}{n_1 + n_2 - 2}$$

В свою очередь дисперсия величины $d = \bar{x} - \bar{y}$ складывается из дисперсий значений \bar{x} и \bar{y} , равных соответственно $\sigma_{\bar{x}}^2 = \sigma^2 / n_1$ и $\sigma_{\bar{y}}^2 = \sigma^2 / n_2$. Тогда

$$S_d^2 = \frac{S^2}{n_1} + \frac{S^2}{n_2} = \frac{n_1 + n_2}{n_1 n_2} \times \frac{\sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{i=1}^{n_2} (y_i - \bar{y})^2}{n_1 + n_2 - 2}$$

Данное выражение является основой для расчёта выборочной статистики Стьюдента, имеющей вид

$$\hat{t} = \frac{d}{S_d} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{i=1}^{n_2} (y_i - \bar{y})^2}{n_1 + n_2 - 2}}} \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

Если объём сравниваемых выборок одинаков, т.е. $n_1 = n_2 = n$ то данное выражение упрощается:

$$\hat{t} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{i=1}^{n_2} (y_i - \bar{y})^2}{n - 1}}} \times \sqrt{n}$$

Величина t имеет распределение Стьюдента с $m_1 = n_1 + n_2 - 2$ степенями свободы.

Если выборочный $|\hat{t}| \geq t_{\alpha, m}$ гипотезу $H_0: \mu_1 = \mu_2$ отвергают.

Если выборочный $|\hat{t}| < t_{\alpha, m}$ гипотезу $H_0: \mu_1 = \mu_2$ принимают.

Описанный t – критерий позволяет построить доверительный интервал для разности двух средних значений для уровня значимости α . Его можно применить для проверки нуль-гипотезы: $\mu_1 = \mu_2$. Разница между μ_1 и μ_2 считается значимой, если интервал

$\left[(\bar{x} - \bar{y}) - t_{\alpha, m} S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}; (\bar{x} - \bar{y}) + t_{\alpha, m} S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right]$ не содержит значение $\mu_1 = \mu_2 = 0$.

ПРИМЕР

В результате исследований скважин было выявлено два ряда глубин, связанных с различной степенью парафинизации нефти. Первый ряд связан с неньютоновскими нефтями, а второй с ньютоновскими:

Таблица 6- Распределение глубин по степени парафинизации нефтей

| Ряды глубин | Глубины (в метрах) | | | | | | | | | |
|-------------|--------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| | 1-ряд | 200 | 400 | 350 | 250 | 200 | 350 | 300 | 400 | 300 |
| 2- ряд | 300 | 500 | 550 | 350 | 400 | 450 | 500 | 500 | 350 | 600 |

Следует ли считать данное различие глубин значимым при уровне значимости 5 % ?

Для проверки воспользуемся критерием Стьюдента. Обозначим значения первого ряда через x , а второй- через y . Тогда получим:

$$\bar{x} = 300; \quad \bar{y} = 450.$$

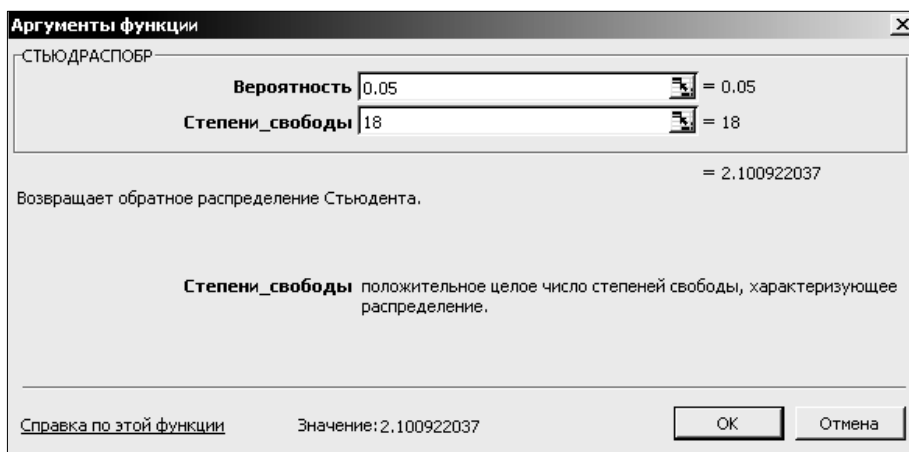
Число степеней свободы $m = 10 + 10 - 2 = 18$.

$$\sum_{i=1}^{10} (x_i - 300)^2 = 50000; \quad \sum_{i=1}^{10} (y_i - 450)^2 = 85000.$$

$$S_d = \sqrt{\frac{10 + 10}{10 * 10} * \frac{50000 + 85000}{10 + 10 + 2}} = 38.73.$$

$$\hat{t} = \frac{85000 - 50000}{38.73} = 903.7.$$

Через Excel-функцию СТЬЮРАСПОБР (0.05, 18) находим максимально допустимое значение для $t_{0.05,18}$, равное 2.101.



Поскольку $903.7 > 2.101$ то следует принять гипотезу о явном различии глубин парафинизации.

1.5.3.2 Неизвестные и неравные дисперсии

Часто у нас нет никаких оснований говорить о равенстве дисперсий для проверки гипотезы о равенстве генеральных средних применяется статистика:

$$\hat{t} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Тогда, для использования $t_{\alpha, m}$ -критерия следует пользоваться изменённым числом степеней свободы, вычисляемом согласно простому выражению:

Других методических отличий от вышеописанного случая нет.

$$m = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 + 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 + 1}} - 2$$

Вопросы для самопроверки:

- 1 Чем отличается постановка задачи о сравнении выборочного среднего с гипотетическим, от сравнения генеральных средних?
- 2 Какие варианты проверки равенства генеральных средних вам известны?
- 3 Какая Excel-функция используется для нахождения критического значения t - критерия?
- 4 В каких обстоятельствах применяется критерий Фишера, и в каких- критерий Стьюдента?

1.6 Непараметрические методы проверки гипотез

Рассмотренные выше классические методы проверки гипотез предполагали нормальность распределения сравниваемых выборок. Однако, часто это предположение не оправдывается, да и, вообще, доказать его не просто. По этой причине в последнее время в статистике бурно развиваются т.н. *непараметрические* методы проверки гипотез. Они не требуют никакого предположения о типе распределения, и единственное условие их применимости формулируется как требование случайного и независимого отбора наблюдений для статистического анализа [6].

Непараметрические критерии принято разделять на две группы по типу решаемых задач- критерии согласия и критерии различия. Из критериев согласия наиболее известным является критерий χ^2 , который чаще всего используется для проверки соответствия выборочного распределения теоретическому.

1.6.1 Проверка распределения по χ^2 -критерию Пирсона

Разработанный Карлом Пирсоном критерий χ^2 служит для проверки принадлежности данной выборки к генеральной совокупности с функцией распределения $F(x)$.

Этот метод состоит в сравнении эмпирического распределения выборки, выраженного через частоты сгруппированного ряда измерений с теоретическим распределением соответствующей генеральной совокупности. При этом проверяются следующие гипотезы:

- H_0 : в основе выборки лежит теоретическое распределение $F(x)$.
- H_1 : выборка принадлежит неизвестному распределению $F(x)$.

Для проверки нуль- гипотезы используется таблица интервальных частот, такая же как и при построении гистограммы.

Если наша выборка объёма n разделена на k классов, то мерой расхождения между эмпирическим и принятым для нулевой гипотезы распределением служит разность наблюдаемых абсолютных частот h_i ($i=1, 2, \dots, k$) и теоретических частот np_i для одного и того же i -го класса или интервала.

Если нулевая гипотеза верна, и мы имеем дело с выборкой из генеральной совокупности, то эмпирические частоты h_i варьируют около теоретических np_i лишь случайно.

Различие между эмпирическим и предполагаемым теоретическим распределением можно охарактеризовать нормированной суммой квадратов отклонений между частотами h_i и np_i так называемой величины «хи-квадрат».

$$\hat{\chi}^2 = \sum_{i=1}^k \frac{(h_i - np_i)^2}{np_i}, \text{ где } k\text{- число интервалов группировки выборки.}$$

Если нулевая гипотеза истина, то случайная величина $\hat{\chi}^2$ приближённо удовлетворяет распределению хи-квадрат с $m=k-1$ степенями свободы. Метод проверки гипотезы H_0 состоит из следующих шагов:

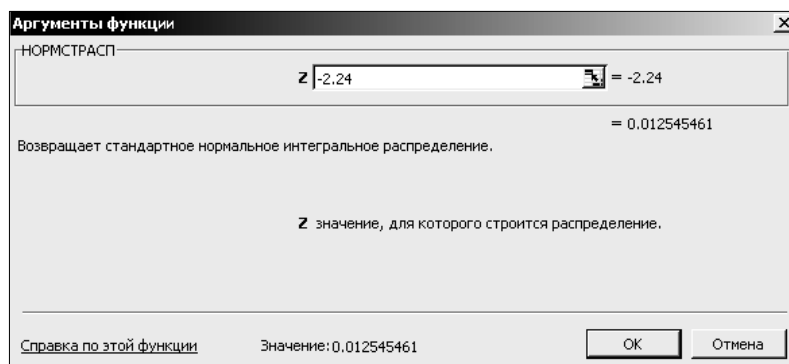
- вычисляем выборочную $\hat{\chi}^2$;
- для каждой нижней и верхней границы интервала t_j и t_{j+1} вычисляют нормированные значения этих границ по формулам:

$$a_i = \frac{t_i - \mu}{\sigma} \quad \text{и} \quad a_{i+1} = \frac{t_{i+1} - \mu}{\sigma};$$

- если известны только оценки \bar{x} и S , то вычисляют:

$$a_i = \frac{t_i - \bar{x}}{S} \quad \text{и} \quad a_{i+1} = \frac{t_{i+1} - \bar{x}}{S};$$

- находим значение $F(a_i)$ и $F(a_{i+1})$;



– вычисляем теоретическую вероятность p_i для каждого интервала группировки по формуле: $p_i = F(a_{i+1}) - F(a_i)$;

– после нахождения всех p_i определяем абсолютную теоретическую частоту попадания в i -ый интервал как $n \times p_i$, где n -объём выборки;

– вычисляем статистику:

$$\hat{\chi}^2 = \sum_{i=1}^k \frac{(h_i - np_i)^2}{np_i},$$

где h_i - эмпирическая частота, т.е. число наблюдений попавших в i -ый интервал группировки;

– рассчитываем теоретическое значение критерия χ^2 при $m=k-1$, если μ и σ не заданы и $m=k-1-2=k-3$, если вместо μ и σ использованы их оценки. С этой целью удобно использовать Excel-функцию ХИ2ОБР (α , М);

– если вычисленное $\hat{\chi}^2 >$ теоретического значения χ^2 , нулевая гипотеза об идентичности сравниваемых генеральных совокупностей отклоняется, а иначе- принимается.

ПРИМЕР:

При построении модели управления процессом первичной переработки нефти необходимо проверить нормальность распределения выходной температуры разделения фракции бензин-авиакеросин.

Было произведено 101 измерение температуры.

Генеральное среднее и стандартное отклонение не известны.

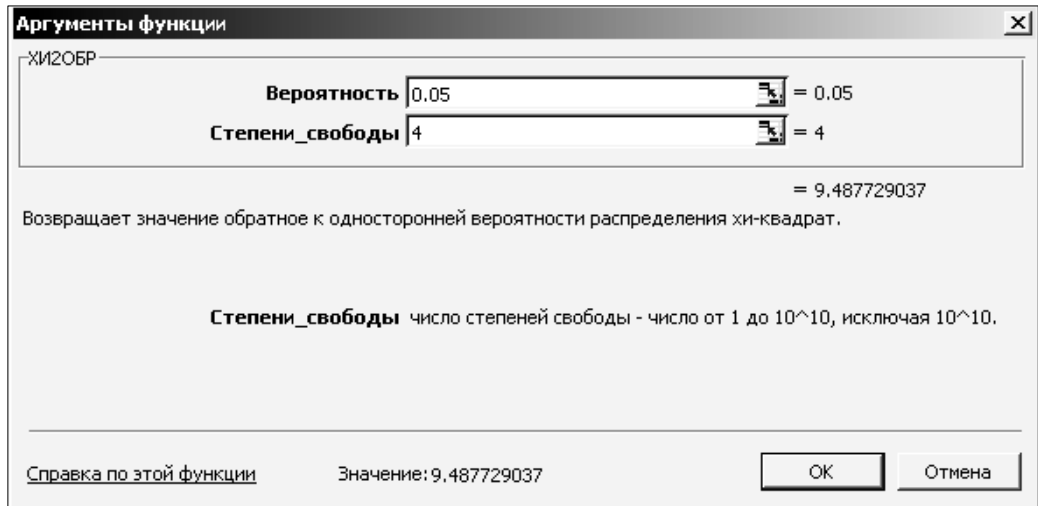
Решение:

– рассчитаем среднее и ст. отклонение: $\bar{x} = 140.05$; $S = 3.125$;

– в качестве оценок параметров μ и σ принимаем \bar{x} и S , соответственно;

– строим вероятностную таблицу (таблица 7);

– вычисляем $\chi_{0.05, 4}$ с помощью Excel-функции ХИ2ОБР(0.05;4). Оно равно 9.49;



- сравниваем теоретическое $\chi^2_{0.05, 4}$ и выборочное $\hat{\chi}^2$. Согласно результатам расчётов, сведённых в таблицу 2 оно равно 3.55;
- поскольку $3.55 < 9.49$, то гипотеза о нормальном законе температуры разделения фракций бензина и авиакеросина принимается.

Таблица 7- Вероятностная таблица для проверки статистического распределения по χ^2 -критерию Пирсона

| Интервалы группировки t_i | Нормированные границы интервалов | | Значение теоретического распределения на концах интервалов | | Теоретическая частота $p_i = F(a_{i+1}) - F(a_i)$ | Абсолютная теоретическая частота np_i | Абсолютная эмпирическая частота h_i | $h_i - np_i$ | $\frac{(h_i - np_i)^2}{np_i}$ |
|--------------------------------|----------------------------------|---|--|--------------|--|--|--|--------------|-------------------------------|
| | $a_i = \frac{t_i - \bar{x}}{S}$ | $a_{i+1} = \frac{t_{i+1} - \bar{x}}{S}$ | $F(a_i)$ | $F(a_{i+1})$ | | | | | |
| 133-135 | -2.24 | -1.60 | 0.0125 | 0.055 | 0.0425 | 4.2925 | 4 | -0.295 | 0.019932 |
| 135-137 | -1.6 | -0.96 | 0.055 | 0.168 | 0.113 | 11.413 | 10 | -1.413 | 0.174938 |
| 137-139 | -0.96 | -0.32 | 0.168 | 0.375 | 0.207 | 20.209 | 17 | •3S07 | 0.730121 |
| 139-141 | 4.32 | 0.32 | 0.375 | 0.625 | 0.250 | 25.25 | 10 | 4.75 | 0.893564 |
| 141-143 | 0.32 | 0.96 | 0.625 | 0.832 | 0.207 | 20.907 | 21 | 0.093 | 0.000414 |
| 143-145 | 0.96 | 1.6 | 0.832 | 0.945 | 0.113 | 11.413 | 12 | 0587 | 0.030191 |
| 145-147 | 1.6 | 2.24 | 0.945 | 0.9875 | 0.0425 | 4.2925 | 7 | 2.707 | 1.707128 |
| $\bar{\chi}^2 = 3.556288$ | | | | | | | | | |

1.6.2 Критерий Вилкоксона

Этот критерий предназначен для проверки гипотез: $H_0: a_1 = a_2$ против альтернатив $H_1: a_1 \neq a_2$, где a_1 и a_2 есть истинные средние для первого и второго объектов.

Являясь ранговой, статистика критерия Вилкоксона нечувствительна к нарушению нормальности распределения исходных геологических данных, а также к наличию «ураганных» содержаний и прочим стохастическим моментам. Единственное условие для 2-х сравниваемых выборок: они должны быть независимыми и подчиняться непрерывным распределениям. Процедура применения критерия Вилкоксона заключается в следующем:

- а) обозначим через m число наблюдений в выборке меньшего объёма, а через n - число наблюдений выборки большего объёма;
- б) составляем из обеих выборок общий вариационный ряд, т.е. сливаем обе выборки в одну;
- в) сортируем новую объединённую выборку по возрастанию значений признака;
- г) припишем её членам номера от 1 до $N = m + n$. Номер в этом ряду представляет собой ранг данного наблюдения;
- д) членам ряда с одинаковыми значениями следует дать один и тот же средний (откорректированный) ранг. Для этого:
 - 1) в общем вариационном ряду выделяются группы рангов, соответствующих одинаковым значениям признака;
 - 2) вычисляется среднее арифметическое значение соответствующих рангов;
 - 3) это значение заносится в скорректированный вариационный ряд;
- е) если значения в общем вариационном ряду не повторяются, то их ранг без изменения заносится в скорректированный вариационный ряд;

ж) вычисляется W-статистика Вилкоксона, представляющая собой сумму рангов, относящихся к членам меньшей по объёму выборки. При этом критические значения W_1 и W_2 вычисляются по двум, различающимся между собой алгоритмам- для выборок с максимальным объёмом более 25 наблюдений и менее 25 наблюдений.

Мы рассмотрим первый случай, когда объём одной из выборок составляет более 25 членов.

ПРИМЕР

Проверяется предположение, что фактор глубинности не влияет на среднюю концентрацию молибдена в гранитах Эльджуртинского массива при 5 % уровне значимости. Общий вариационный ряд, ранги и скорректированные ранги приведены в нижеследующей таблице.

Цветом выделены скорректированные ранги, относящиеся к меньшей по объёму выборке (поверхностной), использованной для подсчёта W-статистики. Она представляет их сумму и равна **213**.

Формулы для подсчёта критических $W_{лев}$ и $W_{прав}$ приводятся ниже.

$$W_{лев} = \frac{m(m+n+1)-1}{2} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{mn(m+n+1)}{12} \left(1 - \frac{\sum_{i=1}^k (t_i^3 - t)}{(m+n+1)(m+n-1)} \right)} = 322.45$$

$W_{прав} = m(m+n+1) - W_{лев} = 487.55 \rightarrow 213 < 322.45 < 487.55 \rightarrow H_0: a_1 = a_2$ отвергается!

В них значение $z_{крит}$ представляет квантиль гауссовского распределения, возвращаемого функцией Excel НОРМСТОБР(ВЕРОЯТНОСТЬ).

В нашем случае в качестве параметра функции использовано значение для двустороннего критерия $= 1 - \alpha / 2 = 1 - 0.025 = 0.975$. Тогда $Z_{крит} = 1.96$.

Поскольку W- статистика (213) не помещается в интервал]322.5-487.55[мы вынуждены отклонить нулевую гипотезу о равенстве средних и принять альтер-

нативную, о том, что распределение концентраций молибдена меняется с глубиной.

Таблица 8- Проверка равенства средних с помощью непараметрического критерия Вилкоксона

| Содержание Мо, г/т | | Ранг в общем вариационном ряду | Скорректированный ранг | Число совпавших величин в группе (t_i) | $t_i^3 - t_i$ |
|----------------------------|--------------------------|--------------------------------|------------------------|--|---------------|
| Граниты поверхности $m=18$ | Граниты скв № 600 $n=26$ | | | | |
| 1 | 2 | 3 | 4 | 5 | 6 |
| 0.40 | | 1.00 | 1.00 | | |
| 0.48 | | 2.00 | 2.00 | | |
| 0.49 | | 3.00 | 3.00 | | |
| 0.54 | | 4.00 | 4.00 | | |
| 0.56 | | 5.00 | 5.00 | | |
| 0.63 | | 6.00 | 6.50 | | |
| 0.63 | | 7.00 | 6.50 | | |
| 0.65 | | 8.00 | 8.50 | 2.00 | 6.00 |
| | 0.65 | 9.00 | 8.50 | | |
| 0.67 | | 10.00 | 10.00 | | |
| 0.71 | | 11.00 | 11.50 | | |
| 0.71 | | 12.00 | 11.50 | | |
| 0.80 | | 13.00 | 13.00 | | |
| | 0.81 | 14.00 | 14.00 | | |
| 0.85 | | 15.00 | 16.50 | 4.00 | 60.00 |
| 0.85 | | 16.00 | 16.50 | | |
| | 0.85 | 17.00 | 16.50 | | |
| | 0.85 | 18.00 | 16.50 | | |
| | 0.86 | 19.00 | 19.00 | | |
| | 0.91 | 20.00 | 20.00 | | |

Продолжение таблицы 8

| 1 | 2 | 3 | 4 | 5 | 6 |
|--------|------|-------|----------|------|---------------------------------------|
| 0.92 | | 21.00 | 21.50 | 2.00 | 6.00 |
| | 0.92 | 22.00 | 21.50 | | |
| 0.98 | | 23.00 | 23.00 | | |
| | 1.13 | 24.00 | 24.00 | | |
| | 1.18 | 25.00 | 25.00 | | |
| 1.20 | | 26.00 | 26.50 | 4.00 | 60.00 |
| 1.20 | | 27.00 | 26.50 | | |
| | 1.20 | 28.00 | 26.50 | | |
| | 1.20 | 29.00 | 26.50 | | |
| | 1.23 | 30.00 | 30.00 | | |
| | 1.25 | 31.00 | 31.00 | | |
| | 1.42 | 32.00 | 32.00 | | |
| | 1.43 | 33.00 | 33.00 | | |
| | 1.45 | 34.00 | 34.00 | | |
| | 1.58 | 35.00 | 35.00 | | |
| | 1.60 | 36.00 | 36.00 | | |
| | 1.66 | 37.00 | 37.00 | | |
| | 1.67 | 38.00 | 38.00 | | |
| | 1.75 | 39.00 | 39.00 | | |
| | 1.83 | 40.00 | 40.00 | | |
| | 1.89 | 41.00 | 41.00 | | |
| | 2.03 | 42.00 | 42.00 | | |
| | 2.05 | 43.00 | 43.00 | | |
| | 2.16 | 44.00 | 44.00 | | |
| СУММЫ: | | | W=213.00 | | $\sum_{i=1}^{44} (t_i^3 - t_i) = 132$ |

Вопросы для самопроверки:

- 1 Назовите основную причину бурного развития непараметрических методов проверки гипотез. Какие варианты проверки равенства генеральных средних вам известны?
- 2 На какие группы разделяются непараметрические критерии проверки статистических гипотез?
- 3 Как рассчитать число степеней свободы при определении теоретического значения критерия χ^2 ?

1.7 Корреляционный и регрессионный анализ при решении прогнозных задач

Как мы уже знаем, наблюдаемые геологические характеристики могут иметь вероятностный характер. В качестве примера можно привести толщины пластов, концентрации металлов, соотношения гранулометрические фракций обломочных пород и многое другое.

Надо сказать, что эффективность геолого-статистического моделирования существенно возрастает при переходе от однопризнаковых статистических моделей к многопризнаковым. Здесь под эффективностью будем понимать успех геологического прогноза, осуществляемого в несколько обязательных этапов. Все вместе они формируют т.н. стандартный граф первичной статистической обработки (ПСО), который включает в себя как минимум:

- установку факта наличия статистической зависимости изучаемых геологических характеристик;
- определение формы этой зависимости;
- определение параметров уравнения связи изучаемых характеристик;
- определение возможности использования найденной закономерности в конкретном регионе;

– получение количественных значений прогнозируемых характеристик.

