Учреждение образования
«Гомельский государственный университет имени Франциска Скорины»

Факультет математики и технологий программирования
кафедра фундаментальной и прикладной математики

| | |
|---|---|
| СОГЛАСОВАНО | СОГЛАСОВАНО |
| Заведующий кафедрой<br>фундаментальной<br>и прикладной математики | Декан<br>факультета математики<br>и технологий программирования<br>С.П. Жогаль |
| _____ Л.Н. Марченко | _____ |
| 27.04.2023г. | 27 04 2023 г. |

УЧЕБНО-МЕТОДИЧЕСКИЙ КОМПЛЕКС
ПО УЧЕБНОЙ ДИСЦИПЛИНЕ

# СТАТИСТИЧЕСКИЕ МЕТОДЫ В ГЕОЛОГИИ

**для специальности переподготовки**
**1-51 02 71 Разработка и эксплуатация нефтяных**
**и газовых месторождений**
**(квалификация: Горный инженер)**

Составители:
Л. Н. Марченко, кандидат технических наук, доцент

Рассмотрено и утверждено
на заседании кафедры фундаментальной и прикладной математики 27.04.2023
Протокол № 9

Рассмотрено и утверждено
на заседании научно - методического совета университета
29.04.2023 г., протокол № 8

## Содержание

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА ..... 3
1. ТЕОРЕТИЧЕСКИЙ РАЗДЕЛ ..... 6
1.1. Тематические планы лекций ..... 6
1.2. Краткий конспект лекций ..... 7
1 Математическое моделирование геологических объектов ..... 7
2 Предварительный анализ данных ..... 8
3 Точечные оценки и интервальные оценки ..... 10
4 Проверка статистических гипотез ..... 16
5 Дисперсионный анализ ..... 24
6 Корреляционный анализ ..... 29
7 Регрессионный анализ ..... 34
8 Кластерный анализ ..... 35
3 ПРАКТИЧЕСКИЙ РАЗДЕЛ ..... 43
3.1 Перечень практических занятий ..... 43
3.2 Задания для выполнения практических работ ..... 43
4 РАЗДЕЛ КОНТРОЛЯ ЗНАНИЙ ..... 46
4.1 Вопросы к зачету учебной дисциплине «Статистические методы в геологии» ..... 46
4.2 Критерии оценок результатов учебной деятельности студентов по учебной дисциплине «Статистические методы в геологии» (на основании письма Министерства образования Республики Беларусь от 28.05.2013 г. № 09- 10/53-ПО) ..... 47
4.3 Самостоятельная работа по дисциплине «Статистические методы в геологии» ..... 51
5 ВСПОМОГАТЕЛЬНЫЙ РАЗДЕЛ ..... 52
5.1 Учебная программа по учебной дисциплине «Статистические методы в геологии» ..... 52

## ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Электронный учебно-методический комплекс (ЭУМК) по учебной дисциплине «Статистические методы в геологии» представляет собой комплекс систематизированных учебных, методических и вспомогательных материалов, предназначенных для использования в образовательном процессе специальности 1-25 02 71 Разработка и эксплуатация нефтяных и газовых месторождений. ЭУМК разработан в соответствии со следующими нормативными документами.

1. Положением об учебно-методическом комплексе на уровне высшего образования, утвержденном постановлением Министерства образования Республики Беларусь от 08.11.2022 № 427.

2. Требованиями Образовательного стандарта высшего образования ОСВО ОСРБ 1-51 02 71-2016. Переподготовка руководящих работников и специалистов, имеющих высшее образование. Специальность: 1-51 02 71 Разработка и эксплуатация нефтяных и газовых месторождений. Квалификация: Горный инженер

3. Учебно-тематическим планом переподготовки в соответствии с типовым учебным планом переподготовки, утвержденным 01.06.2021, регистрационный номер № 110. Специальность (код и наименование) 1-51 02 71 «Разработка и эксплуатация нефтяных и газовых месторождений». Квалификация горный инженер

Данные документы предусматривают получение слушателями теоретических и практических навыков статистического анализа данных в геологии; осмысление показателей, характеризующих многомерные данные; усвоение форм и методов анализа данных, принятие грамотных статистических решений. В документах заложена задача формирования у слушателей профессиональных и личностных компетенций в области анализа данных в геологии.

В Республике Беларусь с развитием цифровой экономики возрастает количественная и качественная потребность в подготовке квалифицированных специалистов, умеющих работать с данными. Слушателям необходимы знания и практические навыки в области анализа данных, их классификации, установления взаимосвязей между различными геологическими показателями.

Целью электронного учебно-методического комплекса «Статистические методы в геологии» является обеспечение теоретической и практической подготовки слушателей, активизации их учебно-познавательной деятельности по вопросам статистического анализа данных в геологии, развитие профессионально-личностных компетенций в данной сфере, совершенствование умений и навыков самостоятельной учебной работы.

ЭУМК способствует успешному осуществлению учебной деятельности, дает возможность планировать и осуществлять работу обучающихся, обеспечивает рациональное распределение учебного времени по темам учебной дисциплины и совершенствование методики проведения занятий.

Задачами учебно-методического комплекса являются:

– раскрытие требований к содержанию учебной дисциплины «Статистические методы в геологии»;

– обеспечение успешного усвоение теоретического материала по дисциплине,

– актуализированные использования традиционных форм и методов контроля знаний и стимулировать инновационные подходы к проверке и оценке знаний, умений и навыков слушателей;

– повышение уровня творческой активности слушателей через выполнение различных видов творческих заданий и проблемно-поисковых задач.

В структурном отношении электронный учебно-методический комплекс «Статистические методы в геологии» включает в себя четыре раздела: теоретический, практический, раздел контроля знаний, вспомогательный.

*Теоретический* раздел содержит основные положения, выносимые на лекции, и включающий в себя 4 темы, предназначенных для аудиторной работы с слушателями (лекции преподавателя – 10 часа). Здесь слушатели получают представление методах многомерного статистического анализа данных, а именно: предварительном анализе данных, проверке статистических гипотез, кластерном анализе, дисперсионном анализе, корреляционном анализе, регрессионном анализе. Материал к лекциям позволяет также сориентировать слушателям в понятийном аппарате статистических методов, осмыслить их применение в практической деятельности. Также приводится материал, предназначенный для самостоятельного изучения слушателями.

В каждой теме представлены наиболее важные положения дисциплины, позволяющие более глубоко познать методы использования статистических методов. При подготовке теоретической части использовалась учебная и научная литература из области статистического анализа данных.

Деятельность в сфере информационных технологий и программирования рассматривается в контексте групповой, а иногда и индивидуальной деятельности субъекта. Подготовка специалиста в области геологии определяется с одной стороны практической значимостью статистического математического аппарата для изучения всех дисциплин специализации, а с другой стороны – той ролью, которую играет математика в совершенствовании общей культуры мышления и формирования мировоззрения. Общепрофессиональный багаж будущий специалист в области прикладной математики может приобрести также через тот материал, который определяет место специалиста в его профессиональной деятельности и формировании для него личностных свойств и качеств как специалиста данной сфере. Такая деятельность требует не только теоретических основ статистического анализа данных, но и практических навыков их использования в различных направлениях, поскольку данные методы являются основой интеллектуального анализа данных и машинного обучения.

*Практический раздел* включает в себя в соответствии с учебным планом дисциплины 8 часов практических занятий. Каждая тема занятия состоит из перечня заданий для письменного контроля знаний (уровень воспроизведения),

заданий творческого характера (уровень применения знаний на практике). Дается список основной и дополнительной литературы для самостоятельного изучения и более глубокой подготовки к практическим занятиям. Для формирования умений и навыков статистического анализа предполагается выполнение индивидуальных домашних заданий (уровень применения умений на практике). Предполагаются различные формы работы на практических занятиях: устный опрос, решение индивидуальных заданий, групповая консультация и другие. Указанные формы работы способствуют не только усвоению знаний и их репредуктивному воспроизведению, но и дать статистические выводы по исследуемым процессам, глубже понять взаимосвязи между рассматриваемыми явлениями.

Раздел *контроля знаний* представлен критериями оценок результатов учебной деятельности слушателей по дисциплине «Статистические методы в геологии» для специальности переподготовки 1-25 02 71 Разработка и эксплуатация нефтяных и газовых месторождений. Здесь приводятся вопросы к зачету. В данном разделе размещен также образец тестовых заданий, предназначенных для проверки уровня академических компетенций слушателей по дисциплине. Они составлены в логической последовательности, и охватывают все темы учебной дисциплины. Представленные задания являются авторскими, и составлены на основе содержащегося в учебно-методическом комплексе материала к лекциям, а также дополнительных источников, рекомендуемых антам для самостоятельного изучения. Задания включают в себя:

а) констатирующую часть, которая описывает ситуацию,

б) процедурную часть, содержащую предложения антам выполнить какие-либо конкретные действия (правильно определить понятие, выбрать правильный ответ при характеристике терминологического понятия и др.,

в) элемент выборы – варианты для выбора правильных ответов, которые формулируются в форме истинных или ложных высказываний.

*Вспомогательный раздел* содержит необходимые элементы учебно-программной документации: учебную программу по дисциплине «Статистические методы в геологии» учреждения образования с пояснительной запиской и содержанием учебного материала. Кроме того, в данном разделе имеется дополнительный материал, который может быть использован при чтении лекций, проведении практических занятий. Этот дополнительный материал включает в себя: список использованных литературных источников.

# 1. ТЕОРЕТИЧЕСКИЙ РАЗДЕЛ

## 1.1. Тематические планы лекций

**Тема 1 Математическое моделирование геологических объектов**

Геологические объекты и их свойства. Свойства геологических объектов. Основные виды математических моделей, применяемых в геологии. Основные этапы и модели статистического анализа.

**Тема 2 Предварительный анализ данных**

Ранжированный ряд. Вариационный ряд. Частота и относительная частота. Полигон частот. Кумулянта. Интервальный ряд. Гистограмма. Ящик с усами.

**Тема 3 Точечные и интервальные оценки**

Меры положения. Медиана. Мода. Среднее значение. Меры разброса. Дисперсия. Среднеквадратическое отклонение. Меры формы. Моменты случайной величины. Асимметрия. Эксцесс.

**Тема 4 Проверка статистических гипотез**

Общая схема проверки гипотез. Возможные ошибки при проверке гипотез. Сравнение математических ожиданий в двух выборках. Сравнение дисперсий в двух выборках. Проверка гипотезы о виде законов распределения вероятности генеральной совокупности. Критерии согласия.

**Тема 5 Дисперсионный анализ**

Уровни фактора. Результативный фактор. Однофакторный дисперсионный анализ. Критерий Фишера. Выборочный коэффициент детерминации. Двухфакторный дисперсионный анализ без повторений и с повторениями. Общая вариация. Критерий Фишера. Непараметрический дисперсионный анализ.

**Тема 6 Корреляционный анализ**

Основы корреляционного анализа. Выборочный парный коэффициент корреляции. Шкала Чеддока. Корреляционная матрица. Значимость парного коэффициента корреляции. Выборочный частный коэффициент корреляции и его значимость. Выборочный множественный коэффициент корреляции и его значимость.

**Тема 7 Регрессионный анализ**

Эндогенные и экзогенные переменные. Основные предпосылки модели множественной линейной регрессии. Метод наименьших квадратов (МНК). Анализ адекватности линейной модели множественной регрессии. Коэффициент детерминации модели. Значимость коэффициентов регрессии. Проверка гипотезы об адекватности модели в целом. $F$-критерий Фишера

**Тема 8 Кластерный анализ**

Измерение близости объектов. Характеристики близости: линейное расстояние, евклидово расстояние, расстояние Чебышева, Манхэттенское расстояние. Расстояние между кластерами. Дендрограмма.

