СТАТИСТИЧЕСКИЕ МЕТОДЫ В ГЕОЛОГОРАЗВЕДОЧНОЙ ПРАКТИКЕ

Методические указания к лабораторным работам для студентов специальности 21.05.02

САНКТ-ПЕТЕРБУРГ
2020

Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное
учреждение высшего образования
Санкт-Петербургский горный университет

Кафедра геологии и разведки
месторождений полезных ископаемых

СТАТИСТИЧЕСКИЕ МЕТОДЫ В ГЕОЛОГОРАЗВЕДОЧНОЙ ПРАКТИКЕ

Методические указания к лабораторным работам
для студентов специальности 21.05.02

САНКТ-ПЕТЕРБУРГ
2020

УДК 55+553 : 519.2 (073)

СТАТИСТИЧЕСКИЕ МЕТОДЫ В ГЕОЛОГОРАЗВЕДОЧНОЙ ПРАКТИКЕ: Методические указания к лабораторным работам / Санкт-Петербургский горный университет. Сост. *Я.Ю. Бушуев*. СПб, 2020. 88 с.

Методические указания содержат требования, предъявляемые к лабораторным работам по курсу «Статистические методы в геологоразведочной практике», указан порядок выполнения работ, приведены примеры оформления работ.

Предназначены для студентов специальности 21.05.02 «Прикладная геология».

Научный редактор проф. *А.В. Козлов*

Рецензент начальник отдела геологических информационных систем *П.П. Учаев* (ПАО «Селигдар»)

ВВЕДЕНИЕ

Предметом исследований геологии являются геологические объекты и геологические процессы. Геологические объекты – различные вещественные геологические тела, слагающие часть земной коры и оконтуренные определёнными границами, которые могут быть естественными и искусственными (условными). Геологические процессы – совокупность физико-химических и биологических природных явлений, между которыми существуют сложные причинно-следственные связи. Поэтому свойства геологических объектов зависят от множества факторов, характеризуются сильной изменчивостью, а сами объекты имеют обычно весьма сложное строение. Сведения о свойствах геологических объектов геолог получает, обрабатывая результаты наблюдений – *геологические данные:* минеральный и химический состав пород и руд, их физические свойства, геологические обстановки их обнаружения; размеры, форма, мощность геологических тел, пространственные координаты проб и т.п.

Огромный объем информации и недоступность большинства геологических объектов и процессов для непосредственного наблюдения (как в пространстве, так и во времени), делают необходимым применение математических методов в геологии. Из всего разнообразия методов математики для этих целей наиболее подходят *математическая статистика* и *математическое моделирование.*

Математическая статистика – это раздел математики, изучающий методы сбора, систематизации, обобщения, наглядного представления и обработки эмпирических данных большого объёма с последующими выводами с целью выявления существующих закономерностей. Статистика позволяет распространить выводы, полученные по ограниченному числу наблюдений (*выборке*) на весь объект изучения (*генеральную совокупность*).

Цель данных лабораторных работ – научить будущих геологов применять некоторые наиболее часто используемые для решения геологических задач методы математической статистики. Для достижения поставленной цели следует научиться правильно подготавливать исходные данные, формулировать геологическую

задачу, выбирать наиболее оптимальные статистические методы для её решения. Кроме того, необходимо владеть элементарными навыками работы на ЭВМ, а также иметь представление о различных компьютерных программах, помогающих проводить обработку данных.

Исходным материалом для выполнения заданий служат массивы данных, предоставляемые преподавателем.

Лабораторное обеспечение: персональные компьютеры, установленные в компьютерном классе кафедры, программные пакеты MS Excel, STATISTICA.

Форма представления результатов: демонстрация файла («**группа_ФамилияИО_лаб№**») с результатами расчётов и выводами.

Работа должна содержать:

- фамилию автора и шифр учебной группы;
- номер лабораторной работы, задания и варианта;
- название темы;
- краткую постановку задачи;
- заголовки этапов решения;
- расчётные таблицы (с "шапкой" и пояснениями);
- результаты расчётов (с пояснениями);
- выводы (по этапам выполнения и для всей работы в целом).

Завершённая работа подлежит защите.

Общие рекомендации

При получении задания исходные данные следует сохранить отдельно. Ход работы комментируйте: надо делать пояснения, чтобы преподаватель в Вашей работе мог разобраться без Вас. Сохраняйте результаты работы чаще. Делайте резервные копии на разных носителях.

Таблицы и рисунки должны быть подписаны. В тексте на них должны быть ссылки. **На графиках не забывайте подписать оси**.

При работе в Excel листы книги должны быть подписаны, а листы, не содержащие какой-либо информации – удалены. При

работе в STATISTICA файлы данных и рабочие книги должны быть названы.

Программы STATISTICA и Excel регулярно обновляются – настоящие методические указания могут не соответствовать версии программы, в которой выполняется лабораторная работа. Поэтому надо следить за всплывающими подсказками и привыкать пользоваться встроенной справкой (F1).

1. РАСЧЁТ СТАТИСТИЧЕСКИХ ХАРАКТЕРИСТИК СЛУЧАЙНОЙ ВЕЛИЧИНЫ

Цель работы: научиться рассчитывать основные выборочные статистики.

Задание 1: По результатам анализов проб (файлы «Массив 1-1» и «Массив 1-2») рассчитать основные статистики (минимальное и максимальное значения, размах, медиану, среднее значение, дисперсию, стандартное отклонение, коэффициент вариации, асимметрию и эксцесс).

Расчёты по массиву 1-1 требуется произвести "вручную[1]" и проверить с помощью стандартных функций, «*Анализа данных*» (“*Data Analyses*”) программы Excel и программы STATISTICA.

Для массива 1-2 расчёты «вручную» не проводить.

Порядок выполнения работы и теоретические основы

1.1. Ознакомиться с предоставленными данными.

Для быстрого знакомства с исходными данными удобно воспользоваться фильтром, либо сортировкой данных в программе Excel.

Всегда надо помнить с какими данными имеешь дело. В данной лабораторной – содержание химических элементов. Если оно в процентах, то его значение не может превышать 100; и не может быть меньше 0 (не зависимо от единиц измерения). Если в данных находятся подобные ошибки (скорее всего – опечатки), то следует обратиться к источнику данных (первичная документация, лаборатория, заказчик работ) за разъяснениями. Кроме опечаток среди содержаний элементов могут присутствовать следующие записи (**Ошибка! Источник ссылки не найден.**).

Таблица 1.1

Основные сокращения в таблице результатов опробования

Запись	Расшифровка	Вариант действий
– ; н.а.	нет анализа	удаляют из дальнейших расчётов

[1] Пользуясь математическими операторами сложения, вычитания, умножения, деления; из встроенных функций Excel использовать только СУММ.

Продолжение табл. 1.1

Запись	Расшифровка	Вариант действий
нпо; нн; <x	ниже порога обнаружения, ниже нижнего	удаляют из расчётов либо заменяют на ½ x
>x	больше верхнего порога обнаружение	удаляют из расчётов либо заменяют на 1,01 x

Данные в геологии очень дороги (в прямом смысле) поэтому любое видоизменение исходных данных следует комментировать.

1.2. Рассчитать статистические характеристики выборки.

1.2.1. Составить таблицу для расчёта статистических характеристик "вручную" (**Ошибка! Источник ссылки не найден.**).

Таблица 1.2

Расчёт статистических характеристик

№ п/п	Исходные данные, x_i	Степени отклонений			
		$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^3$	$(x_i - \bar{x})^4$
1	x_1	√	√	√	√
…	…	…	…	…	…
n	x_n	√	√	√	√
Сумма	√	√	√	√	√
Среднее	√	√	√	√	√
Статистики*	$\bar{x}$	$\mu_1 = 0$	$\mu_2 = S^2_c$	μ_3	μ_4

*рассчитанные средние значения по столбцам будут соответствовать обозначенным в данной строке статистическим характеристикам

Примечания к таблице 1.2:

1) символ «√» означает, что эти ячейки должны быть заполнены рассчитанными данными;

2) μ_1, μ_2, μ_3, μ_4 – центральные моменты случайной величины соответственно первого, второго, третьего и четвёртого порядков. Общая формула расчёта центрального момента порядка k случайной величины x имеет вид:

$$\mu_k = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^k}{n} \qquad (1.1)$$

3) S_c^2 – смещённая дисперсия выборки:

$$S_C^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n} \qquad (1.2)$$

4) «растягивая» введённые формулы по столбцам и строчкам, не забывайте ставить знак **$** перед адресом ячейки с постоянным значением (например, $D25 - закрепление адреса столбца D; D$25 - закрепление строчки 25; D25 - закрепление ячейки D25).

1.2.2. Найти минимальное и максимальное значения; рассчитать размах по формуле:

$$R = x_{\max} - x_{\min} \qquad (1.3)$$

1.2.3. Рассчитать статистические характеристики.

● **Среднее арифметическое значение** (оценка среднего арифметического генеральной совокупности по выборке):

$$\overline{x} = \frac{\sum_{i=1}^{i=n} x_i}{n} \qquad (1.4)$$

где n – объем выборки[1]; x_i – i-тое значение выборки.

Среднее значение характеризует усреднённое значение параметра по выборке (например, среднее значение содержаний химического элемента в минерале, горной породе, полезном ископаемом). По величине среднего содержания полезного компонента в соответствии с действующими кондициями на данный вид сырья руды опробованного участка относят либо к богатым, либо к рядовым, либо к бедным, либо к убогим. Различные усреднённые параметры участвуют в оконтуривании рудных тел и подсчёте запасов.

[1] **Объем выборки** – количество исходных данных (отобранных проб, сделанных замеров, результатов анализов и т.п.), используемых для статистических расчётов.

● **Медиана** – значение признака, которое приходится на середину упорядоченного ряда. Для нахождения медианы нужно расположить все значения в порядке возрастания или убывания и найти член ряда, попадающий в середину. При чётном количестве наблюдений непрерывной[1] случайной величины берут среднее арифметическое между двумя серединными значениями. Если изучается дискретная случайная величина, то при чётном числе наблюдений медиана принимает 2 значения.

● **Мода** – наиболее часто встречающееся значение случайной величины. При прямом расчёте и небольшом количестве исходных данных мода может и не существовать – каждое значение встречается только один раз.

Медиана, мода и среднее значение являются статистическими характеристиками положения – около них группируются измеренные значения случайной величины. ***Для теоретического нормального закона распределения эти параметры совпадут***.

● **Дисперсия выборки** (оценка дисперсии генеральной совокупности по выборке, исправленная дисперсия[2]) – это число, равное среднему квадрату отклонений значений случайной величины от её среднего значения:

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \qquad (1.5)$$

Определить в каких единицах измеряется дисперсия.

---

[1] **Непрерывной** называют такую случайную величину, которая может принимать любые значения из некоторого конечного или бесконечного интервала. **Дискретной** называют такую случайную величину, соседние значения которой не могут отличаться друг от друга на величину, меньше некоторой.

[2] В формуле исправленной дисперсии учитываются, так называемые, степени свободы (n – 1), которые вводятся из-за того, что в формулу дисперсии входит не математическое ожидание, а среднее значение, рассчитанное по той же выборке. Число **степеней свободы** k – это превышение числа наблюдений в выборке над числом оценённых по этой же выборке параметров. Используется в распределениях, зависящих от объёма выборки. При большом объёме выборки $k = n$.

● **Стандартное (среднеквадратическое) отклонение выборки**:

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \qquad (1.6)$$

Определить в каких единицах измеряется стандартное отклонение.

● **Коэффициент вариации:**

$$V = \frac{S}{\bar{x}} \qquad (1.7)$$

Коэффициент вариации измеряется в долях единицы или в процентах, что позволяет сравнивать степени изменчивости величин с различными размерностями.

Дисперсия, стандартное отклонение, коэффициент вариации, а также размах являются ***мерами рассеяния (отклонения)*** случайной величины от среднего значения и характеризуют степень её изменчивости. Чем они больше, тем сильнее изменчивость.

В геологии на основе коэффициента вариации принято разделять распределение компонентов в рудах месторождений по степени изменчивости (**Ошибка! Источник ссылки не найден.**), что, в свою очередь, влияет на классификацию месторождения по сложности строения и выбор системы его разведки.

Таблица 1.3

Группировка оруденения по степени изменчивости (по В.М. Крейтеру)

Коэффициент вариации содержания, %	Распределение компонентов
До 20	Весьма равномерное
20-40	Равномерное
40-100	Неравномерное
100-150	Весьма неравномерное
Более 150	Крайне неравномерное

Требуется определить, какой группе по степени изменчивости соответствуют исследуемые данные.

● **Асимметрия** (коэффициент асимметрии):

$$A = \frac{\sum (x_i - \bar{x})^3}{(n-2) \cdot S^3} \quad \text{или} \quad A = \frac{n}{(n-2)} \frac{\mu_3}{S^3}, \tag{1.8}$$

где μ_3 – центральный момент третьего порядка (см. **Ошибка! Источник ссылки не найден.**).

Коэффициент асимметрии характеризует степень симметричности кривой плотности вероятности случайной величины или гистограммы, построенной по выборке, относительно среднего значения. Для симметричных законов распределения $A = 0$. Если $A < 0$, кривая имеет левую (отрицательную) асимметрию (левая ветвь более длинная, а мода смещена вправо); если $A > 0$ кривая имеет правую (положительную) асимметрию (правая ветвь более длинная, а мода смещена влево).

● **Эксцесс** (коэффициент эксцесса):

$$E = \frac{\sum (x_i - \bar{x})^4}{(n-2) \cdot S^4} - 3 \quad \text{или} \quad E = \frac{n}{(n-2)} \frac{\mu_4}{S^4} - 3, \tag{1.9}$$

где μ_4 – центральный момент четвёртого порядка (см. **Ошибка! Источник ссылки не найден.**).

Коэффициент эксцесса характеризует форму кривой плотности вероятности случайной величины или гистограммы, построенной по выборке. Для теоретического нормального закона распределения $E = 0$. Если $E < 0$, кривая имеет более плоскую вершину; если $E > 0$ кривая имеет более острую вершину, чем у нормального закона.

Для теоретического нормального закона распределения $A = 0$ и $E = 0$. Для логарифмически нормального закона распределения $A > 0$ и $E > 0$ (это необходимое условие, но не достаточное).

1.3. Проверка «ручного счета».

1.3.1. С помощью электронных таблиц Excel:

А) "***Анализ данных***": меню "*Данные*" → «*Анализ данных*»[1] → «*Описательная статистика*» (*Descriptive statistics*). В окне «*Описательная статистика*» необходимо ввести входной интервал с исходными данными и поставить «галочку» в поле «*Итоговая статистика*».

Б) "***мастер функций***" (функции СРЗНАЧ, МЕДИАНА, МОДА, СТАНДОТКЛОН, ДИСП, СКОС (асимметрия), ЭКСЦЕСС, МИН, МАКС, СУММ, СЧЕТ).

Следует учитывать, что Excel рассчитывает некоторые статистические характеристики по более сложным формулам, в связи с этим рассчитанные значения асимметрии и эксцесса могут не совпадать с машинными.

1.3.2. С помощью программы STATISTICA.

Примечание: программный пакет STATISTICA установлен на виртуальной машине (Virtual Box).

Создать в STATISTICA новый файл данных *Spreadsheet*. (*Ctrl+N*). Настройки по умолчанию можно не менять.

По умолчанию создаётся таблица для 10 переменных (Variables) и 10 наблюдений (Cases) по каждой из них. Если количество переменных или значений превышает 10, необходимо расширить исходную таблицу, что можно сделать несколькими способами:

1) кнопки на панели инструментов: *Vars* → *Add…* и *Cases* → *Add…*;

2) главное меню: *Insert* → *Add Variables* или *Add Cases*;

3) вызвать контекстное меню нажатием правой кнопки мыши по заголовкам строк / столбцов;

4) при вставке данных таблица автоматически увеличится.

В Excel выделяется блок данных вместе с заголовками, копируется и вставляется в STATISTICA: правой кнопкой мыши по любой ячейке таблицы → *Paste With Headers* → *Paste With Variable Names*.

---

[1] Если на вкладке "*Данные*" отсутствует строка "*Анализ данных*", надо сперва включить "*Пакет анализа*": «*Файл*» → «*Параметры*» → «*Надстройки*» → «*Надстройки Excel*» перейти «*Пакет анализа*»

Примечание: в первом столбце с номерами строк можно вписывать любую, в том числе и текстовую информацию (например, номера проб). Проще всего сделать это при вставке (*...Paste With Cases Names* или *...Paste With Both*).

При необходимости можно изменить форматирование. В каждом столбце устанавливается необходимый формат ячеек: главное меню *Format* → *Cells…* → *Number*. Количество знаков после запятой также можно установить с помощью кнопки «.00 → .0» на стандартной панели, если формат ячейки задан как *Number*.

С помощью двойного щелчка (или кнопки *Vars* → *Specs...*, или *All Specs…*) в поле имени переменной можно переименовать каждую переменную (столбец).

Расчёт статистических характеристик (описательная статистика):

Выбрать метод через пункт главного меню *Statistics* → *Basic Statistics / Tables* → *Descriptive Statistics*.

В появившемся окне, нажав кнопку *Variables*, выбрать столбцы данных, по которым надо провести расчёт статистических характеристик.

На вкладке «*Advanced*» выбрать (добавить), какие статистические характеристики надо рассчитать (**Ошибка! Источник ссылки не найден.**). Кнопкой «*Select all*» сразу выбираются все параметры.

Таблица 1.4

Английские и русские названия статистических характеристик

Английское наименование	**Перевод на русский**
Valid N	объем выборки
Mean	среднее
Confidence – 95%	нижняя граница доверительного интервала среднего
Confidence 95%	верхняя граница доверительного интервала среднего
Geometric Mean	среднее геометрическое
Harmonic Mean	среднее гармоническое
Median	медиана

Продолжение табл. 1.4

Английское наименование	**Перевод на русский**
Mode	мода
Frequency of Mode	частота модального значения
Sum	Сумма (объем выборки)
Minimum	минимум
Maximum	максимум
Lower Quartile	нижняя квартиль*
Upper Quartile	верхняя квартиль*
Percentile 10	процентиль 10**
Percentile 90	процентиль 90**
Range	размах
Quartile Range	квартильный размах***
Variance	дисперсия
Standard Deviation (Std.Dev.)	стандартное отклонение
Standard Error	стандартная (абсолютная) ошибка среднего
Skewness	асимметрия
Std.Err. Skewness	стандартная ошибка асимметрии
Kurtosis	эксцесс
Std.Err. Kurtosis	стандартная ошибка эксцесса

---

* **Квартили** (от «*кварта*» – четверть) представляют собой значения, которые делят две половины упорядоченной по возрастанию выборки (разбитые медианой) ещё раз пополам. Таким образом, медиана и квартили делят диапазон значений переменной на четыре равные части. Различают **верхний квартиль** (Q_3), который больше медианы и делит пополам верхнюю часть выборки (значения переменной больше медианы), и **нижний квартиль** (Q_1), который меньше медианы и делит пополам нижнюю часть выборки. Нижний квартиль часто обозначают символом 25 %, это означает, что 25 % значений переменной меньше нижнего квартиля. Верхний квартиль обозначают символом 75 %. Нижний и верхний квартили являются соответственно 25-м и 75-му процентилям распределения. Сама мода является средним квартилем (Q_2) или 50-мпроцентилем.