Когда мы говорим о многомерных статистических моделях, появляется возможность сопоставления величин, характеризующих сравниваемые признаки [7]. К примеру, в ряде случаев мы можем заметить, что содержания меди и цинка в выборках изменяются однонаправлено, т.е. чем больше меди, тем больше и цинка. В то же время с повышением концентрации серебра или свинца, количество бария часто уменьшается. На интуитивном уровне понятно, что сравниваемые элементы связаны друг с другом, но эта связь нуждается в доказательстве и объективной оценке. При этом различаются две группы связей или зависимостей – статистическая и функциональная зависимость.

Статистическая взаимосвязь изучается двумя видами анализа: корреляционным и регрессионным.

Задача корреляционного анализа состоит в установлении наличия статистически значимой связи между переменными.

Задача регрессионного анализа заключается в определении параметров статистически значимой связи между переменными.

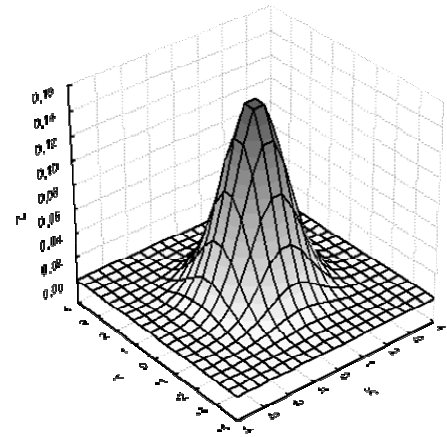
Часто эти два вида анализа рассматриваются совместно и тогда говорят о корреляционно-регрессионном анализе. Методы корреляционного анализа в зависимости от типа используемых данных (количественных, ранговых или качественных), делятся на параметрические и ранговые.

Функциональная связь в отличие от статистической является закономерной, т.е. каждому значению аргумента функции будет однозначно соответствовать её значение. Например, вес бурового инструмента функционально зависит от его длины.

1.7.1 Корреляционный анализ

Рассмотрим задачу этапа (а) графа ПСО, состоящую в выявлении взаимосвязи между двумя признаками. В случае двух переменных выборка представляет собой результаты измерений n пар значений $(x_1, y_1) \dots (x_n, Y_n)$.

Статистическое распределение этих пар иногда можно описать двумерным нормальным распределением, график которого имеет вид колокола- трёхмерной фигуры, в отличие от ранее рассмотренных плоских графиков плотности вероятностей.



Будем для простоты считать, что наш геологический объект описывается количественными признаками, например, мы сравниваем мощности двух пропластков. Для этой цели надо рассчитать коэффициент парной корреляции. Он покажет насколько сильна тенденция к существованию линейной зависимости между двумя мощностями и какой характер эта зависимость носит: прямой (совместное увеличение или уменьшение) или обратный (рост значения мощности одного пропластка сопровождается уменьшением мощности другого).

Коэффициент корреляции рассчитывается следующим образом:

Пусть мы имеем таблицу из n значений двух параметров x и y . Рассчитываются средние значения для каждого параметра:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

Тогда парный коэффициент корреляции равен:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x}) \times (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \times \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Коэффициент корреляции изменяется в диапазоне от -1 до 1. Единичные значения коэффициента корреляции свидетельствуют о *функциональной*, а не о статистической связи сравниваемых переменных.

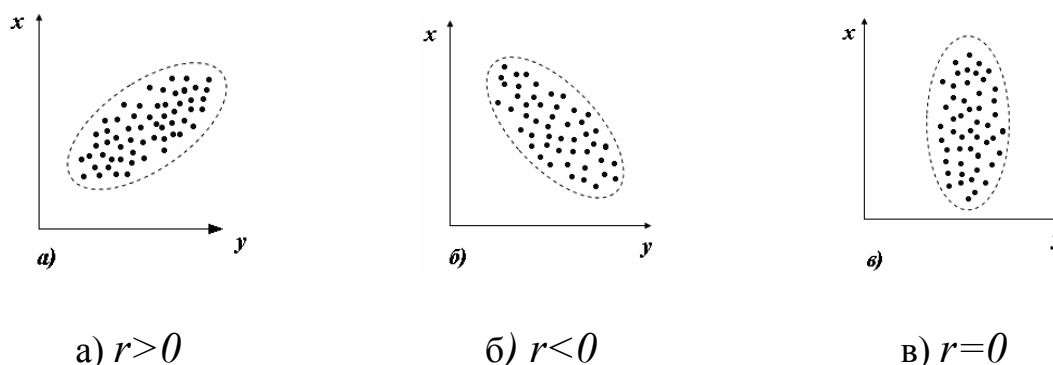


Рисунок 8- Коэффициент корреляции а)- положителен; б)- отрицателен; в)- отсутствует

Коэффициент корреляции является случайной величиной, поскольку вычисляется из случайных величин. Для него можно выдвигать и проверять следующие две гипотезы:

1 Коэффициент корреляции значимо отличается от нуля.

Это означает, что исследуемые случайные величины коррелируют друг с другом.

В этом случае тестовая статистика вычисляется по формуле:

$$\xi = \left(0.5 \ln \left(\frac{1+r}{1-r} \right) - \frac{|r|}{2(n-1)} \right) \sqrt{n-3}$$

и сравнивается с табличным значением коэффициента Стьюдента $t_{\alpha, m}$,

где α - уровень значимости;

m - число степеней свободы.

Если тестовая статистика больше табличного значения, то коэффициент значимо отличается от нуля. По формуле видно, что чем больше измерений n , тем лучше (чем больше тестовая статистика, вероятнее, что коэффициент значимо отличается от нуля).

2 Отличие между двумя коэффициентами корреляции значимо:

Тестовая статистика вычисляется по формуле:

$$\xi = 0.5 \ln \left(\frac{(1+r_1)(1-r_2)}{(1-r_1)(1+r_2)} \right) \frac{1}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}}$$

также сравнивается с табличным значением t - статистики.

Вычисление коэффициента корреляции удобно производить с помощью пакета «Анализ данных», доступного в среде MS Excel через каскад:

Сервис→Анализ данных→Корреляция

Результатом вычислений является корреляционная матрица анализируемых признаков.

Так, на рисунке 9 представлены итоги расчёта корреляционной матрицы выборочных содержаний Cu и Zn на одном из колчеданных месторождений Оренбургской области.

| Cu | Zn |
|-------|--------|
| 37.69 | 30.38 |
| 33.12 | 15.73 |
| 63.19 | 56.65 |
| 65.14 | 64.73 |
| 37.09 | 28.99 |
| 30.89 | 21.82 |
| 11.82 | 24.67 |
| 1.27 | 7.00 |
| 20.68 | 8.84 |
| 41.94 | 5.00 |
| 99.53 | 1.00 |
| 90.86 | 100.16 |
| 98.57 | 90.07 |
| 74.51 | 69.34 |
| 63.88 | 81.54 |
| 88.16 | 86.17 |
| 70.04 | 53.94 |
| 37.58 | 29.35 |
| 81.81 | 68.33 |
| 16.83 | 10.17 |
| 47.88 | 66.06 |
| 90.35 | 83.59 |
| 64.97 | 51.92 |
| 42.97 | 31.67 |
| 70.68 | 79.34 |
| 39.69 | 26.08 |
| 56.15 | 71.45 |
| 20.21 | 36.52 |
| 34.38 | 46.35 |

| | Cu | Zn |
|----|----------|----|
| Cu | 1 | |
| Zn | 0.758563 | 1 |

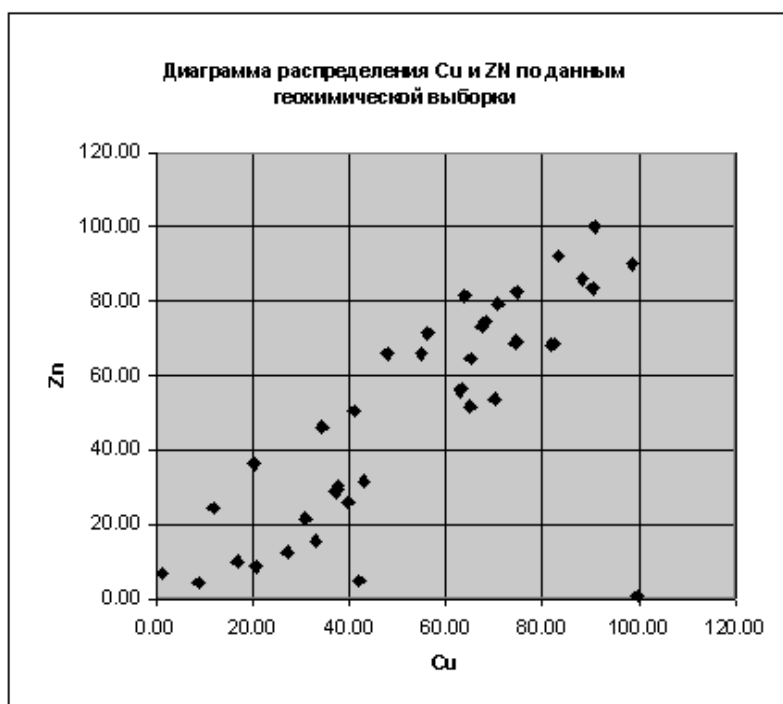


Рисунок 9- Корреляционное облако точек Cu и Zn

Значение коэффициента корреляции меди и цинка здесь составляет более 0.75, что характерно для месторождений этого типа.

При использовании коэффициента корреляции для решения практических задач следует учитывать его особенности.

Во-первых, в условиях распределений признаков отличных от нормального, коэффициент корреляции очень чувствителен к наличию аномальных выбросов.

Во-вторых, значение коэффициента корреляции зависит от системы мер. Сравнимые признаки обязательно должны измеряться в одинаковых шкалах.

1.7.1.1 Корреляционное отношение

Равенство нулю коэффициента корреляции доказывает лишь, что между случайными величинами x и y отсутствует линейная зависимость. Дело в том, что она может быть и нелинейной. Характеристикой, которая указывает на наличие как линейной, так и нелинейной зависимости случайных величин x и y является *корреляционное отношение*. Выборочные варианты этой характеристики, обозначаются так:

η_{yx} - корреляционное отношение Y к X ;

η_{xy} - корреляционное отношение X к Y .

Выборочным корреляционным отношением Y к X называют называется частное деления межгрупповой дисперсии на общую дисперсию признака Y .

$$\eta_{yx} = \frac{S_{yx}^2}{S_y^2}$$

При этом под межгрупповой дисперсией в общем многомерном случае понимается дисперсия групповых средних относительно общей средней, т.е.

$$D_{\text{межгрупп}} = \left(\sum N_j (\bar{x}_j - \bar{x})^2 \right) / n,$$

где \bar{x}_j - групповая средняя по j -ой группе (признаку);

\bar{x} - общая средняя;

N_j - объём j -ой группы (выборки по j -му признаку);

n - сумма объёмов всех групп.

Выборочным корреляционным отношением X к Y называют называется частное деления межгрупповой дисперсии на общую дисперсию признака X.

$$\eta_{xy} = \frac{S_{xy}^2}{S_x^2}$$

Утомительную процедуру расчёта этих видов дисперсий упрощает использование пакета «Анализ данных» MS Excel. Для вычисления корреляционного отношения необходимо:

- выполнить процедуру однофакторного дисперсионного анализа для сравниваемых выборок;
- вычислить межгрупповую дисперсию. Для этого надо просто поделить соответствующие суммы квадратов (SS) на общее число степеней свободы. Оно равно суммарному количеству наблюдений в выборках за вычетом единицы;
- вычислить частное от деления межгрупповой дисперсии на дисперсию соответствующего признака.

В целом, корреляционное отношение ведёт себя так же как коэффициент корреляции. Т.е. если оно равно 0, то связь отсутствует, а если равно 1- то связь функциональна. Главное его преимущество в том, что оно позволяет установить и нелинейные зависимости. Хотя, надо сказать они встречаются довольно редко. Чаще всего при анализе оказывается, что криволинейность связана с полимодальностью выборок.

1.7.1.2 Регрессия

В статистике слово регрессия означает зависимость одной величины (неслучайной) от другой или других (случайных). В неявном виде эта зависимость прослеживается уже при рассмотрении облака рассеяния сравниваемых признаков. Если зависимость между ними линейная, то можно мысленно провести усредняющую прямую через скопление соответствующих точек на плоскости. Эта линия, а вернее её уравнение, и есть функция регрессии. Разумеется, и для нелинейных связей существуют свои уравнения регрессии, но они описываются многочленами более высоких порядков.

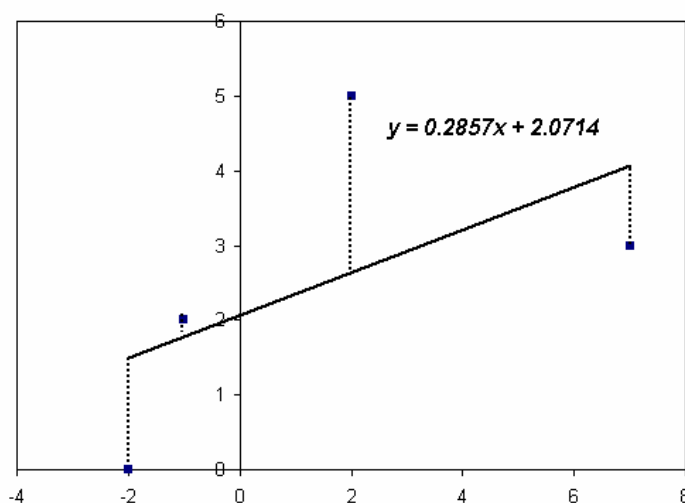
Расчёт уравнения регрессии обычно производится с помощью т.н. метода наименьших квадратов (МНК), который будет нами рассмотрен для простейшего двумерного случая.

Допустим, мы имеем набор из 4-х пар значений признаков:

| | | | | |
|-------|----|----|---|---|
| X_i | -2 | -1 | 2 | 7 |
| Y_i | 0 | 2 | 5 | 3 |

Используем их для построения графика линии, такой, чтобы квадраты расстояний от этих точек до линии были минимальными. Чтобы добиться этого составим систему уравнений:

$$\begin{cases} na + b \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \\ a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i \end{cases}$$



В нашем случае эта система примет вид:

$$\begin{cases} 4a + 6b = 10 \\ 6a + 58b = 29 \end{cases}$$

Отсюда:

$$b = 0.2857;$$

$$a = 2.0714.$$

Именно поэтому уравнение регрессии по четырём приведённым точкам имеет вид $y = 0.2857x + 2.0714$.

В настоящее время, большинство программных комплексов требуют только указать предполагаемый вид зависимости (линейный, парабола и т.д.), а расчёт коэффициентов производится автоматически. Тем не менее, общий порядок действий остаётся таким, как было изложено выше.

Рассмотрим несколько примеров использования уравнений регрессии в геологической практике.

Известно, что в ряде случаев бурение может идти без отбора кернового материала. Это экономит много времени и средств, поскольку бурение, т.н. «сплошным забоем» требует меньше спуско-подъёмных операций, которые делятся иногда целые смены. Однако, хотя керн и не отбирается, геологу необходимо знать важнейшие характеристики геологического разреза разбуриваемой толщи. Это, состав пород и данные о коллекторских свойствах продуктивных пластов.

Дело в том, что так называемые геологические запасы нефтяной залежи напрямую зависят от пористости коллекторов. Флюиды (вода, нефть, газ, конденсат) занимают микроскопические поры горных пород, которые и являются их вместительным.

Проницаемость горных пород- это характеристика обеспечивающая в принципе извлечение флюида из недр. Она контролирует, т.н. извлекаемые запасы. Если поры сообщаются между собой, то проницаемость высока, если нет, то она близка к нулю. В этом случае флюид попадёт в скважину только при разрушении поры буровым снарядом. Каким бы сильным не был буровой насос, нефть из самых ближайших окрестностей скважины не попадёт в ствол, т.к. поры герметически закупорены. Именно по этим причинам определение пористости и проницаемости являются важнейшими задачами интерпретации каротажа. Буквально вся эта работа построена на использовании эмпирически выявленных закономерностей, т.е. уравнениях регрессии. Часто графики этих уравнений для последовательного ряда значений случайной величины отрисованы на планшетах и тогда их называют палетками. Здесь в качестве случайных величин выступают значения физических полей измеренных вдоль ствола скважины с помощью каротажных приборов.

Рассмотрим рисунок 10, иллюстрирующий использование уравнений регрессии при интерпретации каротажа. Здесь в первом столбце каротажной диаграммы изображён график нейтронного гамма каротажа (НГК). Этот метод заключается в радиоактивном облучении пород с последующей регистрацией вызванного нейтронного излучения. Оно пропорционально содержанию атомов водорода в разрезе (при бомбардировке протон превращается в нейтрон), а значит и воды.

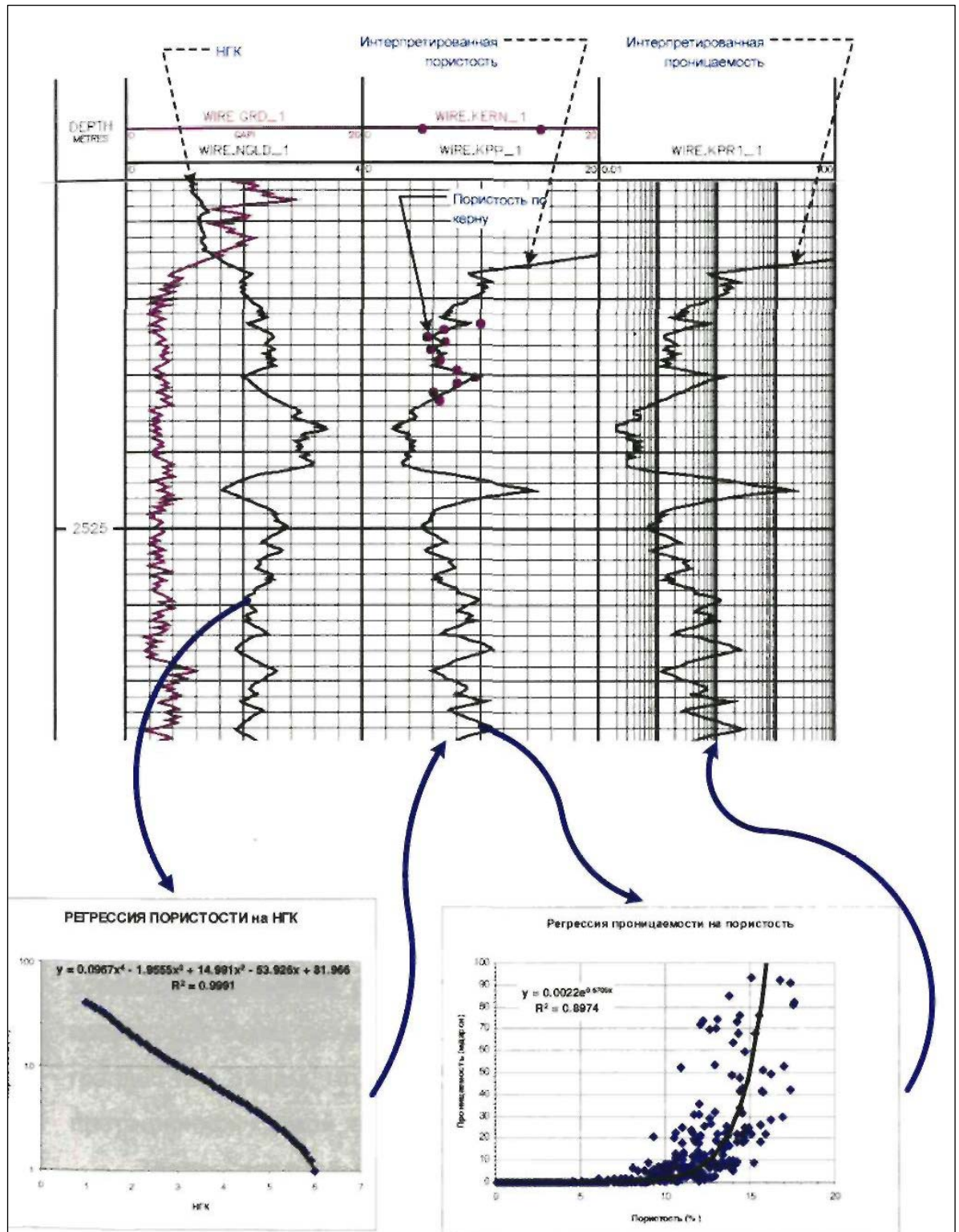


Рисунок 10- Интерпретация НГК с использованием палеток регрессии

Но вода концентрируется в порах горных пород, и логическая цепь рассуждений приводит к мысли о связи результатов НГК с пористостью разреза. При этом понятно, что сами значения НГК являются случайными величинами, поскольку формируются под влиянием множества неизвестных нам процессов. На графике

видно, что они изменяются от 0 до 4 условных единиц. Чтобы связать эти значения с реальной пористостью необходимо иметь данные об истинной пористости керна из скважин нашего региона, определенные в лаборатории физики пласта.

Если мы располагаем этими сведениями, несложно рассчитать уравнение регрессии пористости керна на НГК. Это полином 4-степени, но он очень похож на прямую линию.

Имея это уравнение регрессии, мы подставляем в него вместо x приборные значения НГК и в результате получаем новую кривую, которую назовём интерпретированной пористостью. По сути дела мы прогнозируем значения пористости на основании уравнения регрессии. Прогнозируемая пористость будет уже измеряться в настоящих единицах (от 0 до 20 %) и представлена во втором столбце каротажной диаграммы. К счастью, в этой скважине мы имеем 11 отобранных и проанализированных в лаборатории физики пласта керновых проб. Таким образом, мы имеем возможность проверить правильность нашей интерпретации НГК, на основании использованного уравнения регрессии. С этой целью результаты керновой пористости были вынесены в виде красных кружков на каротажный график. Мы видим, что эти кружки очень близко примыкают к интерпретационной кривой.

Теперь перейдём к определению проницаемости. На основании базы данных лаборатории физики пласта по керновому материалу была рассчитана регрессия проницаемости на пористость. Поскольку мы располагаем данными по интерпретированной пористости всего разреза скважины, то, подставляя их в соответствующую палетку (уравнение регрессии), легко вычисляем проницаемость. Эти данные могут использоваться при подсчёте запасов месторождения.

Приведённый пример- один из наиболее ярких приложений использования уравнений регрессии в геологической практике. а значит наши расчёты были верны.

В рудной геологии уравнения регрессии также применяются чрезвычайно широко. Так, при определении содержаний золота необходимо проведение очень дорогостоящего анализа (т.н. золотометрии). Рядовые анализы в большинстве

своём бесполезны из-за низкой чувствительности на этот металл. Однако, в ряде регионов, установлена тесная корреляция содержания золота и мышьяка, который достаточно просто диагностируется спектральным анализом. Это связано, с концентрацией золота в определённых генерациях арсенопирита.

Таким образом, составляя на основании контрольных выборок уравнение регрессии золота на мышьяк можно вычислять концентрации золота по мышьяку. Конечно, в таких опытах надо знать меру и всегда регулярно проверять результаты регрессии контрольными анализами.