## 1.2. Краткий конспект лекций

### 1 Математическое моделирование геологических объектов

Необходимость применения моделей при описании природных объектов связана с тем, что геологические системы управляются одновременно многими факторами различной физической природы и не поддаются строгому количественному описанию. В отличие от закона, имеющего характер абсолютной истины, модель дает лишь приближенное представление об объекте, точнее, о тех его свойствах, для изучения которых осуществлялось моделирование. Создание геолого-математической модели осуществляется в следующей последовательности:

1) Получение исходных данных об объекте или явлении путем измерения и определения его свойств.

2) Создание геологической модели объекта и формулировка геологической задачи.

3) Выражение поставленной задачи в математической форме. Создание математической модели. При этом может возникнуть необходимость в получении дополнительных данных или в уточнении геологических представлений об объекте.

4) Математические расчеты в соответствии с принятой моделью.

5) Проверка соответствия полученных результатов фактическим данным. Если геологических моделей было несколько (это обычный случай), можно оценить, какая из них лучше соответствует действительности.

Поскольку полученная модель учитывает лишь отдельные свойства объекта, ее можно последовательно усложнять и детализировать. Чем сложнее модель, тем более достоверно она отражает изучаемый объект и позволяет более надежно прогнозировать его свойства. Однако в реальных условиях существует оптимальная степень сложности математических моделей, которая определяется с учетом требований к точности решения поставленной задачи. Степень сложности модели может также ограничиваться возможностями аналитических решений и электронно-вычислительной техники.

Таким образом, в геологии моделируются не сами объекты, а изменчивость их свойств, наблюдаемая на данном уровне изучения объекта. Характер этой наблюдаемой изменчивости зависит не только от природы явления, но и от детальности геологических исследований и методики их проведения. В связи с этим необходимо рассмотреть понятие геологической совокупности.

Под *геологической совокупностью* понимают множество геологических объектов, объединенных каким-либо признаком. Такую геологическую совокупность мы будем называть изучаемой. Понятно, что далеко не вся изучаемая совокупность доступна нам для наблюдения. Геологу чаще всего приходится довольствоваться лишь отдельными обнажениями, характеризующими часть изучаемого объекта. Отсюда ясно, что необходимо различать изучаемую и опробуемую совокупность и всегда отдавать себе отчет

в том, насколько вторая представительна по отношению к первой. В том случае, когда обнаженность объекта позволяет произвольно формировать опробуемую совокупность, объем ее и степень представительности определяется, исходя из имеющихся данных и личного опыта геолога. Однако источник возможных ошибок не ограничивается несовпадением изучаемой и опробуемой совокупности. Последняя также не может быть исследована в полном объеме. Геолог обычно ограничивается определенным количеством образцов, проб, замеров и т.д. Множество всех произведенных над опробуемой совокупностью наблюдений образует *выборочную совокупность*, или просто *выборку*. Очевидно, что выборочная совокупность во много раз меньше опробуемой. В то же время именно по результатам выборочных наблюдений делаются выводы не только по опробуемой, но и по всей изучаемой совокупности. Это обстоятельство всегда надо иметь в виду, делая какие-либо выводы, иначе самые точные вычисления не спасут от ошибки.

К выборочным данным предъявляются следующие требования:

1) выборка должна состоять из наблюдений, полученных в одинаковых условиях;

2) наблюдения должны быть независимы друг от друга.

Возможность распространения выводов, полученных по выборочным данным, на всю изучаемую совокупность обеспечивается применением методов математической статистики.

*Статистика* – это наука, изучающая закономерности, которым подчинены массовые случайные явления. Из этого определения следует, что использование математической статистики для моделирования свойств геологических объектов возможно лишь в том случае, если геологические наблюдения удовлетворяют условию массовости (то есть, их можно многократно повторять при одних и тех же условиях), могут быть представлены в виде схемы случайных событий и выражены случайной величиной. Проведение геологических исследований обычно заключается в замерах значений изучаемого свойства в произвольных точках пространства. Эти замеры можно поэтому рассматривать как серию случайных событий, а получаемые результаты - числовые значения - как случайные величины, поскольку их невозможно предсказать заранее. Замеры эти можно повторять многократно. Следовательно, явления, изучаемые в процессе геологических исследований, могут рассматриваться как случайные и массовые и для них правомерно использование статистических методов.

## 2 Предварительный анализ данных

Основная задача визуального представления данных – это разделение первичных данных, полученных на выборке, на классы или категории с целью получить обобщенную упорядоченную картину, позволяющую их анализировать.

*Таблица исходных данных* – это таблица, строки которой соответствуют объектам, столбцы измеренным признакам. В ячейках таблица находятся данные, соответствующие значениям измеренных признаков.

Пусть дана выборка $X$ объема $n$ из генеральной совокупности, характеризующей некоторый признак. Если выборка объема $n$ содержит $k$ различных элементов $x_1, x_2, x_3, \ldots, x_k$, причем $x_i$ встречается $n_i$ раз, то значение $x_i$ называется вариантой, число $n_i$ – *частотой* варианты $x_i$. Сумма всех частот равна $n$: $\sum_{i=1}^{k} n_i = n$.

Отношение $f_i = n_i / n$ называется *относительной частотой* варианты $x_i$ ($1 \le i \le k$). Сумма всех относительных частот равна единице: $\sum_{i=1}^{k} f_i = 1$.

Таблица – Ряды (таблицы) представления данных

| |
|---|
| ***Несгруппированный ряд*** – упорядочивание значений выборки по величине от минимального до максимального (для данных, измеренных в порядковой, интервальной, абсолютной шкалах). |
| ***Ранжированный ряд*** – присваивание ранга $R$ элементам несгруппированного ряда.<br>*Правила ранжирования:*<br>1 наименьшему числовому значению приписывается ранг 1;<br>2 наибольшему числовому значению приписывается ранг, равный количеству ранжируемых величин;<br>3 если несколько исходных числовых значений несгруппированного ряда оказались равными, то им приписывается ранг, равный среднеарифметическому тех рангов, которые эти величины получили бы, если бы они стояли по порядку друг за другом и были бы не равны;<br>4 общая сумма рангов должна быть равна<br>$S = \frac{n(n+1)}{2}$. |
| ***Вариационный ряд*** – характеризует частоту встречаемости каждого значения в выборке<br>*Правило построения:*<br>1 определяются варианты (различные значения) $x_1$, $x_2$, …, $x_k$ выборки и располагаются в порядке возрастания;<br>2 определяются частоты $n_1$, $n_2$, …, $n_k$, т. е. подсчитывается, сколько раз встречается каждая варианта;<br>3 определяются относительные частоты $f_1 = n_1/n$, $f_2 = n_2/n$, …, $f_k = n_k/n$, указывающие долю наблюдений, приходящихся на то или иное значение признака. |

| |
|---|
| ***Интервальный ряд*** – характеризует частоту встречаемости значений выборки по интервалам (разрядам) значений признака.<br>*Правило построения*:<br>1 определяется размах выборки $R = x_{\max} - x_{\min}$;<br>2 определяется число интервалов $m$, $m$ – целое (на практике обычно принимают $m \approx \sqrt{n}$);<br>3 определяется длина каждого интервала $h = R / m$ и границы интервалов;<br>4 подсчитывается частоты $n_1$, $n_2$, …, $n_m$, т. е. количество значений выборки, попадающих в каждый интервал;<br>5 определяются относительные частоты $f_1 = n_1/n$, $f_2 = n_2/n$, …, $f_m = n_k/n$, указывающие долю наблюдений, попадающих в указанные интервалы.<br>*Можно интервалам присвоить разряды |
| ***Вариационный (интервальный) ряд накопленных частот*** – каждой варианте $x_i$, $1 \le i \le k$, (интервалу) соответствует относительная частота, накопленная к данному значению (интервалу): $\sum_{f_i < x} f_i$ |

Таблица Графики представления данных

| |
|---|
| ***Полигон относительных частот*** – ломаная линия, состоящая из отрезков, соединяющих точки ($x_i$, $f_i$ ), $i = 1, 2, \ldots, k$. Здесь $x_i$ – варианты вариационного ряда. |
| ***Гистограмма относительных частот*** – ступенчатая фигура, состоящая из прямоугольников, построенных на интервалах группировки интервального ряда.<br>*Особенности:*<br>1 основание каждого прямоугольника лежит на оси $Ox$ и равно длине интервала $h$;<br>2 высота каждого прямоугольника равна $f_i / h$, $i = 1, 2, \ldots, m$;<br>3 сумма площадей всех прямоугольников равна единице. |
| |
| ***Кумулянта –*** график накопленных относительных частот. |

Ряды и графики распределения частот дают предварительную информацию о ***форме распределения признака:*** о том, какие значения встречаются реже, а какие чаще, насколько выражена изменчивость признака.

### 3 Точечные оценки и интервальные оценки

Пусть имеется выборка объема $n$ из генеральной совокупности изучаемого признака $X$. Полученные по выборочным данным приближенные характеристики каких-либо свойств изучаемой совокупности называются их *оценками*. Например, в качестве оценки неизвестного среднего значения чаще всего используется среднее арифметическое по выборке, хотя возможны и другие

варианты оценок этого параметра: среднее геометрическое, среднее гармоническое и др. В связи с этим всегда возникает вопрос о выборе из набора возможных вариантов оценок параметров тех из них, которые удовлетворяют некоторым требованиям качества.

Статистические оценки могут быть *точечными* и *интервальными*. При точечной оценке неизвестная характеристика оценивается некоторым числом, а при интервальной оценке указывается некоторый интервал значений, в пределах которого с заданной вероятностью должно находиться истинное значение оцениваемой величины.

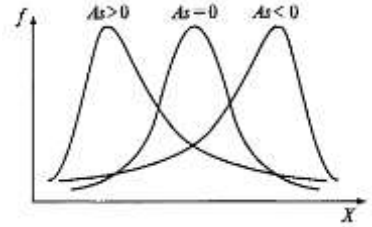
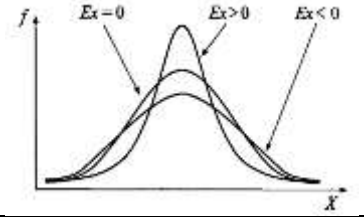
Наиболее важными характеристиками случайной величины являются математическое ожидание и дисперсия.

При *нормальном законе распределения* состоятельной, несмещенной и эффективной оценкой математического ожидания случайной величины является *среднее арифметическое* ($\overline{x}$), полученное по выборочным данным.