** **Процентиль 10** – это значение, ниже которого располагаются 10 % значений упорядоченной по возрастанию выборки. **Процентиль 90** – это значение, ниже которого располагаются 90 % значений переменной.

*** **Квартильный размах** переменных равен разности значений верхнего и нижнего квартилей (75-го и 25-го процентилей). Таким образом, это интервал, содержащий медиану, в который попадает 50 % наблюдений.

Копирование данных из STATISTICA в Excel

Результаты анализа (таблицы, графики) выводятся в рабочую книгу *Workbook*. Переключаться между таблицей исходных данных и рабочей книгой с результатами можно сочетанием клавиш *Ctrl+Tab* либо мышью, предварительно свернув окно таблицы.

В рабочей книге открыть результирующую таблицу (Descriptive Statistics (Spreadsheet1)). Левой кнопкой мыши нажать в верхнюю левую ячейку, оставшаяся таблица станет затемнённой. По затемнённой области нажать правой кнопкой мыши – выбрать *Copy with Headers*. В Excel поставить курсор на свободную ячейку, сочетанием клавиш *Ctrl+Alt+V* вызвать меню специальной вставки – выбрать *HTML*.

Сохранение данных в STATISTICA

В основном результаты анализа достаточно скопировать в Excel, но если требуется сохранить данные в STATISTICA, то надо помнить, что таблицы с данными и рабочие книги сохраняются отдельно. Чтобы сохранить их в требуемой «связке» есть возможность сохранить проект. Меню *File – Save Project.*

2. ОЦЕНКА ЗАКОНОВ РАСПРЕДЕЛЕНИЯ СЛУЧАЙНЫХ ВЕЛИЧИН

Цель работы: научиться оценивать нормальность или логнормальность закона распределения значений случайной величины графическим и аналитическим методами.

Порядок выполнения работы и теоретические основы

Гистограмма – аппроксимация функции плотности вероятности распределения данных.

Как и обычная гистограмма (столбчатый график) статистическая гистограмма – это графический способ отображения табличных данных, но данных строго определённых: по одной оси (обычно оси абсцисс) – значения случайной величины, сгруппированные в интервалы (бины, классы), по другой оси – частоты (частости - относительные частоты) их появления.

1. Последовательность действий при построении гистограммы

1.1. Определить число классов (равных интервалов) группировки (N) (количество столбцов гистограммы). Число интервалов выбирается так, чтобы видны были особенности характера распределения. Существуют разные способы выбора количества классов. Обычно принимается число классов от 6 до 12, но можно брать и до 40. Чем больше исходных данных и чем асимметричнее распределение, тем больше нужно классов (таблица 2.***1***).

Таблица 2.1

Зависимость числа классов в гистограмме от объёма выборки

Объем выборки, n	**Количество классов в гистограмме, *N***
<10	3
10-30	3-4
30-100	4-6
100-500	6-9
500-3000	9-13
>3000	13-18

Существуют и различные формулы для примерного расчёта числа классов, например, правило Стержесса (Стургеса):

$$N = 1 + \log_2 n \approx 3.321 \times lg(n) + 1 \approx 1.44 \times ln(n) + 1 \qquad (2.1)$$

Считается, что формула Стержесса даёт заниженное количество классов при объёме выборки меньше 30 значений и подходит только для выборки с нормальным законом распределения.

Для учёта асимметричности данных предлагается модифицированная формула:

$$N = 1 + \log_2 n + \log_2(1+t_A), \qquad (2.2)$$

где t_A – критерий t для асимметрии см. формулу (2.6.).

Рассчитанное по вышеприведённым формулам количество классов округляется до целого в большую сторону.

1.2. Определить интервал одного класса:

$$h = \frac{x_{max} - x_{min}}{N} \qquad (2.3)$$

Для определения интервала во многих программах, например, Excel, пользуются правилом Скотта:

$$h = \frac{3{,}5\sigma}{\sqrt[3]{n}} \qquad (2.4)$$

1.3. Выбрать границы классов. Границы классов и интервал удобно брать округлёнными. ***Если ваши данные определены с точностью до целого, то брать границы интервалов с точностью до сотых бессмысленно! К тому же всегда надо помнить об удобстве восприятия графика***. Например, по расчёту получились $x_{min} = 0{,}07$, а $h = 0{,}18$, лучше выбрать границы интервалов: 0,0 – 0,2; 0,2 – 0,4; и т.д., а не 0,07 – 0,25; 0,25 – 0,43; или 0,07 – 0,27; 0,27 – 0,47 и т.д.). Границу первого класса выбирают так, чтобы туда обязательно попало минимальное значение, а последующие – простым добавлением округлённого значения интервала h (последний класс должен включать максимальное значение).

1.4. Подсчитать количество наблюдений, попадающих в каждый класс (определить частоты), заполнив столбцы 1-3 (**Ошибка! Источник ссылки не найден.**). Значения, совпадающие с границей класса, относят либо к меньшему, либо к большему классу (т.е. учитывают только один раз).

1.5. Построить гистограмму. По оси X – интервалы классов, по оси Y – частота (число попаданий в классы).

Таблица 2.2

Расчёт эмпирических и теоретических частот для построения гистограммы и соответствующей ей теоретической кривой плотности распределения вероятности

Класс (интервал класса)	**Частота, n_i**	**Накопленная частота**	**Середина класса, $\overline{x}_i$**	**Плотность вероятности, $f(\overline{x}_i)$**	**Теоретическая частота, n_{iT}**	$\frac{(n_i - n_{iT})^2}{n_{iT}}$
1	***2***	***3***	***4***	***5***	***6***	***7***
	n_1	n_1			$f(\overline{x}_i) \cdot h \cdot n$	
	n_2	n_1+n_2				
	n_3	$n_1+n_2+n_3$				

Продолжение табл. 2.2

Класс (интервал класса)	**Частота, n_i**	**Накопленная частота**	**Середина класса, $\overline{x}_i$**	**Плотность вероятности, $f(\overline{x}_i)$**	**Теоретическая частота, n_{iT}**	$\frac{(n_i - n_{iT})^2}{n_{iT}}$
1	***2***	***3***	***4***	***5***	***6***	***7***
		…				
	$\sum=n$					$\sum=\chi^2$*
*χ^2 – критерий Пирсона «хи-квадрат» (см. ниже)						

Примечание. Для ускорения подсчёта частот в классах можно воспользоваться программой Excel или STATISTICA. В Excel: Данные / Анализ данных / Гистограмма – если параметр «Интервал карманов» оставить пустым, то машина сама разобьёт весь ряд наблюдений на классы. В STATISTICA: Statistics / Basic statistics / Frecuency tables - выбрать переменные, перейти на вкладку «Дополнительно» (Advanced) далее в зависимости от задачи.

2. Последовательность действий при построении гистограммы с помощью анализа данных Excel и программного пакета STATISTICA

2.1. С помощью анализа данных Excel: *Данные / Анализ данных / Гистограмма*. Выбираете интервал с исходными значениями, выбираете интервал карманов (верхние границы классов гистограммы). В параметрах вывода поставить все «галочки».

2.2. С помощью программы STATISTICA. Главное меню: *Graphs / Histograms*, на вкладке *Advance* выбирают переменные (*Variables*) и число классов (*Categories*) либо задают границы интервалов (*Boundaries, Specify Boundaries*).

3. Анализ гистограммы

Гистограмма даёт наглядное представление о поведении случайной величины. На ней видны: размах, частоты встречаемости конкретных значений, асимметрия, неоднородность: наличие подозрительных на аномальность значений, полимодальность.

Симметричная колоколообразная гистограмма говорит о возможности нормального распределения.

Правосторонняя асимметрия гистограммы – о возможности логнормального распределения. В этом случае надо исходные значения прологарифмировать и расчёты производить с логарифмами значений.

Сильная неравномерность гистограммы может свидетельствовать о влиянии погрешностей измерений либо о наличии аномальных значений, либо о полимодальности выборки. Чтобы устранить влияние случайных погрешностей, надо увеличить длину интервала классов и построить гистограмму снова.

К подозрительным на аномальность относят единичные значения, значительно удалённые от основного тела гистограммы и отделённые от неё пустыми интервалами. Такие значения следует исключить, вновь разбить оставшуюся часть на классы и построить новую гистограмму.

Если распределение бимодальное, все значения следует разделить на две части по минимуму в гистограмме и исследовать отдельно каждую часть. Проверку на бимодальность можно провести с помощью критерия, рассчитанного по формуле:

$$d = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{S_1^2 + S_2^2}} \ . \qquad (2.5)$$

Если: $\boldsymbol{d \geq 4}$ – гистограммы разделены (выборка неоднородна и состоит из двух совокупностей), при механическом разделении по минимальной частоте погрешность вычислений незначительна;

$\boldsymbol{2 \leq d < 4}$ – гистограммы частично перекрыты и довольно легко разделимы;

$\boldsymbol{0{,}7 \leq d < 2}$ – гистограммы перекрываются, теоретически разделимы, разделение может приводить к значительным погрешностям;

$\boldsymbol{d < 0{,}7}$ – гистограммы не разделимы (возможно, выборка взята из одной совокупности, но нарушена её представительность, например, из-за недостаточного числа данных).

3.1. Написать характеристику выборки и гистограммы. Здесь должно быть указано:

1. Что анализировалось (какой химический элемент или соединение).

2. Объем выборки.

3. На сколько классов и с каким интервалом была разделена выборка для построения гистограммы.

4. Характеристика гистограммы (симметричность, равномерность, однородность (например, бимодальность), наличие значений, подозрительных на аномальность).

3.2. При полимодальности проверить возможность разделения выборки и разделить её, если возможно. Подвыборки снова разделить на классы и построить гистограммы. Охарактеризовать их.

3.3. При наличии аномальных значений исключить их из выборки, оставшуюся часть разделить на классы, построить новую гистограмму и дать её характеристику.

4. Построить теоретическую кривую плотности вероятности нормального закона распределения и совместить её с гистограммой. Для этого надо привести к сопоставимому виду частоту и плотность вероятности. Для удобства расчётов заполняется **Ошибка! Источник ссылки не найден.** (столбцы с 4 по 7). Найти плотность вероятности нормального распределения (столбец 5) можно использовать функцию Excel:

НОРМРАСП(*x;среднее;стандартное_откл;интегральная*) – рассчитывает для случайной величины x в зависимости от значения аргумента «*интегральная*»:

- либо значение функции плотности нормального распределения (*f(x)*), если «*интегральная*» = 0;
- либо значение интегральной функции нормального распределения *F(x),* если «*интегральная*» = 1.

Если аргументам «*среднее*» и «*стандартное_откл*» присвоить соответственно значения 0 и 1, то будут рассчитаны значения стандартной функции нормального распределения для нормализованного значения x.

Сумма значений n_{iT} не должна сильно отличаться от общего числа проб n.

Совместить на одном графике гистограмму частот (по значениям n_i) и кривую нормального закона (по n_{iT}) в одном масштабе. Значения n_{iT} следует совмещать с серединами интервалов

классов. Совмещение можно выполнить или на листе миллиметровки, или в Excel на уже построенной гистограмме, добавив второй ряд с теоретическими частотами.

5. Проверить соответствие эмпирического распределения нормальному закону различными способами:

5.1. Визуально по гистограмме и теоретической кривой нормального закона распределения.

5.2. С помощью критериев t для асимметрии и эксцесса (t_A, t_E):

$$t_A = \frac{A}{\sqrt{\frac{6}{n}}}, \qquad t_E = \frac{E}{\sqrt{\frac{24}{n}}} = \frac{E}{2\sqrt{\frac{6}{n}}} \tag{2.6}$$

Если **и** $|t_A| \le 3$, **и** $|t_E| \le 3$, то распределение данных в выборке не противоречит нормальному закону (нет оснований отвергать гипотезу о нормальности распределения).

Если хотя бы один из критериев окажется больше трёх, то распределение данных в выборке не соответствует нормальному закону с доверительной вероятностью 99,7%.

5.3. С помощью критерии Пирсона χ^2.

Критерий применим для больших выборок ($n > 50$). Необходимо, чтобы количество наблюдений в классе было бы не меньше пяти. Если количество наблюдений оказалось меньше пяти, то такой интервал следует объединить с соседним, (при этом уменьшится число интервалов). Эмпирическое значение критерия рассчитывается по формуле ((2.7) (столбец 7 **Ошибка! Источник ссылки не найден.**):

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - n_{iT})^2}{n_{iT}} \tag{2.7}$$

где n_i – число наблюдений в i-том классе; n_{iT} – теоретическое число наблюдений в этом же классе; m – количество классов (интервалов группирования) после объединения классов с малой частотой.

Эмпирический критерий сравнивают с критическим значением распределения Пирсона χ^2 Число степеней свободы: $\boldsymbol{k = m - 3}$.

Если $\chi^2_{эмпир} < \chi^2_{кр.}$, то гистограмма не противоречит нормальному закону распределения.

Если $\chi^2_{эмпир} > \chi^2_{кр.}$, то гистограмма не соответствует нормальному закону распределения с доверительной вероятностью $p = 1 - \alpha$.

Для нахождения критического значения распределения χ^2 для заданных степеней свободы и уровня значимости можно использовать функцию Excel:

ХИ2ОБР(вероятность; степени_свободы)[1],

где аргументы: **вероятность** – выбранный уровень значимости; **степени_свободы** – число степеней свободы.

5.4. Проверить расчёты с помощью функции Excel:

ХИ2ТЕСТ(фактический_интервал; ожидаемый_интервал), где: **фактический_интервал –** интервал, содержащий фактические частоты (столбец 2, **Ошибка! Источник ссылки не найден.**), **ожидаемый_интервал –** интервал, содержащий теоретические частоты (столбец 6, **Ошибка! Источник ссылки не найден.**).

5.5. Дополнить пункт **3.1**. результатами сравнения эмпирического распределения с теоретическим нормальным.

Задание 1: Сгруппировать данные, построить и проанализировать гистограмму «*вручную*». Построить теоретическую кривую нормального распределения. Проверить соответствие эмпирического закона распределения значений случайной величины нормальному различными способами: а) визуально по гистограмме и теоретической кривой нормального закона; б) с помощью критериев *t* для асимметрии и эксцесса; в) с помощью критерия сходства Пирсона χ^2 [хи-квадрат]. Проверить построения, построив гистограмму с помощью анализа данных Excel и программы STATISTICA. ***Расчёты ведутся для данных из файла Массив 2-1.xlsx, Лист «часть 1».***

Задание 2: Сравнить гистограммы, построенные по данным из файла ***Массив 2-1.xlsx, Лист «часть 2».***

---

[1] Для Excel 2007. **Для других версий синтаксис формулы может быть иной – используйте справку (F1).**

Задание 3: Построить гистограмму по данным из файла ***Массив 2-1.xlsx, Лист «часть 3».*** Проанализировать гистограмму.

3. ПРОВЕРКА ОСНОВНЫХ СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Цель работы: научиться оценивать погрешность основных выборочных статистик, производить проверку статистических гипотез о законе и параметрах эмпирического распределения значений случайной величины.

Порядок выполнения работы

Задание 1. Для массивов данных из файлов «Массив 1-1» и «Массив 1-2» построить гистограммы. По ним **визуально** оценить однородность, закон распределения и наличие аномальных значений.

Задание 2. Аналитически проверит визуальную оценку:

2.1. Проверить (с помощью критериев t_A и t_E) закон распределения.

В случае соответствия эмпирических данных логнормальной модели распределения усложняется оценивание параметров генеральной совокупности, их рассчитывают по следующим формулам:

1) Медиана:

$$x_{me} = e^{\overline{\ln x}}, \tag{3.1}$$

где $\overline{\ln x}$ – среднее из натуральных логарифмов исходных значений. Экспонента среднего арифметического логарифмированных значений также называется ***средним геометрическим***.

2) Мода:

$$x_{mod} = e^{\overline{\ln x} - S_{\ln}^2}, \tag{3.2}$$

где $S_{\ln}{}^2$ – дисперсия натуральных логарифмов исходных значений.

3) Среднее значение (**среднее Сихеля**):

$$\overline{x} = e^{\overline{\ln x} + \frac{S_{\ln}^2}{2}}, \tag{3.3}$$

4) Дисперсия:

$$S^2 = e^{2\overline{\ln x}}\left(e^{2S_{\ln}^2} - e^{S_{\ln}^2}\right). \qquad (3.4)$$

При малом значении дисперсий S^2 и $S_{\ln}{}^2$ кривые плотности распределения вероятностей нормального и логнормального законов близки между собой и в пределе, при стремлении диспрерсии к нулю, совпадают. Поэтому при умеренном объеме выборки в случаях, когда оценка дисперсии логнормального распределения S^2_{lg} менее 0,196 (или $S^2_{\ln}$ менее 1,039), удовлетворительной оценкой математического ожидания при логнормальном законе может служить обычное среднее арифметическое выборочных значений случайной величины.

2.2. Исключить аномальные значения (**провести проверку при помощи всех предложенных критериев**).

Подозрительными на аномальность могут являться краевые (как максимальные[1], так и минимальные) значения упорядоченного ряда, значительно удалённые от остальных. Выборка может не содержать аномальных значений.

Порядок проверки:

1. Исключаем из выборки одно наиболее далеко отстоящее от среднего значение, например, максимальное. Считаем его подозрительным на аномальность (значение $\boldsymbol{x_a}$). Если несколько краевых значений имеют одинаковую величину, их исключают одновременно.

2. Для оставшихся значений рассчитываем среднее и стандартное отклонение.

3. Проверяем исключённое значение по какому-либо критерию (см. ниже).