1.7.1.3 Множественная регрессия

Общее назначение множественной регрессии (этот термин был впервые использован в работе Пирсона – Pearson, 1908) состоит в анализе связи между несколькими независимыми переменными (называемыми также регрессорами или предикторами) и зависимой переменной.

Например, определение временных затрат на проведение геолого-математического моделирования является крайне сложной задачей. Заранее известно, что оно зависит от ряда факторов, таких как:

- количества скважинопластов на месторождении (*сплст*);
- количества анализируемых методов каротажа (*кртж*);
- количества персонала, задействованного на моделировании (*прсн*);
- времени доступа к исходным данным (*дан*);
- заработной платой работников (*зплат*).

Вместе с тем какие-либо утверждённые нормы на проведение этих работ отсутствуют, что не позволяет составить полную смету расходов. Однако, допустим, что у нас есть доступ к информации по уже завершённым проектам моделирования и мы можем выяснить как влияют перечисленные выше характеристики на общие затраты времени, которые также известны. Для этого необходимо составить уравнение множественной регрессии, которое может иметь вид:

$$\text{Время} = 0.5 * \text{сплст} + 0.4 * \text{дан} + 0.3 * \text{прсн} + 0.25 * \text{кртж} + 0.05 * \text{зплат}$$

Как только уравнение множественной регрессии составлено, становится ясен вклад каждой учитываемой позиции (предиктора) в величину зависимой переменной (в нашем случае- это время).

Вычислительная задача, которую необходимо решить для составления уравнения множественной регрессии, так же как и для парной регрессии состоит в расчёте линии тренда по некоторому набору точек. Однако, в многомерном случае, когда имеется более одной независимой переменной, линия регрессии не может быть отображена в двумерном пространстве.

В общем случае, процедуры множественной регрессии будут оценивать параметры линейного уравнения вида:

$$Y = a + b_1 * X_1 + b_2 * X_2 + \dots + b_n * X_n$$

1.7.1.3 Частная корреляция

Регрессионные коэффициенты (или **B**-коэффициенты) представляют независимые вклады каждой независимой переменной в предсказание зависимой переменной. Однако, иногда возникает подозрение, что величина некоторых **B**-коэффициентов складывается не столько под воздействием соответствующих параметров, сколько под согласованным влиянием некоего третьего фактора. Иными словами, если одна величина коррелирована с другой, то это может быть отражением того факта, что они обе коррелированы с третьей величиной или с совокупностью величин. Этот тип корреляции употребляется под названием *частная корреляция*. Вероятно, следующий пример поможет прояснить суть данного понятия.

Кто-то мог бы, вероятно, обнаружить значимую отрицательную корреляцию между длиной волос и ростом (невысокие люди обладают более длинными волосами).

На первый взгляд это может показаться странным; однако, если добавить признак «пол» в уравнение множественной регрессии, корреляция длины волос и роста, скорее всего, исчезнет. Это произойдет из-за того, что женщины, в сред-

нем, имеют более длинные волосы, чем мужчины; при этом они (также в среднем) ниже мужчин.

Частная корреляция представляет собой условную корреляцию между двумя величинами при фиксированных значениях остальных величин

В относительно простом случае, когда число сравниваемых признаков равно трём (например, Fe, W, Mo) частный коэффициент корреляции между W и Mo при фиксированном Fe, будет равен:

$$r_{W, Mo(Fe)} = \frac{r_{W, Mo} - r_{W, Fe} \times r_{Fe, Mo}}{\sqrt{(1 - r_{W, Mo}^2) \times (1 - r_{Fe, Mo}^2)}}$$

Частные коэффициенты корреляции и вся сопутствующая им информация легко вычисляются в пакете STATISTICA в модуле «Множественная регрессия». Они имеют тот же смысл (изменяются от 0 до единицы) и проверяются на значимость так же как и парный коэффициент корреляции.

1.7.1.4 Предположения и ограничения корреляционно-регрессионного анализа

Предположение линейности.

Прежде всего, как это видно уже из названия множественной линейной регрессии, предполагается, что нам известен характер связи, которая чаще всего сводится к линейной. На практике это предположение, в сущности, никогда не может быть подтверждено; хотя к счастью, процедуры множественного регрессионного анализа лишь в незначительной степени подвержены воздействию малых отклонений от этого предположения. В этой связи всегда полезно посмотреть на двумерные диаграммы рассеяния переменных, представляющих интерес. Если нелинейность связи очевидна, то можно рассмотреть или преобразования переменных или явно допустить включение нелинейных членов.

Предположение нормальности.

В множественной регрессии предполагается, что остатки (предсказанные значения минус наблюдаемые) распределены нормально (т.е. подчиняются закону нормального распределения). Но всегда, прежде чем сделать окончательные выводы, стоит рассмотреть статистические распределения представляющих интерес переменных. Вы можете построить гистограммы или нормальные вероятностные графики остатков для визуального анализа их распределения.

Ограничения.

Основное концептуальное ограничение всех методов регрессионного анализа состоит в том, что они позволяют обнаружить только числовые зависимости, а не лежащие в их основе причинные связи.

Например, можно обнаружить сильную положительную связь (корреляцию) между разрушениями, вызванными пожаром, и числом пожарных, участвующих в борьбе с огнем. Следует ли из этого заключить, что пожарные вызывают разрушения?

Конечно, наиболее вероятное объяснение этой корреляции состоит в том, что размер пожара (внешняя переменная, которую забыли включить в исследование) оказывает влияние, как на масштаб разрушений, так и на привлечение определенного числа пожарных (т.е. чем больше пожар, тем большее количество пожарных вызывается на его тушение). Хотя этот пример довольно прозрачен, в реальности при исследовании корреляций альтернативные причинные объяснения часто даже не рассматриваются.

Вопросы для самопроверки:

- 1 В чём состоит отличие статистической зависимости от функциональной?
- 2 В чём заключается задача корреляционного анализа?

- 3 В чём заключается задача регрессионного анализа?
- 4 В каком диапазоне изменяется коэффициент корреляции?
- 5 В каких случаях следует использовать корреляционное отношение?
- 6 В чём состоит суть метода наименьших квадратов?
- 7 В чём состоит основное предназначение метода множественной регрессии?
- 8 Что такое частная корреляция?
- 9 Какие предположения и ограничения корреляционно-регрессионного анализа Вам известны?

1.8 Дисперсионный анализ

1.8.1 Теоретические предпосылки

При создании выборок, описывающих те или иные генеральные совокупности, геолог редко бывает уверен, что наблюдения относятся именно к тому конкретному природному объекту, для изучения которого они предназначены. В одну выборку могут попасть пробы из геохимической аномалии и фоновой совокупности, из горных пород разных стадий магматизма и т.д. Разумеется, такие выборки редко подчиняются нормальному закону статистического распределения, в силу чего использование большинства параметрических критериев их изучения даст неверные результаты.

Возникает вопрос, каким образом выявить выборочную неоднородность на ранних стадиях изучения материала? Ранее мы затрагивали эту тему, говоря о полимодальности статистических распределений. Однако разделить смешанные выборки на основе одних гистограмм нереально, поскольку выборочные средние будут смещены за счёт взаимного влияния. Тем не менее, существует довольно строгий математический аппарат, позволяющий доказательно выявлять неоднород-

ность (однородность) выборок. Он называется дисперсионный анализ. Здесь под однородностью понимается её статистический аналог- равенство средних. Как мы уже все знаем, эта проблема в простых случаях легко решается попарным сравнением выборок с помощью ряда статистических критериев как параметрических, так и ранговых.

Например, выборки могут считаться одинаковыми, если по критерию Вилкоксона доказана нулевая гипотеза о равенстве средних, а затем по критерию Фишера подтверждена нулевая гипотеза о равенстве дисперсий. Однако, надо сознавать, что при изучении множества природных объектов, охарактеризованными выборками, проведение попарного сравнения наталкивается на серьёзные препятствия технического и концептуального характера. Дело в том, что количество комбинаций попарного сравнения может быть очень велико, а равенство или неравенство сравниваемых пар выборок не гарантирует равенства текущей выборки со всеми предыдущими. Так что иногда без дисперсионного анализа обойтись трудно. Однако, следует учесть, что его применение основано на параметрическом критерии Фишера (F-критерии), что налагает на нас обязанность предварительно оценить характер статистического распределения данных, участвующих в анализе. Оно должно соответствовать нормальному. В противном случае применение дисперсионного анализа будет некорректно.

Для примера рассмотрим формулировку типичной геологической задачи, решаемой с его помощью:

Пусть на некоторой территории обнаружены выходы 3-х гранитных массивов, которые после специальных исследований были отнесены к трём последовательно сменяющим друг друга фазам магматизма.

На образцах из этих гранитов проведены измерения радиоактивности. Необходимо решить вопрос- различаются ли средние значения радиоактивности опробованных гранитов. Ответ на него позволил бы выделить некоторый фактор, который можно назвать «приуроченность к определённой фазе магматизма». Зная его, мы сможем в дальнейшем выделять фазы гранитов по радиоактивности, не прибегая к сложным специсследованиям.

1.8.2 Цели и методы дисперсионного анализа

Может показаться странным, что процедура сравнения средних называется дисперсионным анализом. В действительности это связано с тем, что при исследовании статистической значимости различия между средними двух (или нескольких) групп, мы на самом деле анализируем выборочные дисперсии.

Фундаментальная концепция дисперсионного анализа была предложена Фишером в 1920 году. Возможно, более естественным было бы употребление термина «Анализ суммы квадратов» или «Анализ вариаций». Кстати сказать, последнее название принято всем англоязычным миром, где применяется под аббревиатурой ANOVA (от Analysis Of Variance).

Общая цель дисперсионного анализа формулируется как проверка статистической значимости различия между средними для сравниваемых групп [8].

Эта проверка производится с помощью разбиения суммы квадратов на компоненты или части, одна из которых обусловлена случайной ошибкой, а другая с закономерным различием групповых средних. Эта процедура разбиения более подробно излагается ниже.

1.8.2.1 Структура дисперсии и разбиение суммы квадратов

Пусть дана выборка объёмом n наблюдений. Тогда выборочная дисперсия вычисляется как сумма квадратов отклонений от выборочного среднего, делённая на $n-1$. Таким образом, при фиксированном объёме выборки, дисперсия есть функция суммы квадратов отклонений, обозначаемая для краткости SS (от английского *Sum of Squares*).

В основе всего математического аппарата дисперсионного анализа лежит разделение дисперсии на части или компоненты. Процедуру расчёта этих компонент рассмотрим на примере двух групп данных, в каждой из которых содержится по три наблюдения.

Таблица 9- Расчёт структуры выборочной дисперсии

| Наблюдения | Значение признака | |
|--------------|-------------------|----------|
| | Группа 1 | Группа 2 |
| Наблюдение 1 | 2 | 6 |
| Наблюдение 2 | 3 | 7 |
| Наблюдение 3 | 1 | 5 |

Среднее для групп: 2 6

SS для групп: 2 2

Общее среднее: 4

Общая SS: 28

Здесь средние двух сравниваемых групп существенно различны: 2 и 6 соответственно. Тем не менее, сумма квадратов отклонений от среднего значения внутри каждой группы одинакова и равна двум 2. Складывая их, получаем общее среднее, которое равно 4.

Если же теперь повторить вычисление **SS** исходя из общего среднего этих двух выборок, то получим общую величину **SS** равную 28.

Иными словами дисперсия, основанная на внутригрупповой изменчивости, характеризуется значительно меньшими величинами, чем общая дисперсия сравниваемых групп.

Причина этого, очевидно, заключается в существенной разнице групповых средних, что и объясняет существенное различие суммы квадратов.

Для дальнейшей ясности материала нам необходимо немного разобраться с терминологией метода.

1.8.2.2 SS ошибок и эффекта

Внутригрупповая изменчивость называется **SS** ошибки или остаточной компонентой или дисперсией ошибки. Эти слова означают, что при проведении эксперимента, величина данной компоненты дисперсии выборки не может быть

предсказана заранее или объяснена. Её природа объясняется случайным разбросом значений, вызванным множеством неучтённых факторов.

Межгрупповую изменчивость в отличие от предыдущей называют **SS** эффекта. Её можно объяснить различием между средними значениями признака в группах.

1.8.2.3 Проверка значимости

Проверка значимости в дисперсионном анализе основана на сравнении компоненты дисперсии, обусловленной межгрупповым разбросом (**SS** эффекта) и внутригрупповой изменчивостью (**SS** ошибки).

Если верна нулевая гипотеза (равенство средних в двух выборках), то **SS** ошибки должна во-первых, быть невелика и, во-вторых, практически совпадать с общей дисперсией, рассчитанной без учёта межгрупповых различий.

Полученные дисперсии можно сравнить с помощью **F-критерия** для любого уровня значимости. Например, если статистическая гипотеза H_0 : принимается на 5 % уровне значимости, это означает, что только в 5 % наблюдений разница в межгрупповых изменчивостях будет закономерна.

1.8.3 Геологические приложения дисперсионного анализа

1.8.3.1 Однофакторный дисперсионный анализ

В геологии очень часто приходится встречаться с влиянием нескольких причин или факторов на какое-либо явление. Допустим, что мы изучаем влияние глубины на содержание металла в руде. Фактор *глубинности* можно представить его различными уровнями, например, горизонтами 1, 2, 3,...**p**. На каждом горизонте берётся некоторое число проб. Нулевую гипотезу о том, что глубина не влияет на содержание металла можно сформулировать как равенство $a_1 = a_2 = a_3 = \dots = a_p$.

Приведённая схема типична для т.н. однофакторного дисперсионного анализа, который мы произведём сейчас на примере опробования одного из полиметаллических месторождений Урала.

Исходные данные:

- опробовано 7 горизонтов (штолен);
- на каждом горизонте отобрана серия из 7 проб.

Необходимо доказать или опровергнуть нулевую гипотезу о влиянии глубины на состав руд на 5 % уровне значимости.

Аналитические данные сведены в следующую таблицу:

Таблица 10- Результаты погоризонтного опробования месторождения

| Пробы (<i>j</i>) | Горизонты (<i>i</i>) | | | | | | | \bar{x}_j |
|-----------------------|------------------------|-------|-------|-------|-------|-------|-------|-------------------|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| 1 | 2.95 | 2.6 | 2.65 | 2.55 | 2.75 | 2.8 | 2.6 | 2.700 |
| 2 | 2.5 | 2.95 | 2.75 | 2.85 | 2.45 | 2.5 | 2.55 | 2.650 |
| 3 | 2.55 | 2.7 | 2.8 | 2.6 | 2.9 | 2.85 | 2.7 | 2.729 |
| 4 | 2.8 | 2.9 | 2.75 | 2.65 | 3 | 2.95 | 2.7 | 2.821 |
| 5 | 2.8 | 2.65 | 2.6 | 3.1 | 2.5 | 2.95 | 2.95 | 2.793 |
| 6 | 2.6 | 3.25 | 3 | 2.7 | 3 | 2.9 | 2.8 | 2.893 |
| 7 | 2.75 | 2.5 | 3.4 | 3.1 | 3.6 | 3.4 | 3.15 | 3.129 |
| \bar{x}_i | 2.707 | 2.793 | 2.850 | 2.793 | 2.886 | 2.907 | 2.779 | $\bar{X} = 2.816$ |

В этой таблице номера горизонтов (*q*) обозначены номерами $i=1,2,\dots,7$.

Номера проб (*p*) обозначены номерами $j=1,2,\dots,7$.

Таким образом, в данном примере $n_1 = n_2 = n_3 \dots = n_p = q$.

Все формульные выражения для структурирования дисперсии и разбиения суммы квадратов приведены в таблице 11. В ней суммы квадратов Q и Q_I вычисляются непосредственно, а Q_2 получается путём вычитания Q_I из Q .

Для расчёта Q_I требуется:

- найти разность между средним содержанием металла по каждому из 7 гори-

зонтов и средним по всей выборке (2.816);

- возвести её в квадрат;
- найти сумму всех 7-ми квадратов;
- умножить эту сумму на q , т.е. 7.

Для расчёта Q требуется:

- найти разность между каждым из 49 значений содержания металла и средним по всей выборке (2.816);
- возвести её в квадрат;
- найти сумму всех 49-ти квадратов.

Таким образом, процедура решения даже однофакторного дисперсионного анализа весьма трудоёмка.

Таблица 11- Схема однофакторного дисперсионного анализа

| Вид дисперсии | Сумма квадратов отклонений | Степени свободы | Средний квадрат |
|----------------|---|-----------------|------------------------------|
| Между группами | $q \sum_i^7 (\bar{x}_i - \bar{X})^2 = Q_1$ | $p-1$ | $S_1^2 = \frac{Q_1}{p-1}$ |
| Внутри групп | $\sum_{i,j}^7 (x_{ij} - \bar{x}_i)^2 = Q_2$ | $p(q-1)$ | $S_2^2 = \frac{Q_2}{p(q-1)}$ |
| Суммарная | $\sum_{i,j}^7 (x_{ij} - \bar{X})^2 = Q$ | $pq-1$ | $S^2 = \frac{Q}{pq-1}$ |

Вычисленные по этой схеме значения приводятся в таблице 12.

Таблица 12- Результаты однофакторного дисперсионного анализа

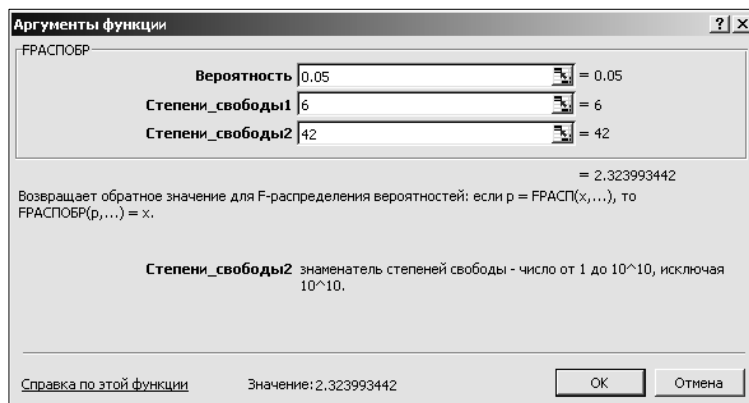
| Вид дисперсии | Сумма квадратов отклонений | Степени свободы | Средний квадрат |
|----------------|----------------------------|-----------------|-----------------|
| Между группами | 0.200510204 | 6 | 0.033418367 |
| Внутри групп | 2.906428571 | 42 | 0.06920068 |
| Суммарная | 3.106938776 | 48 | 0.064727891 |

Вычисляем F-критерий, разделив большую дисперсию (внутригрупповую) на меньшую (межгрупповую): $F = \frac{S_2^2}{S_1^2} = \frac{0.069}{0.033} = 2.07$

С помощью Excel-функции ФРАСПОБР определяем допустимую величину $F_{0.05;6;42}$.

Как видим, она равна 2.3239.

Поскольку вычисленное значение (2.07) меньше допустимого (2.3239), то нуль-гипотезу об отсутствии влияния фактора глубинности на содержание металла в выборках следует принять.



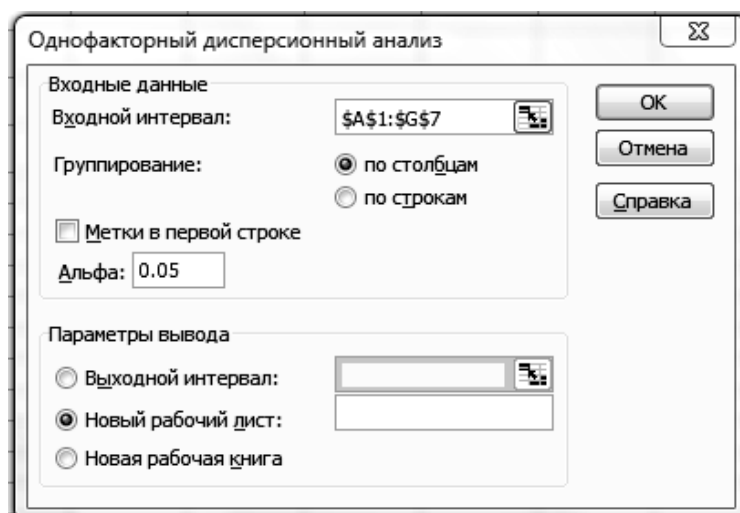
Можно избежать довольно утомительных вычислительных процедур, прибегнув к помощи пакета анализа, входящего в комплект поставки MS Office. Эта возможность становится доступной из меню:

Сервис → **Анализ Данных...** → **Однофакторный дисперсионный анализ.**

Этой компоненте следует только указать на таблицу исходных данных и выбрать уровень значимости (поле «Альфа»).

После его заполнения и запуска модуля все вычисления произведутся автоматически.

Они демонстрируются на рисунке 11, представляющим со-



бой скриншот содержимого результирующего листа MS Excel.

| | A | B | C | D | E | F | G |
|----|------------------------------------|-------------|-------|-------------|-------------|-------------|---------------|
| 1 | Однофакторный дисперсионный анализ | | | | | | |
| 2 | | | | | | | |
| 3 | ИТОГИ | | | | | | |
| 4 | Группы | Счет | Сумма | Среднее | Дисперсия | | |
| 5 | Столбец 1 | 7 | 18.95 | 2.707142857 | 0.026190476 | | |
| 6 | Столбец 2 | 7 | 19.55 | 2.792857143 | 0.066190476 | | |
| 7 | Столбец 3 | 7 | 19.95 | 2.85 | 0.075 | | |
| 8 | Столбец 4 | 7 | 19.55 | 2.792857143 | 0.052857143 | | |
| 9 | Столбец 5 | 7 | 20.2 | 2.885714286 | 0.148928571 | | |
| 10 | Столбец 6 | 7 | 20.35 | 2.907142857 | 0.071190476 | | |
| 11 | Столбец 7 | 7 | 19.45 | 2.778571429 | 0.044047619 | | |
| 12 | | | | | | | |
| 13 | | | | | | | |
| 14 | Дисперсионный анализ | | | | | | |
| 15 | Источник вариации | SS | df | MS | F | P-Значение | F критическое |
| 16 | Между группами | 0.200510204 | 6 | 0.033418367 | 0.482919636 | 0.817257161 | 2.323993797 |
| 17 | Внутри групп | 2.906428571 | 42 | 0.06920068 | | | |
| 18 | | | | | | | |
| 19 | Итого | 3.106938776 | 48 | | | | |

Рисунок 11- Результаты однофакторного дисперсионного анализа

Здесь наименования столбцов в пятнадцатой строке результирующего листа MS Excel означают:

SS - сумму квадратов;

df - число степеней свободы;

MS - средний квадрат, представляющий собой частное от деления **SS/df**;

F - величину выборочного F-критерия, который вычисляется делением межгрупповой дисперсии на внутригрупповую.

Если первая меньше чем вторая, то расчёт F критерия не имеет смысла, поскольку сразу принимается нулевая гипотеза. В русскоязычной литературе алгоритм определения F-критерия несколько иной. В числителе дроби, должна стоять большая дисперсия, а не меньшая как в пакете «Анализ». При таком способе расчёта мы получим значение 2.07.

1.8.3.2 Двухфакторный дисперсионный анализ

Очень часто, значения исследуемого признака в выборках формируются под влиянием 2-х одновременно действующих факторов. Это могут быть природные процессы или субъективные причины, связанные со способом опробования, типом прибора и т.п. При этом рассматриваются 2 варианта этого метода: без повторений и с повторениями. Разница между ними состоит в том, что в первом случае сочетание 2-х факторов описывается только одним наблюдением, а во втором их несколько.