Таблица – Числовые характеристики выборки

| **Название** | **Характеристика** | **Формула** |
|---|---|---|
| *Меры тенденции* | | |
| Медиана, $m_e$ | Значение признака, которое делит ранжированный ряд на две части, равные по числу элементов выборки (серединное значение) | При нечетном $n = 2k + 1$ медиана равна $m_e = x_{k+1}$; при четном $n = 2k$ медиана вычисляется по формуле: $m_e = \frac{x_k + x_{k+1}}{2}$. |
| Квартиль, $Q$ | Значения признака, которые делят вариационный ряд на *четыре* равные группы (по 25% в каждой группе):<br>$Q_1$ – первые 25 %,<br>$Q_2$ – 50 % (медиана),<br>$Q_3$ – 75 %. | |
| Мода, $M_0$ | Значение признака, с наибольшей частотой.<br>Возможны случаи:<br>1) значения признака, стоящие рядом, встречаются одинаковой частотой. В этом случае мода равна среднему арифметическому этих значений признака. Например, в следующем ряду данных: | |

| | | |
|---|---|---|
| | 12, 13, 14, 14, 14, 16, 16, 16, 18, 19<br>*M*о = (14+16)/2= 15.<br>2) два значения, встречаются с одинаковой частотой, но не стоят рядом. В этом случае говорят, что ряд данных имеет две моды, т.е. он бимодальный.<br>3) Если все значения данных имеют одинаковою частоту, то говорят, что ряд не имеет моды. | |
| Выборочное среднее, $\bar{x}$ | значение для наиболее типичного представителя выборки | $\bar{x} = \frac{1}{n}\sum_{i=1}^{m} n_i x_i$ |
| *Меры разброса* | | |
| Размах выборки, *R* | разность максимального и минимального значения признака | $R = x_{\max} - x_{\min}$ |
| Генеральная дисперсия, $\sigma^2$ | Мера изменчивости бесконечного числа измерений (в генеральной совокупности), и есть среднее арифметическое квадратов отклонений наблюдаемых значений признака от их выборочного среднего $\bar{x}$ | $\sigma^2 = \frac{1}{n}\sum_{i=1}^{m} n_i (x_i - \bar{x})^2$ |
| Выборочное среднеквадратическое отклонение, σ | Характеризует изменчивость признака относительно выборочного среднего | $\sigma = \sqrt{\sigma^2}$ |
| Выборочная (исправленная) дисперсия, $S^2$ | Мера изменчивости реально измеренных значений признака, которая используется для оценки дисперсии в генеральной совокупности | $S^2 = \frac{1}{n-1}\sum_{i=1}^{m} n_i (x_i - \bar{x})^2$<br>или<br>$S^2 = \frac{n}{n-1}\sigma^2$ |
| «Исправленное» среднеквадратическое отклонение, *S* | | $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{m} n_i (x_i - \bar{x})^2}$<br>. |
| Коэффициент вариации, *V* | Относительное значение среднеквадратического отклонения, которое служит для сравнения колеблемости несоизмеримых показателей. | $V = \frac{S}{\bar{x}} \cdot 100\%$ |

| | | |
|---|---|---|
| *Мера формы* | | |
| Ассиметрия, *As* | Степень отклонения графика распределения частот от симметричного вида относительно среднего значения $\bar{x}$   | $As = \frac{1}{n}\sum_{i=1}^{m} n_i \frac{(x_i - \bar{x})^3}{\sigma^3}$ |
| Эксцесс, *Ex* | Мера плосковершинности или остроконечности графика распределения измеренного признака   | $Ex = \frac{1}{n}\sum_{i=1}^{m} n_i \frac{(x_i - \bar{x})^4}{\sigma^4} - 3$ |
| Стандартизация, $z_i$ | Перевод измерений в стандартную Z-шкалу со средним $\bar{z} = 0$ и среднеквадратическим отклонением $\sigma_z = 1$ | $z_i = \frac{x_i - \bar{x}}{\sigma}$ |

*Ящик с усами (пяти-числовые сводки)* – прямоугольник, нижняя и верхняя границами которого вычисляются по формулам

$K_1 = Q(0{,}75) - 1{,}5\,(Q(0{,}75) - Q(0{,}25))$

$K_2 = Q(0{,}75) + 1{,}5\,(Q(0{,}75) - Q(0{,}25))$

где $Q(0{,}25)$, $Q(0{,}75)$ – оценки квартилей распределения уровня $p$,

перечная линия соответствует выборочной медиане $x_{med}$.

Значения

$x_{\max}$.=max{ $x_1, x_2, \ldots, x_n$ }, $x_{\min}$ =min{ $x_1, x_2, \ldots, x_n$ } соединяются с границами ящика прямоугольными линиями («усами)».

***Квартильный размах*** охватывает 50% значений выборки. Для компактного и наглядного изображения данных используется ящик с усами.

Ящик с сами имеет вид, представленный на рисунке 2.1.

Алгоритм построения ящика с усами:

1) располагаются данные в порядке возрастания;

2) определяется медиана выборки;

3) определяются сгибы, т есть середины распределений в интервалах от медианы до минимального и максимального значения (1/4 от количества).

Интервал между сгибами изображается в виде прямоугольника («ящика»), а интервалы от сгибов до максимального и минимального значения – в виде прямых линий («усов»).

Фактически на ящике с усами изображены пять характеристик вариационного ряда: $x_{max}$, $x_{min}$, оценка медианы, нижняя и верхняя выборочные квартили.

График служит для выявления аномальных наблюдений и оценки симметричности распределения выборки. Выборочные наблюдения, располагаемые за границами «ящика» интерпретируются как выбросы. В этом случае ящик имеет усы. В противном случае усы отсутствуют. Симметричность частей ящика относительно поперечной линии (медианы) говорит о симметричности распределения выборки.

### *2 Интервальные оценки*

Для того чтобы оценить возможную при этом ошибку, возникающую при замене точного значения $\alpha$ на точечную оценку $\alpha^*$, назначим достаточно большую вероятность $P$ и найдем такое значение $f(\alpha^*, P) = \varepsilon_P$, для которого $P(|\alpha^* - \alpha| \le \varepsilon_P) = P$ или $P(\alpha^* - \varepsilon_P \le \alpha \le \alpha^* + \varepsilon_P) = P$, где $\varepsilon_P$ – допустимая ошибка при вероятности $P$.

При этом диапазон возможных значений $\alpha$ будет $[\alpha^* - \varepsilon_P; \alpha^* + \varepsilon_P]$, а большие по абсолютной величине ошибки, чем $\varepsilon_P$, будут появляться только с малой, практически допустимой вероятностью $q = 1 - P$, называемой *уровнем значимости*, который, как правило, оценивается в процентах (%). Вероятность $P$ называют *доверительной вероятностью*, или коэффициентом доверия, или доверительным уровнем, или статистической надежностью, т.к. она характеризует надежность полученной оценки. При оценке информационных процессов часто принимается доверительная вероятность, равная *0,8 – 0,9*. В технических системах обычно принимают более высокую вероятность, например, в машиностроении эта вероятность принимается равной *0,95 – 0,99*. Интервал $I_P = [\alpha^* - \varepsilon_P; \alpha^* - \varepsilon_P]$ называют *доверительным интервалом*, а границы этого

интервала $\underline{\alpha} = \alpha^* - \varepsilon_P$ и $\bar{\alpha} = \alpha^* + \varepsilon_P$ – *доверительными границами*. На практике обычно для *P* принимают значения *0,9; 0,95* или *0,99* и исходя из этого определяют $I_P$ или $\varepsilon_p$. Для этого используют различные способы, подобно тому, как разными способами оценивают неизвестные параметры распределения по выборке.

Доверительный интервал для математического ожидания нормально распределенной случайной величины имеет вид

$$\bar{x} - \frac{s}{\sqrt{n-1}} t_P \le m_X \le \bar{x} + \frac{s}{\sqrt{n-1}} t_P$$

где значения квантилей $(t_{q/2} = -t_{1-q/2})$ для различных степеней свободы $\nu$ и уровней значимости $q$ находятся по таблицам распределения Стьюдента. На практике и, в первую очередь, при $n \ge 50$ выражение (3.19) можно записать так:

$$\bar{x} - \frac{s}{\sqrt{n}} t_P \le m_X \le \bar{x} + \frac{s}{\sqrt{n}} t_P .$$

При одностороннем интервале для доверительной вероятности $P = 1 - q$, т.е. при $(t_q = -t_{1-q})$, получим:

$m_X \le \bar{x} + \frac{s}{\sqrt{n}} t_P$ – правосторонняя (сверху) оценка;

$m_X \ge \bar{x} - \frac{s}{\sqrt{n}} t_P$ – левосторонняя (снизу) оценка.

***Методика определения доверительного интервала для математического ожидания сводится к следующему:***

1. По заданной выборке определяются оценки математического ожидания $\bar{x}$ и дисперсии $s^2$.
2. Для уровня значимости $q$ и числа степеней свободы $\nu = n - 1$ находится $t$ по таблицам распределения Стьюдента.
3. По оценкам $\bar{x}$ и $s^2$ рассчитывается по формулам двусторонний или по формулам односторонний доверительный интервал на математическое ожидание.

Доверительный интервал генеральной дисперсии $\sigma^2$ имеет вид

$$\frac{\nu \cdot s^2}{\chi^2_{q/2}} \le \sigma^2 \le \frac{\nu \cdot s^2}{\chi^2_{1-q/2}}$$

а односторонние –

$$\sigma^2 \le \frac{\nu \cdot s^2}{\chi^2_{1-q}} \text{ и } \sigma^2 \ge \frac{\nu \cdot s^2}{\chi^2_q},$$

где $\chi^2_{1-q}$ и $\chi^2_q$ критические точки распределения Хи квадрат.

## 4 Проверка статистических гипотез

Пусть $X = \{x_1, x_2, \ldots, x_n\}$ выборка из генеральной совокупности, характеризующей некоторый признак.

Таблица – Статистическая гипотеза

| Статистическая гипотеза<br>– это утверждение, в котором высказывается предположение относительно параметров или вида распределения генеральной совокупности изучаемого признака | |
|---|---|
| Виды статистических гипотез | |
| Простая гипотеза | содержит только одно предположение |
| Сложная гипотеза | состоит из конечного или бесконечного числа простых гипотез |
| Нулевая гипотеза, $H_0$ | основная, выдвинутая гипотеза |
| Альтернативная гипотеза, $H_1$ | гипотеза, противоречащая нулевой гипотезе $H_0$ |

Таблица – Статистический критерий

| Статистическая проверка гипотез<br>– это процедура обоснованного сопоставления выдвинутой гипотезы с имеющейся выборкой, осуществляемая при помощи того или иного статистического критерия | |
|---|---|
| Статистический критерий | правило, по которому принимается решение отклонить или не отклонить нулевую гипотезу $H_0$ |
| Статистический критерий включает в себя | |
| Статистика критерия, $K$ | случайная величина $K(x_1, x_2, \ldots, x_n)$ с множеством значений $W$, с помощью которой принимают решение о принятии или отклонении нулевой гипотезы $H_0$. |
| Уровень значимости, α | вероятность ошибочно опровергнуть нулевую гипотезу $H_0$, когда она верна |
| Число степеней свободы, $df$ | количество возможных направлений изменчивости признака (статистический критерий содержит правило (формулу) определения числа степеней свободы) |
| Статистическая значимость, $p$-value | вероятность получить значение статистики критерия равное наблюдаемому или более нетипичное по сравнению с наблюдаемым при условии, что нулевая гипотеза $H_0$ верна (чем меньше $p$-уровень, тем больше оснований для того, чтобы отклонить $H_0$ в пользу $H_1$):<br>$p = P(K \in W_{кр} \mid H_0$ верна$)$. |
| Критическую область, $W_{кр}$ | совокупность значений статистики $K$, при которых нулевая гипотеза отклоняется $H_0$ на заданном уровне значимости α. Возможны три вида расположения критической области:<br>– правосторонняя критическая область $(W_{кр}, +\infty)$; |

| | |
|---|---|
| | – левосторонняя критическая область $(-\infty, W_{кр})$;<br>– двусторонняя критическая область<br>$(-\infty, W_{кр}) \cup (W_{кр}, +\infty)$. |
| Критические точки, $K_{кр}$ | точки, отделяющие критическую область $W_{кр}$ от области принятия гипотезы $W \setminus W_{кр}$ |

При проверке гипотезы возникают ошибки первого и второго рода.