Если первое значение оказалось аномальным, его исключаем и проверяем следующее за ним значение, т.е. исключаем и его из выборки и опять рассчитываем среднее и стандартное отклонение для оставшейся части. Операции повторяем до тех пор, пока не

[1] Аномально высокие значения содержания называют ***ураганными*** или ураганами. Просто так исключать их нельзя – надо попытаться установить причину их появления. В математике аномальные значения в выборке иногда называют ***выбросами***.

будут удалены все аномальные значения. Затем аналогично проверяем минимальные значения.

Если будет установлено, что выборка содержит аномальные значения, то их исключают, статистические характеристики пересчитывают для оставшейся части, и снова проверяют закон распределения.

Наиболее простые критерии для выявления аномальных значений[1]:

а) критерий, основанный на «правиле трёх сигм». Если выборка не противоречит нормальному закону распределения, а её объем составляет не менее 30 наблюдений:

не аномальные значения лежат в интервале:

$$\bar{x} \pm 3S \qquad (3.5)$$

где $\bar{x}$ и S рассчитаны при исключённых из выборки значениях, подозрительных на аномальность.

Если x_a находится за пределами данного интервала, то оно **является** аномальным с доверительной вероятностью $p = 0{,}997$.

Примечание 1: Можно стандартизовать исходные значения по формуле:

$$t = \frac{x_i - \bar{x}}{S}, \qquad (3.6)$$

где $\bar{x}$ и S рассчитаны при исключённых из выборки проверяемых на аномальность значениях. Аномальными считаются значения t строго больше 3 по модулю (с доверительной вероятностью $p = 0{,}997$).

Примечание 2: Если распределение случайной величины подчиняется логнормальному закону, то правило "трёх сигм" применяется к логарифмам значений.

б) при небольшой выборке, имеющей симметричное распределение, подозрительные на аномальность значения сравниваются с ***t*-критерием Стьюдента**, эмпирическое значение которого рассчитывают по формуле:

---

[1] При проверке гипотез всегда указывают, с помощью какого критерия проводилась проверка, так как разные критерии могут давать неодинаковые результаты.

$$t_{эмпир} = \frac{|x_a - \bar{x}|}{S}, \qquad (3.7)$$

где x_a – подозрительное на аномальность значение выборки, а среднее значение $\bar{x}$ и стандартное отклонение S рассчитаны при исключённых из выборки аномальных значениях.

Эмпирический критерий $t_{эмпир}$ сравнивают с критическим значением, которое находят по таблицам «одностороннего» критерия Стьюдента[1] для $k = n - 1$ степеней свободы[2] и выбранного уровня значимости[3] α (или доверительной вероятности $p = 1 - \alpha$). Для проверки аномальных значений обычно берётся уровень значимости α = 0,003 – 0,005 (0,3 – 0,5 %).

Если $t_{эмпир} < t_{(\alpha,\, k)\ крит.}$, то проверяемое значение не является аномальным (нет оснований относить его к аномальным).

Если $t_{эмпир} > t_{(\alpha,\, k)\ крит.}$, то проверяемое значение является аномальным с доверительной вероятностью $p = 1 - \alpha$.

Примечание 3. При отсутствии таблиц «одностороннего» критерия Стьюдента, можно пользоваться таблицами «двустороннего» критерия Стьюдента. Предельные значения *t*-критерия в этом случае берутся для уровня значимости 2α, а доверительная вероятность остаётся соответствующей α ($p = 1 - \alpha$).

Примечание 4. Для нахождения критического значения распределения Стьюдента для заданных степеней свободы и уровня значимости удобно использовать функцию Excel, возвращающую левостороннее распределение Стьюдента:

---

[1] Точное название: «Таблицы критических точек распределения Стьюдента для односторонней критической области». Распределение Стьюдента похоже на нормальное, но зависит только от объёма выборки. Часто используется в качестве критерия для принятия решения в различных статистических задачах.

[2] **Степень свободы** – превышения объёма выборки над числом оцениваемых по этой выборки параметров. Для каждого критерия рассчитывается по своим правилам, но всегда зависит от объёма выборки.

[3] Уровень значимости α выбирается с учётом риска, который может позволить себе исследователь, в данном случае, ошибившись при решении принять проверяемое значение за аномальное, если на самом деле оно им не является. Чем ответственней задача, например, если речь идёт о жизни людей, тем меньше выбирают уровень значимости (0,001 и меньше).

СТЬЮДЕНТ.ОБР(вероятность; степени_свободы), где: **вероятность** – уровень значимости, соответствующий двустороннему распределению; **степени_свободы** – число степеней свободы. Критическое значение надо взять по модулю.

***Задание 3**:* Оценить погрешность выборочного среднего значения, рассчитать необходимое число проб, которые надо отобрать для доведения погрешности до заданного значения, построить доверительный интервал среднего значения.

Расчёты ведутся для данных из файла «**Массив 1-1**» (если в данных выявлено аномальное значение, то его надо исключить).

Порядок выполнения работы и теоретические основы:

3.1. Рассчитать **стандартную ошибку среднего значения** (абсолютную среднеквадратическую погрешность оценки среднего):

$$\delta = \frac{S}{\sqrt{n}}, \qquad (3.8)$$

где S – выборочное стандартное отклонение; n– число наблюдений.

Определить в каких единицах измеряется эта величина.

3.2. Рассчитать **относительную погрешность оценки среднего** значения:

$$\tau = \frac{\delta}{\bar{x}} \text{ или } \tau = \frac{V}{\sqrt{n}}; \quad \tau = \frac{\delta}{\bar{x}} 100\,\% \text{ или } \tau = \frac{V}{\sqrt{n}} 100\%, \qquad (3.9)$$

где V – коэффициент вариации (см. формулу 1.7).

Определить в каких единицах измеряется эта величина.

3.3. Рассчитать, сколько проб надо отобрать дополнительно для снижения относительной погрешности оценки среднего арифметического значения на ***20 % от имеющейся*** – $\tau_{задан.} = \tau - 0{,}20\ \tau$.

Для решения этого задания следует понимать, что:

- среднее значение и дисперсия для выборки не изменятся;
- число проб может быть только целым (округляют до большего значения).

3.4. Рассчитать доверительные интервалы (интервальную оценку, уровень надёжности) истинного среднего значения *M(x)* измеряемой величины при вероятностях $p = 0{,}95$ и $p = 0{,}99$.

Если из генеральной совокупности взять несколько случайных выборок, то в каждой из них при расчётах среднего значения будут получены в общем случае не равные между собой величины, т.е. ***выборочное среднее значение само является случайной величиной***. При увеличении объёма выборок рассеяние выборочных средних уменьшается. Таким образом, выборочное среднее – это приближенное значение истинного (математического ожидания генеральной совокупности). Истинное среднее значение *M(x)* не известно, но по его оценке $\bar{x}$ и погрешности δ можно определить интервал, в который оно попадает с некоторой вероятностью. Вероятность попадания истинного значения в доверительный интервал задаётся с помощью уровня значимости (α). Чем меньше величина уровня значимости для выбираемого по справочным таблицам *t*-критерия Стьюдента, тем больше вероятность того, что истинное значение действительно находится в пределах данного интервала.

Доверительный интервал строят по следующим формулам:

$$\bar{x} - \lambda < M(x) < \bar{x} + \lambda, \quad \text{где} \quad \lambda = t \cdot \delta = t \cdot \frac{S}{\sqrt{n}}, \qquad (3.10)$$

где t – табличное значение «двустороннего» критерия Стьюдента для $k = n - 1$ степеней свободы и выбранного уровня значимости (0,01; 0,05; 0,1 и т.д.). Например, при $\alpha = 0{,}05$ (5 %) с вероятностью 0,95 (95 %) можно утверждать, что среднее значение параметра в изучаемом объекте больше $\bar{x} - \lambda$, но меньше $\bar{x} + \lambda$.

Примечание 1. При достаточном количестве данных в качестве t можно использовать коэффициенты вероятности нормального распределения, используя таблицы квантилей[1] распределения функции:

---

[1] ***Квантиль*** (p-квантиль) выборки представляет собой число x_p, ниже которого находится p-я часть выборки. Например, квантиль 0,2 для некоторой переменной – это такое значение x_p, ниже которого находится 20 % значений переменной.

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-t}^{t} e^{-\frac{t^2}{2}} dt$$

(вероятность попадания случайной величины в симметричный интервал от - *t* до + *t*). Следует запомнить эти коэффициенты при трёх значениях вероятности (таблица 3.1):

Таблица 3.1

Коэффициенты вероятности нормального закона распределения

Коэффициент вероятности *t*	1	2	3
Вероятность *p=Φ(t)*	0,6827	0,9545	0,9973

Примечание 2. Рассчитать доверительный интервал (λ) для выборки достаточного объёма можно с помощью «*Мастера функций*»: Статистические → ДОВЕРИТ. В качестве *t* (формула (3.10)) функция ДОВЕРИТ использует коэффициенты вероятности нормального распределения.

Для закрепления понятия ***квантиль*** для всех массивов данных построить график кумулятивной гистограммы (по оси Y – относительная частота) и график квантиль-квантиль (способ построения найти самостоятельно).

Задание 4*:* Проверка статистических гипотез. Сравнить выборочное среднее с заранее заданным, а также средние значения и дисперсии двух выборок друг с другом.

Расчёты ведутся для данных из файла «**Массив 1-1**».

Теоретические основы и порядок выполнения работы

Для решения любых статистических вопросов принято формулировать, так называемые, статистические гипотезы, а затем проверять их с помощью статистических критериев.

Статистическая гипотеза – это любое высказывание о генеральной совокупности, проверяемое по выборке.

При построении статистических гипотез в ответ на любой вопрос формулируются два ответа: *нулевая* (основная) гипотеза H_0 и *альтернативная* (конкурирующая) H_1. Причём нулевая гипотеза всегда несёт отрицание, а альтернативная должна быть довольно

общей. Проверка нулевой гипотезы относительно альтернативы проводится на основании выборки из *n* элементов и представляет собой расчёты, так называемых, эмпирических критериев и их сравнение с известными критическими значениями различных статистических критериев.

В результате проверки мы принимаем или отвергаем H_0. При этом мы можем ошибочно отвергнуть верную нулевую гипотезу. Вероятность такой ошибки называют уровнем значимости и обозначают чаще всего α. Например, уровень значимости $\alpha = 0{,}05$ означает, что, в 1 случае из 20 мы ошибочно отвергнем H_0. Обычно уровень значимости задают до применения критерия.

4.1. Сравнить среднее значение, рассчитанное по выборке, с заданным (***$x_{задан}$ на 10 % больше среднего***). Предполагается, что выборка отобрана из нормальной совокупности, а дисперсия генеральной совокупности неизвестна.

Для решения данной задачи сравнения надо вычислить эмпирическое значение критерия по формуле:

$$t_{эмпир} = \frac{\left| x_{задан} - \overline{x} \right| \cdot \sqrt{n}}{S} \qquad (3.11)$$

Обратите внимание, что фактически мы проверяем попадает ли заданное значение в доверительный интервал среднего (формула 1.20).

Величина *t* имеет распределение Стьюдента с $k = n - 1$ степенями свободы.

Дальнейшее решение зависит от поставленной задачи.

Задача 1. Проверить, что выборка взята из совокупности, имеющей заданное среднее значение. Например, надо выяснить, соответствует ли среднее значение содержаний кремнезёма в пробах его среднему значению в диорите.

Для решения такой задачи формулируется нулевая гипотеза $H_0 : \overline{x} = x_{задан}$ (т.е. вычисленное среднее **не** отличается от заданного) при множестве двусторонних альтернатив $H_1 : \overline{x} \neq x_{задан}$, т.е. нас не интересует, больше наше среднее значение, чем заданное, или меньше.

Правило 1 для принятия решения. По таблицам критических точек распределения Стьюдента для двухсторонней критической области найти при заданных уровне значимости α и числе степеней свободы *k* критическое значение $t_{(\alpha,\ k)\ крит.}$ и сравнить с $t_{эмпир.}$.

Если $t_{эмпир} < t_{(\alpha,\ k)\ крит.}$, то нет оснований отвергать нулевую гипотезу. Следовательно, можно принять, что с некоторой (неизвестной) вероятностью среднее значение выборки существенно не отличается от заданного.

Если $t_{эмпир} > t_{(\alpha,\ k)\ крит.}$, то среднее значение выборки отличается от заданного с доверительной вероятностью $p = 1 - \alpha$.

4.2. Для задач 2 и 3 рассчитать интервалы значений, в которых ошибка 1-го рода (α) находится в пределах от 5 % до 0,001 %.

Задача 2. Есть задачи, в которых нам не безразлично, больше наше среднее значение, чем заданное, или меньше. Так при подсчёте запасов минерального сырья вопрос промышленного использования полезного ископаемого на изучаемом участке месторождения решается путём сравнения полученных оценок среднего содержания полезного компонента с минимальным промышленным (т. е., если $\bar{x} > x_{min\ пром.}$, то это промышленная руда, и наоборот). В этом случае используется «односторонний» («правосторонний») критерий, построенный для проверки нулевой гипотезы $H_0 : \bar{x} \le x_{задан}$ при множестве односторонних альтернатив $H_1 : \bar{x} > x_{задан}$.

Задача 3. Если же для данного вида сырья существуют ограничения на содержание вредных примесей, то полученные оценки средних содержаний вредных компонентов должны быть сравнены с максимально допустимым значением. В этом случае используют «левосторонний» критерий Стьюдента, построенный для проверки нулевой гипотезы $H_0 : \bar{x} \ge x_{задан}$ при множестве односторонних альтернатив $H_1 : \bar{x} < x_{задан.}$.

Примечание. Для расчёта *t* при одностороннем критерии Стьюдента модуль в формуле (1.21) не нужен.

Для понимания сути принятия гипотез надо графически представлять распределение вероятности (см. рис. 3.**1**).

Из рисунка видно, что при одинаковой доверительной вероятности левосторонний и правосторонний критерии равны по модулю. Значение t одностороннего критерия совпадёт с двусторонним, но вероятность ошибки первого рода будет отличаться в 2 раза.

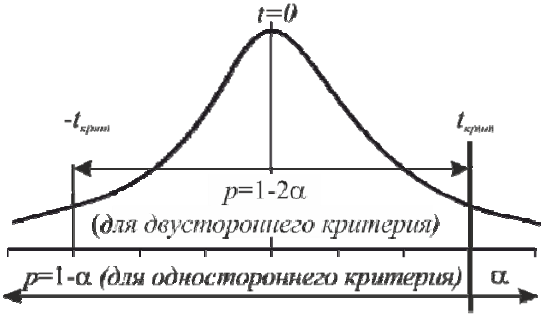

Рис. 3.1. График плотности распределения Стьюдента: площади под кривой – вероятности допустить ошибку 1-го рода: справа от $t_{крит}$ для правостороннего критерия; слева от $t_{крит}$ для левостороннего критерия; справа от $t_{крит}$ и слева от $-t_{крит}$ – для двустороннего критерия

4.3. Проверить существенность различий значений дисперсий двух выборок с помощью критерия Фишера. Расчёты ведутся для файла «Массив 1-1».

Для решения задач сравнения дисперсий и средних двух выборок (п. 3.3 и 3.4) в качестве первой выборки берётся свой вариант, второй выборки – следующий вариант анализов проб на тот же элемент в том же файле.

Для сравнения двух дисперсий надо рассчитать эмпирический критерий Фишера:

$$F_{эмпир} = \frac{S_1^2}{S_2^2} \quad \text{при} \quad S_1^2 > S_2^2 \qquad (3.12)$$

Число степеней свободы при поиске табличного критического значения критерия F в таблице Фишера (таблице

критических точек распределения F Фишера-Снедекора): $k_1 = n_1 - 1$; $k_2 = n_2 - 1$.

Если $F_{эмпир} < F_{(k,\alpha)крит.}$, то нет оснований считать дисперсии этих выборок различными (дисперсии двух выборок существенно не различаются).

Если $F_{эмпир} > F_{(k,\alpha)крит.}$, то дисперсии двух выборок различаются с доверительной вероятностью $p = 1 - \alpha$.

Примечание. Для нахождения критического значения распределения Фишера для заданных степеней свободы и уровня значимости можно использовать функцию Excel:

F.ОБР(вероятность; степени_свободы1; степни_свободы2),

где аргументы: вероятность – выбранный уровень значимости; степени_свободы1 – это число степеней свободы для числителя в отношении дисперсий; степени_свободы2 – это число степеней свободы для знаменателя.

Проверить результат с помощью функции F.ТЕСТ(массив1; массив2).[1]

4.4. Проверить существенность различий средних значений двух выборок. Выборки берутся те же, что и для п. 3.3.

А) Если средние значения и исправленные дисперсии найдены по двум большим ($n_1 > 20$-30 и $n_2 > 20$-30) независимым выборкам из нормальных совокупностей (предполагается, что генеральные дисперсии σ_1 и σ_2 известны и неравны), то для проверки нулевой гипотезы $H_0 : \bar{x}_1 = \bar{x}_2$ (т.е. средние существенно не различаются) вычисляют эмпирический критерий по формуле:

$$z_{эмпир} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \qquad (3.13)$$

[1] Для MicrosoftOffice 2010.

где $\bar{x}_1$, S_1^2 и n_1 – соответственно: среднее значение, дисперсия и объем первой выборки; $\bar{x}_2$, S_2^2 и n_2 – соответственно: среднее значение, дисперсия и объем второй выборки.

Проверка проводится в зависимости от выдвигаемой альтернативной гипотезы:

Правило 1. При двусторонней альтернативной гипотезе $H_1 : \bar{x}_1 \neq \bar{x}_2$ по таблицам функции Лапласа находят критическое значение $z_{кр.}$ для выбранной α из равенства:

$$\Phi(z_{кр.}) = 1 - \alpha \qquad (3.14)$$

Если $z_{эмпир} < z_{кр.}$, то нет оснований считать средние значения двух выборок различными.

Если $z_{эмпир} > z_{кр.}$, то средние значения двух выборок различаются с доверительной вероятностью $p = 1 - \alpha$.

Правило 2. При односторонней альтернативе $H_1 : \bar{x}_1 > \bar{x}_2$ (или $H_1 : \bar{x}_2 > \bar{x}_1$) по таблицам функции Лапласа[1] находят критическое значение $z_{кр.}$ из равенства:

$$\Phi(z_{кр.}) = 1 - 2\alpha \qquad (3.15)$$

Если $z_{эмпир} < z_{кр.}$, то нет оснований считать средние значения двух выборок различными (средние значения двух выборок существенно не различаются).