Двухфакторный дисперсионный анализ без повторений

Применяется, для доказательства нулевых гипотез о влиянии каждого фактора на значения выборочных данных. Результаты наблюдений удобно записывать в виде следующей таблицы.

Таблица 13- Исходные данные по радиометрии гранитоидов для дисперсионного анализа без повторений

| Бригада радиометрической съёмки | Радиоактивность (в усл. единицах) по массивам гранитоидов | | |
|---------------------------------|--|----------|--------|
| | Зурбаган | Туманный | Вампир |
| Александр Плохов | 19.00 | 25.00 | 17.00 |
| Эльвира Халикова | 22.00 | 19.00 | 19.00 |
| Павел Сапожников | 26.00 | 23.00 | 22.00 |
| Андрей Цанцингер | 18.00 | 26.00 | 20.00 |
| Евгений Кувшинов | 21.00 | 22.00 | 21.00 |

В ней сводятся данные по естественной радиоактивности гранитоидов 3-х массивов (1-ый пространственный фактор) и по специалистам, производившим замеры (2-ой личностный фактор). Метод двухфакторного анализа применяется для того, чтобы выяснить:

- различаются ли распределения замеров по различным массивам;
- различаются ли распределения замеров, произведённые различными операторами?

Если на первый вопрос ответ может быть любой, то от ответа на второй вопрос зависит оценка правильности всей методики проведения радиометрической съёмки. В самом деле, если команда пользовалась одинаковой приборной базой, то личность съёмщика не должна оказывать влияния на статистические распределения замеров.

При этом двухфакторный дисперсионный анализ предполагает, по меньшей мере, три источника наблюдаемой, т.е. общей дисперсии:

- изменчивость, обусловленная или зашифрованная в строках (личностный или субъективный фактор);
- изменчивость, зашифрованная в столбцах и объяснимая принадлежностью к экспериментальной группе (территориальный или природный фактор);
- погрешность или случайная ошибка, связанная с внутригрупповой изменчивостью.

Отметим, что существует ещё один возможный источник изменчивости-*взаимодействие факторов*.

Необходимые для данного метода дисперсионного анализа алгебраические выкладки слишком громоздки, чтобы их вам приводить. Всё решение происходит в автоматическом режиме при запуске пакета Анализ из меню Сервис. Для этого необходимо:

- исходные данные по радиометрическим наблюдениям, приведённые в таблице 1 следует перенести на рабочий лист приложения MS Excel;
- запустить пакет «Анализ данных»;
- выбрать из списка «Инструменты анализа» пункт «Двухфакторный анализ без повторений» и соответствующим образом заполнить появившееся диалоговое окно.

Результаты представлены на рисунке 12. Они позволяют сопоставить значение F-критерия с его критической величиной. Если значение в столбце «F» ниже приведённых в столбце «F критическое» (в нашем случае это именно так) можно считать доказанным, что для 5 % -го уровня значимости на статистическое распределение выборок не влияет ни природный фактор ни личностный. Другими словами, средняя радиоактивность гранитов 3-х массивов не различается между собой, а члены бригады одинаково хорошо (или плохо) производили её замеры.

| | A | B | C | D | E | F | G |
|----|---|-------------|--------------|----------------|------------------|-------------------|----------------------|
| 1 | Двухфакторный дисперсионный анализ без повторений | | | | | | |
| 2 | | | | | | | |
| 3 | ИТОГИ | <i>Счет</i> | <i>Сумма</i> | <i>Среднее</i> | <i>Дисперсия</i> | | |
| 4 | Александр Плохов | 3 | 61 | 20.33333333 | 17.33333333 | | |
| 5 | Эльвира Халикова | 3 | 60 | 20 | 3 | | |
| 6 | Павел Сапожников | 3 | 71 | 23.66666667 | 4.333333333 | | |
| 7 | Андрей Цанцингер | 3 | 64 | 21.33333333 | 17.33333333 | | |
| 8 | Евгений Кувшинов | 3 | 64 | 21.33333333 | 0.333333333 | | |
| 9 | | | | | | | |
| 10 | Зурбаган | 5 | 106 | 21.2 | 9.7 | | |
| 11 | Туманный | 5 | 115 | 23 | 7.5 | | |
| 12 | Вампир | 5 | 99 | 19.8 | 3.7 | | |
| 13 | | | | | | | |
| 14 | | | | | | | |
| 15 | Дисперсионный анализ | | | | | | |
| 16 | <i>Источник вариации</i> | <i>SS</i> | <i>df</i> | <i>MS</i> | <i>F</i> | <i>P-Значение</i> | <i>F критическое</i> |
| 17 | Строки | 24.66666667 | 4 | 6.166666667 | 0.837104072 | 0.538418393 | 3.837853355 |
| 18 | Столбцы | 25.73333333 | 2 | 12.86666667 | 1.746606335 | 0.234743825 | 4.458970108 |
| 19 | Погрешность | 58.93333333 | 8 | 7.366666667 | | | |
| 20 | | | | | | | |
| 21 | Итого | 109.3333333 | 14 | | | | |

Рисунок 12- Результаты двухфакторного дисперсионного анализа (без повторений)

Данный пример иллюстрирует случай, когда некоторые факторы не влияют на исследуемый объект. Ниже приводится иная ситуация, когда влияние определённых факторов на изменчивость выборки существенно.

Двухфакторный дисперсионный анализ с повторениями

Этот тип анализа предполагает более сложную схему, когда для каждой комбинации факторов произведено не одно, а несколько наблюдений. Поскольку мы уже привыкли к постановке задачи об использовании естественной радиоактивности для классификации гранитоидов, будем использовать ту же терминологию и тех же действующих лиц. Однако, сейчас каждый участник нашего эксперимента произведёт несколько замеров, что отражается в следующей таблице:

Таблица 14- Исходные данные по радиометрии гранитоидов для дисперсионного анализа с повторениями

| Бригада радиометрической съёмки | Радиоактивность (в усл. единицах) по массивам гранитоидов | | |
|---------------------------------------|---|----------|--------|
| | Зурбаган | Туманный | Вампир |
| Александр Плохов | 0 | 0 | 3 |
| | 1 | 3 | 4 |
| | 5 | 4 | 5 |
| | 6 | 9 | 20 |
| Эльвира Халикова | 0 | 9 | 8 |
| | 6 | 10 | 12 |
| | 7 | 10 | 14 |
| | 7 | 11 | 14 |
| Павел Сапожников | 6 | 7 | 10 |
| | 7 | 8 | 13 |
| | 9 | 11 | 13 |
| | 10 | 11 | 18 |
| Андрей Цанцингер | 6 | 8 | 8 |
| | 9 | 9 | 9 |
| | 10 | 12 | 10 |
| | 11 | 15 | 13 |
| Евгений Кувшинов | 9 | 8 | 7 |
| | 9 | 11 | 9 |
| | 11 | 12 | 14 |
| | 11 | 13 | 18 |

Каждый член полевого отряда на каждом из гранитных массивов произвёл 4 замера. Нам интересно, повлияли ли изменившиеся условия эксперимента на результаты дисперсионного анализа с повторениями? С какими источниками вариации мы будем иметь дело в данном случае? Будем называть их так, как они обозначены в выходной расчётной таблице. Их всего 4:

- 1 Выборка. Дисперсия между измерениями, произведёнными участниками радиометрической съёмки. В предыдущем случае (без повторений) её аналогом были строки исходной таблицы.
- 2 Столбцы. Природная изменчивость гранитов из 3-х массивов.
- 3 Взаимодействие. Изменчивость вносимая совместным воздействием факторов на результаты замеров.
- 4 Внутри. То же самое, что и погрешность, т.е. внутригрупповая дисперсия или SS ошибок.

Запуск инструментария производится обычным образом, но из списка «Инструменты анализа» следует выбрать «Двухфакторный дисперсионный анализ с повторениями». После запуска модуля надо указать диапазон входных данных, включая заголовок и первый столбец, где обозначены факторы, влияние которых на выборку мы намерены изучить.

Результаты анализа представлены на рисунке 13.

Прежде всего, следует обратить внимание на то, что в первых двух строках таблицы «Источник вариации», приведённой на этом рисунке, значения в столбце *F* более чем в два раза превышают значения в столбце *F критическое*. Это означает, что личностный фактор для 5 % уровня значимости явно влияет на статистическое распределение изучаемого числового материала.

Таким образом, сравнивая значения *F*-критерия для выборок, столбцов и взаимодействий мы имеем право сделать следующие выводы:

- 1 Личность радиометриста значимо влияет на результаты съёмки. Сравнение средних по 1-ой таблице позволяет конкретизировать нарушителя. Это Александр Плохов, данные которого сильно занижены по сравнению со средними по массивам. Разброс данных для остальных людей невелик.

- 2 Средние по массивам значимо отличаются друг от друга.
- 3 Фактор взаимодействия не проявился, поскольку значения F -критерия в строке *взаимодействия* ниже приведённых для F критического ($0.79 < 2.15$).

| | A | B | C | D | E | F | G |
|----|---|-----------|-----------|-----------|----------|-------------------|----------------------|
| 1 | Двухфакторный дисперсионный анализ с повторениями | | | | | | |
| 2 | | | | | | | |
| 3 | ИТОГИ | | | | Итого | | |
| 4 | <i>Александр Плохов</i> | | | | | | |
| 5 | Счет | 4.00 | 4.00 | 4.00 | 12.00 | | |
| 6 | Сумма | 12.00 | 16.00 | 32.00 | 60.00 | | |
| 7 | Среднее | 3.00 | 4.00 | 8.00 | 5.00 | | |
| 8 | Дисперсия | 8.67 | 14.00 | 64.67 | 28.91 | | |
| 9 | | | | | | | |
| 10 | <i>Эльвира Халикова</i> | | | | | | |
| 11 | Счет | 4.00 | 4.00 | 4.00 | 12.00 | | |
| 12 | Сумма | 20.00 | 40.00 | 48.00 | 108.00 | | |
| 13 | Среднее | 5.00 | 10.00 | 12.00 | 9.00 | | |
| 14 | Дисперсия | 11.33 | 0.67 | 8.00 | 14.91 | | |
| 15 | | | | | | | |
| 16 | <i>Павел Сапожников</i> | | | | | | |
| 17 | Счет | 4.00 | 4.00 | 4.00 | 12.00 | | |
| 18 | Сумма | 32.00 | 37.00 | 54.00 | 123.00 | | |
| 19 | Среднее | 8.00 | 9.25 | 13.50 | 10.25 | | |
| 20 | Дисперсия | 3.33 | 4.25 | 11.00 | 11.11 | | |
| 21 | | | | | | | |
| 22 | <i>Андрей Цанцингер</i> | | | | | | |
| 23 | Счет | 4.00 | 4.00 | 4.00 | 12.00 | | |
| 24 | Сумма | 36.00 | 44.00 | 40.00 | 120.00 | | |
| 25 | Среднее | 9.00 | 11.00 | 10.00 | 10.00 | | |
| 26 | Дисперсия | 4.67 | 10.00 | 4.67 | 6.00 | | |
| 27 | | | | | | | |
| 28 | <i>Евгений Кувшинов</i> | | | | | | |
| 29 | Счет | 4.00 | 4.00 | 4.00 | 12.00 | | |
| 30 | Сумма | 40.00 | 44.00 | 48.00 | 132.00 | | |
| 31 | Среднее | 10.00 | 11.00 | 12.00 | 11.00 | | |
| 32 | Дисперсия | 1.33 | 4.67 | 24.67 | 9.09 | | |
| 33 | | | | | | | |
| 34 | <i>Итого</i> | | | | | | |
| 35 | Счет | 20.00 | 20.00 | 20.00 | | | |
| 36 | Сумма | 140.00 | 181.00 | 222.00 | | | |
| 37 | Среднее | 7.00 | 9.05 | 11.10 | | | |
| 38 | Дисперсия | 11.79 | 12.47 | 21.67 | | | |
| 39 | | | | | | | |
| 40 | | | | | | | |
| 41 | Дисперсионный анализ | | | | | | |
| 42 | <i>Источник вариации</i> | <i>SS</i> | <i>df</i> | <i>MS</i> | <i>F</i> | <i>P-Значение</i> | <i>F критическое</i> |
| 43 | Выборка | 270.60 | 4.00 | 67.65 | 5.77 | 0.00 | 2.58 |
| 44 | Столбцы | 168.10 | 2.00 | 84.05 | 7.17 | 0.00 | 3.20 |
| 45 | Взаимодействие | 74.40 | 8.00 | 9.30 | 0.79 | 0.61 | 2.15 |
| 46 | Внутри | 527.75 | 45.00 | 11.73 | | | |
| 47 | | | | | | | |
| 48 | Итого | 1040.85 | 59.00 | | | | |

Рисунок 13- Результаты двухфакторного дисперсионного анализа (с повторениями)

Анализ многофакторного взаимодействия (когда число факторов >2) весьма трудоёмок и не вошёл в пакет анализа. Для этого существует специальный аппарат факторного анализа, который будем изучаться нами несколько позже.

Вопросы для самопроверки:

- 1 Как Вы понимаете структуру дисперсии?
- 2 Чем отличается внутригрупповая дисперсия от межгрупповой дисперсии?
- 3 Каковы задачи однофакторного дисперсионного анализа?
- 4 Какие типы двухфакторного дисперсионного анализа Вам известны?
- 5 Какой инструментарий является минимально необходимым для автоматизации работ по дисперсионному анализу?

1.9 Кластерный анализ

Главной особенностью геологических исследований является недоступность объектов недр для чувственного восприятия. Геологи вынуждены судить о них по косвенным признакам, добываемым при бурении, доставляемым телеметрической аппаратурой или специальными приборами изучения физических полей. Опосредованный характер таких контактов существенно затрудняет геологическую диагностику и почти всегда сводит её к задачам математического распознавания образов или формальной классификации. Как правило, они не имеют однозначных решений, но при добросовестном и методически правильном подходе могут способствовать выявлению новых причинно-следственных связей или разумному объяснению уже выявленных закономерностей.

1.9.1 Основная цель и терминология

Один общий вопрос объединяет исследователей всех областей естествознания. Он состоит в том, как *организовать* наблюдаемые данные в относительно однородные структуры, закономерно отличающиеся друг от друга. Эти структуры называются классификациями, а их элементы- таксонами. Правильная классификация предметной области (т.е. самодостаточная и логически непротиворечивая) сама по себе является крупным научным открытием.

Однако, далеко не все классификации так бесспорны как периодическая таблица Д. И. Менделеева, буквально раскрывающая тайны строения материи. Например, в биологии насекомые отличаются от животных по совершенно формальному признаку- они имеют шесть ног. Это означает, что паук не является насекомым, хотя целиком состоит из хитина, как и мухи, которыми он питается.

Вообще говоря, любая классификация, по сути, является формальным произведением, но только одна из них полностью содержательна, т.е. соответствует внутренней природе вещей, непротиворечиво описывая конкретную предметную область. Эта идея лежит в основе автоматического создания искусственных классификаций для их последующего осмысления и содержательного толкования. Понятно, что число вариантов подобных классификаций бесконечно и на их качество, помимо методики, влияет набор признаков лежащих в основе таксономии.

Наиболее известные процедуры искусственной классификации связаны с т.н. кластерным анализом. Название этого метода происходит от англоязычного термина *cluster*, что означает скопление, пучок, группа элементов, характеризующихся каким-либо общим свойством. Собственно кластерный анализ объединяет множество методов интеграции явлений любой природы в однородные группы [9]. Эта однородность должна пониматься как относительная близость сравниваемых объектов в многомерном пространстве выбранных признаков, что само по себе не гарантирует истинность формальных классификаций. Смена признаковых координат, почти наверняка приведёт к новым кластерным решениям.

1.9.2 Область применения

Техника кластеризации применяется в самых разнообразных областях знаний. Например, в медицине кластеризация симптомов ведёт к автоматической диагностике вызвавших их заболеваний.

Задачи геологии в некотором смысле напоминают медицинские. Геологи имеют дело с внешними проявлениями природных процессов, которые происходят или происходили в особом организме планетарного масштаба. Изучение этих проявлений даёт единственную возможность типизации и разграничения геологических образований на объекты разного организационного уровня, т.е. их *классификацию*.

Представим, что петрограф диагностировал некоторый образец горной породы как гранит. Опытный специалист должен понимать, что для решения практических задач, возникающих в ходе особо сложных исследований (геологической съёмки, например) этого явно недостаточно. Необходимо отнести этот гранит к одной из разновидностей гранитов, известных на этой территории и, если этого сделать не удаётся, ввести в сводный геологический разрез новую гранитную разновидность. Следует иметь в виду, что два образца гранита, даже отобранные из одного обнажения могут несколько отличаться друг от друга. Игнорировать ли эти отличия или, напротив, использовать их в качестве разделительной линии между таксонами- дело чрезвычайно ответственное. Однако, без этой работы (по сути- классификации) в принципе невозможно понять особенности строения недр данного региона.

Обычно исследователь решает одну из двух задач классификации:

- типизация, состоящая в разбиении множества наблюдений на сравнительно небольшое число групп. При этом элементы в группах должны быть более близки друг другу, чем сами группы между собой;
- выявление естественного расслоения исходных наблюдений и объектов на чётко выраженные кластеры.

1.9.3 Процедура кластеризации

Под классификацией можно понимать группировку объектов по их схожести. Попытки разработать методы автоматической классификации выявили необходимость количественной оценки этой схожести. Её стали связывать с понятием межпризнаковых расстояний или метрик, разделяющих классифицируемые объекты и сами группы этих объектов.

В качестве первого приближения и для осознания принципов формальной группировки полезно рассмотреть случай, когда в качестве метрики используются обычные евклидовы расстояния на плоскости.

Представим себе городскую площадь в новогодние праздники. Люди на ней образуют некоторые мало заметные скопления, и мы предполагаем, что не все из них формируются совершенно случайно. Можно ли по плану распределения народа сделать какие-либо выводы о том- кто эти люди, чем они занимаются, какие у них отношения?

На первый взгляд это кажется совершенно невозможным, но именно с информацией такого типа чаще всего имеют дело геологи. Изучаемые объекты недр обычно недоступны для непосредственного наблюдения и их характеристика возможна лишь по косвенным данным. В нашем случае мы можем оперировать только линейным расстоянием между людьми, применяя его как единственную объективную метрику.

Организуем работу так, чтобы объединять людей в группы, если они находятся друг к другу ближе некоторой критической дистанции. Тогда, постоянно её увеличивая, мы сумеем генерировать новые группировки до тех пор пока не образуется одна-единственная, объединяющая всех.

Эти действия иллюстрируются рисунком 14, на котором показаны изолинии равных расстояний (изодистанты) вокруг каждого человека. Шкала подобрана таким образом, что с увеличением дистанции изолинии становятся темнее. Некоторые устойчивые группы обведены по периметру

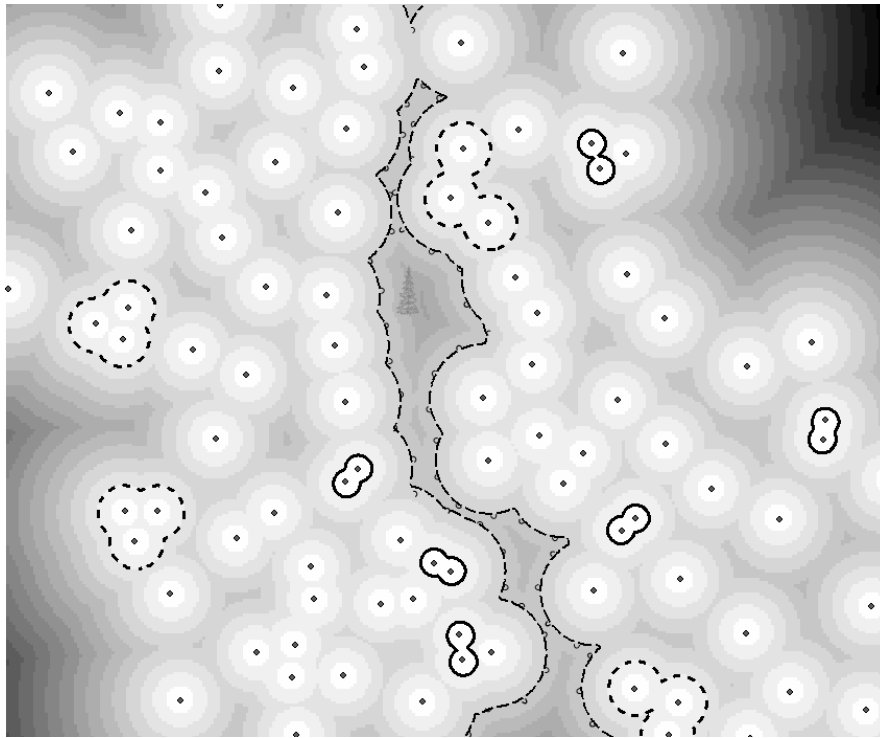


Рисунок 14- Группировка объектов по метрике Евклидовых расстояний

Результаты группировки сведены в таблицу 15

Таблица 15- Таблица группировки объектов по метрике евклидовых расстояний

| Критическое расстояние | Группировка |
|------------------------|--|
| 1 см | Выделяется 92 группы по одному человеку в каждой. Все люди обособлены друг от друга. |
| 50 см | Выделилось 6 групп (обведены сплошной линией) |
| 1 метр | Выделяется 4 группы (обведены пунктиром) |
| 3 метра | Выделяется две меридиональные группы (пунктир с бергштрихами) |
| 10 метров | Выделяется одна большая группа |

Попытаемся дать более или менее правдоподобное объяснение событиям, сопровождающим нашу понятную, но довольно примитивную классификацию:

- на первом этапе сколько-нибудь заметной группировки не происходит. Расстояние между людьми больше чем 1 см и, поэтому каждый человек (а всего их на площади 92) представляет собой обособленную группу;

- с возрастанием критического расстояния до полуметра начинают образовываться кластеры из двух человек. Они держатся парами, на удалении от толпы. Есть подозрение, что они равнодушно друг к другу;
- с увеличением критической дистанции до 1 метра возникают кластеры из трёх человек. Не исключено, что это мужские компании;
- дальнейшее увеличение критического расстояния проявляется формированием двух, приблизительно равных по численности групп. Они разделены узкой полосой, скорее всего недоступной для пешеходов. Может быть, это ледяная стена или иное препятствие;
- увеличение критической дистанции до десяти метров приводит к объединению всех людей в единую группу.

1.9.4 Типы расстояний

В приведённом выше примере в качестве меры сходства объектов использовалось простое геометрическое расстояние в двумерном пространстве. Оно может быть легко распространено на трёхмерные случаи и гиперпространства с числом измерений больше трёх. Однако, в различных ситуациях имеет смысл использовать разные способы измерения расстояний, из которых мы рассмотрим только два, вероятно, самых известных и универсальных.

1.9.4.1 Евклидово расстояние

Евклидово расстояние представляет собой длину гипотенузы в пространстве n -ой мерности, вычисляемую по теореме Пифагора. В пространстве размерностью n оно вычисляется следующим образом:

$$\text{Расстояние}(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Это означает, что если нам необходимо определить евклидово расстояние

между точками в геохимическом пространстве с координатами Cu, Zn, Co то оно будет равно: $\sqrt{(Cu_1 - Cu_2)^2 + (Zn_1 - Zn_2)^2 + (Co_1 - Co_2)^2}$

1.9.4.2 Расстояние городских кварталов (Манхэттенское расстояние)

Улицы центра Нью-Йорка (остров Манхэттен) пересекаются под прямым углом, что и послужило поводом для такого названия.