Таблица – Ошибки первого и второго рода

| | | |
|---|---|---|
| Ошибка первого рода | будет отклонена правильная нулевая гипотеза $H_0$. Вероятность ошибки первого рода равна уровню значимости α:<br>$\alpha = P(\text{отклонить } H_0 \mid H_0 \text{ верна})$. | |
| Ошибка второго рода | будет принята неправильная нулевая гипотеза $H_0$ с вероятностью β:<br>$\beta = P(\text{принять } H_0 \mid H_0 \text{ неверна})$. | |
| *Результаты проверки гипотезы* | | |
| | *$H_0$ верна* | *$H_1$ неверна* |
| *Гипотеза $H_0$ отклоняется* | Ошибка первого рода | Правильный вывод |
| *Гипотеза $H_0$ не отклоняется* | Правильный вывод | Ошибка второго рода |
| Мощность критерия | вероятность попадания статистики $K$ в критическую область $W_{кр}$ при условии, что справедлива альтернативная гипотеза $H_1$:<br>$\gamma = 1 - \beta = P(\text{отклонить } H_0 \mid H_1 \text{ верна})$,<br>то есть это вероятность принять нулевую гипотезу $H_0$, когда она верна. Чем больше мощность критерия, тем меньше вероятность принятия неверной гипотезы. Мощность критерия позволяет выбрать оптимальную статистику $K$ для проверки гипотезы среди возможных статистик критерия. | |

Таблица – Правило принятия статистического вывода

| | |
|---|---|
| 1 | На основании выборочных данных выдвигаются нулевая гипотеза $H_0$ и альтернативная ей гипотеза $H_1$ |
| 2 | Выбирается уровень значимости α (на практике α = 0,1; 0,05; 0,025; 0,01; 0,005; 0,001). |
| 3 | Выбирается статистика $K$, имеющая известный закон распределения. |
| 4 | По выборочным данным вычисляется наблюдаемое значение статистики $K_{набл}$. |
| 5 | Определяется вид критической области из условия<br>$P(K \in W_{кр} \mid H_0 \text{ верна}) = \alpha$<br>и область принятия гипотезы в зависимости от формулировки альтернативной гипотезы. |
| 6 | Принимается статистическое решение: если $K_{набл}$ попадает в критическую область $W_{кр}$, то нулевая гипотеза $H_0$ отклоняется. |
| 7 | На основе $p$- value возможен один из статистических выводов:<br>– если $p$-value < α, то гипотеза $H_0$ отклоняется на уровне значимости α, на имеющихся данных,<br>– если $p$-value > α ⇒ гипотеза $H_0$ не отклоняется на уровне значимости α, на имеющихся данных.<br>*Важно*: в выводе относительно отклонения/неотклонения нулевой гипотезы необходимо указывать уровень значимости α, так как от этого зависит результат.<br>Например, если $p$-value = 0,02, то есть основания отвергнуть нулевую гипотезу на уровне значимости 5 % (0,02 < 0,05), и нет оснований отвергнуть ее на уровне значимости 1 % (0,02 > 0,01). |

Таблица – Критерий Стьюдента (Student) для независимых выборок

| | | |
|---|---|---|
| **1** | *Область применения критерия*: позволяет установить сходство-различие средних значений в двух независимых выборках из генеральных совокупностей нормально распределенных случайных величин $\xi_X \sim N(m_x, \sigma_x^2)$, $\xi_Y \sim N(m_y, \sigma_y^2)$ | |
| **2** | *Статистическая модель*: даны независимые выборки $X = \{ x_1, x_2, \ldots, x_{n_x} \}$ и $Y = \{ y_1, y_2, \ldots, y_{n_y} \}$ объемами $n_x$, и $n_y$ соответственно из генеральных совокупностей, имеющих нормальные распределения с неизвестными и, но равными дисперсиями: $D_x^2 = D_y^2 = \sigma^2$, и математическими ожиданиями $m_x$, и $m_y$. | |
| **3** | *Нулевая и альтернативные гипотезы*: | |
| | **3а** | $H_0$: $m_x, = m_y$ (средние значения признака равны). |
| | | $H_1$: $m_x, \neq m_y$ (средние значения признака не равны). |
| | **3б** | $H_0$: $m_x, \leq m_y$. |
| | | $H_1$: $m_x, > m_y$. |
| **4** | Задается уровень значимости α. | |

| | | |
|---|---|---|
| **5** | Для проверки гипотезы используется статистика $$T=\frac{\bar{x}-\bar{y}}{\sqrt{(n_x-1)S_x^2+(n_y-1)S_y^2}}\sqrt{\frac{n_x n_y(n_x+n_y-2)}{n_x+n_y}},$$ которая при условии истинности гипотезы $H_0$ имеет распределение Стьюдента с числом степей свободы $df = n_x + n_y - 2$.<br>Здесь<br>$\bar{x}=\frac{1}{n_x}\sum_{i=1}^{n_x}x_i$, $\bar{y}=\frac{1}{n_y}\sum_{i=1}^{n_y}y_i$ – выборочные средние,<br>$Sx^2 = \frac{1}{n_x-1}\sum_{i=1}^{n_x}(x_i-\bar{x})^2$, $S_y^2 = \frac{1}{n_y-1}\sum_{i=1}^{n_y}(y_i-\bar{x})^2$ – выборочные дисперсии. | |
| **6** | Расчет по выборочным данным *наблюдаемого значения* статистики $T$ ($T_{\text{набл}}$) | |
| **7** | *Определение критической области* | |
| | **7а** | На уровне значимости α критическое значение $T_{\text{кр}} = t_{1-\alpha/2}(df)$ находится как квантиль распределения Стьюдента, критическая область $(-\infty, T_{\text{кр}}) \cup (T_{\text{кр}}, +\infty)$. |
| | **7б** | На уровне значимости α критическое значение $T_{\text{кр}} = t_{1-\alpha}(df)$ находится по таблицам распределения Стьюдента, критическая область $(T_{\text{кр}}, +\infty)$. |
| **8** | *Принятие статистического решения* | |
| | **8а** | Если $\lvert T_{\text{набл}} \rvert < T_{\text{кр}}$, то нет оснований для отклонения гипотезы $H_0$ об отсутствии различий.<br>Если $\lvert T_{\text{набл}} \rvert \geq T_{\text{кр}}$, то гипотеза $H_0$ отклоняется в пользу гипотезы $H_1$. |
| | **8б** | Если $T_{\text{набл}} < T_{\text{кр}}$, то нет оснований для отклонения гипотезы $H_0$ об отсутствии различий.<br>Если $T_{\text{набл}} \geq T_{\text{кр}}$, то гипотеза $H_0$ отклоняется в пользу гипотезы $H_1$. |

**Таблица – Критерий Фишера (Fisher)**

| | | |
|---|---|---|
| **1** | *Область применения критерия:* предназначен для оценки равенства дисперсий двух независимых выборок их генеральных совокупностей нормально распределенных случайных величин $\xi_X \sim N(m_x, \sigma_x^2)$ и $\xi_Y \sim N(m_y, \sigma_y^2)$ | |
| **2** | *Статистическая модель*: даны независимые выборки $X = \{ x_1, x_2, \ldots, x_{n_x} \}$ и $Y = \{ y_1, y_2, \ldots, y_{n_y} \}$ объемами $n_x$, и $n_y$ соответственно из генеральных совокупностей, имеющих нормальные распределения с неизвестными дисперсиями $\sigma_x^2$ и $\sigma_y^2$ и математическими ожиданиями $m_x$ и $m_y$. | |
| **3** | *Нулевая и альтернативные гипотезы*: | |
| | **3а** | $H_0$: $\sigma_x^2 = \sigma_y^2$ (дисперсии признака равны). |
| | | $H_1$: $\sigma_x^2 \neq \sigma_y^2$ (дисперсии значения признака не равны). |
| | **3б** | $H_0$: $\sigma_x^2 \leq \sigma_y^2$. |
| | | $H_1$: $\sigma_x^2 > \sigma_y^2$. |
| **4** | Задается уровень значимости α. | |
| **5** | Для проверки гипотезы используется статистика | |

| | | |
|---|---|---|
| | | $$F = \frac{\max\{\sigma_x^2;\sigma_y^2\}}{\min\{\sigma_x^2;\sigma_y^2\}} = \frac{\sigma_1^2}{\sigma_2^2},$$ которая при условии истинности гипотезы $H_0$ асимптотически имеет распределение Фишера с $df_1$ и $df_2$ степенями свободы.<br>Здесь $\sigma_x^2 = \frac{1}{n_x}\sum_{i=1}^{n_x}(x_i - \bar{x})^2$, $\sigma_y^2 = \frac{1}{n_y}\sum_{i=1}^{n_y}(y_i - \bar{x})^2$ – выборочные дисперсии,<br>$df_1$ – число степеней большей дисперсии ($df_1 = n_x - 1$ или $df_1 = n_y - 1$),<br>$df_2$ – число степеней меньшей дисперсии ($df_2 = n_y - 1$ или $df_2 = n_x - 1$). |
| **6** | | Расчет по выборочным данным *наблюдаемого значения* статистики $F_{набл}$. |
| **7** | | *Определение критической области* |
| | **7а** | На уровне значимости α критические значение $F_{кр1} = F_{1-\alpha/2}(df_1, df_2)$ и $F_{кр2} = F_{\alpha/2}(df_1, df_2)$ находятся как квантиль распределения Фишера, критическая область $(-\infty, F_{кр1}) \cup (F_{кр2}, +\infty)$. |
| | **7б** | На уровне значимости α критическое значение $T_{кр} = t_{1-\alpha}(df)$ находится по таблицам распределения Стьюдента, критическая область $(F_{кр}, +\infty)$. |
| **8** | | *Принятие статистического решения* |
| | **8а** | Если $F_{кр1} < F_{набл} < F_{кр}$, то нет оснований для отклонения гипотезы $H_0$ об отсутствии различий. Иначе гипотеза $H_0$ отклоняется в пользу $H_1$. |
| | **8б** | Если $F_{набл} < F_{кр}$, то нет оснований для отклонения гипотезы $H_0$ об отсутствии различий. Иначе гипотеза $H_0$ отклоняется в пользу гипотезы $H_1$. |

**Критерий Пирсона ($\chi^2$)**

| | | |
|---|---|---|
| **1** | | *Область применения критерия*: позволяет сравнить эмпирическое и теоретическое распределение |
| **2** | | *Статистическая модель*: выборка, состоящая из независимых выборочных значений $X = \{x_1, x_2, \ldots, x_n\}$ получена из генеральной совокупности, имеющей распределение $F_0(x)$, зависящей от $k$ параметров, из которых $k_1$ параметров неизвестно. |
| **3** | | Сформулировать гипотезы: |
| | $H_0$ | Полученное эмпирическое распределение не отличатся от теоретического распределения. |
| | $H_1$ | Полученное эмпирическое распределение отличатся от теоретического распределения. |
| **4** | | Выборочные данные представить в виде: |
| | **а** | Вариационного ряда распределения частот (для нормального распределения): |

| Варианты, $x_i$ | $x_1$ | $x_2$ | .... | $x_m$ | |
|---|---|---|---|---|---|
| Частота, $k_i^*$ | $n_1^*$ | $n_2^*$ | .... | $n_m^*$ | $\sum_{i=1}^{m} n_i^* = n$ |

**б** Интервального ряда распределения частот (для равномерного распределения, для нормального распределения):

| Интервалы, $[x_i, x_{i+1})$ | $[x_0; x_1)$ | $[x_1; x_2)$ | .... | $[x_{m-1}; x_m]$ | |
|---|---|---|---|---|---|
| Частота, $k_i^m$ | $n_1^*$ | $n_2^*$ | .... | $n_m^*$ | $\sum_{i=1}^m n_i^* = n$ |

Здесь $x_0 = -\infty$, $x_m = +\infty$

Частота для каждой ячейки таблицы должна быть больше 4.