Если $z_{эмпир} > z_{кр.}$, то средние значения двух выборок различаются с доверительной вероятностью $p = 1 - \alpha$.

Примечание: при отсутствии таблиц, можно сравнивать с коэффициентами вероятности нормального распределения (таблица 3.***1***).

Б) Если имеется две небольшие ($n_1 < 30$ и $n_2 < 30$) независимые выборки из нормальных совокупностей, по которым найдены средние значения и исправленные выборочные дисперсии

[1] Функция Лапласа характеризует вероятность попадания случайной величины (при стандартизированном нормальном распределении) в симметричный интервал от $-z$ до $+z$.

(генеральные дисперсии σ_1 и σ_2 хотя и неизвестны, но предполагаются одинаковыми), то для проверки нулевой гипотезы $H_0 : \bar{x}_1 = \bar{x}_2$ (т.е. средние существенно не различаются) вычисляют эмпирический критерий по формуле:

$$t_{эмпир} = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}} \qquad (3.16)$$

Число степеней свободы при поиске критического (табличного) значения критерия $t_{кр.}$ в таблице Стьюдента $k = n_1 + n_2 - 2$.

Сформулируйте самостоятельно правила проверки гипотез при двусторонней и односторонних альтернативах.

Если на основании проверок гипотез получен вывод, что дисперсии и средние значения двух выборок существенно не различаются, то с некоторой неизвестной вероятностью можно предположить, что наблюдения в двух выборках отобраны из одной генеральной совокупности.

4.5. Сравнение средних графическим способом («ящик с усами»).

Скопировать сравниваемые массивы в STATISTICA. Меню Graphs / 2D Graphs / Box Plots. На вкладке *Advanced* выбрать:

- тип графика (Graph Type) – Box-Whiskers, Multiple;
- переменные (Variables) – зависимые (Dependent) – выбрать оба массива;
- центральная точка (Middlepoint) – среднее (Mean);
- ящик (Box) – доверительный интервал (Conf. interval), вероятность 0,95;
- усы (Whisker) – разброс значений (Min-Max).

Если доверительные интервалы не перекрываются, то средние значения надёжно (с доверительной вероятностью 95 %) различаются между собой. Если перекрываются, то средние статистически не различимы.

4. ДИСПЕРСИОННЫЙ АНАЛИЗ (ДА)

Цель работы: научиться оценивать влияние **выбранного исследователем** фактора (однофакторный ДА) или комбинации факторов (двухфакторный ДА) на изменчивость изучаемой случайной величины.

Теоретические основы

Изменчивость случайной величины вызывается одновременным действием целого ряда причин – факторов. Это внешние условия, влияющие на какое-либо явление. В геологии такими факторами могут являться: глубина образования геологических объектов; вмещающие породы; тип метасоматоза; метод опробования, вес пробы; метод анализа, оборудование; интенсивность процессов выветривания; мощность залежи; пространственное расположение опробуемого участка относительно какой-либо структуры (оси складки, разрывного нарушения, интрузива, расстояние по простиранию рудного тела, по напластованию или поперёк пород) и т.п.

Во многих случаях возникает необходимость оценить меру влияния отдельных факторов и их взаимодействия на изменения исследуемой величины.

Дисперсионный анализ – раздел статистики, изучающий влияние факторов на изменчивость случайной величины. Главной его задачей является выделение тех факторов и их сочетаний, которые оказывают существенное влияние на изменение изучаемой величины.

В зависимости от количества учитываемых факторов различают однофакторный, двухфакторный и многофакторный ДА.

Каждый фактор представляет собой переменную величину, изменяющуюся дискретно или непрерывно, количественно или качественно. Точечные значения дискретно изменяющихся факторов и интервальные значения непрерывных факторов называют уровнями факторов. **Уровни факторов исследователь выбирает сам**, разделяя общее воздействие этой величины на некоторые уровни. Если количество наблюдений изучаемой случайной величины на всех уровнях, по всем факторам одинаково,

дисперсионный анализ называют равномерным, если различное – неравномерным.

Чтобы разделить суммарное влияние всех факторов (как учтённых, так и неучтённых), надо общую дисперсию (σ^2) случайной величины X (на которую действуют взаимно независимые факторы) представить в виде суммы дисперсий σ_i^2, обусловленных влиянием изучаемых (учтённых) факторов и остаточной дисперсии $\sigma^2_{ост}$, вызванной неучтёнными факторами:

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2 + \sigma^2_{ост} \qquad (4.1)$$

Суждение о влиянии (или не влиянии) определённого фактора (или комбинации факторов) на изменчивость изучаемой случайной величины основано на группировке её замеров по факторам и их уровням и проверке гипотезы о равенстве двух дисперсий: дисперсии, обусловленной данным фактором, и остаточной дисперсии, вызванной неучтёнными факторами, т.е. $H_0: \sigma_1^2 = \sigma^2_{ост}$; $\sigma_2^2 = \sigma^2_{ост}$ и т.д.; и $H_1: \sigma_1^2 > \sigma^2_{ост}$; $\sigma_1^2 > \sigma^2_{ост}$ т.д.

Если нулевая гипотеза отвергается, то делается вывод, что выбранный нами фактор оказывает существенное влияние на изменение изучаемого свойства геологического объекта.

I. Однофакторный дисперсионный анализ

Однофакторный дисперсионный анализ применяется для изучения влияния только одного фактора на изменение случайной величины. Например, влияние глубины от поверхности на содержание металла в руде. Содержание металла замерено на разных горизонтах горных выработок (*1, 2, ..., m*). Следовательно, разные горизонты – это уровни фактора «глубина». На каждом горизонте отобрано некоторое число проб *n*. Количество проб на каждом горизонте может быть одинаковым или нет. В зависимости от этого используют равномерный или неравномерный дисперсионный анализ.

Задание 1: Оценить влияние указанного в условии задачи фактора (файл «Массив 4-1») на изменение изучаемого параметра и определить доли влияния учтённых и неучтённых факторов.

Порядок выполнения работы

1. Разбить совокупность из N замеров изучаемой случайной величины на m групп по изучаемому фактору (А) с n_i наблюдениями в каждой группе, составив матрицу наблюдений (таблица 4.1).

Таблица 4.1

Матрица наблюдений с расчётом средних значений

Уровни фактора *A*	**Наблюдения**						**Сумма по строкам**	**Среднее по строкам**
	1	**2**	**…**	***j***	**…**	***n***		
1	x_{11}	x_{12}	…	x_{1j}	…	x_{1n}	Σx_1	$\overline{x_1}$
2	x_{21}	x_{22}	…	x_{2j}	…	x_{2n}	Σx_2	$\overline{x_2}$
…	…	…	…	…	…	…	…	…
i	x_{i1}	x_{i2}	…	x_{ij}	…	x_{in}	Σx_i	$\overline{x_i}$
…	…	…	…	…	…	…	…	…
m	x_{m1}	x_{m2}	…	x_{mj}	…	x_{mn}	Σx_m	$\overline{x_m}$

Примечание к таблицам. Не следует объединять ячейки при работе в Excel, т.к. могут возникнуть ошибки при расчётах с помощью «Анализа данных»

2. Рассчитать средние арифметические значения, по уровням фактора A и общее среднее.

2.1. Среднее арифметическое, полученное по n_i наблюдениям на i-том уровне фактора:

$$\overline{x_i} = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i}, \qquad (4.2)$$

где x_{ij} – j-тое значение наблюдения на i-том уровне фактора A (значения наблюдений в каждой ячейке); n_i – количество наблюдений на i-том уровне.

2.2. Среднее арифметическое всех значений (общее среднее):

$$\overline{x} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} x_{ij}}{N} = \frac{\sum_{i=1}^{m} n_i \overline{x}_i}{N}, \qquad (4.3)$$

где N – общее количество наблюдений (сумма всех n_i).

3. Найти суммы квадратов отклонений (графа 2 таблицы 4.2). Следует помнить, что

$$Q_{общ} = Q_A + Q_{ост},$$

где $Q_{общ}$ – общая сумма квадратов отклонений отдельных наблюдений x_{ij} от общего среднего; Q_A – сумма квадратов отклонений между группами (т.е. расхождение между уровнями, или рассеивание по факторам); $Q_{ост}$ – сумма квадратов отклонений внутри групп (расхождения между наблюдениями *i*-того уровня – остаточное рассеивание за счёт неучтённых факторов).

Таблица 4.2

Схема вычислений при однофакторном дисперсионном анализе

Вид дисперсии	**Сумма квадратов отклонений**	**Число степеней свободы**	**Оценка дисперсии**
1	***2***	***3***	***4***
Межгрупповая по фактору *A*	$Q_A = \sum_{i=1}^{m} n_i(\overline{x}_i - \overline{x})^2$ или при равномерном дисперсионном анализе $Q_A = n\sum_{i=1}^{m}(\overline{x}_i - \overline{x})^2$	$m - 1$	$S_A^2 = \frac{Q_A}{m-1}$
Внутригрупповая (остаточная)	$Q_{\text{ост}} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(x_{ij} - \overline{x}_i\right)^2$	$N - m$	$S_{\text{ост}}^2 = \frac{Q_{\text{ост}}}{N-m}$
Общая	$Q_{\text{общ}} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(x_{ij} - \overline{x}\right)^2$	$N - 1$	$S_{\text{общ}}^2 = \frac{Q_{\text{общ}}}{N-1}$

Примечание к таблице. Для расчётов удобно создать промежуточные таблицы вычислений различных квадратов отклонений.

4. Рассчитать число степеней свободы (графа 3, таблица 4.***2***).

5. Оценить общую дисперсию $S^2_{общ}$, дисперсии между группами S^2_A (уровнями фактора *A* – где каждая группа наблюдений на уровне выступает как одно усреднённое наблюдение) и дисперсию внутри групп $S^2_{ост}$ (графа 4, таблица 4.***2***).

6. Оценить влияние выбранного фактора *A* на изучаемое свойство при помощи *F*-критерия Фишера, эмпирическое значение которого вычисляется по формуле:

$$F_A = \frac{S_A^2}{S_{ост}^2}, \qquad (4.4)$$

Если $F_A > F_{(\alpha,\ m\text{-}1,\ N\text{-}m)\ кр.}$, то нулевая гипотеза отвергается, а исследуемый фактор рассматривается как влияющий на изучаемую случайную величину с доверительной вероятностью $p = 1 - \alpha$.

Если $F_A < F_{(\alpha,\ m\text{-}1,\ N\text{-}m)\ кр.}$, тогда нет основания считать, что фактор А влияет на исследуемую величину.

7. Проверить расчёты с помощью "*Анализа данных*" программы Excel: «*Сервис*» → «*Анализ данных*» → «*Однофакторный дисперсионный анализ*»; в окне дисперсионного анализа параметр "*Группирование*" поставить "*по строкам*". Обозначения в результирующей таблице: *SS* – сумма квадратов отклонений (*Q*); *df* – число степеней свободы; *MS* – оценка дисперсии; *F* – эмпирический критерий Фишера; *Fкритическое* – критическое значение критерия; *p-уровень* – вероятность ошибиться при отклонении нулевой гипотезы.

8. Определить доли, приходящиеся на остаточную дисперсию и дисперсию, обусловленную влиянием фактора *A*, приняв $Q_{общ}$ за 100 %.

Если при проверке гипотезы был сделан вывод о влиянии фактора на изучаемую величину, необходимо охарактеризовать, как влияет фактор на изменение изучаемого свойства геологического объекта, и сделать предположение о причинах такого влияния. ***Сопроводить графиком средних с доверительными интервалами.***

II. Двухфакторный дисперсионный анализ

При двухфакторном дисперсионном анализе изначально определяют 2 фактора, влияние которых на изменчивость измеряемого признака предполагается проверить.

При этом сумма квадратов отклонений от общего среднего разделяется на компоненты, соответствующие двум

предполагаемым факторам A и B и остаточное рассеивание за счёт неучтённых факторов.

Если по фактору A выделено m уровней, а по фактору B – l уровней, то получаем таблицу в m строк и l столбцов. Если в каждой ячейке одно наблюдение, то это так называемый *двухфакторный дисперсионный анализ без повторений*.

Если же для каждого сочетания факторов A_i и B_j произведено n наблюдений (так называемый *двухфакторный дисперсионный анализ с повторением*), то в каждую ячейку этой таблицы на пересечении строки фактора A_i и столбца фактора B_j помещается n значений.

Задание 2. Оценить влияние указанных в условии задачи факторов и их сочетания на изменение изучаемого параметра (Массив 4-2).

Порядок выполнения работы

1. По таблице исходных данных (таблица 4.***3***) рассчитать средние значения для каждой ячейки по формуле:

$$\bar{x}_{ij*} = \frac{\sum_{k=1}^{n} x_{ijk}}{n}, \qquad (4.5)$$

где x_{ijk} – k-тое значение наблюдения на i-том уровне фактора A и j-том уровне фактора B (каждое значение наблюдений); n – количество наблюдений на i-том уровне фактора A и j-том уровне фактора B.

Общее количество ячеек (групп) ml. Общее количество наблюдений $N=nml$.

2. Рассчитать средние значения по уровням факторов A и B и общее среднее.

3. Рассчитать суммы квадратов отклонений, степени свободы и оценить дисперсии (таблица 4.4)

$Q_{ост}$ можно также рассчитать по формуле:

$$Q^2_{ост} = Q^2_{общ} - Q^2_A - Q^2_B - Q^2_{AB} \qquad (4.6)$$

4. Для проверки гипотезы о влиянии на изменчивость изучаемого свойства каждого фактора в отдельности и их сочетания (совместного влияния) рассчитываются эмпирические значения критериев Фишера (таблица 4.***5***) и сравниваются с табличными для соответствующих комбинаций степеней свободы.

5. Проверить расчёты с помощью "*Анализа данных*" программы Excel: *Сервис → Анализ данных → Двухфакторный дисперсионный анализ*. Таблицу исходных данных предварительно надо транспонировать (рис. 4.1) (Копировать → Ctrl+Alt+V (специальная вставка) → Транспонировать). В окне дисперсионного анализа в качестве входного интервала указывается вся таблица с названиями факторов (см. рис. 4.1); в параметре "*Число строк для выборки*" указывается количество замеров на одном уровне факторов (*n*).

Таблица 4.3

Исходные данные с вычисленными средними

		Уровни фактора B						**Сумма и среднее по строкам (по фактору *A*)**	
		B_1	B_2	…	B_j	…	B_l	**Сумма**	**Среднее**
Уровни фактора A	A_1	$x_{111}, x_{112,\ldots}, x_{11k,\ldots}, x_{11n}$	$x_{121}, x_{122,\ldots}, x_{12k,\ldots}, x_{12n}$	…	$x_{1j1}, x_{1j2,\ldots}, x_{1jk\ldots}, x_{1jn}$	…	$x_{1l1}, x_{1l2,\ldots}, x_{1lk\ldots}, x_{1ln}$	Σx_{1*}	$\overline{x_{1*}}$
	A_2	$x_{211}, x_{212,\ldots}, x_{21k,\ldots}, x_{21n}$	$x_{211}, x_{212,\ldots}, x_{22k,\ldots}, x_{21n}$	…	$x_{2j1}, x_{2j2,\ldots}, x_{2jk\ldots}, x_{2jn}$	…	$x_{2l1}, x_{2l2,\ldots}, x_{2lk\ldots}, x_{2ln}$	Σx_{2*}	$\overline{x_{2*}}$
	…	…	…	…	…	…	…	…	…
	A_i	$x_{i11}, x_{i12,\ldots}, x_{i1k,\ldots}, x_{i1n}$	$x_{i21}, x_{i22,\ldots}, x_{i2k,\ldots}, x_{i2n}$	…	$x_{ij1}, x_{ij2,\ldots}, x_{ijk\ldots}, x_{ijn}$	…	$x_{2l1}, x_{2l2,\ldots}, x_{ilk,\ldots}, x_{2ln}$	Σx_{i*}	$\overline{x_{i*}}$
	…	…	…	…	…	…	…	…	…
	A_m	$x_{m11}, x_{m12,\ldots}, x_{m1k,\ldots}, x_{m1n}$	$x_{m21}, x_{m22,\ldots}, x_{m2k,\ldots}, x_{m2n}$	…	$x_{mj1}, x_{mj2,\ldots}, x_{mjk,\ldots}, x_{mjn}$	…	$x_{ml1}, x_{ml2,\ldots}, x_{mlk,\ldots}, x_{mln}$	Σx_{m*}	$\overline{x_{m*}}$
Сумма по столбцам (по фактору *B*)		Σx_{*1}	Σx_{*2}	…	Σx_{*j}	…	Σx_{*l}	Общая сумма Σx_{ijk}	
Среднее по столбцам (по фактору *B*)		$\overline{x_{*1}}$	$\overline{x_{*2}}$	…	$\overline{x_{*j}}$	…	$\overline{x_{*l}}$		Общее среднее $\bar{x}$

Таблица 4.4

Схема вычислений при двухфакторном дисперсионном анализе

Вид дисперсии	**Сумма квадратов отклонений**	**Число степеней свободы**	**Оценка дисперсии**
Межгрупповая по фактору *A*	$Q_A = nl\sum_{i=1}^{m}(\overline{x_{i*}} - \bar{x})^2$	$m-1$	$S_A^2 = \frac{Q_A}{m-1}$
Межгрупповая по фактору *B*	$Q_B = nm\sum_{j=1}^{l}(\bar{x}_{*j} - \bar{x})^2$	$l-1$	$S_B^2 = \frac{Q_B}{l-1}$
Взаимодействия факторов	$Q_{AB} = n\sum_{i=1}^{m}\sum_{j=1}^{l}(\bar{x}_{ij*} - \bar{x}_{i*} - \bar{x}_{*j} + \bar{x})^2$	$(m-1)(l-1)$	$S_{AB}^2 = \frac{Q_{AB}}{(m-1)(l-1)}$
Внутригрупповая (остаточная)	$Q_{ост} = \sum_{i=1}^{m}\sum_{j=1}^{l}\sum_{k=1}^{n}(x_{ijk} - \bar{x}_{ij*})^2$	$ml(n-1) = N - ml$	$S_{ост}^2 = \frac{Q_{ост}}{N-ml}$
Общая	$Q_{общ} = \sum_{i=1}^{m}\sum_{j=1}^{l}\sum_{\kappa=1}^{n}(x_{ijk} - \bar{x})^2$	$N-1$	$S_{общ}^2 = \frac{Q_{общ}}{N-1}$

Таблица 4.5

Расчёт эмпирических критериев Фишера

Эмпирический критерий Фишера	$F_A = \frac{S_A^2}{S_{ост}^2}$	$F_B = \frac{S_B^2}{S_{ост}^2}$	$F_{AB} = \frac{S_{AB}^2}{S_{ост}^2}$
Число степеней свободы для выбора табличного критерия Фишера			
k_1	$m - 1$	$l - 1$	$(m - 1)(l - 1)$
k_2	$ml(n - 1) = N - ml$		

Факторы	A1	A2
B1	1.1	1.5
	1.2	1.3
	1.4	1.4
B2	1.3	1.2
	1.2	1.4
	1.1	1.6

Рис. 4.1. Вид исходных данных и область входного интервала для проведения двухфакторного дисперсионного анализа в Microsoft Excel

Если при проверке гипотезы был сделан вывод о существенном влиянии фактора или сочетания факторов на изучаемую величину, необходимо охарактеризовать, как они влияют на изменение изучаемого свойства геологического объекта, и сделать предположение о причинах такого влияния. Противоположный случай (отсутствие влияния) так же следует объяснить.