Вычисление расстояния «городского квартала» между двумя объектами заключается во взятии модуля разности между ними по каждому измерению, а затем в суммировании этих разностей.



$$\text{Расстояние}(x, y) = \sum_{i=1}^n |x_i - y_i|$$

Название метрики интерпретируется достаточно просто: если бы два объекта были зданиями в городе, то путь от одного к другому пролегал бы вдоль городских кварталов, пересекающихся под прямым углом (что отличает эту метрику от прямого кратчайшего маршрута, представляющего евклидову метрику).

Влияние на результат вычисления отдельных больших разностей в этом случае по сравнению с евклидовым расстоянием уменьшается, поскольку они не возводятся в квадрат.

1.9.5 Методы объединения в кластеры

Как было показано выше, основной целью кластерного анализа является выделение естественно возникающих групп объектов на основе их сходства. Эта задача имеет различные способы решения в соответствии с правилами формирования кластеров. В зависимости от поставленных целей классификации исследователь может применять разнообразные методы кластерного анализа, среди которых явно доминируют два: иерархический и неиерархический методы.

При иерархической кластеризации объекты (отдельные наблюдения или кластеры), попавшие в кластер, остаются объединенными на всех последующих этапах кластеризации. Рассмотренный ранее пример с городской площадью иллюстрирует типичный иерархический подход к классификации объектов.

Методы кластеризации различаются по способам оценки расстояния между кластерами.

1.9.5.1 Иерархические методы

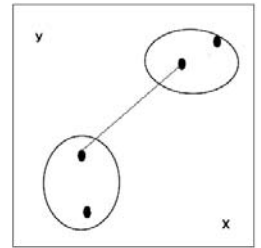
Иерархические методы кластеризации различаются между собой по способам оценки расстояния между кластерами (их «близости») при формировании кластеров. Например, если есть два кластера, содержащие по два объекта в каждом, то:

- в случае, если в качестве расстояния между ними рассматривается дистанция между центрами тяжести кластеров, имеет место метод «центроида» (или средних значений);
- если в качестве расстояния рассматривается дистанция между ближайшими точками из разных кластеров, то мы имеем дело с методом «ближайших соседей»;
- если напротив, рассматривается дистанция между самыми удалёнными точками сравниваемых кластеров, то речь идет о кластеризации методом «самого далекого соседа»;
- если исследователь предпочитает явным образом включать в вычисление расстояния все объекты из кластера, он может остановиться на методе межгруппового среднего связывания, при котором расстояние оценивается для всех возможных пар наблюдений из двух разных кластеров и затем берется среднее значение;
- метод Варда создает кластеры, комбинируя те из них, которые приводят к наименьшим внутрикластерным суммам квадратов.

Рассмотрим более подробно принципы вычисления наиболее употребительных межкластерных расстояний и особенности создаваемых при этом кластеров.

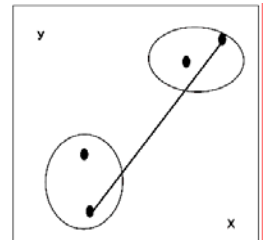
Метод ближайшего соседа (простого связывания):

- тяготеет к созданию удлинённых, «колбасообразных» кластеров, вытягивающихся за счёт присоединения ближайшей точки;
- слабо чувствителен к выбросам.



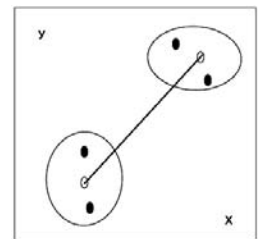
Метод самого дальнего соседа:

- слабо чувствителен к выбросам.



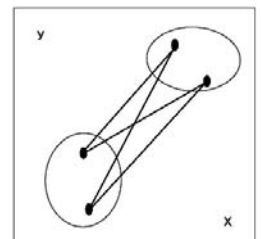
Метод центроида:

- хорошо подходит для кластеризации данных из гетерогенных «засорённых выборок»;
- слабо чувствителен к выбросам.



Метод межгруппового среднего связывания:

- хорошо работает с «засоренными» данными;
- чувствителен к выбросам;
- по мнению специалистов, наиболее универсальный метод, подходящий для разнообразных условий.



Основным результатом иерархических кластеризаций является особый график, который называется дендрограммой (от греческого *dendron*- дерево). Он схематично отражает процесс агломерации, т.е. слияния отдельных наблюдений в

единый окончательный кластер.

Попытаемся построить, а затем проинтерпретировать кластерную дендрограмму на реальном примере из геологической практики, связанном с классификацией подземных вод нефтяного месторождения.

В гидрохимическую выборку были включены пробы, отобранные с помощью герметичного испытателя пластов с глубин около 4000 м. В выборку были также включены, отобранные желонкой пробы из относительно мелких скважин. Цель исследований состояла в выяснении возможности распознавания подземных вод т.н. подсолевого водоносного комплекса от вод верхнепермского возраста, по данным их катионно-анионного состава.

В качестве инструмента кластеризации использовалась программная система STATISTICA v.6.

Для проведения кластерного анализа была выбрана евклидова метрику в восьмимерном пространстве катионов, удельного веса, общей минерализации и метод сходства наблюдений по методу «самый дальний сосед».

Исходные данные по гидрохимической выборке из 18 проб представлены на рисунок 15.

| Глубина отбора | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|----------------|-------|--------|---------|---------|--------|---------|-------|------|
| | УдВес | Минер | Na+K | Ca | Mg | Cl | SO4 | HSO3 |
| 4015.63 м | 1.170 | 241.50 | 81.02 | 4.04 | 4.27 | 141.72 | 1.56 | 0.15 |
| 210.1 м | 1.170 | 241.50 | 3480.21 | 201.92 | 351.09 | 3997.80 | 32.51 | 2.40 |
| 4083.5 м | 1.170 | 241.50 | 43.21 | 2.48 | 4.32 | 49.13 | 0.40 | 0.03 |
| 4031.38 м | 1.150 | 249.50 | 61.89 | 24.85 | 4.74 | 152.48 | 0.57 | 0.10 |
| 220.5 м | 1.150 | 249.50 | 2657.97 | 1240.99 | 390.01 | 4301.49 | 11.93 | 1.70 |
| 4022.71 м | 1.150 | 249.50 | 30.69 | 14.33 | 4.51 | 49.68 | 0.14 | 0.02 |
| 4059.28 м | 1.185 | 263.60 | 64.99 | 27.61 | 3.77 | 159.90 | 0.25 | 0.08 |
| 200.0 м | 1.185 | 263.60 | 2782.94 | 1378.59 | 310.01 | 4482.60 | 5.28 | 1.30 |
| 4088.64 м | 1.185 | 263.60 | 30.89 | 15.38 | 3.44 | 49.76 | 0.06 | 0.02 |
| 4043.76 м | 1.185 | 286.50 | 72.71 | 28.73 | 5.61 | 178.90 | 0.50 | 0.06 |
| 185.35 м | 1.185 | 286.50 | 3161.09 | 1433.80 | 461.50 | 5045.00 | 10.39 | 1.00 |
| 4073.02 м | 1.185 | 286.50 | 31.26 | 14.18 | 4.56 | 49.89 | 0.10 | 0.01 |
| 4034.07 м | 1.216 | 256.50 | 123.96 | 27.66 | 4.13 | 128.68 | 0.57 | 0.27 |
| 300.5 м | 1.216 | 256.50 | 5319.60 | 1380.00 | 340.00 | 3628.00 | 11.80 | 4.40 |
| 4024.37 м | 1.216 | 256.50 | 49.79 | 12.92 | 3.18 | 33.96 | 0.11 | 0.04 |
| 4091.57 м | 1.174 | 309.61 | 84.06 | 25.05 | 8.51 | 101.43 | 0.41 | 0.15 |
| 235.40 м | 1.174 | 309.61 | 3460.87 | 1250.00 | 700.00 | 5400.00 | 8.47 | 2.40 |
| 4072.7 м | 1.174 | 309.61 | 31.98 | 11.55 | 6.47 | 49.40 | 0.08 | 0.02 |

Рисунок 15- Исходные данные гидрохимической выборки

Результирующая дендрограмма представлена на рисунке 16.

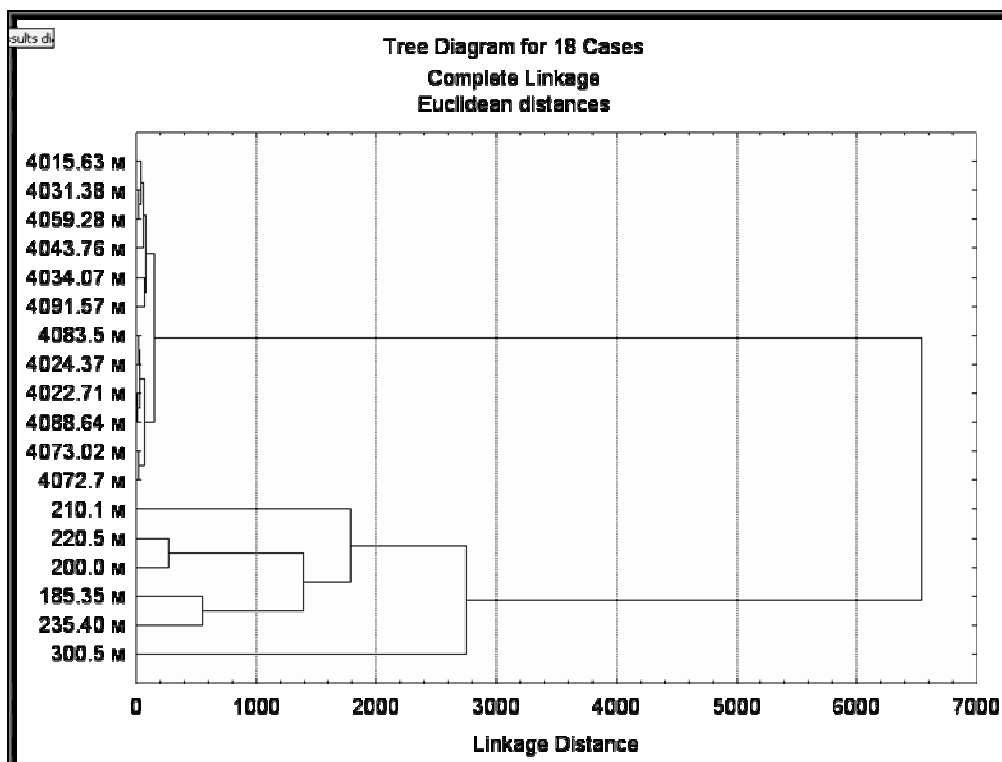


Рисунок 16- Дендрограмма по данным гидрохимической выборки

Для интерпретации дендрограммы будем двигаться вдоль оси расстояний (Linkage Distance) слева направо, обращая внимание на число и состав кластеров для наиболее характерных точек графика:

- а) при расстоянии 0 каждая гидрохимическая проба является отдельным кластером, поскольку вертикаль к оси дистанций здесь пересекает 18 ветвей дендрограммы;
- б) при расстоянии 1000 ед. выделяются 5 кластеров, объединяющих следующие пробы, отобранные с глубин:
 - 1) 300.5;
 - 2) 235.4; 185.35;
 - 3) 200.0; 220.5;
 - 4) 210.1;
 - 5) 4072.7; 4073.02; 4088.64; 4022.71; 4024.37; 4083.5; 4091.57; 4034.07; 4043.76; 4059.28; 4031.38; 4015.63;
- в) при расстоянии 2000 ед. выделяются 3 кластера, объединяющих следующие пробы, отобранные с глубин:
 - 1) 300.5;

2) 235.4; 185.35; 200.0; 220.5; 210.1;

3) 4072.7; 4073.02; 4088.64; 4022.71; 4024.37; 4083.5; 4091.57; 4034.07;
4043.76; 4059.28; 4031.38; 4015.63;

г) далее до расстояния 6500 на дендрограмме представлено только 2 кластера:

1) 300.5; 235.4; 185.35; 200.0; 220.5; 210.1;

2) 4072.7; 4073.02; 4088.64; 4022.71; 4024.37; 4083.5; 4091.57; 4034.07;
4043.76; 4059.28; 4031.38; 4015.63.

Таким образом, дендрограмма на этапе (г) показала явное разделение геохимической выборки минимум на две чётко обособленные группы (т.е. задача классификации решена):

- первая группа включает пробы надсолевого комплекса, отобранные с глубин не превышающих 300.5 метра;
- вторая группа объединила пробы исключительно подсолевого водоносного комплекса (глубины 4015.63-4088.64 м). Удивительная компактность этой группы (пробы находятся друг от друга на расстоянии менее 200 ед.) объясняется застойным гидродинамическим режимом на глубинах свыше 4 км.

1.9.5.2 Неиерархические методы

Неиерархические методы кластеризации не требуют, чтобы два объекта, попавшие в один и тот же кластер, оставались там и впоследствии. Другими словами они накладывают менее строгие ограничения на структуру данных, чем иерархические методы. Самой популярной техникой в этом классе методов является метод (или алгоритм) k-средних. Буква «k» в названии метода связана с тем, что при каждом обращении к методу исследователь должен сам выбирать (предлагать) число (k) образуемых кластеров. Слово «средних» в названии означает, что каждый кластер определяется средним значением (или центром тяжести) своих объектов.

Из практики применения метода известно, что кластеризация методом k-

средних эффективна, когда центры тяжести исходных кластеров достаточно удалены друг от друга и на больших файлах метод работает гораздо быстрее чем иерархические. При этом, однако, следует иметь в виду, что:

- прежде чем остановиться на каком-либо решении, требуется, как правило, произвести несколько попыток;
- поскольку этот метод неиерархический, для него нельзя построить дендрограмму, весьма полезную при оценивании кластерных решений.

Таким образом, для файлов данных содержащих менее ста наблюдений некоторые иерархические методы (полного связывания, межгруппового среднего связывания и др.) работают очень хорошо, и результаты можно представить в виде дендрограммы. Для больших файлов (много сотен или тысяч наблюдений) наиболее эффективным, а с точки зрения доступных системных ресурсов единственно возможным методом будет алгоритм k-средних.

Для сопоставления рассмотренных методов произведём кластерный анализ методом k-средних нашей гидрохимической выборки из 18 проб (рисунок 17). Будем изначально ориентироваться на выделение двух кластеров, рассчитывая, что они отразят деление вод по их принадлежности к надсолевому и подсолевому комплексам.

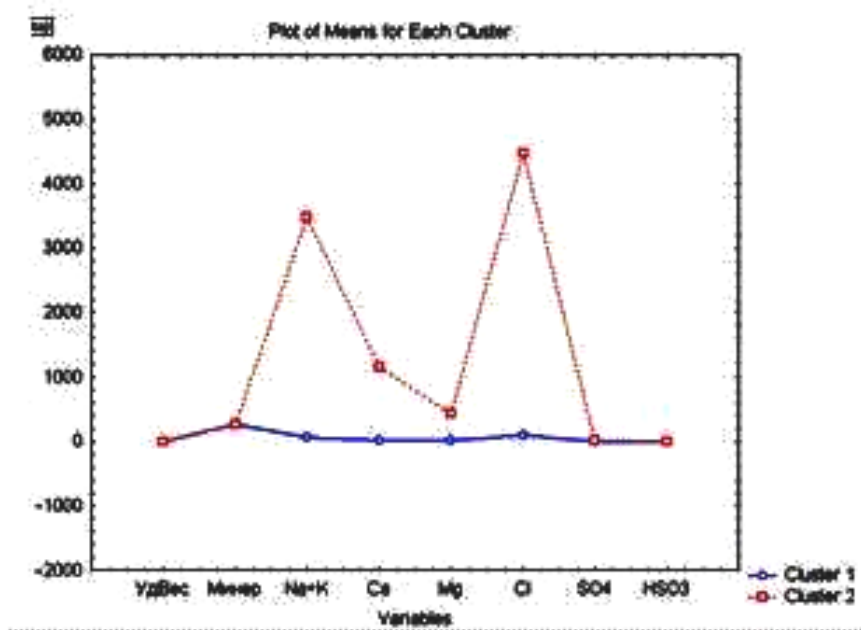


Рисунок 17- Результаты кластеризации методом k-средних

Выделенные кластеры представляются в виде двух ломаных линий. Одна из них сплошная, почти горизонтальная связана с пробами вод подсолевого комплекса, почти неотличимых друг от друга по химизму.

Пунктирная линия обозначает кластер проб из надсолевого комплекса, с существенной разницей в расстоянии (солевом составе) вод.

Таким образом, оба подхода (иерархический и неиерархический) в применении к одним и тем же данным привели к одинаковым результатам.

Вопросы для самопроверки:

- 1 Чем отличается евклидово расстояние от Манхэттенского расстояния?
- 2 Какие способы оценки расстояния между кластерами Вам известны?
- 3 Что представляет собой кластерная дендрограмма?
- 4 Каким образом производится интерпретация кластерных дендрограмм?
- 5 Чем отличаются иерархические методы кластеризации от неиерархических?

1.10 Факторный анализ

Под факторным анализом понимается совокупность статистических моделей, описывающих и объясняющих наблюдаемые данные с помощью небольшого числа скрытых (латентных) факторов, которые могут быть сконструированы с помощью определенных математических методов [10]. Модели факторного анализа применяются при решении следующих задач:

- сокращение числа переменных (редукция данных);
- определение структуры взаимосвязей между переменными, т.е. классификация переменных.

Основным объектом преобразований в факторном анализе является корреляционная матрица из коэффициентов корреляции Пирсона (иногда – дисперсионно-ковариационная матрица).

Существуют две основных модели факторного анализа: собственно факторный анализ и метод главных компонент (ГК). Оба метода предназначены для объяснения изменчивости выборки из случайных значений. Основное различие методов состоит в том, что:

- при анализе главных компонент предполагается наличие главных факторов или компонент, влияющих на изменчивость всех переменных, т.е. объясняющих всю дисперсию выборки;
- при собственно факторном анализе предполагается наличие общих факторов, влияющих на все переменные и специфических, влияющих на каждую переменную в отдельности. В терминах факторного анализа доля дисперсии отдельной переменной, принадлежащая общим факторам (и разделяемая с другими переменными) называется *общностью*.

К предпосылкам применения этих методов можно отнести:

- наличие сильно коррелированных признаков, приводящих к дублированию информации;
- слабую информативность ряда признаков;
- возможность и целесообразность агрегирования нескольких признаков.

Таким образом, сокращение размерности обоими методами факторного анализа означает отказ от описания исследуемых явлений через исходный набор результативных признаков. Они заменяются меньшим числом наиболее информативных (с точки зрения их влияния на результат) искусственных переменных (факторов), представляющих линейные комбинации подмножества исходных признаков.

Очевидно, обе схемы факторного анализа имеют право на жизнь и в большинстве случаев приводят к близким результатам. Однако, анализ главных компонент часто более предпочтителен как метод редукции данных, в то время как анализ главных факторов лучше применять с целью определения структуры данных и их классификации.

Алгоритм анализа для всех методом факторно-аналитической обработки

состоит из одних и тех же основных этапов:

- подготовка исходной матрицы данных;
- вычисление матрицы взаимосвязей признаков;
- собственно факторизация (с указанием количества выделяемых факторов);
- вращение факторов- преобразование факторов, облегчающее их интерпретацию;
- подсчёт факторных значений по каждому наблюдению;
- интерпретация результатов.

В ходе выполнения факторного анализа некоторые из указанных этапов можно опустить. Например, в качестве исходной матрицы данных можно сразу использовать корреляционную или иную матрицу взаимосвязей признаков, рассчитанную заранее и тогда работа начинается со второго этапа. Выполнение четвёртого и пятого этапов также не всегда являются обязательными.

1.10.1 Подготовка исходных данных

Практически всегда в качестве исходных данных для факторного анализа используются матрицы. Матрица- это прямоугольная таблица чисел, в которой строки соответствуют наблюдениям (объектам), а столбцы- переменным. Что считать объектами, а что переменными исследователь решает сам.

Наиболее распространён случай, когда целью исследований является анализ неких признаков (например, содержание химических элементов), характеризующих каждое наблюденное событие. Тогда эти события (например, результаты анализа) должны соответствовать строкам матрицы, а значения конкретных элементов- столбцам. Если же нас интересует не связь признаков, а связь результатов химического анализа, т.е. наблюдений, то матрицу данных надо транспонировать- записать столбцы как строки, а строки автоматически получатся записанными в столбцы.

1.10.2 Вычисление матрицы взаимосвязей признаков

Процедура факторного анализа начинается с вычисления матрицы взаимосвязей переменных между собой. Это квадратная матрица, размер которой равен количеству переменных.

Наиболее распространённой мерой взаимосвязи является корреляционная связь. Чаще всего корреляционная матрица заполняется коэффициентами корреляции Пирсона, но может вычисляться и коэффициент ранговой корреляции. В этом случае говорят о непараметрическом факторном анализе. Коэффициент корреляции удобен тем, что это стандартизованная мера взаимосвязи, не зависящая ни от единиц измерения, ни от диапазона значений переменных. В тех многочисленных случаях, когда переменные измеряются в неодинаковых единицах производится их нормировка, согласно выражению:

$$X_s = (X - X_{cp}) / \sigma$$

где X_s - стандартизованное значение переменной X ,

X_{cp} - среднее значение,

σ - стандартное отклонение.

1.10.3 Факторизация

Важная теорема матричной алгебры гласит, что матрицы, удовлетворяющие определённым условиям, могут быть диагонализированы, т.е. преобразованы в матрицу, на главной диагонали которой стоят числа, а во всех остальных позициях- нули. Матрицы взаимосвязей относятся именно к такому типу.

Само преобразование производится по формуле: $L=V'RV$ (1)

т.е. диагонализация матрицы R выполняется умножением её сначала (слева) на транспонированную матрицу V , обозначаемую как V' , а потом (справа) на саму матрицу V .

Столбцы в матрице V называются собственными векторами, а числа на главной диагонали матрицы L - собственными числами. Первый собственный вектор соответствует первому собственному числу и т.д.

Поскольку целью факторного анализа является обобщение матрицы взаимосвязей посредством как можно меньшего количества векторов и каждая собственная величина соответствует разным потенциально возможным факторам, обычно в расчёт принимаются только факторы с большими собственными значениями.

Предположим, мы исследуем взаимосвязь четырёх переменных, характеризующих условия для выбора места летнего отдыха:

- стоимость путёвки;
- комфортабельность комплекса;
- температура воздуха;
- температура воды.