**5** Для каждой варианты (интервала) вычислить теоретические частоты $k_i$ (частоты, которые следует ожидать, когда гипотеза $H_0$ справедлива) по формуле: $k_i = n \cdot p_i$, где $p_i$ :

**5.1** Для нормального распределения

$$p_i = P(x_i < X \le x_{i+1}) = \Phi\left(\frac{x_{i+1} - \bar{x}}{\sigma}\right) - \Phi\left(\frac{x_i - \bar{x}}{\sigma}\right),\ i = 0, 1, \ldots, m.$$

Значения $\Phi(z)$ интеграла вероятностей.

**5.2** Для распределения случайной величины по равномерному закону, если исходные выборочные данные сгруппировать и представить в виде:

**а** вариационного ряда распределения частот $p_i = P(X = x_i) = 1 / m$;

**б** интервального ряда распределения частот

$$p_i = P(x_i < X \le x_{i+1}) = \frac{x_{i+1} - x_i}{x_{\max} - x_{\min}}.$$

**6** Вычислить значение $\chi^2_{\text{набл}}$ по формуле:

$$\chi^2_{\text{набл}} = \sum_{i=1}^m \frac{(n_i^* - n_i)^2}{n_i}.$$

**7** Определить количество степеней свободы по формуле:

**а** для нормального распределения $df = n - 3$;

**б** для распределения случайной величины по равномерному закону $df = n - 1$. Если $df = 1$, то внести поправку на «непрерывность».

**8** Определить критические значения $\chi^2_{1\text{кр}}$ и $\chi^2_{2\text{кр}}$, которые отвечают уровням значимости в 5 % и 1 % по таблицам

**9** Расположить эмпирическое значение критерия $\chi^2_{\text{набл}}$ и критические значения $\chi^2_{1\text{кр}}$ и $\chi^2_{2\text{кр}}$ на оси значимости.

**10** Если $\chi^2_{\text{набл}}$ находится в зоне незначимости, то принимается гипотеза $H_0$ об отсутствии различий. Если $\chi^2_{\text{набл}}$ находится в зоне значимости, то гипотеза об отсутствии различий $H_0$ отклоняется и принимается гипотеза $H_1$ о наличии различий. Если $\chi^2_{\text{набл}}$ находится в зоне неопределенности, то существует вероятность принятия ложного решения.

| Критерий Пирсона<br>позволяет сравнить одно эмпирическое распределение с другим эмпирическим распределение |
|---|
| *Критерий Пирсона включает следующие этапы:* |

<table>
<tr><td>1</td><td colspan="2">Определить признак, содержащий $k$ разрядов, участвующий в сопоставлении (значения признака должны быть представлены не ниже порядковой шкалы).</td></tr>
<tr><td>2</td><td colspan="2">Произвести выборку двух групп респондентов.</td></tr>
<tr><td>3</td><td colspan="2">Провести две серии наблюдений на двух независимых выборках респондентов объема $n_1$, и $n_2$:<br>$x_1, x_2, \ldots x_i, \ldots x_{n_1}$;<br>$y_1, y_2, \ldots, y_j, \ldots, y_{n_2}$,<br>где случайная переменная $x$ характеризует состояние изучаемого свойства в одной из рассматриваемых совокупностей, а случайная переменная $y$ – состояние того же свойства во второй совокупности (число членов в обеих выборках должно быть в сумме больше 40, т. е. $n_1 + n_2 > 40$)</td></tr>
<tr><td rowspan="3">4</td><td colspan="2">Сформулировать гипотезы:</td></tr>
<tr><td>$H_0$</td><td>Законы распределения случайных величин $X$ и $Y$ одинаковы в обеих рассматриваемых совокупностях.</td></tr>
<tr><td>$H_1$</td><td>Законы распределения случайных величин $X$ и $Y$ различны в обеих рассматриваемых совокупностях.</td></tr>
<tr><td>5</td><td colspan="2">Распределить элементы каждой выборки объемов $n_1$ и $n_2$ на $k$ категорий, соответствующие разрядам исследуемого признака.</td></tr>
<tr><td>6</td><td colspan="2">На основе полученных результатов составить таблицу вида:
<table>
<tr><td rowspan="2">выборки</td><td colspan="4">разряды исследуемого признака</td><td rowspan="2">$\sum$</td></tr>
<tr><td>1</td><td>2</td><td>…</td><td>$k$</td></tr>
<tr><td>1</td><td>$m_{11}$</td><td>$m_{12}$</td><td>…</td><td>$m_{1k}$</td><td>$n_1$</td></tr>
<tr><td>2</td><td>$m_{21}$</td><td>$m_{22}$</td><td>…</td><td>$m_{2k}$</td><td>$n_2$</td></tr>
<tr><td>$\sum$</td><td>$m_{11} + m_{21}$</td><td>$m_{12} + m_{22}$</td><td>…</td><td>$m_{1k} + m_{2k}$</td><td>$n_1 + n_2$</td></tr>
</table>
Не рекомендуется использовать критерий для проверки гипотез, если хотя бы одно из значений $m_{ij}$ меньше 5.</td></tr>
<tr><td>7</td><td colspan="2">Вычислить значение $\chi^2_{набл}$ по формуле: $\chi^2_{набл} = \frac{1}{n_1 \cdot n_2} \sum_{i=1}^{k} \frac{(n_1 \cdot m_{2i} - n_2 m_{1i})^2}{m_{1i} + m_{2i}}$.</td></tr>
<tr><td>8</td><td colspan="2">Определить критические значения $\chi^2_{1кр}$ и $\chi^2_{2кр}$, которые отвечают уровням значимости в 5 % и 1 % по таблице приложения. Количество степеней свободы при этом определяется по формуле: $df = k - 1$. Если $df = 1$, то внести поправку на «непрерывность».</td></tr>
<tr><td>9</td><td colspan="2">Расположить эмпирическое значение критерия $\chi^2_{набл}$ и критические значения $\chi^2_{1кр}$ и $\chi^2_{2кр}$, на оси значимости.</td></tr>
<tr><td>10</td><td colspan="2">Если $\chi^2_{набл}$ находится в зоне незначимости, то принимается гипотеза $H_0$ об отсутствии различий. Если $\chi^2_{набл}$ находится в зоне значимости, то гипотеза об отсутствии различий $H_0$ отклоняется и принимается гипотеза $H_1$ о наличии различий. Если $\chi^2_{набл}$ находится в зоне неопределенности, то существует вероятность принятия ложного решения.</td></tr>
</table>

Таблица – Критерий Колмогорова о законе распределения случайной величины $\xi$

| | |
|---|---|
| **1** | *Область применения критерия*: критерий предназначен для проверки согласия эмпирического распределения с заданным непрерывным теоретическим распределением. Критерий позволяет найти точку, в которой сумма накопленных расхождений между двумя распределениями является наибольшей, и оценить достоверность этого расхождения. Если различия между двумя распределениями существенны, то в какой-то момент разность накопленных частот достигнет критического значения, и мы сможем признать различия статистически достоверными. |
| **2** | *Статистическая модель*: выборка, состоящая из независимых выборочных значений $\{ x_1, x_2, \ldots, x_n \}$, полученных из генеральной совокупности, распределение которой предполагается непрерывным |
| **3** | *Нулевая и альтернативные гипотезы*: |
| | $H_0$: $F^*(x) = F_0(x)$ (экспериментальные данные $\{ x_i \}$ согласуются с известным распределением $F_0(x)$). |
| | $H_0$: $F^*(x) \neq F_0(x)$ |
| **4** | Задается уровень значимости α. |
| **5** | Для проверки гипотезы используется статистика $$\lambda = \sqrt{n}\, D,$$ которая при условии истинности гипотезы $H_0$ асимптотически имеет распределение Колмогорова $$P(\sqrt{n}\, D < x) \to K(x),\ n \to \infty.$$ Здесь $K(x)$ – функция Колмогорова, $D = \max\limits_{-\infty < x < \infty} \lvert F^*(x) - F_0(x) \rvert$ – случайная величина, имеющая распределение Колмогорова и характеризующая максимальное отклонение эмпирической функции распределения $F^*(x)$ от теоретической $F_0(x)$; |
| **6** | Расчет по выборочным данным *наблюдаемого значения* статистики λ ($\lambda_{\text{набл}}$) |
| **7** | *Определение критической области* |
| | На уровне значимости α критическое значение $\lambda_{\text{кр}} = \lambda(\alpha)$ находится по таблице распределения Колмогорова , критическая область $(\lambda_{\text{кр}}, +\infty)$. |
| **8** | *Принятие статистического решения* |
| | Если $\lambda_{\text{набл}} < \lambda_{\text{кр}}$, то нет оснований для отклонения гипотезы $H_0$ и экспериментальные данные $\{ x_i \}$ согласуются с известным распределением $F_0(x)$.<br>Если $\lambda_{\text{набл}} \geq \lambda_{\text{кр}}$, то гипотеза $H_0$ отклоняется в пользу гипотезы $H_1$. |

## 5 Дисперсионный анализ

Однофакторный дисперсионный анализ позволяет статистически обосновать степень влияния на результативный признак $Y$ одного фактора $A$ для различных уровней (групп) $A_1, A_2, \ldots, A_k$. В таблице 1 представлены выборочные значения результирующего фактора $Y$ для различных ей фактора $A$.

*Таблица* Данные наблюдений для однофакторного дисперсионного анализа

| Номера наблюдений | Группы (уровни) фактора $A$ | | | |
|---|---|---|---|---|
| | $A_1$ | $A_2$ | … | $A_k$ |
| 1 | $y_{11}$ | $y_{12}$ | … | $y_{1k}$ |
| 2 | $y_{21}$ | $y_{22}$ | … | $y_{2k}$ |
| … | … | … | … | … |
| $n_j$ | $y_{n_1 1}$ | $y_{n_2 2}$ | … | $y_{n_k k}$ |
| Групповая средняя | $\bar{y}_1$ | $\bar{y}_2$ | … | $\bar{y}_k$ |

Здесь $n_1, n_2, \ldots, n_k$ – количество наблюдений в группах $A_1, A_2, \ldots, A_k$ соответственно, при этом $n = n_1 + n_2 + \ldots + n_k$.

Требуется оценить значимость влияния фактора $A$ на результативный признак $Y$.

Пусть

$Q = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(y_{ij} - \bar{y}\right)^2$ – общая сумма квадратов отклонений (общая вариация) отдельных наблюдений $y_{ij}$ от общей средней $\bar{y}$, вызванная влиянием на $Y$ фактора $A$ и случайных неучтенных (остаточных) факторов;

$Q_ф = \sum_{j=1}^{k}\left(\bar{y}_j - \bar{y}\right)^2 n_j$ – сумма квадратов отклонений между группами (межгрупповая вариация), вызванная влиянием фактора $A$ на $Y$ и характеризующая рассеяние групповых средних около общей средней;

$Q_o = \sum_{j=1}^{k}\sum_{i=1}^{n_i}\left(y_{ij} - \bar{y}_j\right)^2$ – сумма квадратов отклонений внутри групп (остаточной вариации), вызванная влиянием на $Y$ остаточных факторов и характеризующая рассеяние отдельных наблюдений группы около ее средней.