Сопроводить графиками средних с доверительными интервалами.

5. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Цель работы: научиться оценивать наличие, тесноту и направленность связи между значениями двух случайных величин.

При решении различных геологических задач часто необходимо совместно рассмотреть несколько случайных величин и проследить за изменением одного признака с изменением другого. В одних случаях изучаемые свойства геологических объектов проявляются независимо друг от друга, а в других между ними могут быть выявлены более или менее отчётливые взаимосвязи. Изменение свойств геологических объектов под действием различных факторов имеет, как правило, статистический характер и практически всегда отличается от функционального. Для их изучения и описания используются двумерные и многомерные статистические модели. Зависимость между признаками может быть **линейной** и **нелинейной** (параболической, экспоненциальной, степенной, синусоидальной и т.д.). В математической статистике взаимосвязь явлений и их признаков изучают методом корреляции.

I. Парная линейная корреляционная зависимость

Для выявления линейной корреляционной зависимости необходимо иметь хотя бы 2 ряда сопряжённых наблюдений случайной величины: 2 признака определяются в одном объекте (например, образце, пробе и т.д.). *Поэтому при выявлении корреляционной зависимости ни в коем случае нельзя сдвигать данные относительно друг друга.*

Графически такую систему двух случайных величин представляют в виде корреляционного графика (диаграммы рассеяния) – облака точек в двумерном пространстве координат, по которым откладываются значения измеренных свойств. Чем ближе точки расположены к прямой линии, тем теснее линейная зависимость. Беспорядочно расположенные точки, образующие более или менее изометричное облако, говорят об отсутствии зависимости между свойствами. Если точки аппроксимируются кривой линией, зависимость нелинейная. Отдельные удалённые от облака точки могут указывать на аномальные значения в выборке. Если точки распадаются на несколько облаков, это обычно

свидетельствует о наличие нескольких совокупностей, которые следует изучать отдельно.

Линейная корреляционная зависимость может быть прямая и обратная. **Прямая корреляция** характеризует такую статистическую зависимость, когда при возрастании одной случайной величины и другая будет в среднем возрастать. При **обратной корреляции** возрастание одной случайной величины приводит в среднем к убыванию другой.

Задание 1. Выявить зависимость между двумя свойствами геологических объектов (файл «Массив 5-1»).

Расчёты требуется произвести "вручную" и проверить с помощью стандартных функций программы Excel. При "ручном" счёте заполняется таблица 5.***1***.

Таблица 5.1

Расчёт коэффициента корреляции

№ п/п	Исходные данные		Степени отклонений и их произведения				
	x_i	y_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
1	x_1	y_1	√	√	√	√	√
…	…	…	…	…	…	…	…
n	x_n	y_n	√	√	√	√	√
∑	√	√	√	√	√	√	√
Сред-нее	$\bar{x}$	$\bar{y}$	$\mu_{1x} = 0$	$\mu_{2x} \approx S_x^2$	$\mu_{1y} = 0$	$\mu_2 \approx S_y^2$	$\mu_{11} \approx K_{xy}$

Примечания: 1) за счёт округлений μ_{1x} и μ_{1y} могут отличаться от нуля; 2) μ_{11} – смешанный центральный момент (**ковариация** без ограничений степеней свободы).

Порядок выполнения

1. Построить (в программах Excel и STATISTICA) точечный график зависимости между свойствами. По графику оценить однородность выборки, линейность зависимости, качественно определить наличие и тесноту связи (тесная, средняя, слабая, отсутствует) и её характер (прямая, обратная).

1.1. Построить (в программах Excel и STATISTICA) диаграмму рассеяния с учётом категорий (точечный график с различными маркерами для разных групп значений). Категории для

такого типа графика выбираются также как фактор и уровни фактора в дисперсионном анализе – самостоятельно исследователем на основе его знаний и опыта.

2. Рассчитать статистические характеристики системы двух случайных величин (среднеарифметические значения, дисперсии, стандартные отклонения, ковариацию, коэффициент корреляции).

Обозначим: x_i – значения одной случайной величины, y_i – значения другой случайной величины.

2.1. Корреляционный момент, или **ковариация**:

$$\mathrm{cov}(x,y) = K_{xy} = \frac{\sum_{i=1}^{n}((x_i - \bar{x})(y_i - \bar{y}))}{n-1} \qquad (5.1)$$

Вывести единицу измерения ковариации.

2.2. Коэффициент парной корреляции (r_{xy} или r):

$$r_{xy} = \frac{K_{xy}}{S_x S_y} \quad \text{или} \quad r_{xy} = \frac{\sum_{i=1}^{n}((x_i - \bar{x}) \cdot (y_i - \bar{y}))}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (5.2)$$

Коэффициент корреляции – мера линейной взаимосвязи между двумя случайными величинами. *Для характеристики нелинейной зависимости непригоден*. Это безразмерная величина, измеряемая в долях единицы. Коэффициент линейной корреляции изменяется в пределах от –1 до +1 и не зависит от точки начала отсчёта и единиц измерения.

По величине коэффициента корреляции судят о силе связи:

- если $r_{xy} = 0$ – линейная зависимость отсутствует, но это не означает, что связи нет, т.к. может быть нелинейная зависимость. При расчёте коэффициента корреляции для реальных случайных величин никогда не получим точно нулевое значение. В связи с этим необходимо проводить проверку значимости отличия коэффициента корреляции от нуля (см. далее п. 4);

- если $|r_{xy}| = 1$ – функциональная линейная зависимость;

- если r_{xy} близок к 1 – тесная прямая (положительная) линейная статистическая зависимость (прямая корреляция);

- если r_{xy} близок к -1 – тесная обратная (отрицательная) линейная статистическая зависимость (обратная корреляция).

3. Проверить полученный коэффициент корреляции с помощью «*Мастера функций*» программы Excel (функция КОРРЕЛ).

4. Убедиться в том, что коэффициент корреляции значимо отличается от нуля (т.е. проверить надежность корреляции).

Как и при проверке других статистических величин формулируется отрицательная нулевая гипотеза: H_0: $r_{x,y} = 0$, (т.е. «коэффициент корреляции статистически не отличается от нуля или, иными словами, значимой линейной корреляции между изучаемыми величинами нет»), при множестве альтернатив H_1: $r_{x,y} \neq 0$.

Проверка осуществляется при помощи критерия Стьюдента:

$$t_{эмпир} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad (5.3)$$

где r – рассчитанный коэффициент парной линейной корреляции.

Для сравнения используют предельные (табличные) значения двустороннего t-критерия Стьюдента с числом степеней свободы $\boldsymbol{k = n-2}$ и выбранным уровнем значимости α.

Если $|t_{эмпир}| > t_{(\alpha,\, k)\, табл}$, связь считается доказанной с доверительной вероятностью $p = 1 - \alpha$. В противном случае линейная зависимость считается не установленной.

При значительном объёме выборки можно в качестве $t_{крит}$ брать значения коэффициента вероятности нормального распределения (см. таблицу 1.3).

Чаще действуют в обратном порядке: при выбранном $t_{крит}$ сразу рассчитывают значимый коэффициент корреляции по формуле:

$$r_{знач} = \frac{t_{крит}}{\sqrt{n-2+t_{крит}^2}} \qquad (5.4)$$

где, задавшись необходимой доверительной вероятностью (уровнем значимости), выбирают $t_{крит}$, либо из таблиц двустороннего

критерия Стьюдента с числом степеней свободы $k = n - 2$, либо из таблиц коэффициентов вероятности нормального распределения.

Так, например, при $t_{крит} = 3$ ($p = 0{,}997$), $r_{знач} = \frac{3}{\sqrt{n+7}}$.

Все коэффициенты корреляции будут значимы с вероятностью $p = 0{,}997$, если $|r| \geq r_{знач.}$. Коэффициенты корреляции меньше $r_{знач.}$ считаются неотличимыми от нуля (приравниваются к нулю).

Формулой (5.4) удобно пользоваться при проверке значимости сразу нескольких коэффициентов корреляции, например, при составлении матрицы коэффициентов корреляции.

Вывести из формулы зависимость $r_{знач.}$ от объёма выборки.

О силе связи судят, разбив интервал от $r_{знач.}$ до 1 на три части. Однако значение коэффициента корреляции меньшее 0,5 лучше считать проявлением слабой зависимости.

5. Обобщить результаты, указать возможные геологические причины выявленных закономерностей.

Задание 2. Рассчитать матрицу коэффициентов корреляции между всеми парами свойств (файл «Массив 5-2») и построить граф (схему) ветвящихся связей для значимых зависимостей; ***указать возможные геологические причины выявленных закономерностей***.

Порядок выполнения работы

1. Рассчитать матрицу коэффициентов корреляции с помощью «*Анализа данных*» программы Excel: *Сервис → Анализ данных → Корреляция*. В окне «Входной интервал» ввести адреса ячеек всей таблицы, включая «шапку»; поставить «галочку» в окне «Метки в первой строке»; в строке «Группировать по» задать «по столбцам».

2. Рассчитать матрицу коэффициентов корреляции с помощью программы STATISTICA: Statistics → Basic Statistics → *Correlation matrices.*

3. Рассчитать величину значимого коэффициента корреляции. Определить диапазоны значений для слабой, средней и сильной связи (равномерно от $r_{знач}$ до 1).

4. Построить схемы ветвящихся связей для переменных со значимыми положительными (а) и отрицательными (б) коэффициентами парной корреляции. Для построения можно воспользоваться онлайн сервисом построения графов, например, https://graphonline.ru/, но не запрещается сделать это вручную.

II. Выявление зависимости между полуколичественными признаками

Корреляция рангов используется в случае, когда признак может быть упорядочен (порядковая шкала измерений).

Если пронумеровать объекты, упорядоченные по какому-либо признаку, то такая совокупность будет называться ранжированной.

содержание:	нет	очень мало	мало	много	очень много
ранг:	1	2	3	4	5

Если несколько наблюдений обладают одинаковым показателем, то их предварительно располагают друг за другом, а затем каждому присваивают исправленный ранг, равный среднему арифметическому их предварительных рангов (таблица 5.2).

Аналогичным образом поступают со вторым признаком, замеренном в том же объекте.

Таблица 5.2

Расчёт исправленных рангов

Содержание	Предварительный ранг	Исправленный ранг
нет	1	(1+2+3)/3=2
нет	2	2
нет	3	2
очень мало	4	4
мало	5	(5+6)/2=5,5
мало	6	5,5
много	7	7
очень много	8	(8+9)/2=8,5
очень много	9	8,5

Затем рассчитывают разность рангов d в каждом наблюдении (таблица 5.3) и собственно ранговый коэффициент корреляции:

$$\rho = 1 - \frac{6 \cdot \sum d^2}{(n^2 - 1) \cdot n} \qquad (5.5)$$

где d – разность между рангами соответствующих признаков; n – количество проб, в которых замерены оба признака.

Ранговый коэффициент корреляции изменяется в пределах от -1 до +1.

Таблица 5.3

Расчёт рангового коэффициента корреляции

Содержание элементов		**Ранг (предварительный)**		**Исправленный ранг**		**Разность рангов $R'_x - R'_y$**	**d^2**
Признак x	**Признак y**	**R_x**	**R_y**	**R'_x**	**R'_y**	**d**	
–	Следы	1	3	1,5	4,0	–2,5	6,25
Следы	Следы	3	4	4,0	4,0	0	0
< 0,001	< 0,001	6	6	6,5	6,5	0	0
–	–	2	1	1,5	1,5	0	0
Следы	–	4	2	4,0	1,5	2,5	6,25
< 0,001	< 0,001	7	7	6,5	6,5	0	0
Следы	Следы	5	5	4,0	4,0	0	0
Сумма	–	–	–	–	–	–	12,50

Оценку значимости рангового коэффициента корреляции проводят, рассчитывая значение коэффициента Стьюдента:

$$t_{эмпир} = \rho \cdot \sqrt{n-1} \qquad (5.6)$$

и сравнивают его с табличным, для $k = n - 1$ и выбранным уровнем значимости α.

Если $|t_{эмпир.}| > t_{(\alpha,\ k)\ табл.}$ – связь между свойствами значимая с вероятностью $p = 1 - \alpha$. В противном случае зависимость считается не установленной.

Задание 3. Выявить зависимость между двумя свойствами геологического объекта (файл «Массив 5-3»).

Порядок выполнения

1. Присвоить признакам предварительные ранги.

2. Рассчитать значения исправленных рангов для признаков.

3. Рассчитать разность между рангами в каждом наблюдении.

4. Вычислить ранговый коэффициент корреляции.

5. Оценить значимость рангового коэффициента корреляции.

6. Найти в программе STATISTICA способ расчёта рангового коэффициента корреляции; проверить с его помощью пп. 4 и 5.

7. Обобщить результаты.

III. Выявление зависимости между качественными признаками

Коэффициент взаимной сопряжённости используется для качественных признаков, которые нельзя упорядочить по какому-либо критерию (синий-красный-зелёный, гранит-известняк, девочки-мальчики, рудный-безрудный и т.п.).

Обозначим: A_1, A_2, …, A_m – первый признак, где m – количество градаций признака A; B_1, B_2, …, B_l – второй признак; где l – количество градаций признака B; n – общее число наблюдений.

Прежде, чем вычислять коэффициент взаимной сопряжённости следует рассчитать эмпирический критерий Пирсона χ^2, по которому сразу же определяют значимость зависимости:

$$\chi^2_{эмпир} = \sum_{i=1}^{n} \frac{(n_{ij} - \tilde{n}_{ij})^2}{\tilde{n}_{ij}} \qquad (5.7)$$

где n_{ij} – эмпирическая частота; $\tilde{n}_{ij}$ – теоретическая частота, рассчитанная по формуле:

$$\tilde{n}_{ij} = \frac{\sum_{i=1}^{l} n_i \cdot \sum_{j=1}^{m} n_j}{n} \qquad (5.8)$$

Оценка производится путём сравнения $\chi^2_{эмпир}$ с табличным критическим значением для уровня значимости α и числа степеней свободы $k = (m - 1)(l - 1)$, где m – количество градаций признака A,

l – количество градаций признака B. Если $\chi^2_{эмпир} > \chi^2_{(\alpha,\ k)\ табл.}$, то связь считается значимой с вероятностью $p = 1 - \alpha$.

Коэффициент взаимной сопряжённости рассчитывается по формуле:

$$K = \sqrt{\frac{\chi^2_{эмпир}}{n\sqrt{(m-1)(l-1)}}} \qquad (5.9)$$

где n – количество проб; m – количество градаций признака A; l – количество градаций признака B; χ^2 – вычисленный ранее критерий Пирсона.

Задание 4. Выявить зависимость между двумя качественными свойствами геологического объекта (файл «Массив 5-4»).

1. Составить таблицу комбинаций признаков (таблица 5.4). Количество проб в ячейках – эмпирические частоты. Для составления удобно воспользоваться сводной таблицей Excel или в программе STATISTICA:

Statistics → Basic Statistics…→ Tables and banners.

2. Рассчитать теоретические частоты (таблица 5.**5**). Теоретические частоты округляют так, чтобы их суммы по строкам и столбцам совпадали с эмпирическими.

Таблица 5.4

Распределение эмпирических частот по комбинациям признаков

		Признак A_j (цвет)			Сумма частот, $\sum n_i$
		A_1 массивный	A_2 слоистый	$...A_m...$ оолитовый	
признак B_i (тип породы)	B_1 известняк	25	5	2	35
	B_2 мергель	5	20	0	25
	B_3 боксит	5	10	15	30
	$...B_l...$ песчаник	5	25	0	30
Сумма частот, $\sum n_j$		40	60	20	120

Таблица 5.5

Расчёт теоретических частот

		Признак A_j (текстура)			Сумма частот, $\sum \tilde{n}_i$
		A_1 массивный	A_2 слоистый	$\ldots A_m \ldots$ оолитовый	
признак B_i (тип породы)	B_1 известняк	$\tilde{n}_{11} = \frac{40 \cdot 35}{120}$	$\tilde{n}_{12} = \frac{60 \cdot 35}{120}$	≈5,8	35
	B_2 мергель	$\tilde{n}_{12} = \frac{40 \cdot 25}{120}$	$\tilde{n}_{12} = \frac{60 \cdot 25}{120}$	≈4,2	25
	B_3 боксит	10	15	≈5	30
	$\ldots B_l \ldots$ диабаз	10	15	≈5	30
Сумма частот, $\sum \tilde{n}_j$		40	60	20	120

3. На основе попарных разностей соответствующих эмпирических и теоретических частот, по формуле (5.7) рассчитать показатель критерия Пирсона χ^2 и оценить значимость будущего коэффициента взаимной сопряжённости.

4. Вычислить коэффициент взаимной сопряжённости.

5. Обобщить результаты, указать возможные геологические причины выявленных закономерностей.

6. Предложить вариант графического оформления результата.

7. Найти ошибки в таблицах 5.4 и 5.5.

IV. Оценка нелинейной зависимости

Корреляционное отношение η характеризует степень нелинейной зависимости между случайными переменными x и y. Оно изменяется в пределах от 0 до 1. При $\eta = 0$ – никакой связи нет. При линейной зависимости $\eta = |r|$.