Имея статистически значимую выборку результатов опроса населения о важности перечисленных условий, мы используем уравнение (1) для расчёта матрицы собственных значений допустив, что мы располагаем необходимыми сомножителями для этого. При этом наложим ограничение на количество выделяемых факторов, определив его равным двум:

$$L = \begin{bmatrix} -.28 & .18 & .66 & .68 \\ .65 & -.69 & .25 & .21 \end{bmatrix} \times \begin{bmatrix} 1 & -.95 & -.06 & -.13 \\ -.94 & 1 & -.09 & -.04 \\ -.06 & -.09 & 1 & 0.99 \\ -.13 & -.04 & 0.99 & 1 \end{bmatrix} \times \begin{bmatrix} -.28 & .65 \\ .18 & -.69 \\ .66 & .25 \\ .68 & .21 \end{bmatrix} = \begin{bmatrix} 2.00 & 0.00 \\ 0.00 & 1.91 \end{bmatrix}$$

Сведём полученные собственные векторы и соответствующие им собственные числа в таблицу 16:

Таблица 16- Таблица соответствия собственных векторов и собственных чисел

| Собственный вектор 1 | Собственный вектор 2 |
|------------------------|------------------------|
| -0.28 | 0.65 |
| 0.18 | -0.69 |
| 0.66 | 0.252 |
| 0.68 | 0.207 |
| Собственное значение 1 | Собственное значение 2 |
| 2.00 | 1.91 |

Расчёты собственных величин и векторов весьма трудоёмки и обычно выполняются в фоновом режиме в среде специализированных программных средств. К числу наиболее используемых относятся пакеты SPSS и СТАТИСТКА v.6.

Зная матрицу собственных значений, мы можем приступить к вычислению матрицы факторных нагрузок, которое производится согласно выражению:

$$A = V\sqrt{L} \quad (2)$$

В нашем примере:

$$A = \begin{bmatrix} -.28 & .65 \\ .18 & -.69 \\ .66 & .25 \\ .68 & .21 \end{bmatrix} \times \begin{bmatrix} \sqrt{2.00} & 0 \\ 0 & \sqrt{1.91} \end{bmatrix} = \begin{bmatrix} -.40 & .90 \\ .25 & -.95 \\ .93 & .35 \\ .96 & .27 \end{bmatrix}$$

Матрица факторных нагрузок является матрицей взаимосвязей (интерпретируемых коэффициенты корреляций) между факторами и переменными. Первый столбец- это корреляция между первым фактором и каждой переменной по очереди (таблица 17)

Таблица 17- Матрица факторных нагрузок

| Переменная | Фактор 1 | Фактор 2 |
|-----------------------------|----------|----------|
| Стоимость путёвки | -0.4 | 0.9 |
| Комфортабельность комплекса | 0.25 | -0.95 |
| Температура воздуха | 0.93 | 0.35 |
| Температура воды | 0.96 | 0.27 |

Интерпретация каждого фактора производится на основе сильно с ним связанных (т.е. имеющих к нему высокие нагрузки) переменных. Так, первый фактор в нашем примере можно назвать «климатическим», а второй «экономическим».

Интерпретируя эти факторы, следует обратить внимание на то, что переменные, имеющие высокие нагрузки по первому фактору (температура воды и температура воздуха) связаны положительно, тогда как переменные с высокими нагрузками по второму фактору (стоимость путёвки и комфортабельность ком-

плекса) взаимосвязаны отрицательно. Действительно, от дешёвого курорта нельзя ожидать большой комфортабельности.

Знаки факторных нагрузок имеют только относительное значение, поскольку выбор знака происходит во время вычислений случайным образом. Другими словами замена всех знаков на противоположные решения не меняет.

Обычно фактор легче интерпретируется, если с ним сильно связана только небольшая часть переменных.

1.10.3.1 Методы факторизации

К самым распространённым методам факторизации относятся метод главных компонент и метод главных факторов. Оба метода основаны на вычислении набора ортогональных компонент или факторов, по которым можно воспроизвести матрицу взаимосвязей признаков.

Метод главных компонент

Метод главных компонент ориентирован на выделение малого набора ортогональных компонент таким образом, чтобы они объясняли максимум дисперсии анализируемого набора данных. Первая главная компонента- это линейная комбинация наблюдаемых переменных, которая в максимальной степени разделяет наблюдения, максимизируя дисперсию их компонентных значений. Вторая компонента формируется из остаточных показаний взаимосвязей (корреляций) и представляет собой линейную комбинацию наблюдаемых переменных, объясняющую максимум изменчивости, не связанной с первой компонентой. Далее аналогично выделяются следующие компоненты, также объясняющие максимум остаточной изменчивости и ортогональные для всех компонент, выделенных на предыдущих шагах.

Главные компоненты упорядочены, причём первая объясняет наибольшую долю дисперсии, а последняя наименьшую. Решение является единственным, и если сохранить все компоненты, то можно в точности воссоздать наблюдаемую

матрицу взаимосвязей исходных признаков.

Для лучшего понимания вопроса рассмотрим геометрическую интерпретацию метода главных компонент.

Представим себе простейший случай, когда имеются только две переменные x_1 и x_2 . Такие данные легко изобразить на плоскости (рисунок 18).

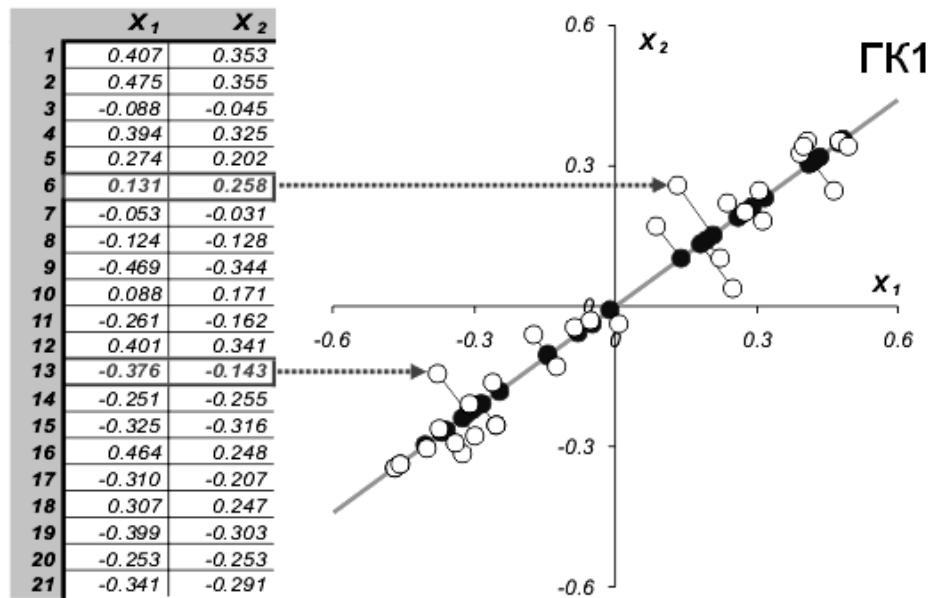


Рисунок 18- Графическая интерпретация главной компоненты двумерных данных

Здесь строкам исходной таблицы соответствуют точки на плоскости, обозначенные пустыми кружками. Проведем прямую тренда, чтобы вдоль нее происходило максимальное изменение данных, которая и будет называться осью первой главной компоненты или ГК1.

После определения ГК1 спроецируем на неё все исходные точки (закрашены). Предположим, что изначально все наши экспериментальные точки и должны были лежать на этой новой оси, но по неизвестным причинам отклонились от правильного положения, а мы вернули их на место. Тогда все отклонения от новой оси могут быть просто шумом, но в этом надо убедиться. Для этого найдём ось – ГК2, перпендикулярную и ГК1 объясняющую большую часть оставшейся выборочной дисперсии. В нашем случае, когда переменных только две, эллипсоид рассеяния плоский и потому две главные компоненты позволяют полностью описать изменчивость данных. Если переменных много, будем повторять эти дейст-

вия до тех пор, пока шум уже не станет действительно шумом, т.е. случайным хаотическим набором величин. При этом остаточное облако примет изометричную форму без выраженной оси симметрии.

В результате, мы переходим от большого количества переменных к новому представлению, размерность которого значительно меньше. Часто удается упростить данные на порядки и при этом ничего не исчезает - все переменные учитываются. В то же время несущественная для сути дела часть данных отделяется, превращается в шум. Найденные главные компоненты и дают нам искомые скрытые переменные, управляющие устройством данных.

Метод главных факторов (собственно факторный анализ)

Допустим, что в каждой пробе из геологического объекта мы измерили четыре характеристики, которые обусловлены действием двух факторов F_1 и F_2 . Фактор F_1 действует на все четыре характеристики объекта, а фактор F_2 действует лишь на два признака X_2 и X_3 (рисунок 19).

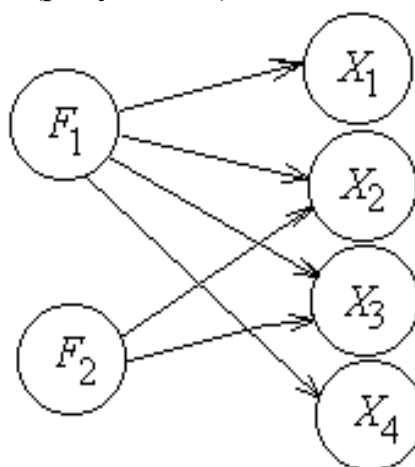


Рисунок 19- Принципиальная модель собственно факторного анализа

Таким образом, значения признаков X_1 и X_4 определяются только фактором F_1 , а признаки X_2 и X_3 определяются совокупным действием фактором F_1 и F_2 .

Это типичная модель собственно факторного анализа, который предназначен для изучения только общей дисперсии, присущей сразу нескольким наблюдаемым переменным. В неё не включаются дисперсия ошибок измерений и специфическая дисперсия конкретной переменной.

В факторном анализе, как и в методе главных компонент, анализируемая дисперсия есть сумма величин, стоящих на главной диагонали корреляционной матрицы. Однако, в методе главных компонент здесь расположены единицы и анализируется дисперсия, соответствующая количеству первичных переменных. Буквально вся дисперсия распределяется по главным компонентам, включая дисперсию ошибок измерения и дисперсию, специфическую для каждой наблюдаемой переменной. Поэтому, если для дальнейшего анализа сохраняются все компоненты, то они в точности воспроизводят наблюдаемую матрицу взаимосвязей.

В модели факторного анализа в отличие от модели главных компонент присутствует дополнительное слагаемое, в состав которого входит характерный фактор, влияющий только на некоторые признаки. На рисунке 19 ему соответствует фактор F_2 .

Такой подход основан на главном постулате факторного анализа, что эти дисперсии ухудшают общую картину изучаемого явления. Сама общая дисперсия оценивается общностями, занимающими главную диагональ матрицы взаимосвязей и принимающими значения в диапазоне от 0 до 1. Факторное решение выбирается на основе переменных с высокими общностями. В этой связи полного воспроизведения исходной матрицы взаимосвязей по главным факторам не произойдёт, поскольку часть изменчивости в них просто не учитывается!

1.10.3.2 Число выделяемых факторов

И при анализе ГК и при собственно факторном анализе необходимо решать- сколько факторов достаточно для описания природной изменчивости анализируемой выборки?

Наиболее часто с этой целью используются два достаточно условных критерия- критерий Кайзера и критерий каменистой осыпи.

Критерий Кайзера.

Этот критерий предложен Кайзером (Kaiser, 1960), и содержит одну очень

простую рекомендацию- сохранять только те факторы, собственные значения которых превышают единицу. По существу, это означает, что если фактор не выделяет дисперсию, эквивалентную, по крайней мере, дисперсии одной переменной, то он опускается..

Критерий каменистой осыпи.

Этот критерий является графическим методом, впервые предложенным Кэттелем (Cattell, 1966). Вы можете изобразить собственные значения, представленные в таблице ранее, в виде простого графика. Кэттель предложил найти на этом графике такое место, где убывание собственных значений слева направо максимально замедляется (рисунок 20). Предполагается, что справа от этой точки находится только "факториальная осыпь". Здесь слово "осыпь" применяется в геологическом смысле, обозначая буквально обломки горных пород, скапливающиеся в нижней части скалистого склона.

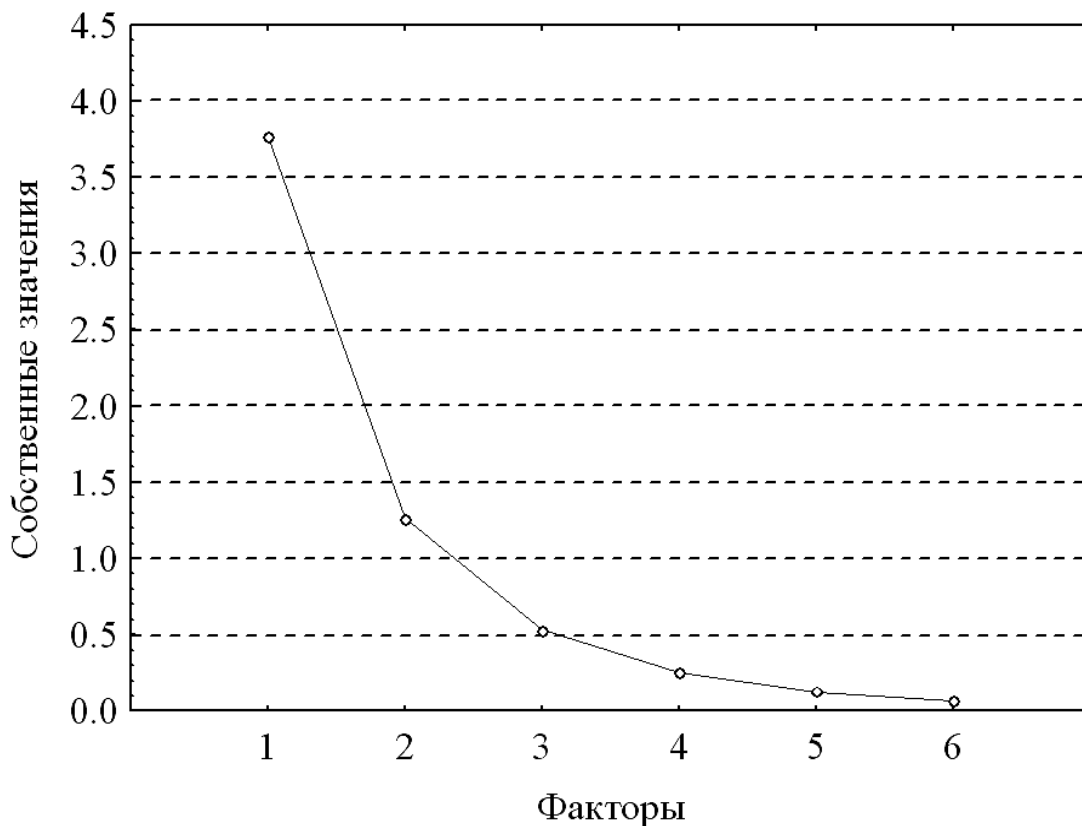


Рисунок 20- График собственных значений факторов

На вышеприведённом рисунке осыпь начинается с 4-го фактора.

1.10.4 Вращение факторов

Выделенные в результате факторизации факторы обычно трудно однозначно интерпретировать. Чтобы получить более осмысленное решение, рекомендуется использовать процедуру т.н. *вращения* факторов.

Подобно тому, как различные методы факторизации при хорошем наборе данных в большинстве случаев дают сходные результаты, различные методы вращения при верном определении структуры взаимосвязей, как правило, приводят к близким решениям.

При выборе вращения, прежде всего, следует определиться с его видом. Вращение может быть ортогональное или косоугольное. В первом случае факторы остаются ортогональными друг к другу, т.е. независимыми. Обычно ортогональное решение легче поддаётся описанию и интерпретации, но если независимость факторов не предполагается заранее, то такое вращение может исказить реальность.

Если же исследователь уверен, что базисные факторы взаимосвязаны друг с другом, он обязан использовать косоугольное вращение.

1.10.4.1 Ортогональное вращение

Наиболее часто используется вариант ортогонального вращения, который называется варимакс-вращение. Цель вращения варимакс- выбор наиболее простого факторного решения путём максимизации дисперсии факторных нагрузок по каждому фактору. При этом высокие факторные нагрузки ещё более повышаются, а низкие понижаются.

Кроме того, при варимакс-вращении доли дисперсий, объясняемые факторами, перераспределяются в сторону выравнивания.

1.10.4.2 Косоугольное вращение

Из-за того, что взаимосвязи между неортогональными факторами бесконечно вариативны, виды косоугольных вращений весьма многообразны. При косоугольном вращении взаимосвязь между факторами измеряется как коэффициент корреляции.

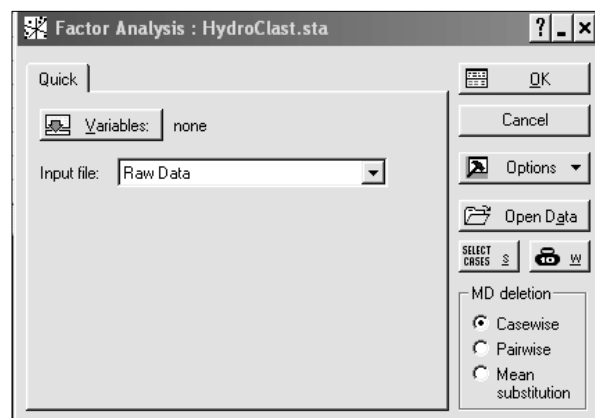
Главная задача вращения – получение т.н. «простой структуры». Она достигается в случае, если новые оси (факторы) проходят вблизи точек скопления анализируемых переменных на факторных диаграммах. После достижения простой структуры необходимо решить – какое количество факторов достаточно для достижения цели исследования?

Пример факторного решения

В качестве примера выполним факторное решение методом главных компонент применительно к уже известной нам гидрохимической выборке из 18 проб (рисунок 15). При этом будем использовать превосходный инструментальный пакет STATISTICA 6 [11]

Собственно анализируемая выборка представляет собой набор гидрохимических проб, отобранных с разных глубин из скважин, пробуренных в пределах нефтяного месторождения. Задачей анализа является выяснение главных процессов становления наблюдаемого в выборке солевого состава подземных вод. При этом предполагается, если не полное тождество, то тесная связь между этими процессами и главными компонентами, которые должны определиться в ходе факторного решения.

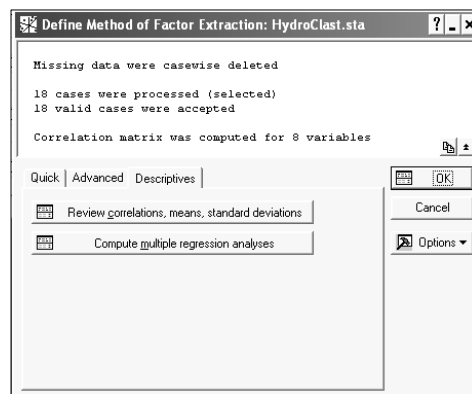
Для выбора метода анализа из главного меню пакета STATISTICA выполняем каскадный вызов пункта:



Статистика → Многомерные исследовательские методы → Анализ фактора.

В результате этих манипуляций откроется начальное диалоговое окно Factor Analysis.

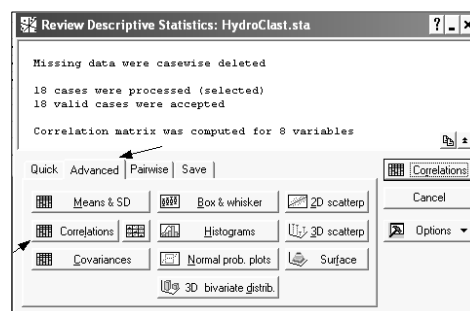
Нажмём кнопку <Variables>, позволяющую нам выбрать исходные переменные для проведения анализа, выберем их все (т.е. 8 переменных от УдВес до HSO3).



Нажав кнопку <ОК> мы перейдём к следующему диалогу, который определяет метод факторного анализа, а так же обзор разнообразных статистик выборки.

Выберем в нём вкладку *Descriptives*, а в ней кнопку <Review correlations, means, standart deviations>

Откроется новое диалоговое окно описательной статистики. Откроем в нём вкладку *Advanced* и нажмём кнопку <Correlations>, чтобы посмотреть корреляционную матрицу, построенную по нашим исходным данным (рисунок 21).



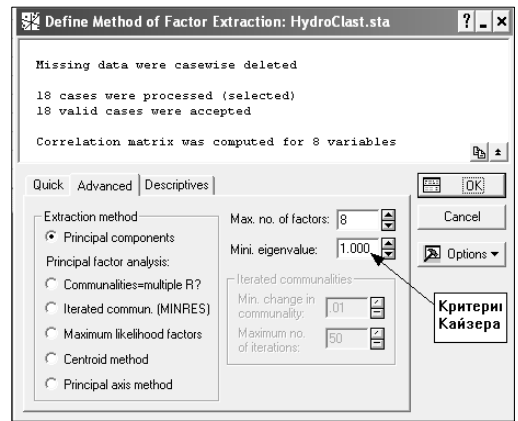
По диагонали корреляционной матрицы стоят единицы. Значения коэффициентов корреляций в ней имеют как положительные, так и отрицательные значения.

| Variable | УдВес | Минер | Na+K | Ca | Mg | Cl | SO4 | HSO3 |
|----------|-------|-------|-------|------|-------|-------|-------|-------|
| УдВес | 1.00 | 0.08 | 0.15 | 0.08 | -0.04 | -0.03 | -0.09 | 0.20 |
| Минер | 0.08 | 1.00 | -0.01 | 0.12 | 0.19 | 0.08 | -0.22 | -0.07 |
| Na+K | 0.15 | -0.01 | 1.00 | 0.87 | 0.88 | 0.92 | 0.75 | 0.95 |
| Ca | 0.08 | 0.12 | 0.87 | 1.00 | 0.87 | 0.92 | 0.44 | 0.74 |
| Mg | -0.04 | 0.19 | 0.88 | 0.87 | 1.00 | 0.97 | 0.66 | 0.77 |
| Cl | -0.03 | 0.08 | 0.92 | 0.92 | 0.97 | 1.00 | 0.71 | 0.78 |
| SO4 | -0.09 | -0.22 | 0.75 | 0.44 | 0.66 | 0.71 | 1.00 | 0.71 |
| HSO3 | 0.20 | -0.07 | 0.95 | 0.74 | 0.77 | 0.78 | 0.71 | 1.00 |

Рисунок 21- Корреляционная матрица гидрохимической выборки

Нажав кнопку <Cancel> мы снова перейдём к предыдущему диалогу *Define Method of Factor Extraction*, откроем вкладку *Advanced* и в блоке *Extraction method* активизируем радиокнопку *Principal components*.

В качестве первого приближения укажем максимально возможное число выделяемых компонент- 8, но применение критерия Кайзера должно несколько сократить это число.

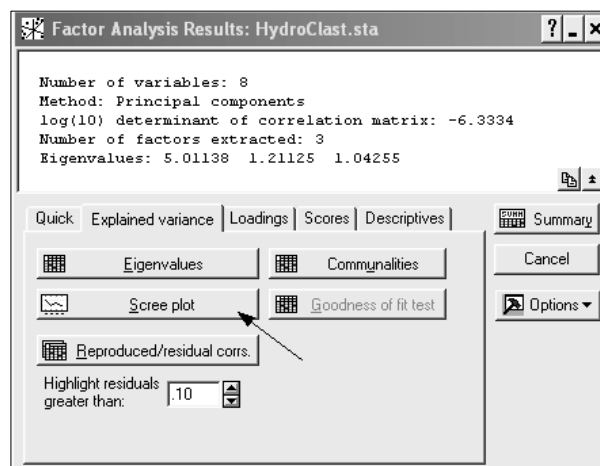


После нажатия кнопки <OK> возникает диалог результатов факторного анализа, из которого мы выберем вкладку *Explained variance* (объяснённая изменчивость) и нажмём кнопку <Eigenvalues>, что означает <собственные значения>.

| Value | Eigenvalue | % Total variance | Cumulative Eigenvalue | Cumulative % |
|-------|------------|------------------|-----------------------|--------------|
| 1 | 5.011385 | 62.64231 | 5.011385 | 62.64231 |
| 2 | 1.211247 | 15.14059 | 6.222632 | 77.78291 |
| 3 | 1.042547 | 13.03184 | 7.265180 | 90.81475 |

Как и ожидалось, критерий Кайзера сократил число определяемых факторов до 3-х. Все они характеризуются собственными числами (первый столбец матрицы), значение которых больше единицы. Хотелось бы проверить это число с помощью дополнительного критерия «каменистой осыпи».