Здесь

$\bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}$ – групповые средние ($j = 1, \ldots, k$),

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{k}\bar{y}_j n_j = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} y_{ij}$$ – общее среднее.

В основе дисперсионного анализа лежит **правило сложения дисперсий:** *общая дисперсия равна сумме межгрупповой и средней из внутригрупповых дисперсий:*

$Q = Q_ф + Q_о$.

По выборочным данным результативного признака $Y$ находятся четыре несмещенные оценки

$$S^2 = \frac{1}{n-1}Q,\ S_ф^2 = \frac{Q_ф}{k-1},\ S_о^2 = \frac{Q_о}{n-k}$$

Причем оценка $S_о^2$ всегда является несмещенной оценкой, а $S^2$, $S_ф^2$ только при выполнении гипотезы $H_0$: $a_1 = a_2 = \ldots = a_k$, т. е. только в том случае, когда фактор $A$ не влияет на результативный признак $Y$.

Изменчивость значений $y_{ij}$ от одного уровня к другому объясняется влиянием фактора $A$ на результативный фактор $Y$, а изменчивость тех же значений в пределах одного уровня характеризуется влиянием неучтенных случайных (остаточных) факторов.

Однофакторная модель дисперсионного анализа аддитивного типа имеет следующий вид

$$y_{ij} = \bar{y}_j + F_j + \xi_{ij}$$

где $y_{ij}$ – значения результативного признака $Y$ (отклики) для $i$-го объекта наблюдения при $j$-м уровне фактора;

$\bar{y}_j$ – групповая средняя для $j$-го уровня фактора $A$;

$F_j$ – эффект, обусловленный влиянием $j$-го уровня фактора;

$\xi_{ij}$ – случайная компонента, обусловленная влиянием неконтролируемых факторов.

1 Равенство нулю математического ожидания случайной компоненты: $M(\xi_{ij}) = 0$.

2 Дисперсия случайной компоненты постоянна : $D(\xi_{ij}) = \sigma^2$.

3 Случайная компонента $\xi_{ij}$ , а следовательно, и $y_{ij}$ имеют нормальный закон распределения.

4 Число уровней фактора $A$ должно быть не менее трех, $k \geq 3$.

5 Случайные ошибки взаимно независимы и нормально распределены:

$\xi_{ij} \sim N(0,\sigma^2)$.

При невыполнении последнего условия (т. е. при наличии взаимосвязей) отношение дисперсий может оказаться сильно искаженным.

*Алгоритм проведения однофакторного дисперсионного анализа*

<u>Шаг 1 Проверяются предпосылки ОДА</u>

– нормальный закон распределения результативного признака $Y$ с

помощью критерия $\chi^2$;

– проверка дисперсий уровней фактора А на однородность (гомогенность) с помощью критерия Бартлетта.

Выдвигается гипотеза

$H_0$: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$

при альтернативной

$H_1$: $\sigma_j^2 \neq \sigma_l^2$ $(j \neq l)$.

Для проверки гипотезы используется статистика Бартлетта:

$$V = \frac{1}{C} \cdot \left( \sum_{j=1}^{k} (n_j - 1) \ln \frac{\bar{S}^2}{S_j^2} \right),$$

имеющей распределение близкое к $\chi^2$-распределению с $df = k - 1$ степенями свободы.

Здесь $n_j$ – количество наблюдений на $j$-м уровне ($j = 1, \ldots, k$),

$S_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$ – выборочная дисперсия $j$-ой группы ($j = 1, \ldots, k$),

$$\bar{S}^2 = \frac{\sum_{j=1}^{k} (n_j - 1) S_j^2}{\sum_{j=1}^{k} (n_j - 1)},$$

$$C = 1 + \frac{1}{3(k-1)} \left( \sum_{j=1}^{k} \frac{1}{n_j - 1} - \frac{1}{n_1 + n_2 + \ldots + n_k - k} \right).$$

При заданном уровне значимости α находится критическая точка $V_{кр} = \chi_\alpha^2(k - 1)$, определяющая правостороннюю критическую область ($V_{кр}$; + ∞). По выборочным данным вычисляется наблюдаемое значение статистики $V_{набл}$. Если $V_{набл} < V_{кр}$, то гипотеза $H_0$ согласуется с экспериментальными данными. Тогда можно приступать непосредственно к процедуре однофакторного дисперсионного анализа.

**Шаг 2** Проведение однофакторного дисперсионного анализа

Выдвигается гипотеза

$H_0$: $a_1 = a_2 = \ldots = a_k$

при альтернативной

$H_1$: $a_j \neq a_l$ $(j \neq l)$,

где $a_1, a_2, \ldots, a_k$ – математические ожидания результативного признака $Y$ соответственно при уровнях $A_1, A_2, \ldots, A_k$. Если для различных групп фактора $A$ математические ожидания не изменяются, то считается, что результативный признак $Y$ не зависит от фактора $A$, в противном случае такая зависимость имеется.

Проверка нулевой гипотезы $H_0$ основывается на сравнении оценок $S_ф^2$, $S_о^2$. Для этого используется статистика

$$F = \frac{S_{\text{ф}}^2}{S_{\text{о}}^2} = \frac{\frac{1}{k-1}\sum_{j=1}^{k}\left(\bar{y}_j - \bar{y}\right)^2 n_j}{\frac{1}{n-k}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(y_{ij} - \bar{y}_j\right)^2},$$

имеющая распределение Фишера с $df_1 = k - 1$ и $df_2 = n - k$ степенями свободы.

По статистическим данным вычисляется наблюдаемое значение статистики $F_{\text{набл}}$. Для заданного уровня значимости α по таблицам критических точек $F$-распределения находится критическая точка $F_{\text{кр}} = F_{\alpha}(k - 1;\ n - k)$ и строится правосторонняя критическая область $(F_{\text{кр}};+\infty)$. Если $F_{\text{набл}} < F_{\text{кр}}$, то нулевая гипотеза согласуется с экспериментальными данными, и в этом случае говорят, что влияние фактора $A$ на признак $Y$ не подтвердилось выборочными наблюдениями.

**Шаг 3** Если в процессе анализа выявлено влияние фактора $A$ на результативный признак $Y$, то степень данного влияния измеряется с помощью ***выборочного коэффициента детерминации***

$$R^2 = \frac{Q_{\text{ф}}}{Q},$$

показывающего какая доля вариации $Q$ результативного признака $Y$ объясняется влиянием фактора $A$.

**Замечание**. По результатам дисперсионного анализа строиться дисперсионная таблица (таблица 2).

| *Источник вариации* | SS | df | MS | F | p-знач | F-крит |
|---|---|---|---|---|---|---|
| Между группами | $Q_{ф} = \sum_{j=1}^{k} \left(\bar{y}_j - \bar{y}\right)^2 n_j$ | $k - 1$ | $\frac{1}{k-1} Q_{ф}$ | $F_{набл} = \frac{\frac{1}{k-1} Q_0}{\frac{1}{n-k} Q_{ф}}$ | $P\left(F_{набл} > F_{кр}\right)$ FРАСП | $F_{кр} = F_{\alpha}(k - 1; n - k)$ FРАСПОБР |
| Внутри групп | $Q_{о} = \sum_{j=1}^{k} \sum_{i=1}^{n_i} \left(y_{ij} - \bar{y}_j\right)^2$ | $n - k$ | $\frac{1}{n-k} Q_{о}$ | | | |
| Итого | $Q = Q_{ф} + Q_{о}$ | | | | | |

Однофакторный непараметрический дисперсионный анализ

В случаях, когда нарушаются предпосылки однофакторного дисперсионного анализа, то используются непараметрические критерии проверки влияния фактора *A* на результативный признак *Y*.

***Алгоритм однофакторного непараметрического дисперсионного анализа с помощью критерия Краскала-Уоллиса.***

**Шаг 1** Ранжируются все значения результативного признака *Y* по возрастанию от 1 до *n*, где $n = \sum_{j=1}^{k} n_j$ – объем всей выборки, *k* – количество уровней исследуемого фактора *A*.

**Шаг 2** Составляется таблица данных, где вместо значений фактора *Y* записаны их ранги.

**Шаг 3** Рассчитываются суммы $R_1$, $R_2$, ..., $R_k$ рангов для каждого уровня анализируемого фактора *A*.

**Шаг 4.** Вычисляется наблюдаемое значение критерия Краскала-Уоллиса по формуле

$$H = \frac{12}{n(n+1)}\left(\frac{R_1^2}{n_1} + \ldots + \frac{R_k^2}{n_k}\right) - 3(n+1).$$

**Шаг 5.** Сравнивается полученное значение $H_{\text{набл}}$ с критическим значением $H_{\text{кр}}$ для заданного уровня значимости $\alpha$.

При достаточно большом объеме выборки, когда количество наблюдений по каждому уровню превышает 5, значение $H_{\text{кр}}$ определяется по таблицам распределения $\chi^2$ для числа степеней свободы $df = k - 1$, где *k* – количество уровней исследуемого фактора.

Для малых выборок критические значения критерия Краскала-Уоллиса определяются по специальным таблицам (приложение 1).

Если $H_{\text{набл}} \geq H_{\text{кр}}$, то гипотеза об отсутствии влияния анализируемого фактора на изменчивость изучаемого свойства отвергается.

## 6 Корреляционный анализ

Корреляционный анализ имеет своей задачей количественное определение тесноты связи между изучаемыми явлениями. Корреляционная связь – эта связь, при которой определенному значению факторного признака соответствует среднее значение результирующего признака. Изучение тесноты и направления связи является важной задачей математической статистики. Оценка тесноты связи между признаками предполагает определение меры соответствия вариации результирующего признака от одного (при изучении парных зависимостей) или нескольких (множественных) факторных признаков.

*Выборочный коэффициент корреляции* между двумя случайными величинами *X* и *Y* был впервые введен Пирсоном, поэтому его часто *называют коэффициентом корреляции Пирсона*. В теории разработаны и на практике применяются различные модификации формулы расчета данного коэффициента. Приведем некоторые из них:

$$r = \frac{\overline{(x-\bar{x})(y-\bar{y})}}{\sigma_x \sigma_y} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sigma_x \sigma_y}.$$

Используя преобразование, получают

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

или через дисперсии

$$r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{xy}^2}{2\sigma_x^2 \sigma_y^2}.$$

При большом объеме выборки выборочный коэффициент корреляции будет приближаться к корреляционному моменту генеральной совокупности $\rho$, который определяется как

$$\rho = \frac{M[(x - m_x)(y - m_y)]}{\sigma_x \sigma_y}$$

Если r = 0, то величины X и Y независимы, а при r = 1 зависимость между X и Y является функциональной. Коэффициент корреляции тесно связан с коэффициентом регрессии:

$$r = b\frac{\sigma_x}{\sigma_y},$$

где b – коэффициент регрессии в уравнении вида yi = a + bxi;

$\sigma_x$ – среднеквадратичное отклонение x;

$\sigma_y$ – среднеквадратичное отклонение y.

Значимость коэффициента корреляции проверяется на основе критерия Стьюдента. При проверке этой гипотезы вычисляется t-статистика:

$$t_{рас} = \sqrt{\frac{r^2(n-2)}{1-r^2}} = \frac{|r|}{\sqrt{1-r^2}}\sqrt{n-2}$$

Расчетное значение сравнивается с табличным значением tq,ν=n – 2. Если расчетное значение больше табличного, это свидетельствует о значимости коэффициента корреляции, а, следовательно, и о статистической существенности зависимости между X и Y .