Для расчёта выборочного корреляционного отношения всё множество значений переменной X разбивается на m групп по количеству одинаковых значений x_i или по интервалам, аналогично построению гистограммы. Для каждой группы рассчитывается своё среднее $\bar{y}_i$ и стандартное отклонение $S_{\bar{y}i}$. Корреляционное отношение рассчитывают по формуле:

$$\eta_{эмпир} = \frac{S_{\bar{y}_i}}{S_y} = \frac{\sqrt{\frac{1}{n} \cdot \sum_{i=1}^{m} n_i (\bar{y}_i - \bar{y})^2}}{S_y}, \qquad (5.10)$$

где n_i – число наблюдений в i-ой группе; m – число групп (интервалов); n – общее число наблюдений; S_y– стандартное отклонение величины Y для всей выборки, рассчитываемое по обычной формуле исправленного выборочного стандартного отклонения.

Проверка значимости корреляционного отношения

Способ 1. По аналогии с коэффициентом корреляции, статистическая значимость отличия корреляционного отношения от нуля может быть проверена с помощью критерия Стьюдента:

$$t_{эмпир} = \frac{\eta \sqrt{n-2}}{\sqrt{1-\eta^2}}, \qquad (5.11)$$

Эмпирическое значение сравнивают с табличным критическим значением двустороннего t-критерия Стьюдента, для $k = n - 2$ и выбранным уровнем значимости α.

Если $|t_{эмпир.}| > t_{(\alpha,\ k)крит.}$ – связь между свойствами значимая с вероятностью $p = 1 - \alpha$. В противном случае зависимость считается не установленной.

Способ 2.

$$t_{эмпир} = \frac{\eta^2_{эмпир}(n-m-2)}{(1-\eta^2_{эмпир})(m-2)} \sqrt{\frac{(m-2)(n-m-4)}{2(n-4)}}, \qquad (5.12)$$

где m – количество классов группирования.

Эмпирическое значение критерия сравнивается с критическим по таблицам функции нормированного нормального распределения.

Например: если $t_{эмпир} > 3$ связь является значимой с доверительной вероятностью 0,997; если $t_{эмпир} > 2$ связь является значимой с доверительной вероятностью 0,954.

Проверка значимости отличия корреляционного отношения от коэффициента линейной корреляции

Для проверки значимости отличия корреляционного отношения от коэффициента линейной корреляции можно рассчитать эмпирическое значение критерия по формуле:

$$t_{эмпир} = \frac{(\eta^2 - r^2)\sqrt{n}}{2\sqrt{\eta^2 - r^2 + (\eta^2 - r^2)^2(2 - \eta^2 - r^2)}}, \qquad (5.13)$$

где r – эмпирический коэффициент линейной корреляции для этих же данных.

Рассчитанный эмпирический критерий сравнивают с табличным критическим значением двустороннего t-критерия Стьюдента, для $k = n - 2$ и выбранным уровнем значимости α.

Если $|t_{эмпир.}| > t_{(\alpha,\ k)крит.}$ – корреляционное отношение значимо отличается от коэффициента корреляции с вероятностью $p = 1 - \alpha$. В противном случае отличие считается не установленным.

Задание 5. Рассчитать корреляционное отношение между двумя свойствами геологического объекта (файл «Массив 5-5») и сделать выводы о характере и силе связи. Для наглядности расчетов корреляционного отношения заполнить таблицу 5.6.

Порядок выполнения

1. Построить точечный график зависимости между свойствами. По графику оценить однородность выборки; тип, характер и тесноту связи.

2. Разбить все наблюдения на группы (равные интервалы) по свойству X.

3. Заполнить таблица 5.6.

Таблица 5.6

Расчёт корреляционного отношения

Группа (интервал)	Частота, n_i	Среднее значение в группе, $\bar{y}_i$	$n_i(\bar{y}_i - \bar{y})^2$
√	n_1	√	√
…	…	…	…
√	n_m	√	√
Сумма	n	-	√

5. Рассчитать корреляционное отношение (формула (5.10)).

6. Проверить значимость эмпирического корреляционного отношения двумя способами.

7. Проверить отличие корреляционного отношения от коэффициента линейной корреляции.

8. Сделать выводы о характере и силе связи.

6. РЕГРЕССИОННЫЙ АНАЛИЗ

Цель работы: научиться рассчитывать уравнения связи (регрессии) между значениями двух независимых случайных величин.

Проведение регрессионного анализа зависимостей значений переменных *X, Y, Z* и т.д. обычно делится на три этапа:

1. Выбор вида функциональной зависимости (вида уравнения).

2. Вычисление коэффициентов выбранного уравнения.

3. Оценка достоверности полученного уравнения.

В качестве вида функциональной зависимости могут быть использованы линейная, параболическая второго порядка, синусоидальная, показательная функции и другие.

Линейная регрессия

Задание 1. Рассчитать уравнения регрессии y на x и x на y, оценить погрешности и качество полученных уравнений. Расчёты ведутся для данных из файла «Массив 6-1».

Порядок выполнения работы и теоретические основы

1. Построить точечный график зависимости и визуально оценить однородность выборки, направление, тесноту и форму связи.

2. Составить уравнения регрессии y на x и x на y.

Если коэффициент корреляции значим и близок к корреляционному отношению, а график эмпирической зависимости

близок к прямой линии, то зависимость между ними линейная и, в общем виде, выражается уравнением: $y = ax + b$.

Поскольку переменных две, то обычно строят две линии зависимости $x = f(y)$ и $y = f(x)$ (можно просто поменять местами столбцы с данными) и таким образом имеем комбинацию из двух уравнений регрессии, описывающую имеющиеся у нас взаимоотношения между переменными:

$$y = a_1 x + b_1 \quad \text{(регрессия } y \text{ на } x\text{)} \tag{6.1}$$

$$x = a_2 y + b_2 \quad \text{(регрессия } x \text{ на } y\text{)} \tag{6.2}$$

Эти линии близки, но все-таки несколько отличаются друг от друга. При отсутствии зависимости они перпендикулярны друг другу.

Найти уравнения таких прямых можно разными способами.

Способ 1. Уравнения линейной регрессии можно выразить через статистические характеристики системы из двух случайных величин:

$$y = \bar{y} + r\frac{S_y}{S_x}(x - \bar{x}), \quad x = \bar{x} + r\frac{S_x}{S_y}(y - \bar{y}) \tag{6.3}$$

Способ 2. Коэффициенты в уравнении регрессии можно найти, используя метод наименьших квадратов. Для расчёта коэффициентов a_1 и b_1 уравнения регрессии y на x применение этого метода даст следующие формулы корней уравнений:

$$a_1 = \frac{\Sigma x_i y_i \cdot n - \Sigma x_i \cdot \Sigma y_i}{\Sigma x_i^2 \cdot n - (\Sigma x_i)^2}; \quad b_1 = \frac{\Sigma x_i^2 \cdot \Sigma y_i - \Sigma x_i \cdot \Sigma x_i y_i}{\Sigma x_i^2 \cdot n - (\Sigma x_i)^2} \tag{6.4}$$

Для расчёта коэффициентов a_2 и b_2 уравнения регрессии x на y, необходимо просто поменять местами x и y в предыдущей паре формул корней уравнения.

3. Проверить расчёты с помощью программы Excel: на графике активизировать точки, подведя курсор к любой из них и нажать правую кнопку мыши; в появившемся окне выбрать пункт «*Добавить линию тренда*»; в окне «*Линия тренда*» выбрать

линейный тип тренда, а на вкладке параметры поставить галочку в строке «*показать уравнение на диаграмме*».

4. Проверить расчёты с помощью «Мастера функций» Excel: 1) выделить интервал для выходных данных размером 2×5; 2) в «Мастере функций» выбрать Статистические → ЛИНЕЙН; 3) в окне функции ввести интервалы значений y и x, аргумент «конст» оставить пустым, аргумент «Статистика» = ИСТИНА. Для выполнения этой функции нажать клавишу F2, а затем — клавиши CTRL+SHIFT+ENTER (одновременно). Результаты будут записаны в виде таблица 6.1.

Таблица 6.1

Расположение результатов расчёта функции «ЛИНЕЙН»

Коэфф. a в уравнении регрессии	Коэфф. b в уравнении регрессии
Стандартная ошибка коэффициента a	Стандартная ошибка коэффициента b
Коэффициент детерминированности	Ошибка прогнозирования
Эмпирический F-критерий	Степень свободы
Регрессионная сумма квадратов	Остаточная сумма квадратов

5. Оценить, насколько точно полученное уравнение описывает наши эмпирические наблюдения. Это можно сделать разными способами.

5.1. Определить величины вкладов остаточной и закономерной дисперсий в общий разброс значений. Для этого надо вначале оценить, насколько велики отклонения эмпирических точек y_i от теоретических $\tilde{y}_i(y(x_i))$, рассчитанных по полученному уравнению.

● По уравнению $y = a_1 x + b_1$ рассчитать теоретические значения $\tilde{y}_i$ для каждого x_i .

● Для каждого $\tilde{y}_i$ рассчитать квадрат разности $(\delta_i)^2$ с соответствующим ему эмпирическим y_i: $(\delta_i)^2 = (y_i - \tilde{y}_i)^2$

● Рассчитать случайную составляющую дисперсии случайной величины y (остаточную дисперсию – S^2_δ):

$$S^2_\delta = \frac{\sum_{i=1}^{n} (y_i - \tilde{y}_i)^2}{n-1} \qquad (6.5)$$

Эта составляющая дисперсии характеризует рассеяние значений случайной величины около линии регрессии (тренда), т.е. разброс точек, необъяснимый нашим уравнением регрессии.

Обратите внимание, что при линейной регрессии дисперсии S^2_δ и S^2_y связаны между собой соотношением:

$$S^2_\delta = S^2_y (1 - r^2) \qquad (6.6)$$

Поскольку | r | не больше 1, остаточная дисперсия всегда меньше общей S^2_y.

● Рассчитать закономерную составляющую дисперсии $S^2_{зак}$. Закономерная составляющая – это дисперсия, поглощённая линией регрессии (тренда).

Так как общую дисперсию случайной величины y (S^2_y) можно разложить на две составляющие: случайную S^2_δ и закономерную $S^2_{зак}$, то между ними справедливо соотношение:

$$S^2_y = S^2_\delta + S^2_{зак} \qquad (6.7)$$

Следовательно, закономерная составляющая дисперсии составит:

$$S^2_{зак} = S^2_y - S^2_\delta \qquad (6.8)$$

● Определить величины вкладов остаточной и закономерной дисперсий в общий разброс значений. Для этого, приняв S^2_y за 100 %, рассчитать доли, приходящиеся на $S^2_{зак}$ и S^2_δ. Чем больший процент приходится на закономерную составляющую, тем лучше зависимость.

5.2. Определить величину достоверности аппроксимации R^2 (**коэффициент детерминированности**). Он показывает, насколько хорошо наше уравнение описывает эмпирическую зависимость. Чем больше его величина, тем точнее уравнение.

$$R^2 = 1 - \frac{S^2_\delta}{S^2_y} = 1 - \frac{\sum_{i=1}^{n} (y_i - \tilde{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \qquad (6.9)$$

Если величину R^2 умножить на 100 %, то это будет процентный вклад дисперсии, объясняемой уравнением регрессии ($S^2_{зак}$), в общую дисперсию.

Для проверки расчётов можно воспользоваться функцией «ЛИНЕЙН» (см. п. 4) или графиком зависимости, где на вкладке «*Параметры*» окна «*Линия тренда*» поставить галочку в строке «*поместить на диаграмму величину достоверности аппроксимации R^2*».

5.3. Проверить значимости уравнения регрессии с помощью критерия Фишера.

Проверка проводится на основе суммы квадратов отклонений исходных данных от их среднего и суммы квадратов отклонений исходных данных от данных, рассчитанных по уравнению:

$$F_{эмпир} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}(n-2) \qquad (6.10)$$

Степени свободы при поисках табличного критического значения критерия Фишера $k_1 = 1$; $k_2 = n - 2$.

5.4. Качество полученного уравнения также можно оценить при помощи критерия разностного ряда. При этом, как и при расчёте остаточной дисперсии, вычисляют отклонения $\delta_i = \tilde{y}_i - y_i$ для каждой пары $\tilde{y}_i$ и y_i, и рассчитываются среднее значение отклонения $\bar{\delta}_i$ и стандартное отклонение отклонений $S(\delta_i)$. Далее рассчитывается $t_{эмпир}$ и сравнивается с табличным для $k = n - 1$ степеней свободы и выбранным уровнем значимости по таблицам двустороннего критерия Стьюдента:

$$t_{эмпир} = \frac{\overline{\delta_i} \cdot \sqrt{n}}{S(\delta_i)} \qquad (6.11)$$

Если $|t_{эмпир.}| < t_{таб(\alpha,\, k)}$, различие между эмпирическими и рассчитанными значениями несущественно.

Если $|t_{эмпир.}| > t_{таб(\alpha,\, k)}$, различие между эмпирическими и рассчитанными значениями существенно с вероятностью $p = 1 - \alpha$.

Если различие несущественно, то уравнение можно использовать для определения значений Y по значениям X. Если же критерий показывает существенное различие, то, возможно, необходимо выбрать другой тип зависимости.

5.5. Рассчитать **погрешность прогнозирования** (погрешность уравнения регрессии).

Основное назначение уравнения регрессии – прогноз. Так как зависимость носит статистический характер, прогнозирование по уравнению регрессии будет сопровождаться погрешностью (ε_y), которую можно вычистить по формуле (6.12) и добавить к уравнению регрессии:

$$\varepsilon_y = tS_\delta = tS_y\sqrt{1-r^2}\ , \qquad (6.12)$$

В качестве коэффициента t используют табличные значения двустороннего t-критерия Стьюдента с числом степеней свободы $k = n - 1$ и выбранным уровнем значимости α.

При значительном объёме выборки можно в качестве t брать значения коэффициента вероятности нормального распределения (таблица 3.1), ***которые следовало запомнить***.

Таким образом, полностью, уравнение линейной регрессии, например, Y на X, используемое для прогноза, будет иметь вид:

$$y=ax + b \pm \varepsilon_y \qquad (6.13)$$

6. Рассчитать корреляционное отношение $\eta = R$

$$\eta = \sqrt{1-\frac{S_\delta^2}{S_y^2}} = \sqrt{1-\frac{\sum_{i=1}^{n}(y_i-\tilde{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \qquad (6.14)$$

Обратите внимание: **корреляционное отношение — это корень квадратный из коэффициента детерминированности**.

Кроме того, при линейной регрессии корреляционное отношение равно коэффициенту корреляции.

Чем больше коэффициент корреляции по абсолютной величине, тем меньше погрешность прогнозирования. Считается, что для надёжного прогноза, следует использовать только такие зависимости, для которых коэффициент корреляции больше 0,87.

Необходимо понимать, что ***любые уравнения зависимости надёжно работают только в области имеющихся эмпирических данных***. При использовании уравнений для прогноза данных на

такую область значений X, для которой нет экспериментальных значений Y, погрешность уравнения и вероятность ошибки резко возрастают. Наличие аномальных значений сильно искажает уравнение линейной регрессии, что, в свою очередь, сказывается на точности прогноза. Следовательно, необходимо исключать аномальные значения.

Если выборка неоднородна, то возможно наилучшее решение – разделить выборку на однородные совокупности и подобрать для каждой из них отдельное уравнение.

7. Построить графики и проверить расчёты в программе STATISTICA.

Нелинейная регрессия

Задание 2. Подобрать уравнение нелинейной зависимости между двумя свойствами (файл «Массив 6-2»).

1. Построить точечный график зависимости и визуально оценить однородность выборки, направление, тесноту и форму связи. При наличии аномальных значений – удалить их из выборки. При явной неоднородности выборки уравнение подбирается для каждой однородной совокупности.

2. Подбор уравнения нелинейной зависимости выполнить с помощью программы Excel: на графике активизировать точки, подведя курсор к любой из них и нажав правую кнопку мыши; в появившемся окне выбрать пункт «*Добавить линию тренда*»; в окне «*Линия тренда*» выбрать подходящий тип тренда, а на вкладке параметры поставить галочки в строках «*показать уравнение на диаграмме*» и «*поместить на диаграмму величину достоверности аппроксимации R^2*». Наилучшему уравнению будет соответствовать наибольший R^2.

Промежуточные построения и расчёты сохранять в качестве обоснования принимаемых решений.

3. Построить графики и проверить расчёты в программе STATISTICA.

7. КОНТРОЛЬ ОПРОБОВАНИЯ

Цель работы: научиться выявлять случайные и систематические ошибки химического анализа проб.

Теоретические основы

При геологических работах отбираются различные пробы. Пробы служат для изучения качества полезного ископаемого, вмещающих пород, концентратов, хвостов обогатительных фабрик и т.п., оконтуривания рудных тел, подсчёта запасов и т.д. Опробование включает несколько операций: 1) отбор проб; 2) обработку проб; 3) испытание проб (анализ). На каждой из этих операций могут возникать погрешности, которые подразделяют на: 1) случайные; 2) систематические; 3) промахи (грубые ошибки, например, при нумерации проб, описки при переписывании данных и т.п.). При обнаружении грубых ошибок, по возможности, следует заново отбирать пробы, либо исключать такие данные из обработки и дальнейших расчётов.

1. Случайные погрешности не устранимы, имеют разный знак и величину, при вычислении среднего содержания они обычно компенсируются, следовательно, их влияние на величину среднего значения невелико, но при оконтуривании могут сильно исказить контур рудных тел.

Случайные погрешности химического анализа оцениваются с помощью так называемого **внутреннего контроля** лаборатории. Он заключается в том, что вместе с основными пробами в ту же лабораторию направляются контрольные пробы обязательно в зашифрованном виде. Число контрольных проб должно составлять 3-10 % от числа основных, но не менее 20 – 30 проб. Одновременный анализ в той же лаборатории обеспечивает равноточность основных и контрольных проб, что является необходимым условием внутреннего контроля.

Для выявления случайной погрешности находят абсолютную и относительную погрешности анализов. Оценка **абсолютной среднеквадратичной случайной погрешности** анализов (случайной погрешности, стандартной ошибки) выполняется по формуле:

$$\delta_{сл.} = \sqrt{\frac{\sum_{i=1}^{n}(x-y)^2}{2n}}, \qquad (7.1)$$

где x – содержания в основных пробах; y – содержания в контрольных пробах; n – число контрольных (или основных) проб.

Зная среднеквадратичную погрешность, можно определить **относительную случайную погрешность** химического анализа:

$$\tau_{сл.} = \frac{2\delta_{ан}}{\bar{x}+\bar{y}}100\% \qquad (7.2)$$

Величина допустимой относительной случайной погрешности регламентирована инструкцией ГКЗ (Государственной комиссии по запасам). Если случайная погрешность превысит допустимую, результаты химических анализов непригодны для оконтуривания рудных тел и подсчёта запасов руд. Анализы следует выполнить вновь по всем основным пробам.