Для этого в том же диалоговом окне нажмём кнопку <Scree plot> .



После этих действий на экране появится график «Каменистой осыпи» (рисунок 22).

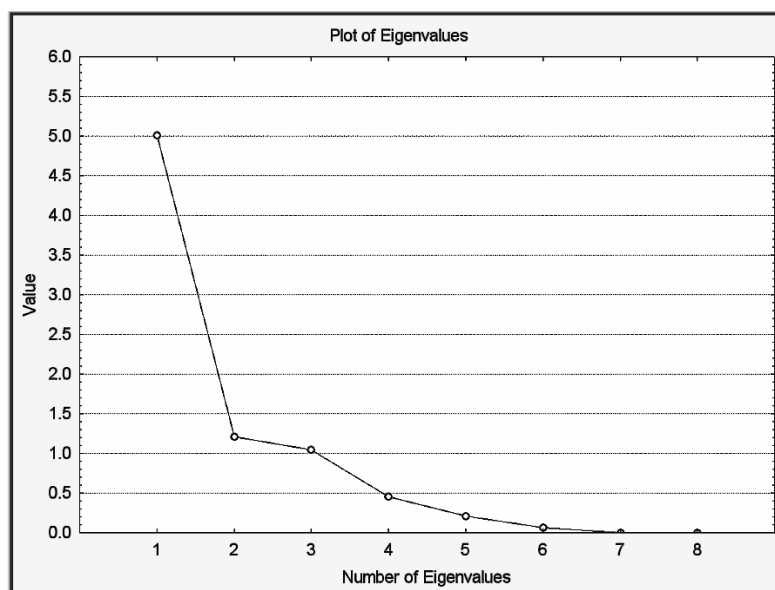
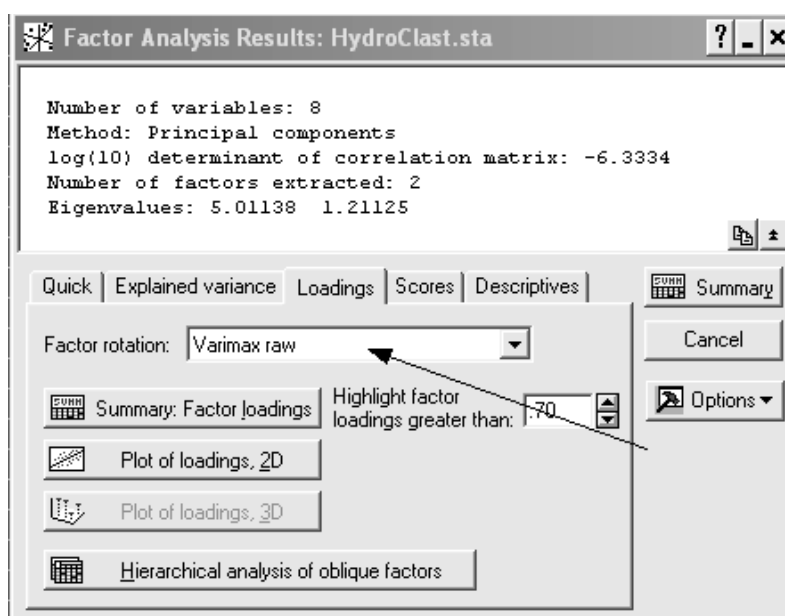


Рисунок 22- График критерия «Каменистая осыпь»

Хорошо заметно, что первый перегиб графика связан с фактором №2, т.е. для описания изменчивости выборки достаточно только двух главных компонент.

С учётом полученных сведений можно повторить анализ, задав в качестве предельного числа выделяемых компонент только две, причём для лучшей их интерпретации зададим их вращение типа *varimax*.

Для этого в диалоговое окно результатов факторного анализа заполним по образцу.



После выполнения вращения рассмотрим график исходных переменных в системе координат 2-х главных компонент (рисунок 23)

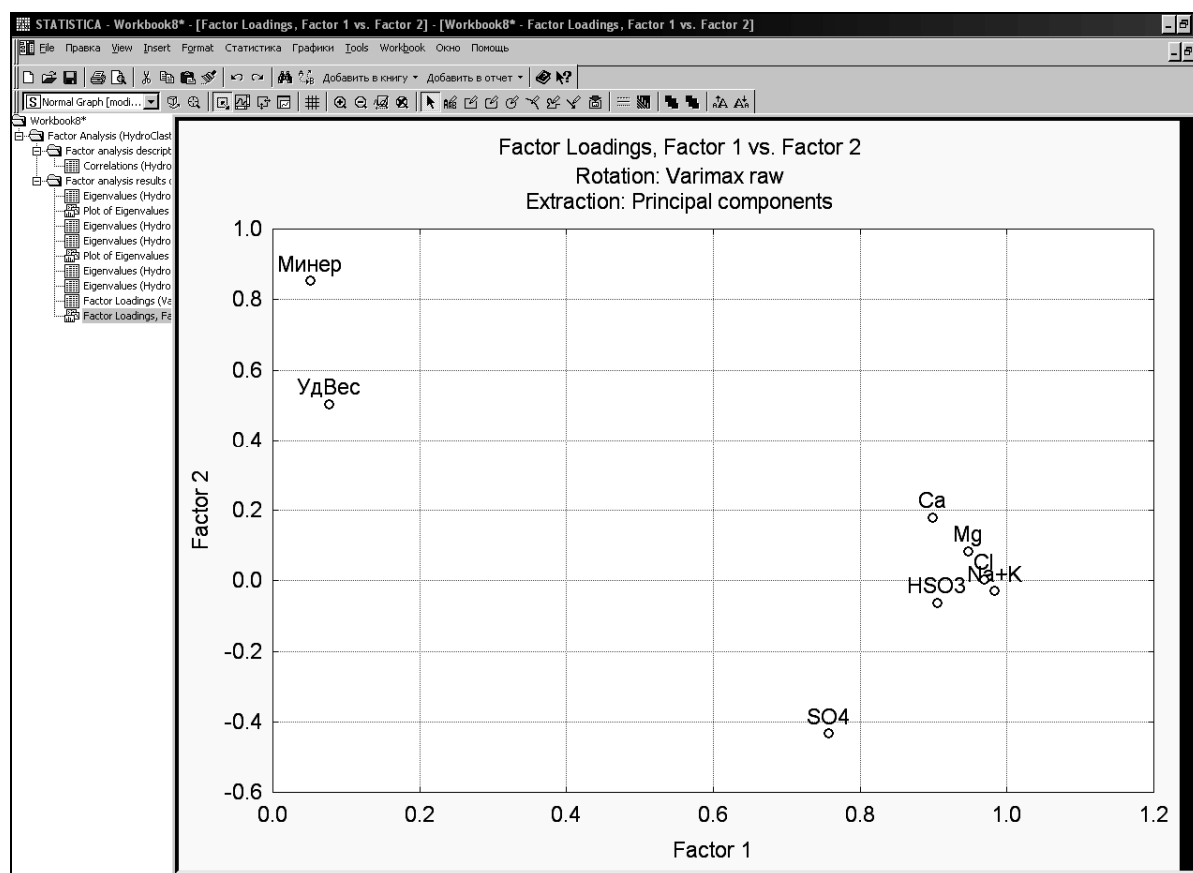


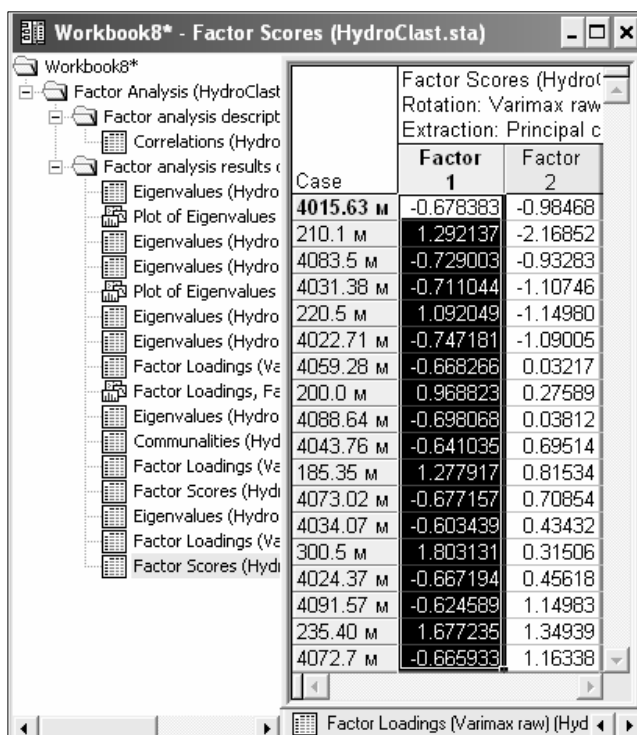
Рисунок 23- Диаграмма «простой» факторной структуры матрицы вариаций

На рисунке 23 хорошо заметно, что с фактором 2 сильнее всего коррелируют две переменные- Минер и УдВес. По сути дела это близкие величины, отличающиеся разницей между сухим остатком и растворёнными летучими соединениями. Эта связь имеет «сквозной» характер, пожалуй, для любых текучих вод, и потому его можно проинтерпретировать как фактор растворимости. Их факторные нагрузки по первому фактору близки к нулю, что говорит о независимости (ортогональности) главных компонент.

Остальные переменные уверенно связаны с первой главной компонентой. Их объединяет явная связь с надсолевым водоносным комплексом, химизм вод которого сформировался во многом за счёт контакта с галогенными отложениями кунгурского яруса, экранирующими газонефтяные залежи.

Свидетельством тому могут служить значения главных компонент, расчи-

танные для каждой из проб (рисунок 24).



The screenshot shows a SPSS window titled "Workbook8* - Factor Scores (HydroClast.sta)". The window contains a table of factor scores. The table has three columns: "Case", "Factor 1", and "Factor 2". The "Case" column lists 15 cases with their respective values. The "Factor 1" and "Factor 2" columns show the scores for each case. The scores for Factor 1 range from approximately -0.67 to 1.80, and for Factor 2 from approximately -1.16 to 1.16.

| Case | Factor 1 | Factor 2 |
|-----------|-----------|----------|
| 4015.63 M | -0.678383 | -0.98468 |
| 210.1 M | 1.292137 | -2.16852 |
| 4083.5 M | -0.729003 | -0.93283 |
| 4031.38 M | -0.711044 | -1.10746 |
| 220.5 M | 1.092049 | -1.14980 |
| 4022.71 M | -0.747181 | -1.09005 |
| 4059.28 M | -0.668266 | 0.03217 |
| 200.0 M | 0.968823 | 0.27589 |
| 4088.64 M | -0.698068 | 0.03812 |
| 4043.76 M | -0.641035 | 0.69514 |
| 185.35 M | 1.277917 | 0.81534 |
| 4073.02 M | -0.677157 | 0.70854 |
| 4034.07 M | -0.603439 | 0.43432 |
| 300.5 M | 1.803131 | 0.31506 |
| 4024.37 M | -0.667194 | 0.45618 |
| 4091.57 M | -0.624589 | 1.14983 |
| 235.40 M | 1.677235 | 1.34939 |
| 4072.7 M | -0.665933 | 1.16338 |

Рисунок 24- Значения главных компонент в пробах гидрохимической выборки

Как видим, максимальные величины первой компоненты тяготеют к относительно неглубоким замерам, что подтверждает правомочность наших выводов. В этой связи первую главную компоненту можно назвать *надсолевой*.

Такова в общих чертах непростая процедура факторного анализа, который относится к числу самых мощных статистических методов. Вместе с тем его результаты во многом зависят как от качества исходного статистического материала, так и от уровня профессиональной подготовки самого исследователя.

Вопросы для самопроверки:

- 1 Для решения каких задач целесообразно применение факторного анализа?
- 2 В чём состоит основное различие метода главных компонент от метода собственно факторного анализа?
- 3 Приведите пример матрицы взаимосвязей переменных.

- 4 Объясните смысл термина *общность*.
- 5 Какие способы определения оптимального числа выделяемых факторов Вам известны?
- 6 Для чего применяется процедура вращения факторов?
- 7 Какие типы вращения факторов Вам известны?

2 Статистический анализ геологических данных

В основе методологий исследования числовых полей геологической природы лежит понятие аппроксимации данных. Здесь под аппроксимацией подразумевается описание некоторой, порой заданной неявно, зависимости или множества представляющих её данных с помощью другой более простой или единообразной зависимости. В геологических предметных областях потребность в аппроксимации фактического материала особенно высока. Это связано с тем, что исследуемые геологические объекты и процессы, редко доступны для непосредственного чувственного восприятия. Их свойства диагностируются по набору характеристик, значения которых зависят от множества случайных факторов, маскирующих полезный сигнал. Именно его выявление является целью практически всех преобразований исходных геологических сведений, предназначенных для анализа и прогноза состояния недр, воплощаемых в особых информационных структурах: горно-геометрических моделях (ГГМ).

Горно-геометрической модель представляет собой упрощённое представление исследователя о характере геологических полей в изучаемом блоке земной коры.

Горно-геометрическая модель может быть задана с помощью некоторого аналитического выражения, числового массива или в графическом виде.

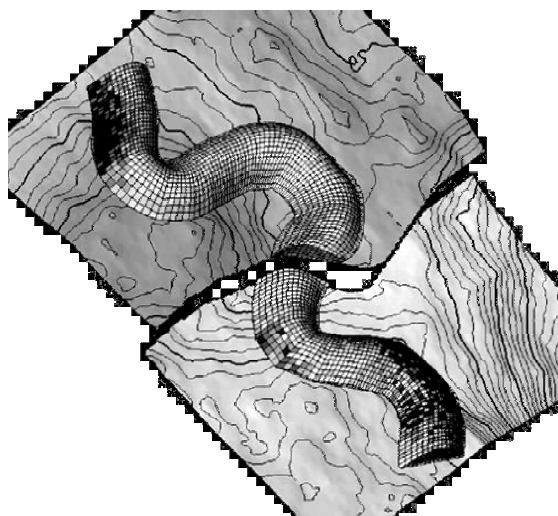


Рисунок 25- Горно-геометрическая модель «шнурковой» залежи нефти

В основе построения большинства ГГМ лежит идея распространения фактических данных, характеризующих отдельные точки наблюдений на их окрестности. В методическом плане данный процесс выполняется в рамках двух не сильно различающихся между собой подходов- интерполяции и экстраполяции данных.

Распространение по определённым правилам значений на участки между точками наблюдений называется интерполяцией.

Распространение данных за пределы крайних точек наблюдений называется экстраполяцией или прогнозом.

Поскольку определить крайние точки наблюдений в пределах территории исследований очень непросто, часто весь процесс называется экстраинтерполяцией. Она является основой т.н. *восстановления* геологического поля в местах неохарактеризованных фактическим материалом.

2.1 Восстановление геологического поля

Для лучшего понимания вопроса восстановления геологических полей полезно провести параллель с методологиями, многократно описанными в произведениях детективного жанра. Как следователь по отдельным уликам восстанавливает картину преступления, так и геолог, по крайне недостаточным и неполным фактическим данным должен воссоздать целостное представление о цифровых полях изучаемых геологических признаков.

Существуют два вида восстановления геологических полей- восстановление без сглаживания исходных данных и восстановление полей со сглаживанием [6].

В первом случае воссоздаются горно-геометрические модели, обеспечивающие неизменность значений числового поля признака в исходных точках после экстраинтерполяции, а во втором – модели, где эти значения несколько искажены.

Надо заметить, что восстановление без сглаживания представляет собой

весьма трудоёмкую процедуру и, несмотря на академическую строгость, часто нет смысла. Это особенно ясно, когда сам интерполируемый признак измеряется с некоторой ошибкой, которая вполне может превышать разницу сглаживания. Есть серьёзные теоретические основания полагать, что модели, построенные с методически верными параметрами сглаживания предпочтительнее в смысле точности, чем точные модели без сглаживания. Это очень интересно- оказывается, чтобы характеристики модели были точнее в пространстве между точками наблюдений, необходимо снизить её точность в самих исходных точках.

Поверхности, ограничивающие горно-геометрические модели нам заранее неизвестны. Часто они вычисляются программным образом в ходе специальных исследований, связанных с *тренд-анализом*.

Математический аппарат тренд- анализа весьма разнообразен и сложен, поскольку допускает множество решений и полученные результаты порой носят спорный характер. Выбор конкретного метода во многом субъективен и зависит от опыта и мнения геолога.

В основе тренд-анализа лежит предположение, что любое из наблюдаемых значений z может быть представлено в виде суммы двух компонент, одна из которых F рассматривается как неслучайная функция от координат, а другая (φ) случайная.

Для одномерного (профильного) случая функция тренда определяется выражением $Z(x)=F(x)+(\varphi)$.

Для двумерного (площадного) варианта приходится учитывать ещё одну координату, в связи с чем формула принимает вид $Z(x,y)=F(x,y)+\varphi(x,y)$.

В зависимости от смысла решаемой геологической задачи внимание исследователя обычно сосредоточено на одном из двух вопросов:

- выявление региональной составляющей (тенденции или тренда) в изменении признака z ;
- обособлении локальной составляющей (поиск положительных и отрицательных аномалий).

2.1.1 Выделение региональной составляющей

Представим, что мы летим в самолёте над тайгой и делаем аэрофотосъёмку. В тайге нет двух в точности одинаковых по высоте деревьев, а нам необходимо построить топографическую карту земной поверхности, на которой они произрастают. Деревья очень мешают и многие студенты уже отчаялись, мечтая о посадке, поскольку решили, что всё равно у нас ничего не выйдет. Однако, выход есть, и он состоит в тренд-анализе верхней кромки леса.

В общих чертах такая же проблема стоит перед всеми геологами. Необходимо отвлечься от частных, случайных колебаний измеряемого признака и выделить только его закономерную составляющую- тренд. Для этого разработано много методов, основанных на сглаживании наблюдаемых значений.

Перечислим основные из них:

- сглаживание числовых полей методами скользящего среднего;
- аппроксимация числовых полей алгебраическими полиномами (Фурье-анализ);
- аппроксимация числовых полей гармониками;
- аппроксимация числовых полей сплайнами.

2.1.1.1 Методы скользящего среднего

Методы скользящего окна употребляются наиболее часто, поскольку не требуют изощрённого программного обеспечения и относительно просты в понимании. В их основе лежит следующая общая процедура или алгоритм. Для первых m членов (m -нечётно) сглаживаемого ряда объёмом n наблюдений ($m < n$) строится полином степени P ($P \leq m-1$), после чего вычисляется его значение для точки $k=(m-1)/2$. Затем вновь берётся m членов, начиная со второго, и все расчёты выполняются заново. Другими словами на каждом шаге происходит сдвиг на одно наблюдение. В простейшем случае, когда $P=1$ сглаживание выполняется простым

усреднением значений z :

$$Z_k = \frac{1}{m} \sum_{t=k-(m-1)/2}^{t=k+(m-1)/2} cZ_t,$$

где m - число точек сглаживания (окно);

Z_t - исходное значение признака;

Z_k - вычисленное значение признака;

C – весовой коэффициент. При простом осреднении он равен единице.

Для двумерного случая указанное выражение превращается в:

$$\bar{z}(x_i, y_i) = \alpha \sum_1^m cz(x, y),$$

где x и y координаты исходных точек наблюдения;

α - некий масштабный множитель.

На рисунке 26 представлены два графика одних и тех же учебных данных, представляющих собой классическую синусоиду осложнённую случайным шумом, полученным с помощью генератора случайных чисел.

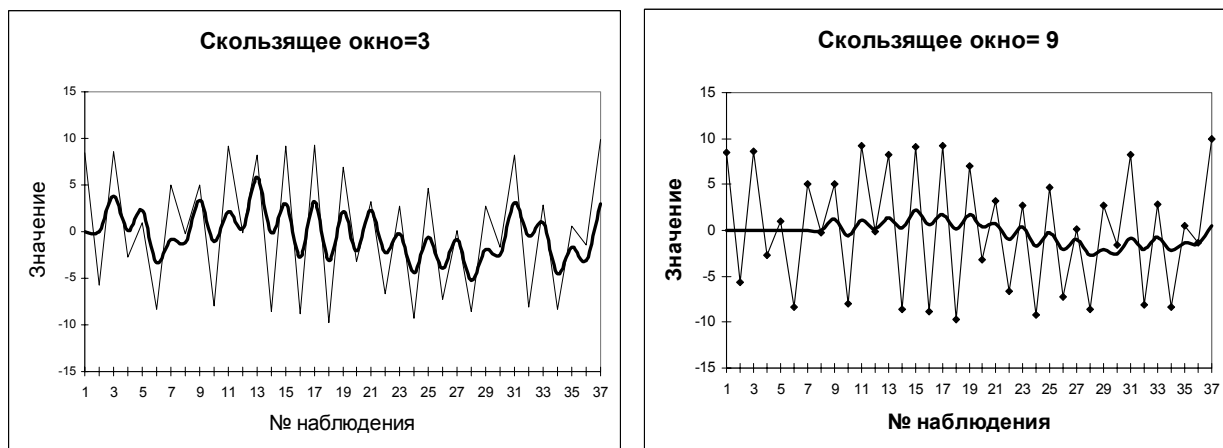
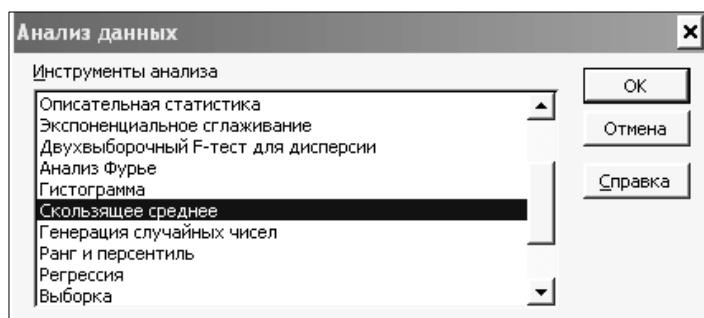


Рисунок 26- Сглаживание данных методом скользящего среднего.

Для построения соответствующих диаграмм использован Инструмент «Скользящее среднее» модуля «Анализ данных» электронной



таблицы MS Excel.

Произведено усреднение с окнами размером 3 и 9 наблюдений соответственно. Хорошо заметно, что при малых размерах окна сглаженные кривые практически повторяют ход изменения данных, а при больших- проявляется синусоидальный тренд, правда, осложнённый «шумовыми» гармониками.

Кроме того, в левой части графиков заметно отсутствие сглаженных значений, что связано с тем, что часть окна осреднения выходит за поле данных.