При большом числе наблюдений (n > 100) используется следующая формула t-статистики:

$$t_{рас} = \frac{|r|}{\sqrt{1-r^2}}\sqrt{n}$$

*Множественный коэффициент корреляции* рассчитывается при наличии линейной связи между результирующим признаком *Y* и несколькими

факторными признаками, а также между парой факторных признаков. Множественный коэффициент корреляции вычисляется по формуле

$$R_{y/x_1,x_2} = \sqrt{\frac{\delta^2}{\sigma^2}} = \sqrt{1-\frac{\sigma^2_{ост}}{\sigma^2}},$$

где $\delta^2$ – дисперсия теоретических значений признака *Y*, рассчитанная по уравнению множественной регрессии;

$\sigma^2$ – общая дисперсия признака *Y*;

$\sigma^2_{ост}$ – остаточная дисперсия.

В случае оценки связи между результирующим признаком *Y* и двумя факторными признаками $x_1$ и $x_2$ множественный коэффициент корреляции можно определить по формуле

$$R_{y/x_1,x_2} = \sqrt{\frac{r^2_{yx_1} + r^2_{yx_2} - 2r_{yx_1} r_{yx_2} r_{x_1x_2}}{1-r^2_{x_1x_2}}},$$

где $r$ – парные коэффициенты корреляции между признаками.

Множественный коэффициент корреляции изменяется в пределах от 0 до 1. Приближение *R* к единице свидетельствует о сильной зависимости между признаками. При небольшом числе наблюдений величина множественного коэффициента корреляции, как правило, завышается. В этом случае множественный коэффициент корреляции корректируется:

$$\tilde{R}_{y/x_1,x_2...x_k} = \sqrt{1-(1-R^2)\frac{n-1}{n-k-1}},$$

где $\tilde{R}$ – скорректированное значение;

*n* – число наблюдений;

*k* – число факторных признаков.

Корректировка *R* не производится при условии, если $(n-k)/k \geq 20.$

Проверка значимости множественного коэффициента корреляции осуществляется по критерию Фишера:

$$F_{рас} = \frac{1/2 R^2_{y/x_1,x_2}}{\frac{1}{n-3}(1-R^2_{y/x_1,x_2})}.$$

Множественный коэффициент корреляции считается значимым, если $F_{рас} > F_{q,v_1=2,v_2=n-3}$.

*<u>Вычисление непараметрических коэффициентов корреляции</u>*

Если признаки подчиняются различным (отличным от нормального) законам распределения, то можно рассчитать непараметрические коэффициенты корреляции. Эти коэффициенты могут быть использованы для определения тесноты связи как между количественными, так и между

качественными признаками при условии, если их значения упорядочить или проранжировать по степени убывания или возрастания признака.

Среди непараметрических методов оценки тесноты связи наибольшее распространение получили следующие:

*1. Коэффициент ранговой корреляции Спирмана*

$$\rho_{x/y} = 1 - \frac{6\sum d_i^2}{n(n^2-1)},$$

где $d_i^2$ – квадрат разности рангов;

$n$ – число наблюдений (число пар рангов).

Коэффициент Спирмана принимает значения от *–1* до *1*. Значимость коэффициента проверяется на основе *t*-критерия Стьюдента. Расчетное значение *t*-критерия определяется по формуле

$$t_{рас} = \rho_{x/y}\sqrt{\frac{n-2}{1-\rho_{x/y}^2}}.$$

Значение коэффициента корреляции считается статистически существенным, если $t_{рас} > t_{q,v=n-2}$ .

*2. Коэффициент ранговой корреляции Кенделла*

$$\tau = \frac{2(P-Q)}{n(n-1)} = \frac{2S}{n(n-1)},$$

где $n$ – число наблюдений;

S – разность между числом последовательностей и числом инверсий по второму признаку.

Расчет коэффициента Кенделла производится в такой последовательности:

1. значения *X* ранжируются в порядке возрастания или убывания;
2. значения *Y* располагаются в порядке, соответствующем значениям *X*;
3. для каждого ранга *Y* определяется число следующих за ним значений рангов, превышающих его по величине. Суммируя эти числа, определяем величину *P* (число последовательностей) – меру соответствия последовательностей рангов *X* и *Y*;
4. для каждого ранга *Y* определяется число следующих за ним рангов, меньших его величины. Суммируя величины, получаем величину *Q* (число инверсий);
5. определяется разность по всем членам ряда $S = P - Q$ и вычисляется $\tau$.

Связь между признаками можно признать статистически значимой, если значение коэффициента корреляции $\tau > 0{,}5$.

4. *Коэффициент конкордации.* Для определения тесноты связи между несколькими ранжированными признаками применяют *множественный коэффициент ранговой корреляции – коэффициент конкордации*, который

вычисляется по формуле

$$W = \frac{12S}{m^2(n^3 - n)},$$

где $m$ – количество факторов;

$n$ – число наблюдений;

$S$ – отклонение суммы квадратов рангов от среднего квадрата суммы рангов.

В случае, если корреляционная связь имеет нелинейный характер, она может существовать и при $r$=0. В этой ситуации необходимо вычисление ***корреляционного отношения*** **(η).** Корреляционное отношение показывает, какую долю от общей дисперсии составляет дисперсия, учтенная уравнением регрессии (закономерная составляющая дисперсии):

$$\eta_{y/x} = \frac{S_{\bar{y}_i}}{S_y}; \qquad \eta_{x/y} = \frac{S_{\bar{x}_i}}{S_x}.$$

Незакономерная, случайная составляющая дисперсии характеризует разброс значений вокруг линии регрессии. Таким образом:

$$S^2_y = S^2_{\bar{y}_i} + S^2_{\text{случ.}}$$

Отсюда ясно, что, чем меньше случайная составляющая, т. е., чем меньше разброс значений от линии регрессии, тем выше значение корреляционного отношения. Закономерная составляющая дисперсии рассчитывается по формулам:

$$S_{\bar{y}_i} = \sqrt{\frac{1}{N}\sum_{i=1}^{m}(\bar{y}_i - \bar{y})^2 \cdot n_i} \qquad S_{\bar{x}_i} = \sqrt{\frac{1}{N}\sum_{i=1}^{m}(\bar{x}_i - \bar{x})^2 \cdot n_i}$$

Выборочные данные при этом разбиваются на группы, для каждой из которых подсчитываются $\bar{x}_i$ (или $\bar{y}_i$ ). В формуле $N$ - общее число наблюдений, $m$ - число групп, $n_i$ - число наблюдений в $i$ группе.

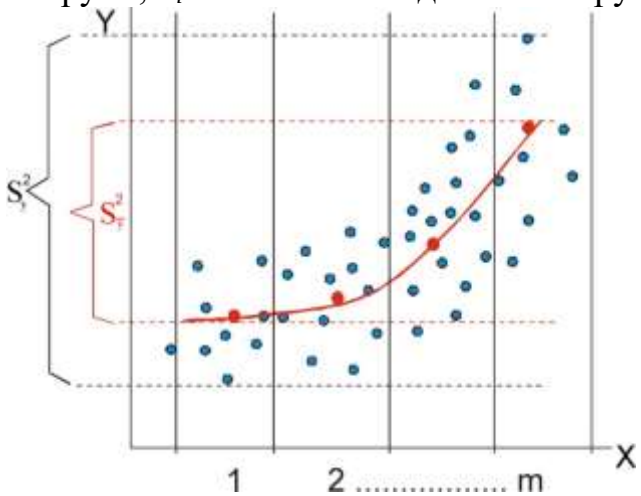



Рисунок – Соотношение закономерной составляющей дисперсии и полной дисперсии

Значения $\eta$ изменяются от 0 (связь отсутствует) до 1 (связь функциональная). В случае линейного характера взаимосвязи:

$$\eta_{x/y} = \eta_{y/x} = |r|.$$

Таким образом, коэффициент корреляции можно рассматривать как частный случай корреляционного отношения. Закономерная составляющая дисперсии в этом случае связана с коэффициентом корреляции соотношением:

$$S^2_{\overline{y_i}} = S^2_y \cdot r^2$$

Значимость отличия $\eta$ от нуля проверяется по критерию:

$$\Theta_y = \frac{\eta^2_{y/x}(N-m-2)}{(1-\eta^2_{y/x})(m-2)} \cdot \sqrt{\frac{(m-2)(N-m-4)}{2(N-4)}}$$

Если $\eta = 0$, то $\Theta_y$ распределена по нормальному закону с параметрами (0,1). Следовательно, если вычисленное значение $\Theta_y$ превысит 3 (при доверительной вероятности 0,99) или 2 (при доверительной вероятности 0,95), считаем, что корреляционная связь существует.

Поскольку уравнения линейной регрессии являются наиболее простыми, на практике всегда необходимо выяснять причины нелинейности взаимосвязи величин и, по возможности, устранять их. К примеру, нелинейность связи может быть обусловлена неоднородностью выборочных данных, специфическими условиями эксперимента (оконтуривание рудных тел по заданной минимальной мощности) и т.д. В любом случае, вначале необходимо выяснить, линейный или нелинейный характер имеет установленная взаимосвязь. Критерий для этой цели предложен Фишером. Он основан на сравнении значений $\eta$ и $r$, которые в случае линейного характера связи должны быть равны.

$$F = \frac{\eta^2_{y/x} - r^2}{1-\eta^2_{y/x}} \cdot \frac{N-m}{m-2}.$$

Обозначения те же. Если вычисленное значение превысит $F_{q,\ m-2,\ N-m}$, взятое из таблиц распределения Фишера, то связь признается ***нелинейной***.

## 7 Регрессионный анализ

Уравнение множественной регрессии можно записать в следующем виде

$$y = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n = a_0 + \sum_{i=1}^{n} a_ix_i$$

Найденное уравнение позволяет наилучшим образом оценить совместное влияние многих параметров на переменную $y$. По значениям коэффициентов

$a_i$ можно также судить, каково влияние на $y$ каждого отдельного параметра.

Задача регрессионного анализа сводится к нахождению коэффициентов уравнения множественной регрессии. Они определяются из соотношения:

$$a_i = \frac{Sy}{Si} \sum r_{ij} (r_{ij})^{-1},$$

где $S_y$ - стандартное отклонение зависимой переменной;

$S_i$ - стандартное отклонение $i$ -й независимой переменной (значения $S_i$ находятся по диагонали ковариационной матрицы);

$r_{ij}$ - парная корреляция между $y$ и $i$-.й независимой переменной; $(r_{ij})^{-1}$- обратная величина парной корреляции между независимыми переменными ( $r_{ij}$ берется из корреляционной матрицы).

Свободный член уравнения вычисляется по формуле:

$$a_0 = \overline{y} - \sum a_i \overline{x}_i .$$

Поскольку уравнение регрессии есть смысл отыскивать лишь в том случае, если корреляция между $y$ и набором переменных $x_i$ существует, то предварительно следует вычислить коэффициент множественной корреляции:

$$R = \sqrt{1 - \frac{|L|}{a_{11}|L'|}}$$

где $|L|$ - определитель ковариационной матрицы;

$|L'|$ - определитель ковариационной матрицы без первого столбца и первой строки.

## 8 Кластерный анализ

Кластер-анализ основан на объединении объектов в группы таким образом, чтобы «расстояние» между объектами в группе было минимальным, а между группами – максимальным. В качестве объединяемых объектов могут выступать как *переменные*, так и *наблюдения*. Соответственно, при этом выделяются либо ассоциации ***элементов*** (минералов), либо ассоциации ***проб*** с близкими характеристиками.