Согласно некоторым методическим указаниям рекомендуется выделять классы точности анализов в зависимости от отношения z вычисленной случайной погрешности к допустимой (таблица 7.1). Для подсчёта запасов главных компонентов руд пригодны только анализы I-III классов точности.

Таблица 7.1

Классы точности анализов

Класс точности анализов	Значение коэффициента z	Примечание
I	Меньше 0,33	Количественные анализы
II	0,33-0,5	Количественные анализы
III	0,5-1	Количественные анализы
IV	1-2	Количественные анализы
V	Больше 2	Полуколичественные анализы

2. Систематические погрешности постоянны по знаку и величине в каждой пробе, следовательно, вносят ошибку при вычислении среднего содержания и, в дальнейшем, в подсчёт запасов. Эти ошибки обусловлены постоянным влиянием какого-

либо фактора на результаты опробования, в ряде случаев их можно выявить и устранить.

Для выявления систематической погрешности химических анализов наиболее распространён **внешний контроль**, когда контрольные анализы выполняются в другой лаборатории более точным методом, чем основные. На внешний контроль направляют по 20 – 30 зашифрованных проб, обычно это составляет 3-5 % от основных анализов. Обработка и анализ основных и контрольных проб проводится в одно время, т.к. некоторые руды со временем окисляются с изменением состава.

Выявление систематической погрешности в работе химических лабораторий проводится на основе уравнения регрессии.

Обозначим: x – данные одной (основной) лаборатории; y – данные другой (контрольной) лаборатории, полученные в результате проведения внешнего контроля анализов.

Исследование результатов внешнего контроля анализов лучше начать с построения точечного графика зависимости контрольных анализов от основных. В случае совпадения данных точки будут расположены на биссектрисе угла XOY, т.е. вдоль линии y = x. Однако под влиянием случайной погрешности анализов точки группируются около биссектрисы, образуя эллипс рассеяния. При наличии систематической погрешности эллипс смещается от биссектрисы в ту или иную сторону.

Способ 1. Чтобы решить эту задачу аналитически, находят уравнение регрессии $y = ax + b$ и с помощью критерия t оценивают степень отличия коэффициента a от единицы, а коэффициента b от нуля:

$$t_a = \frac{|a-1|}{S_a},\ t_b = \frac{|b|}{S_b},\ \text{где}\ S_a = \frac{S_y}{S_x}\sqrt{\frac{1-r^2}{n-2}},\ S_b = S_a\sqrt{S_x^2 + \bar{x}^2} \qquad (7.3)$$

Если хотя бы один из критериев будет больше трёх, то систематическое расхождение установлено с вероятностью 0,997.

Способ 2. Известен и другой более простой, но менее эффективный способ решения этой же задачи, основанный на сравнении средних содержаний в основных и контрольных пробах по формуле:

$$t = \frac{|\bar{x} - \bar{y}|}{S}, \text{ где } S = \sqrt{\frac{S_x^2 + S_y^2 - 2rS_xS_y}{n}} \qquad (7.4)$$

Если $t > 3$, то систематическое расхождение установлено с вероятностью 0,997.

При установлении систематической погрешности, если есть возможность, основные анализы должны быть выполнены вновь, в противном случае к результатам анализов вводят поправки путем пересчета основных анализов по уравнению регрессии на уровень контрольных анализов или применяют поправочные коэффициенты с предварительной группировкой анализов по классам содержания.

Задание 1: По результатам анализов основных и контрольных проб (Файл «Массив 7-1») оценить случайную погрешность химического анализа, сравнить относительную случайную погрешность с допустимым значением и сделать вывод о пригодности результатов анализа для оконтуривания рудных тел и подсчёта запасов руд.

Порядок выполнения работы

1. Построить точечный график, по которому оценить визуально наличие случайной погрешности. На график добавить линию $y = x$; оси сделать равными по длине, цене деления и крайним значениям.

2. Найти абсолютную случайную погрешность по формуле (7.1).

3. Найти относительную случайную погрешность по формуле (7.2).

4. Сравнить относительную случайную погрешность с допустимой, определить класс точности и сформулировать выводы.

Задание 2: По результатам анализов основных и контрольных проб (Файл «Массив 7-2») определить наличие систематической погрешности химического анализа, оценить ее величину, при необходимости составить уравнение регрессии и пересчитать результаты анализов основных проб на уровень контрольных.

Порядок выполнения работы

1. Построить точечный график, по которому оценить визуально наличие систематической погрешности. На график добавить линию $y = x$; оси сделать равными по длине, цене деления и крайним значениям.

2. Составить уравнение регрессии (любым способом).

3. Оценить значимость погрешности разными способами.

4. При наличии систематической погрешности составить уравнение зависимости основных проб от контрольных и пересчитать с его помощью результаты анализов основных проб на уровень контрольных.

5. Сформулировать выводы.

6. Предложить альтернативные варианты способов внутреннего и внешнего контроля опробования и способов их анализа.

8. ПРИМЕНЕНИЕ МНОГОМЕРНОЙ РЕГРЕССИИ ДЛЯ ОПРЕДЕЛЕНИЯ ЭЛЕМЕНТОВ ЗАЛЕГАНИЯ РАЗРЫВНОГО НАРУШЕНИЯ

Цель работы: научиться рассчитывать уравнения зависимости между несколькими случайными величинами.

Теоретические основы

Во многих случаях приходится изучать зависимость одной величины от нескольких других, так называемую многофакторную, или множественную зависимость. Уравнение, устанавливающее зависимость между признаком X_1 (функцией) и несколькими другими X_2, X_3, …, X_m (аргументами), называется уравнением множественной регрессии. В общем случае его можно записать в

виде $X_1 = f(X_2, X_3, \ldots, X_m)$. Различают линейную и нелинейную множественную регрессию. В случае линейной зависимости уравнение регрессии имеет вид $X_1 = a_1x_2+a_2x_2+\ldots+a_mx_m+c\pm\varepsilon$, где x_2, $x_2,\ldots x_m$ – переменные аргументы; a_1, a_2, $\ldots a_m$ и c – постоянные коэффициенты, которые требуется найти. ***Вспомните, что такое ε.*** Этому уравнению соответствует, так называемая, гиперплоскость, т.е. плоскость *m*-мерного пространства.

Для трёх переменных *X*, *Y*, и *Z* зависимость *Z* от независимых *X* и *Y* можно записать в виде уравнения $z = ax + by + c$. Геометрически это уравнение выражает обычную плоскость в трёхмерном пространстве. Коэффициенты уравнения надо вычислить 3 способами.

Расчёт коэффициентов уравнения регрессии с помощью программного пакета STATISTICA

1. Открыть программу STATISTICA.

2. Скопировать исходные данные (координаты *X*, *Y*, *Z*).

3. В главном меню выбрать *Statistics* → *Multiple Regression*.

4. Выбрать (кнопка *Variables*) зависимую *Dependent* (*Z*) и независимые *Independent* (*X*, *Y*) переменные.

5. В окне *Multiple Regression Results* выбрать кнопку *Regression Summary*. В появившемся окне (рис. 8.**1**) взять необходимые для построения уравнения регрессии значения: *R* (*MultipleR*) – коэффициент множественной корреляции; *Std. (standard) Error of Estimate* – погрешность уравнения регрессии (ε); в столбце *B*: в строке *Intercept* – свободный член уравнения (*c*); в строке VAR1 – коэффициент *a* при переменной *X*; в строке VAR2 – коэффициент *b* при переменной *Y*.

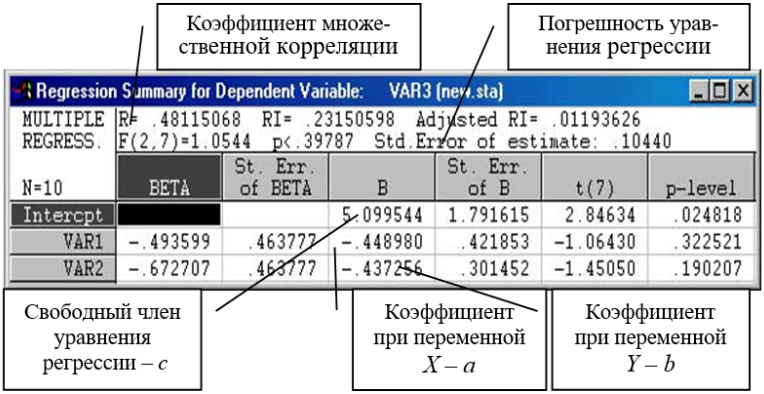

Рис. 8.1. Окно Multiple Regression Results программного пакета STATISTICA

Расчёт коэффициентов уравнения регрессии с помощью электронных таблиц Excel

1. Выбрать: *Данные* → *Анализ данных* → *Регрессия.*

2. В появившемся окне поставить галочку в поле «*Метки*» (“*Labels*”); в поле «*Входной интервал Y*» (“*Input Y Range*”) ввести интервал значений зависимой переменной (*Z*); в поле «*Входной интервал X*» (“*Input X Range*”) – интервал значений независимых переменных (*X* и *Y*), ***интервалы выделять вместе с названиями переменных,***→ OK.

3. Из первой таблицы берутся значения коэффициента множественной корреляции *R* (*Multiple R*) и погрешности уравнения регрессии (*ε*) (*Standard Error*); из третьей таблицы в первом столбце в строке *Y-пересечение* (*Intercept*) – свободный член уравнения (*c*); во второй строке (*X* или *Переменная X1* (*X Variable1*)) – коэффициент *a* при переменной *X* в уравнении регрессии; в третьей строке (*Y* или *Переменная X2* (*X Variable2*)) – коэффициент *b* при переменной *Y*.

Расчёт коэффициентов уравнения регрессии вручную

1. Уравнение линейной регрессии для двух независимых переменных *X* и *Y* можно записать в виде уравнения:

$$z - \bar{z} = A(x - \bar{x}) + B(y - \bar{y}), \qquad (8.1)$$

где $A = \frac{r_{xz} - r_{xy} r_{yz}}{1 - r_{xy}^2} \cdot \frac{S_z}{S_x}$; $B = \frac{r_{yz} - r_{xy} r_{xz}}{1 - r_{xy}^2} \cdot \frac{S_z}{S_y}$ (8.2)

2. Подстав значения A, B и средние значения в уравнение (8.1) и раскрыв скобки, получить уравнение вида $z = ax + by + c \pm \varepsilon$.

3. Найти коэффициент множественной корреляции (см. корреляционное отношение Лаб. раб. 6).

4. Вычислить погрешность уравнения (см. Лаб. раб. 6).

***Задание 1**:* Разрывное нарушение пересечено скважинами в точках с координатами X, Y, Z (файл «Массив 8-1»). Так как скважины искривляются в процессе бурения и замеры искривлений сопровождаются погрешностями, то измеренные точки пересечения скважин с разрывным нарушением не лежат в одной плоскости, а колеблются около неё. Этому способствует так же естественные неровности поверхности нарушений. Требуется аппроксимировать поверхность разрывного нарушения плоскостью, определить её элементы залегания, построить и структурный план разрывного нарушения.

Порядок выполнения работы

1. Рассчитать коэффициенты в уравнении регрессии $z = ax + by + x + \varepsilon$ разными способами.

2. По полученным данным составить уравнение зависимости высотных отметок Z разрывного нарушения от координат местности X и Y.

3. С помощью найденного уравнения $z = ax + by + c$ рассчитать теоретические значения координаты Z.

4. На листе миллиметровки или чертёжной бумаги нанести скважины по координатам X и Y в масштабе, указанном в задании (X – направление на Север; Y – на Восток). Рядом с каждой скважиной указать эмпирические и теоретические значения координаты Z (например, в виде дроби). Не забыть указать масштаб (численный и линейный) и сделать условные обозначения.

5. По теоретическим значениям Z определить элементы залегания разрывного нарушения.

6. Построить структурный план разрывного нарушения (план в стратоизогипсах). Масштаб плана и сечение стратоизогипс указаны в варианте задания.

9. РАСЧЁТ ИНФОРМАТИВНЫХ СВОЙСТВ В УРАВНЕНИИ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

Цель работы: научиться определять среди всех измеренных свойств наиболее значимые (информативные) для описания многомерной зависимости.

Теоретические основы

Так как влияние свойств, а, следовательно, и коэффициентов в уравнении регрессии, на выбранную зависимую переменную может быть различной, следует выявить из независимых свойств наиболее информативные (значимые). Для этого рассчитывают по формуле (9.1) остаточную дисперсию S_m^2 (дисперсию отклонений эмпирических наблюдений от линии регрессии) с учётом степеней свободы $k = m + 1$, где m – количество свойств в уравнении множественной регрессии:

$$S_m^2 = S_\delta^2 \frac{n}{n - m - 1} \qquad (9.1)$$

При увеличении числа учитываемых свойств дисперсия S_m^2 вначале убывает, потом достигает минимума и далее начинает расти. Когда дисперсия достигнет минимума, информативные свойства определены. Дальнейшее увеличение числа случайных величин (измеренных свойств) приведёт к росту дисперсии и внесёт искусственный «шум» в результаты прогнозирования по уравнению регрессии.

Задание 1. По данным файла «Массив 9-1» изучить влияние свойств на содержание указанного в задании зависимого компонента, выступающего в роли функции *y*, и выбрать среди независимых свойств наиболее информативные.

Порядок выполнения работы

1. Рассчитать, с учётом всех признаков, коэффициенты уравнения множественной регрессии, коэффициент множественной корреляции и погрешность прогнозирования любым способом.

2. Составить уравнение регрессии; рассчитать для него остаточную дисперсию по формуле (6.5) и остаточную дисперсию с учётом степеней свободы по формуле (9.1).

3. Рассчитать дисперсию зависимого компонента y.

4. Составить матрицу коэффициентов парной корреляции.

5. Выбрать компонент x_1, имеющий самый высокий по модулю коэффициент парной корреляции r_1 с прогнозируемым компонентом y. Этот компонент будет являться наиболее информативным.

6. Рассчитать остаточную дисперсию S^2_δ (дисперсию отклонений), не учтённую линией регрессии y на компонент x_1, по формуле:

$$S^2_\delta = S^2_y(1 - r_1^2) \qquad (9.2)$$

7. Рассчитать остаточную дисперсию с учётом степеней свободы по формуле (9.1), для $m = 1$, т.е. учитывая только одно свойство.

8. К ведущему свойству x_1 поочерёдно по одному присоединить другие компоненты и рассчитать уравнения регрессии y на каждые два компонента (x_1 и x_2; x_1 и x_3 ;…; x_1 и x_m). По всем полученным уравнениям найти теоретические значения y, дисперсии отклонений S^2_δ и остаточную дисперсию с учётом степеней свободы S^2_m при $m = 2$. Из всех полученных дисперсий $S^2_{m=2}$ выбрать наименьшую и сравнить её с остаточной дисперсией $S^2_{m=1}$. Если наименьшая из $S^2_{m=2}$ окажется меньше $S^2_{m=1}$, то второй по силе влияния признак найден. В противном случае вычисления прекращают.

9. Если второй по силе влияния независимый признак найден, то к наиболее информативной паре свойств поочерёдно по одному присоединить оставшиеся компоненты и расчёты повторить для трёх независимых переменных. Операцию повторять до тех пор, пока остаточная дисперсия не начнёт возрастать.

10. Найти уравнение регрессии, учитывающее только информативные свойства, погрешность прогнозирования для него и коэффициент множественной регрессии.

11. Сравнить полученный результат с уравнением множественной регрессии, рассчитанным в STATISTICA и Excel.

12. Сформулировать выводы.

10. КЛАСТЕРНЫЙ АНАЛИЗ

Цель работы: научиться классифицировать (разделять на однородные группы) объекты с помощью кластерного анализа.

Теоретические основы

Кластерный анализ служит для классификации многомерных объектов в более или менее однородные группы, т.е. для разделения их на группы (кластеры) со сходными характеристиками. При этом предполагается, что степень сходства объектов, объединяемых в один класс, должна быть существенней сходства между объектами различных классов.

Пусть имеется n объектов, у которых измерено m свойств. Требуется иерархически их расклассифицировать.

Множество данных образует матрицу $n \times m$. Геометрическим выражением кластеров являются облака точек в m-мерном признаковом пространстве. Минимальное количество объектов в кластере принимают равным двум. При построении облаков необходимо выбрать наиболее оптимальный масштаб по координатным осям (стандартизовать или нормализовать данные) – это позволяет учитывать каждую переменную с одинаковым весом.

Между каждой парой объектов вычисляется некоторая мера сходства. Это может быть или коэффициент корреляции r_{ij}, или различные дистанционные коэффициенты d_{ij} (меры расстояний), например, стандартизованное (взвешенное) m-мерное евклидово расстояние, которое учитывает не только модуль расстояния, но и характер пространства. Внутри отдельного кластера эта мера сходства максимальна, т.е. объекты имеют наибольшие положительные коэффициенты корреляции r_{ij} или наименьшие расстояния d_{ij}.

Составляется матрица сходства: либо матрица расстояний между всеми точками (симметричная матрица, размером $n \times n$), либо матрица коэффициентов корреляции, размером $m \times m$.

Задача кластерного анализа сводится к разбиению матрицы мер сходства изучаемых объектов на группы (кластеры) таким образом, чтобы внутри кластеров объединялись объекты с наивысшими значениями характеристик сходства, а разобщённые группы оставались бы при этом максимально изолированы. Для этого надо найти самое маленькое расстояние между двумя объектами (если используется мера расстояния) и объединить их в кластер. Далее их рассматривают как один объект (чаще всего заменяя средним значением). Матрица вычисляется снова, при этом число переменных уменьшается (матрица уменьшается на 1 столбец за счёт вычисления расстояния от усреднённого значения до всех других точек). Снова выбираются объекты с самыми маленькими расстояниями, группируются и т.д. Повторяя эту операцию многократно можно свести всю матрицу в 1 столбец.

Объединять всё в единый кластер бессмысленно – при резком увеличении расстояния объединения следует прекратить кластеризацию данных.

Если в качестве мер сходства используются коэффициенты корреляции (обычно для классификации свойств), то наибольшее сходство будет при самом высоком положительном значении r. В качестве граничного значения можно выбрать $r_{знач}$.

Результаты кластерного анализа изображают в виде древовидного графа – **дендрограммы** (рис. 10.1), в которой по одной оси располагают символические обозначения объектов исследования, например, химических элементов или номеров проб, а по другой оси – значения мер сходства (коэффициентов корреляции или мер расстояний).