Аппроксимация скользящим средним, подавляя высокочастотную компоненту, сохраняет в то же время общую конфигурацию крупных пиков, соответствующих региональной составляющей.

Среди многочисленных модификаций метода скользящего окна, используемых в геологии выделяется метод «ближайших точек». Он иллюстрирует работу алгоритма восстановления поля без сглаживания и имеет следующие особенности:

- размер и форма окна заранее не устанавливаются;
- исходные точки могут располагаться на карте неравномерно;
- число «ближайших точек», участвующих в сглаживании постоянно.

Чтобы построить тренд методом «ближайших точек», вся картируемая территория покрывается регулярной прямоугольной сеткой. Далее вычисляются сглаженные значения z , соответствующие узлам этой сетки. Для этого необходим такой алгоритм:

Шаг 1. Отыскиваются m точек, ближайших очередному узлу (x_i, y_i) нашей вспомогательной сети.

Шаг 2. Рассчитываются расстояния D между узлом сети и каждой из точек: $D = \sqrt{(x_i - x)^2 + (y_i - y)^2}$

Шаг 3. Вычисляется нормирующий множитель $\alpha_{ij} = \frac{1}{\sum_{j=1}^m \frac{1}{D_j}}$

Шаг 4 Производится сглаживание согласно общему выражению двумерного осреднения:

$$\mathcal{E}(x_i, y_i) = \frac{1}{\sum_{j=1}^m \frac{1}{D_j}} \sum_{j=1}^m \frac{z(x_i, y_i)}{D_j}$$

Таким образом, исходные точки учитываются с весами обратно пропорциональными их расстояниям до узла координатной сетки.

2.1.1.2 Аппроксимация алгебраическими полиномами

Прежде всего, вспомним, что такое полином. Полиномом кой степени называется выражение: $A(x) = a * x^0 + b * x^1 + c * x^2 + d * x^3 + \dots + h * x^k$

Смысл полиномиальной аппроксимации состоит в подборе подходящей степени полинома для сглаживания наблюдаемых значений геологического признака.

Полиномиальная аппроксимация связана с решением двух серьёзных проблем. Во- первых, отсутствуют чёткие критерии выбора степени полинома, а во- вторых- это достаточно трудоёмкая вычислительная задача. В общем случае, чем степень полинома выше, тем более его график приближается к исходным данным, т.е. тем меньше они сглаживаются. Это хорошо видно на рисунке 27 для полиномов 1-ой, 2-ой и 5-ой степеней.

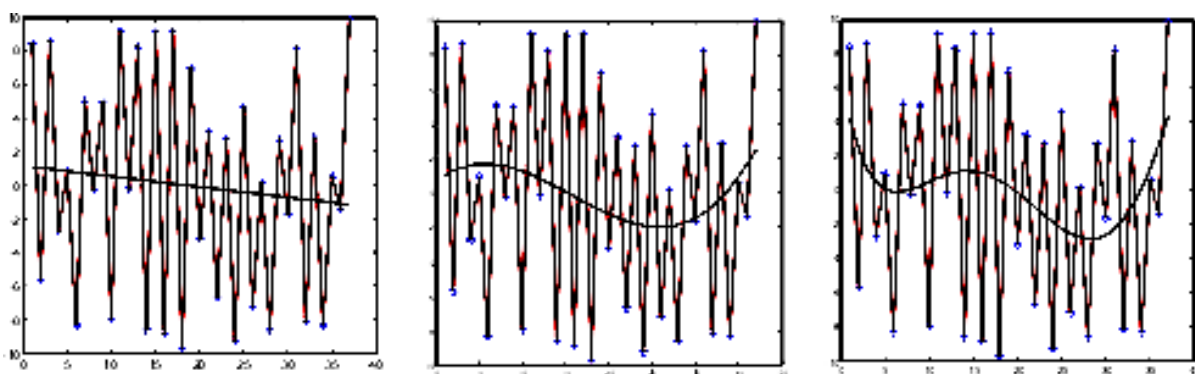


Рисунок 27- Графики аппроксимации данных полиномами разных степеней

Для вывода соответствующих графиков использовалось программная система Matlab 6.5 [12]. Ей на вход подавались в точности те же данные, что и при сглаживании методом скользящего среднего, но разница в результатах очень заметна.

Рабочая панель Matlab 6.5 для полинома степени 3 показана на рисунке 28

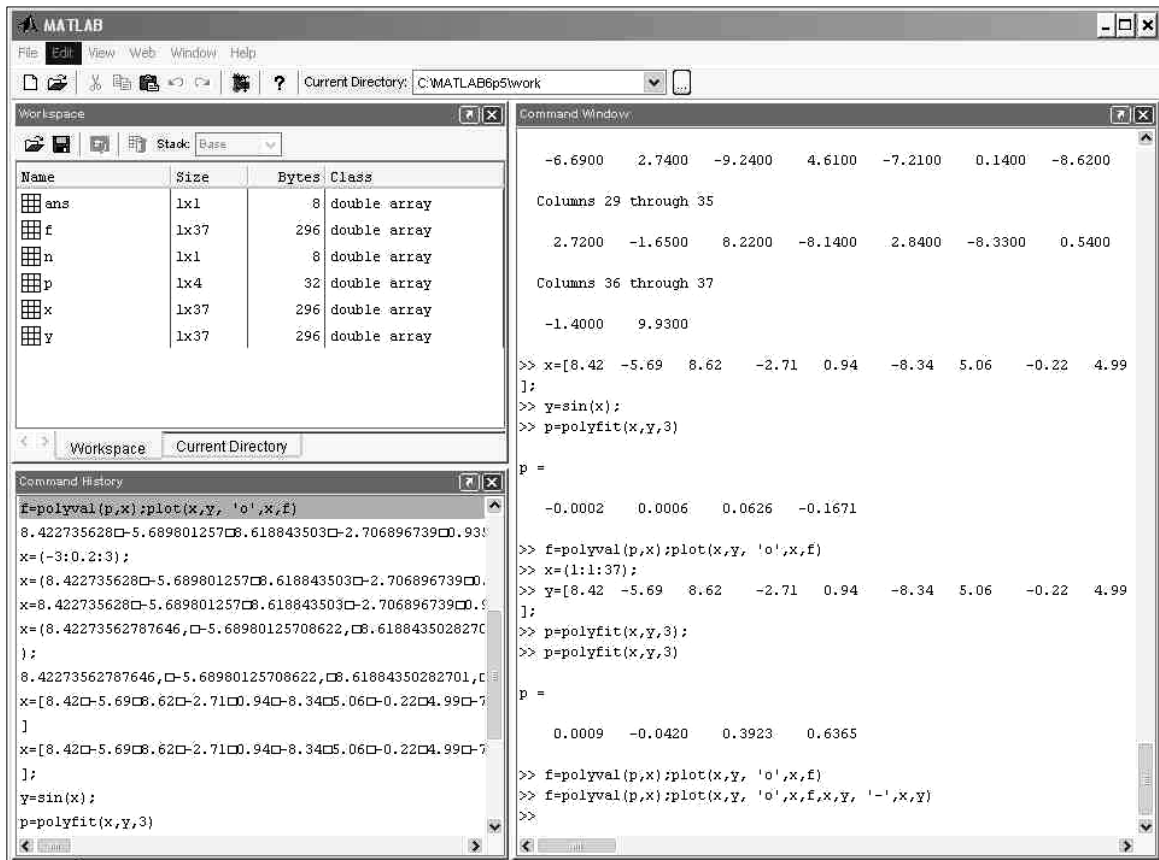


Рисунок 28- Панель Matlab 6.5 для расчёта полинома 3-ей степени

Таким образом, определение оптимальной для конкретной аппроксимации степени полинома требует привлечения дополнительных данных о геолого-геометрической модели, а она, как правило, отсутствует. В любом случае мы должны решить- какой минимальный пространственный размер возмущений геологического поля нас интересует. Только на этом основании можно подобрать полином подходящей степени, не осложнённый случайными флуктуациями.

В целом использовать полиномиальную аппроксимацию надо осторожно в виду следующих обстоятельств:

- с повышением степени полинома на диаграммах тренда начинают преобладать эффекты, связанные с действием локальных возмущений числовых полей;
- по краям полиномиальных карт и графиков возможно появление неконтролируемых отклонений аппроксимирующей поверхности от фактических данных. Это так называемые «краевые эффекты» выражающиеся в появлении

необоснованно высоких или, напротив, низких значений сглаженного признака в областях отсутствия данных.

2.1.1.3 Аппроксимация гармониками

В тех случаях, когда предполагается периодический характер изменения геологического признака в пространстве, может использоваться аппроксимация гармониками с помощью аппарата ФУРЬЕ-анализа. Такую цикличность можно ожидать в ситуациях, характеризующихся более или менее регулярной повторяемостью в пределах изучаемого блока земной коры некоторых геологических условий, что в свою очередь обеспечивает периодичность изменения величины геологических признаков.

Наиболее часто или даже в обязательном порядке преобразования Фурье применяются во всех геологических приложениях, требующих интерпретации любых волновых картин. Самой яркой иллюстрацией этого служат, конечно, современный граф обработки сейсморазведочных данных.

Математический смысл преобразования Фурье состоит в возможности представления любого сигнала $y(x)$ в виде бесконечной суммы синусоид вида $F(v) * \sin(v * x)$. Функция $F(v)$ называется преобразованием Фурье, интегралом Фурье или Фурье-спектром сигнала. Её аргумент v - представляет собой частоту этой составляющей сигнала.

В своё время эта гениальная идея Фурье категорически отвергалась такими знаменитыми математиками как Эйлер, Лагранж и Даламбер. Великие учёные ошибались, но даже самые горячие поклонники таланта Фурье не могли предположить, какое широкое применение получит его работа в результате изобретения вычислительных машин.

Фурье-спектр представляет собой синусоидальный тренд, отражая разнопорядковые гармоники, образовавшиеся при сложении синусоидальных сигналов.

Рассмотрим рисунок 29, на котором представлены результаты Фурье-преобразований данных, полученных искусственным путём.

На верхнем графике отображена исходная кривая, возникшая при сложении трёх синусоид, согласно приведённым на рисунке выражениям. Задача состоит в восстановлении характеристик этих элементарных синусоид из действительно сложного сигнала. Учебный характер данных позволяет однозначно оценить результаты решения путём его сопоставления с исходными слагаемыми (на рисунке подчёркнуты). С этой целью средствами пакета MATHCAD 2001 получен график Фурье-спектра или прямого преобразования [13].

```

N := 128
xMAX := 100
i := 0..N - 1
Δ := xMAX / N
xi := i · Δ
yi := sin(2 · π · 0.05 · xi) + 0.5 · sin(2 · π · 0.1 · xi) + 0.25 · sin(2 · π · 0.4 · xi)
F := fft(y)
Ωi := (i + 1) ·  $\frac{1}{xMAX}$ 

```

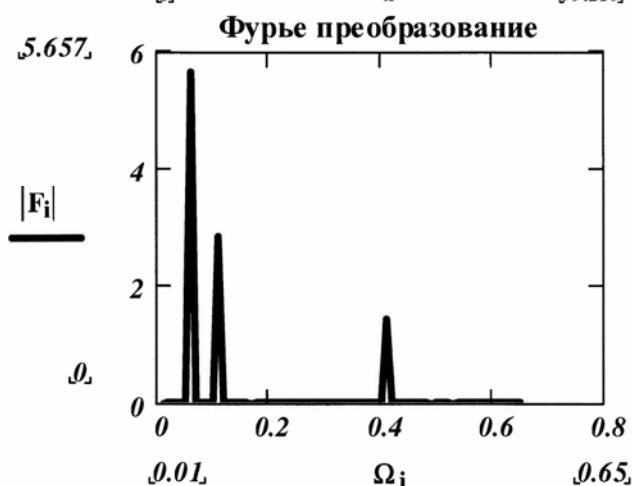
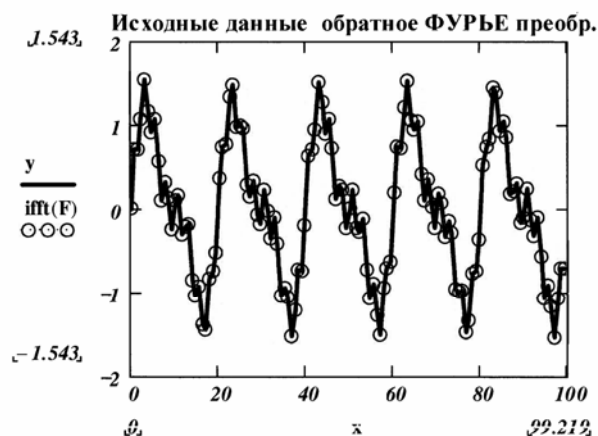


Рисунок 29- Иллюстрация Фурье- преобразований

Обратите внимание на три особенности положение пиков графика Фурье-

преобразования:

- на графике Фурье-спектра присутствует 3 пика- столько же сколько складывается синусоид;
- x - координата (частота) каждого любого равна коэффициенту при x_j в аргументе соответствующей синусоиды, а именно: **0.05**, **0.1** и **0.4**;
- высота каждого пика равна половине предыдущего, что объясняется коэффициентами при синусоидах, а именно: **1**, **0.5** и **0.25**.

Первый пик объясняет гармонику (или тренд) самого низкого порядка. Второй- осложняет первую гармонику, а третий пик- вторую.

Зная Фурье-спектр, мы можем восстановить исходные значения, оперируя известной вам формулой Фурье $F(v) * \sin(v * x)$. Это называется обратным преобразованием Фурье. Рассчитанные восстановленные точки очень точно легли на график.

2.1.1.4 Аппроксимация сплайнами

Термин «сплайны» ввели в обращение средневековые кораблестроители. Так назывались длинные стальные линейки, которыми они размечали корпуса будущих каравелл и бригантин. Если эти линейки не слишком изогнуты, то полученные с их помощью кривые являются самыми гладкими из всех возможных лекал. Именно это обстоятельство стало особо привлекательным для изучения явлений гладкости математическими методами. В результате была разработана теория сплайнов, широко эксплуатируемая в машино- и авиастроении, словом везде, где требуются особо гладкие обводы корпусов.

Свойство сглаживать «острые углы» очень полезно и в геологии. Они позволяют устранить ряд недостатков, свойственных полиномиальной аппроксимации (например, снизить искажения, вносимые краевыми эффектами).

Сплайн-интерполяция относится к методам кусочно-полиномиальной

аппроксимации.

В понятийном плане применение сплайнов в геологических задачах площадного картирования, например, описывается следующей процедурой:

- вся картируемая территория разбивается на относительно небольшие, непересекающиеся участки треугольной формы или прямоугольники (локусы);
- в вершинах локусов должны находиться точки наблюдений;
- для каждого локуса в отдельности строится свой полином. При этом надо так подобрать коэффициенты полиномов, чтобы на стыках смежных локусов не было разрывов, т.е. общее полотно геологического поля должно быть непрерывным и гладким.

При таком подходе отсутствие разрывов гарантируется общими смежными точками наблюдений, в которых, разумеется, значение признака должно быть одинаковым. Гладкость аппроксимирующей кривой обеспечивается непрерывностью первых и вторых производных в точках склеивания её фрагментов.

На рисунке 30 представлен профиль, построенный с помощью сплайн-функции, программно реализованной в пакете MatLab 6.5. Значения в исходных точках наблюдений генерировались согласно выражению $y=3*\cos(x)$.

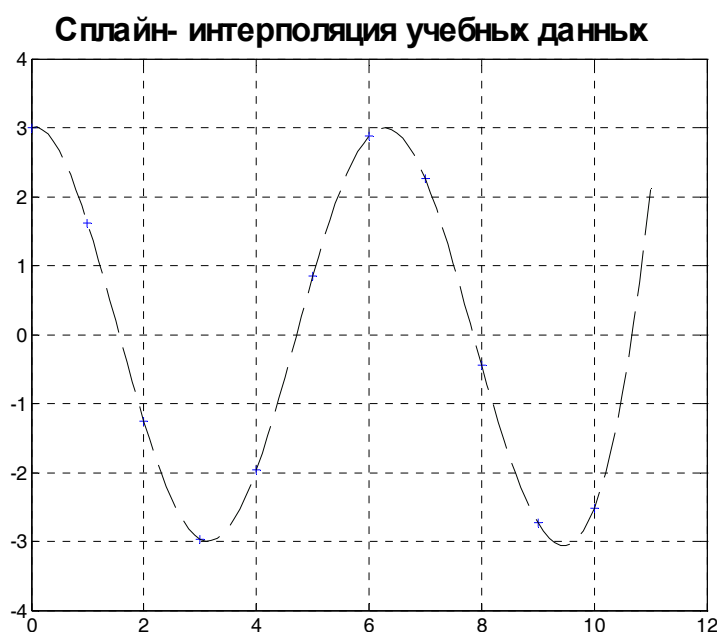


Рисунок 30- Сплайн-интерполяция учебных данных значений функции: $y=3*\cos(x)$

На рисунке хорошо заметно, что произошло не сглаживание, а именно интерполяция геологического признака между точками наблюдений, через которые абсолютно точно прошла сплайн-кривая. Другими словами, сплайн аппроксимация является методом восстановления поля *без сглаживания* (т.е. без осреднения) и для расчёта тренда в чистом виде не подходит. Эта задача решается комплексированием нескольких методов, например, скользящего окна и сплайн-функций.

2.1.2 Обособление локальной составляющей

Часто, особенно при работах повышенной детальности, необходимо отвлечься от низкочастотных составляющих геологического поля. Иногда геолога интересуют только его локальные вариации. Например, нам надо выяснить подчиняется ли их статистическое распределение нормальному закону, т.е. случайны ли они? Если да, то их в расчёт можно не принимать, а если нет- то некоторые из них могут иметь поисковое значение.

Комплексируя методы тренд-анализа с формально-логическими приёмами можно добиться исключения трендовой составляющей из геологического поля. После этой процедуры осложняющие тренд вариации будут представлены в чистом виде.

Для решения данной задачи необходимо постулировать вид трендового поля, которое может быть аппроксимировано функциями различного типа и сложности. Если он известен априори, то достаточно выполнить следующие шаги:

Шаг 1. Исходя из априорной информации о тренде рассчитать уравнение регрессии.

Шаг 2. Вычесть в каждом наблюдении трендовую составляющую. Эти операции иллюстрируются рисунком 31.

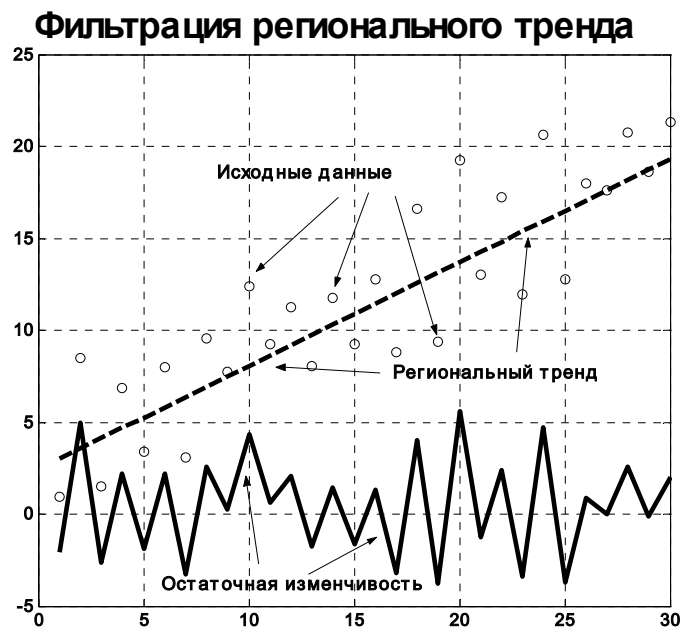


Рисунок 31- Определение локальной изменчивости вычитанием регионального тренда

Вопросы для самопроверки:

- 1 В чём состоит отличие восстановления числовых полей без сглаживания и восстановление со сглаживанием?
- 2 К какому типу восстановления числовых полей относится сплайн-аппроксимация?
- 3 Чем отличается интерполяция от экстраполяции?
- 4 Каким образом размер скользящего окна влияет на подавление локальной составляющей числового поля?
- 5 Какие недостатки аппроксимации числовых полей полиномами Вам известны?
- 6 Что происходит с возрастанием степени полинома?
- 7 В каких случаях целесообразно применение аппарата Фурье-анализа для аппроксимации числовых полей?
- 8 Каким образом можно восстановить исходные значения числового поля по Фурье-спектру и для чего это нужно?
- 9 В чём состоит смысл сплайн-аппроксимации?
- 10 Каким образом можно выделить локальную составляющую числового поля?

Список использованных источников

- 1 **Вентцель, Е. С.** Теория вероятностей : учеб. для вузов / Е. С. Вентцель.- 7-е изд., стер. - М. : Высш. шк., 2001. - 575 с. - ISBN 5-06-003650-2
- 2 **Каждан, А. Б.** Методологические основы разведки полезных ископаемых /А. Б. Каждан-М. :Недра, 1974. - 272 с.
- 3 **Вуколов, Э.А.** Основы статистического анализа. Практикум по статистическим методам и исследованию операций с использованием пакетов STATISTICA и EXCEL/ Э. А. Вуколов- М. : Форум, 2008. -464 с.
- 4 **Гусейнзаде, М. А.** Методы математической статистики в нефтяной и газовой промышленности / М. А. Гусейнзаде, Э. В. Калинина, М. В. Добкина. - М. : Недра, 1979. - 401 с.
- 5 **Каждан, А. Б.** Математические методы в геологии: учебник для вузов / А. Б. Каждан, О. И. Гуськов. -М. : Недра, 1990. -251 с. – ISBN 5-247-00857-X
- 6 **Родионов, Д. А.** Статистические решения в геологии / Д. А. Родионов.- М. :Недра, 1981. -232 с.
- 7 **Бугаец, А. Н.** Математические методы при прогнозировании месторождений полезных ископаемых / А. Н. Бугаец, Л. Н. Дуденко.- Л. : Недра, 1976. - 270 с.
- 8 **Родионов, Д. А.** Статистические методы разграничения геологических объектов по комплексу признаков / Д. А. Родионов; М. : Недра, 1968. - 158 с.
- 9 **Дуброва, Т. А.** Кластерный анализ с использованием ППП "SPSS": учеб. пособие для вузов / Т. А. Дуброва, М. Ю. Архипова, П. М. Стрелкова. - М. : МЭСИ, 2001. – 54 с.
- 10 **Белонин, М. Д.** Факторный анализ в геологии / М. Д. Белонин, В. А. Голубева, Г. Т. Скублов. - М. : Недра, 1982. - 269 с.
- 11 **Салин, В.Н.** Практикум по курсу 'Статистика' в системе STATISTICA: Учеб. пособие для студентов / В.Н. Салин, Э.Ю. Чурилова. - М. : Социальные отношения: Перспектива, 2002. - 188 с
- 12 **Дьяконов, В.П.** MATLAB 6/6.1/6.5 + Simulink 4/5. Основы применения

[Текст] : полное руководство пользователя / В. П. Дьяконов . - М. : СОЛОН-Пресс, 2004. - 768 с. - ISBN 5-98003-007-7

13 **Берков, Н. А.** Применение пакета MathCad: практикум / Н. А. Берков, Н. Н. Елисеев . - Москва : МГИУ, 2006. - 132 с. - ISBN 5-276-00960-0.