По дендрограмме визуально выделяют кластеры, выбрав некоторое значение меры сходства. Кроме вышеприведённых рекомендаций при выборе этого значения имейте в виду, что получившуюся классификацию надо объяснить с геологической точки зрения.

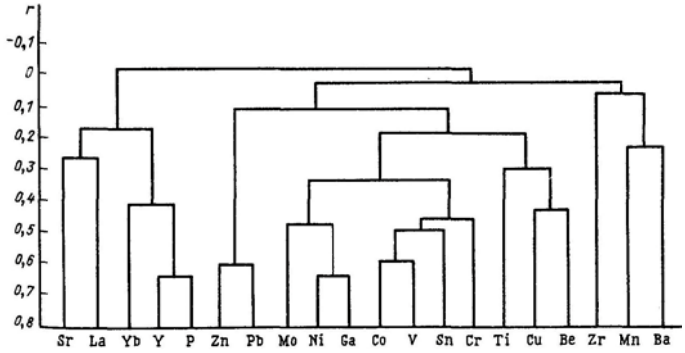

Рис. 10.1. Дендрограмма

Задание 1. Провести иерархическую классификацию данных (файл «Массив 9-1») с помощью кластерного анализа.

Порядок выполнения работы

1. Скопировать исходные данные в STATISTICA.

2. В главном меню выбрать: *Statistics* → *Multivariate Exploratory Techniques* → *Cluster Analysis.*

3. В окне *Clustering Method* выбрать *Joining (tree clustering)*.

4. В окне *Cluster Analysis: Joining…* задать переменные (кнопка *Variables*); в списке *Input File* выбрать *Raw data.*

5. На вкладке *Advanced* в списке *Cluster* выбрать *Variables (columns)* – классификация свойств (переменных); в списке *Amalgamation (linkage) rule* (правило связывания) выбрать любое и посмотреть, чем отличаются; в списке *Distance measure* (мера сходства) выбрать любое[1] и посмотреть, чем отличаются; в переключателе *MD deletion* (удаление ошибочных данных) выбирается способ исправления ошибочных данных либо *Casewise* (удаление), либо *Mean Substitution* (замена на среднее значение).

6. Построить дендрограмму (кнопка *OK*).

[1]При кластеризации химических компонентов часто в качестве меры сходства берут коэффициент корреляции. В программе STATISTICA ему соответствует мера сходства (*"Distance measure")* «*1 - Pearsonr*». Но «Эвклидово расстояние» («*Euclidean distances*») часто тоже даёт хорошие результаты. ***Главное – суметь объяснить преподавателю особенность каждой.***

7. По дендрограмме определить в какие группы объединяются свойства объектов (выделить ассоциации химических элементов).

8. Вернуться в окно *Cluster Analysis: Joining*. На вкладке *Advanced* в списке *Cluster* выбрать *Cases (rows)* – классификация наблюдений (проб) и снова построить дендрограмму.

9. По дендрограмме выделить кластеры проб (группы проб, схожие по уровням содержаний химических элементов).

10. Для этих групп рассчитать основные статистики. При достаточном объёме провести корреляционный анализ.

11. Сравнить выделенные группы проб; в том числе графически.

12. Сформулировать выводы. Объяснить полученную классификацию с точки зрения геологии.

11. ФАКТОРНЫЙ АНАЛИЗ

Цель работы: научиться проводить факторный анализ методом главных компонент и интерпретировать полученные результаты.

Теоретические основы

Факторным анализом называют некоторое множество вычислительных процедур для обработки многомерных статистических данных. В геологии обычно используют его частный случай – метод главных компонент.

Проведение факторного анализа имеет смысл тогда, когда объем исходных данных весьма значителен и использование простейших статистических процедур не позволяет быстро разобраться в существующих внутри них закономерностях. Считается, что минимальными требованиями являются: число переменных (свойств) – не менее 5, число наблюдений (проб) – не менее 30-50. Число наблюдений обязательно должно быть больше числа переменных.

С увеличением количества переменных возрастает размерность признакового пространства. Следовательно, возрастают

трудности изучения геологических объектов, и возникает проблема замены многочисленных наблюдаемых признаков меньшим их числом без существенной потери полезной информации.

При проведении факторного анализа заранее предполагается, что в наборе многомерных наблюдений имеется скрытая простая структура, которая выражается через дисперсии и ковариации переменных, а также с помощью мер сходства между наблюдениями. А все многообразие корреляционных связей можно объяснить действием нескольких обобщённых факторов (причин), являющихся функциями исследуемых свойств. При этом сами обобщённые факторы могут быть и неизвестны, однако их можно выразить через исследуемые свойства. Факторы, выраженные через исходные свойства, можно интерпретировать как некоторые существенные внутренние характеристики объектов.

Кроме того, факторный анализ может быть применён в задачах группировки и классификации объектов. Он позволяет группировать объекты со сходными сочетаниями свойств и группировать свойства с общим характером изменения от объекта к объекту.

Для проведения факторного анализа информация должна быть представлена в виде двумерной таблицы чисел размерностью $n \times m$. Строки этой матрицы должны соответствовать объектам наблюдений (i=1, 2,…, n), столбцы – свойствам (j=1, 2,…, m). Каждое свойство является как бы статистическим рядом, в котором значения варьируют от объекта к объекту. Если признаки, характеризующие объект наблюдения, имеют различную размерность, матрицу исходных данных обычно нормируют, вводя единый масштаб. Это устраняет влияние размерности и обеспечивает сопоставимость свойств.

Математический аппарат факторного анализа разработан для выборок с нормальным распределением.

Разновидностью факторного анализа является **метод главных компонент** или компонентный анализ, который реализует модель вида:

$$t_j = a_{j1}F_1 + a_{j2}F_2 + ... + a_{jm}F_m \qquad (11.1)$$

где t_j– j-й признак (величина случайная); F_1, F_2, ..., F_m – общие факторы (величины случайные, имеющие нормальный закон распределения); a_{j1}, a_{j2}, ..., a_{jm} – факторные нагрузки, характеризующие существенность влияния каждого фактора (параметры модели, подлежащие определению), m – количество признаков (измеренных свойств).

Факторные нагрузки характеризуют величину влияния того или иного общего фактора на вариации данного признака. Основная задача факторного анализа – определение факторных нагрузок.

В данной модели каждое из наблюдаемых свойств линейно зависит от m некоррелированных между собой новых компонент (факторов) F_1, F_2, ..., F_m.

Результатом компонентного анализа, так же как и факторного, является **матрица факторных нагрузок**. Поиск факторного решения – это ортогональное преобразование матрицы исходных переменных, в результате которого каждый признак может быть представлен линейной комбинацией найденных m факторов, которые называют главными компонентами. Т.е. исходные данные преобразуются в главные компоненты, которые не зависят друг от друга и несут определённую смысловую геологическую информацию. Преобразование носит линейный характер и сводится к переносу и вращению систем координат в многомерном признаковом пространстве. Центр тяжести облака точек переносится в начало координат, а поворот производится таким образом, чтобы оси многомерного эллипсоида, охватывающего облако точек, совпали с осями координат. Оси эллипсоида ранжируются по длине, и та координатная ось, которая совпала с наиболее длинной осью эллипсоида, называется первой, следующая по длине – второй и т.д. Новые координаты точек облака после переноса и поворота системы координат и называются **значениями главных компонент** (значениями факторов, главными компонентами).

В процессе вращения происходит перераспределение дисперсий свойств, но сумма дисперсий остаётся постоянной. Первая из компонент (первый фактор) должна учитывать максимум суммарной дисперсии признаков; вторая (второй фактор) – не

коррелировать с первой и учитывать максимум оставшейся дисперсии, и так до тех пор, пока вся дисперсия не будет учтена. Сумма учтённых всеми компонентами (всеми факторами) дисперсий равна сумме дисперсий исходных свойств.

Если для дальнейшего анализа оставить все найденные *m* компонент, то тем самым будет использована вся информация, заложенная в корреляционной матрице. Однако это нецелесообразно. Максимальная дисперсия оказывается сосредоточена в нескольких (двух-трёх) первых главных компонентах, которые тем самым несут основную информацию, и по которым делаются все дальнейшие выводы. Минимальной дисперсией обладают последние компоненты, они несут малую информацию, и ими можно пренебречь. Происходит как бы сосредоточение информации в небольшом числе независимых переменных величин – главных компонент. Существуют различные критерии для оценки числа оставляемых компонент. Чаще всего оставляют столько компонент, чтобы суммарная дисперсия, учитываемая ими, составляла заранее установленное число процентов (обычно не менее 70 %).

В таблице факторных нагрузок факторы располагают в порядке убывания их вклада в общую дисперсию. Обычно удаётся объяснить природу лишь первых двух-трёх факторов. При объяснении факторов надо понимать, что на изучаемые величины (чаще всего содержания элементов) могут влиять факторы, связанные с пробоотбором и анализом.

Задание 1. Провести факторный анализ методом главных компонент для массива данных (файл «Массив 9-1»).

Порядок выполнения работы

Подготовка данных

1. Убедиться в отсутствии пустых, нулевых и текстовых ячеек в таблице исходных данных.

2. Проверить соответствие распределений всех признаков нормальному закону. Если переменные имеют логнормальное распределение, прологарифмировать их.

3. Проверить и исключить аномальные значения. Аномальные значения можно выявить, проведя факторный анализ для всего массива данных (см. ниже). Затем исключить их и провести повторно факторный анализ для оставшейся части.

4. Скопировать данные в STATISTICA.

Факторный анализ

1. Главное меню *Statistics* → *Multivariate Exploratory Techniques* → *Factor Analysis*.

2. Выбирать переменные для факторного анализа (клавиша *Variables*). Остальные параметры окна: *Input file* = *Row Data*; *MD deletion* = *Casewise* → *OK*.

3. В окне *Define Method of Factor Extraction* в графе *Maximum no*. *of factor* необходимо указать количество выделяемых факторов (соответствует числу свойств). В графе *Minimum eigenvalue* указать минимальное значение рассчитываемых собственных чисел (задаём 0.000, чтобы получить собственные числа для каждого фактора).

4. На вкладке *Advanced* указать, каким методом будет проводиться факторный анализ (метод главных компонент – *Principal components*) → *OK*.

5. Через окно *Factor Analysis Results* получаем результаты факторного анализа:

5.1. На вкладке *Quick* или *Explained Variance* клавиша *Eigenvalues* выдаёт окно *Eigenvalues* (**собственные числа**), в котором результаты приведены в колонках: *Value* – номер фактора в порядке убывания значимости; *Eigenvalue* – значение собственного числа по каждому фактору; *% Total Variance* – доля (процент) общей изменчивости переменных, приходящаяся на данный фактор; *Cumulative Eigenvalue* – накопленные значения собственных чисел; *Cumulative %* – накопленный процент общей изменчивости.

5.2. На вкладке *Explained Variance* клавиша *Screeplot* позволяет построить график изменения значений собственных чисел от наиболее значимых факторов к наименее значимым.

5.3. На вкладке *Quick* или *Loadings* клавиша *Summary. Factor loadings* выдаёт **значения факторных нагрузок** на каждую переменную, которые отражают степень и направленность влияния

каждого фактора на свойства (*Factor rotation* оставить в режиме *Unrotated*).

5.4. На вкладке *Quick* клавиша *Plot of factorloadings, 2D* или на вкладке *Loadings* клавиша *Plot of loadings, 2D* выводит графическую модель действия факторов на каждое свойство. Необходимо построить графики в координатах выбранных наиболее информативных факторов. Например, если значимых факторов 3, то надо построить три графика в следующих координатах: *Factor 1 - Factor 2; Factor 1 - Factor 3; Factor 2 - Factor 3*.

5.5. На вкладке *Scores* клавиша *Factor scores* выводит координаты точек (проб) в новом, факторном пространстве (**таблица значений факторов**, или значений главных компонент).

5.6. Построить графики распределения точек (проб) в координатах выбранных наиболее значимых факторов: например, 1-го и 2-го; 1-го и 3-го; 2-го и 3-го и т.д. (допускается в Excel).

Интерпретация результатов

1. Выбрать наиболее значимые факторы, те, которые сильно влияют на изменчивость исходных данных. По таблице собственных чисел выбирают факторы, суммарный вклад которых в дисперсию исходных данных составляет не менее 70 % (70-80 %). Обычно такой вклад приходится на 2-3, реже 4, иногда 5-6 первых фактора. Хотя существуют геологические объекты и ситуации, когда, несмотря на относительно невысокую нагрузку, наиболее логически объяснимым может оказаться и любой из последних факторов. Выбрать значимые факторы можно и по графику собственных чисел (так называемый, метод «каменной осыпи»). Здесь наиболее значимые факторы выбираются по точке резкого перегиба.

2. Таблица и графики факторных нагрузок показывают вклад каждой из исходных переменных в выделенный фактор. При интерпретации результатов таблицы факторных нагрузок каждый фактор рассматривается независимо от других. Из всех коэффициентов вектора факторной нагрузки рассматриваемого фактора выбираются максимальные по абсолютному значению. Каждый фактор указывает, что есть какая-то причина, которая способствует одновременному накоплению в пробах тех свойств, у которых эти коэффициенты положительные, и удалению тех

свойств, у которых эти коэффициенты отрицательные. Иногда для осознания геологической сущности фактора приходится знаки “+” и “–“ менять местами; стоит учитывать наибольшие по абсолютному значению коэффициенты. Необходимо чётко понимать, что результатами факторного анализа являются числа, а выводы исследователь должен сделать сам. Геолог, исходя из знания природы изучаемого объекта, сам решает, что может привести к подобной комбинации свойств или наблюдений и на основе этого может дать геологическую интерпретацию тому или иному фактору.

Диаграммы факторных нагрузок строят в координатах значимых факторов (например, первого и второго, первого и третьего, второго и третьего и т.д.). Точками на таких диаграммах являются исходные свойства (например, химические элементы). Поскольку расчёты производятся на основе корреляционной или ковариационной матрицы, то переменные, точки которых находятся рядом друг с другом на такой диаграмме, обладают высокими положительными коэффициентами парной корреляции между собой. И наоборот, элементы, точки которых наиболее удалены друг от друга, обладают высоким отрицательным коэффициентом корреляции. Таким образом, могут выделяться характерные ассоциации химических элементов (литофильные, халькофильные и т.п.)

3. Таблица значений факторов позволяет изобразить размещение точек изучаемых проб в новых координатах. Для этого выбирают два каких-либо ведущих фактора (например, первый и второй, первый и третий, второй и третий), по значениям которых строят точечные графики. Таблица и графики значений факторов показывают насколько серьёзно выделенный фактор влияет на положение точки конкретного наблюдения (пробы) в общем облаке данных. Возможно и обратное влияние: резко выделяющаяся аномальная точка может создать крайне высокую дисперсию и, как следствие, высокие факторную нагрузку на переменную и значение фактора для этой пробы. На построенных графиках точки нередко образуют скопления, что создаёт предпосылки для разделения объектов на группы. Это позволяет предполагать наличие нескольких различных по составу групп проб (наблюдений).

Наблюдения, точки которых находятся рядом на этих диаграммах, обладают примерно одинаковыми уровнями значений исходных переменных. Например, в случае обработки результатов химического анализа, соседние точки соответствуют пробам с примерно одинаковыми содержаниями тех элементов, у которых нагрузки по этим факторам максимальные (по абсолютному значению).

На построенных графиках необходимо выделить группы проб, определить их номера, усреднить их значения и сравнить друг с другом (аналогично лабораторной работе 10). Если отдельные облака не выделяются, но наблюдается одно вытянутое в каком-либо направлении облако, то в разных его частях можно проследить изменение свойств (например, состава) под действием факторов.

4. Сравнить результаты классификации кластерным и факторным анализами.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

Основная литература:

1. Поротов Г.С. Математические методы моделирования в геологии / Г.С. Поротов. СПб: Изд-во Санкт-Петербургского горного института, 2006. 223 с.

2. Смоленский В.В. Статистические методы обработки экспериментальных данных. Учебное пособие / В.В. Смоленский. СПб:. Изд-во Санкт-Петербургского горного института, 2003. 101 с.

Дополнительная литература:

3. Боровиков В. STATISTICA. Искусство анализа данных на компьютере: для профессионалов */ В. Боровиков*. СПб.: Питер, 2003. 688 с.

4. Дэвис Дж. Статистический анализ данных в геологии. В 2 книгах / пер. с англ. В.А. Голубевой. – М.: Недра, 1990. Книга 1 – 319 с. Книга 2 – 427 с.

5. Каждан А.Б. Математические методы в геологии / А.Б. Каждан, О.И. Гуськов. М.: Недра, 1990. 251 с.

6. Макарова Н.В. Статистика в Excel: Учебное пособие / Н.В. Макарова, В.Я. Трофимец. М.: Финансы и статистика, 2006. 368 с.

7. Миллер Р.Л. Статистический анализ в геологических науках / Р.Л. Миллер, Дж.С. Канн. М.: Мир, 1965. 482 с.

8. Родионов Д.А. Статистические решения в геологии / Д.А. Родионов. М.: Недра, 1981. 231 с.

9. Справочник по математическим методам в геологии. / Д.А. Родионов, Р.И. Коган, В.А. Голубева и др. – М.: Недра, 1987. 335 с.

10. Шарапов И.П. Применение математической статистики в геологии / И.П. Шарапов. М.: Недра, 1971. 248 с.

11. Шестаков Ю.Г. Математические методы в геологии: Учебное пособие для студентов геологических специальностей / Ю.Г. Шестаков. Красноярск: Изд-во Красноярского университета, 1988. 208 с.

СОДЕРЖАНИЕ

СТАТИСТИЧЕСКИЕ МЕТОДЫ В ГЕОЛОГОРАЗВЕДОЧНОЙ ПРАКТИКЕ

Методические указания к лабораторным работам для студентов специальности 21.05.02

Сост. *Я.Ю. Бушуев*

Печатается с оригинал-макета, подготовленного кафедрой геологии и разведки месторождений полезных ископаемых

Ответственный за выпуск *Я.Ю. Бушуев*

Лицензия ИД № 06517 от 09.01.2002

Подписано к печати 26.10.2020 . Формат 60×84/16.
Усл. печ. л. 5,1. Усл.кр.-отт. 5,1. Уч.-изд.л. 4,8. Тираж 75 экз. Заказ 748.

Санкт-Петербургский горный университет
РИЦ Санкт-Петербургского горного университета
Адрес университета и РИЦ: 199106 Санкт-Петербург, 21-я линия, 2