При обработке геохимических данных, в том и другом случае, исходные многомерные данные должны быть ***нормированы.*** Это обусловлено тем, что в кластеры объединяются не только коррелирующиеся между собой элементы, но и те, которые имеют близкий уровень содержаний, то есть, близкие кларки.

Нормирование исходных данных можно выполнить по одной из формул:

$$t = \frac{x - x_{\min}}{x_{\max} - x_{\min}} ; \qquad t = 1 - 2\frac{x - x_{\min}}{x_{\max} - x_{\min}} ; \qquad t = \frac{x - \overline{x}}{S}$$

Число кластеров задается исходя из конкретных условий моделирования. Пробы, относящиеся к разным классам, оконтуривают в

пространстве области со специфическими геохимическими параметрами. По каждому из выделенных классов рассчитывается геохимический спектр, что позволяет дать их содержательную интерпретацию.

В кластерном анализе для количественной оценки сходства объектов вводится понятие *метрики.* Сходство или различие между классифицируемыми объектами устанавливается в зависимости от метрического расстояния между ними. Если каждый объект описывается *k* признаками, то он может быть представлен как точка в k-мерном пространстве и сходство с другими объектами будет определяться как соответствующее расстояние. В кластерном анализе используются различные меры расстояния между объектами (*метрики)*:

1) евклидово расстояние: $d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$

2) взвешенное евклидово расстояние: $d_{ij} = \sqrt{\sum_{k=1}^{m} w_k (x_{ik} - x_{jk})^2}$

3) расстояние *city-block:* $d_{ij} = \sum_{k=1}^{m} \left| x_{ik} - x_{jk} \right|$

4) расстояние *Минковского:* $d_{ij} = \left( \sum_{k=1}^{m} \left| x_{ik} - x_{jk} \right|^P \right)^{1/P}$

где $d_{ij}$ – расстояние между *i*-м и *j*-м объектами;
$x_{ik}$, $x_{jk}$ – значения *k*-й переменной соответственно у *i*-го и *j*-го объектов;
$w_k$ – вес, приписываемый *k*-й переменной.

Из всех методов кластерного анализа самыми распространенными являются иерархические агломеративные методы. Сущность этих методов заключается в том, что на первом шаге каждый объект выборки рассматривается как отдельный кластер. Процесс объединения кластеров происходит последовательно: на основании матрицы расстояний или матрицы сходства объединяются наиболее близкие объекты. Если матрица сходства первоначально имеет размерность $m \times m$ т, то полностью процесс кластеризации завершается за m – 1 шагов, в итоге все объекты будут объединены в один кластер. Последовательность объединения легко поддается геометрической интерпретации и может быть представлена в виде графа-дерева (дендрограммы). На дендрограмме указываются номера объединяемых объектов и расстояние (или иная мера сходства), при ко-тором произошло объединение.

Дендрограмма на рисунке показывает, что в данном случае на первом шаге были объединены в один кластер объекты n2 и n3. Расстояние между ними 0,15. На втором шаге к ним присоединился объект п1. Расстояние от первого объекта до кластера, содержащего объекты n2 и n3, было 0,3 и т. д.

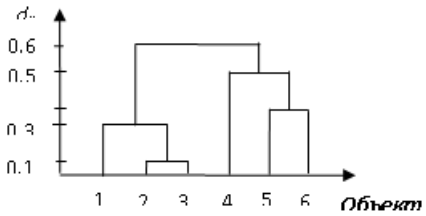



Рисунок – ример дендрограммы иерархического агломеративного кластерного анализа

Множество методов иерархического кластерного анализа различаются не только используемыми мерами сходства (различия), но и алгоритмами классификации. Из них наиболее распространены: метод одиночной связи, метод полных связей, метод средней связи, метод Уорда.

Метод одиночной связи. Алгоритм образования кластеров следующий: на основании матрицы сходства (различия) определяются два наиболее схожих или близких объекта, они и образуют первый кластер. На следующем шаге выбирается объект, который будет включен в этот кластер. Таким объектом будет тот, который имеет наибольшее сходство хотя бы с одним из объектов, уже включенных в кластер. Например, имеется матрица евклидовых расстояний между объектами.

При совпадении данных на основании одинаковых мер сходства (различия) будет идти образование сразу нескольких кластеров.

К достоинствам этого метода следует отнести нечувствительность алгоритма к преобразованиям исходных переменных и его простоту. Недостатками являются необходимость постоянного хранения матрицы сходства и невозможность определения по результатам кластеризации, сколько же кластеров можно образовать в исследуемой совокупности объектов.

*Метод полных связей.* Включение «нового» объекта в кластер происходит только в том случае, если расстояние между всеми объектами не больше некоторого заданного уровня С.

*Метод средней связи.* Для решения вопроса о включении нового объекта в уже существующий кластер вычисляется среднее значение меры сходства, которое затем сравнивается с заданным пороговым уровнем. Если речь идет об объединении двух кластеров, то вычисляют расстояние между их центрами и сравнивают его с заданным пороговым значением. Рассмотрим геометрический пример с двумя кластерами.

*Метод Уорда.* Данный метод предполагает, что на первом шаге каждый кластер включает один объект. Первоначально объединяются два ближайших кластера. Для них определяются средние значения каждого признака и рассчитывается сумма квадратов отклонений $V_k$:

$$V_k = \sum_{i=1}^{n_k} \sum_{j=1}^{p} (x_{ij} - \overline{x_{jk}})^2$$

где $k$ – номер кластера;

$i$– номер объекта;

$j$ – номер признака;

$n_k$ – количество объектов в $k$-*м* кластере;

$p$ – количество признаков, характеризующих каждый объект.

В дальнейшем на каждом шаге работы алгоритма объединяются те объекты или кластеры, которые дают наименьшее приращение величины $V_k$. Метод Уорда приводит к образованию кластеров приблизительно равных размеров с минимальной внутрикластерной вариацией. В итоге все объекты оказываются объединенными в один кластер.

Алгоритм иерархического кластерного анализа можно представить в виде последовательности процедур.

*Шаг 1.* Значения исходных переменных нормируются одним из способов, указанных в разд. 8.2.2.

*Шаг 2.* Рассчитывается матрица расстояний или матрица мер сходства.

*Шаг 3.* Находится пара самых близких кластеров. По выбранному алгоритму объединяются эти два кластера. Новому кластеру присваивается меньший из номеров объединяемых кластеров.

*Шаг 4.* Шаги 2, 3 и 4 повторяются до тех пор, пока все объекты не будут объединены в один кластер, или до достижения заданного «порога» сходства.

Мера сходства для объединения двух кластеров в шаге 3 определяется четырьмя методами.

*Метод «ближайшего соседа»* – степень сходства оценивается по степени сходства между наиболее схожими (ближайшими) объектами этих кластеров.

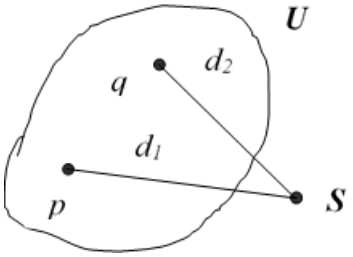



Рисунок Определение расстояния между кластерами методом «ближайшего соседа»

Пусть $d_1$ и $d_2$ – евклидовы расстояния, тогда если $d_1 > d_2$, то $S$ войдет в кластер $U$ по $d_2$.

*Метод «дальнего соседа»* – степень сходства оценивается по степени сходства между наиболее отдаленными (несхожими) объектами кластеров, тогда $S$ войдет в кластер $U$ по $d_1$.

*Метод средней связи* – степень сходства оценивается как средняя величина степеней сходства между объектами кластеров. В этом случае $S$ войдет в кластер $U$ по $d=1/2(d_1+d_2)$.

*Метод медианной связи* – расстояние между любым кластером $S$ и новым кластером, который получился в результате объединения кластеров $p$ и $q$, определяется как расстояние от центра кластера $S$ до середины отрезка, соединяющего центры кластеров $p$ и $q$.

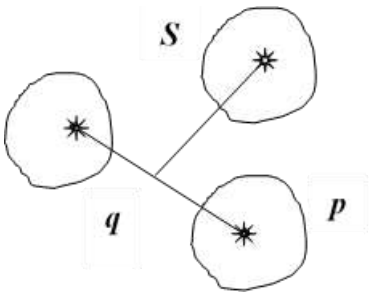



Рисунок – Определение расстояний между кластерами методом медианной связи

Использование различных алгоритмов объединения в иерархических агломеративных методах риводит к различным кластерным структурам и сильно влияет на качество проведения кластеризации. Поэтому алгоритм должен выбираться с учетом имеющихся сведений о существующей структуре совокупности наблюдаемых объектов или с учетом требований оптимизации математических критериев.

*Метод К-средних*

Наряду с иерархическими методами классификации существует многочисленная группа так называемых *итеративных методов* кластерного анализа. Сущность их заключается в том, что процесс классификации начинается с задания некоторых начальных условий (количество образуемых кластеров, порог завершения процесса классификации и т. д.). Итеративные методы в большей степени, чем иерархические, требуют от пользователя интуиции при выборе типа классификационных процедур и задания начальных условий разбиения, так как большинство этих методов очень чувствительны к изменению задаваемых параметров. Например, выбранное случайным образом число кластеров может не только увеличить трудоемкость процесса классификации, но и привести к образованию «размытых» или мало наполняемых кластеров. Поэтому целесообразно сначала провести классификацию по одному из иерархических методов или на основании экспертных оценок, а затем уже подбирать начальное разбиение и статистический критерий для работы итерационного алгоритма. Как и в иерархическом кластерном анализе, в итерационных методах существует проблема определения числа кластеров. В общем случае их число может быть неизвестно. Не все итеративные методы требуют первоначального задания числа кластеров. Но для окончательного решения вопроса о структуре изучаемой совокупности можно испробовать несколько алгоритмов, меняя либо число образуемых кластеров, либо установленный порог близости для объединения объектов в кластеры. Тогда появляется возможность выбрать наилучшее разбиение по задаваемому критерию качества.

Пусть имеется $n$ наблюдений, каждое из которых характеризуется $p$ признаками $X_1$ $X_2$,..., $X_p$. Эти наблюдения необходимо разбить на $k$ кластеров. Для начала из $n$ точек исследуемой совокупности отбираются случайным

образом или задаются исследователем исходя из каких-либо априорных соображений $k$ точек (объектов). Эти точки принимаются за эталоны. Каждому эталону присваивается порядковый номер, который одновременно является и номером кластера. На первом шаге из оставшихся $(n - k)$ объектов извлекается точка $X_i$ с координатами $(x_{i1}, x_{i2}, ..., x_{ip})$ и проверяется, к какому из эталонов (центров) она находится ближе всего. Для этого используется одна из метрик, например евклидово расстояние:

$$d_{il} = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{lj})^2}$$

Проверяемый объект присоединяется к тому центру (эталону), которому соответствует min $d_{il}$ ($l$ = 1,..., $k$). Эталон заменяется новым, пересчитанным с учетом присоединенной точки, и вес его (количество объектов, входящих в данный кластер) увеличивается на единицу. Если встречаются два или более минимальных расстояния, то $i$-й объект присоединяют к центру с наименьшим порядковым номером. На следующем шаге выбирается точка $X_{i+1}$ и для нее повторяются все процедуры. Таким образом, через $(n - k)$ шагов все точки (объекты) совокупности окажутся отнесенными к одному из $k$ кластеров, но на этом процесс разбиения не заканчивается. Для того чтобы добиться устойчивости разбиения по тому же правилу, все точки $X_1, X_2, ..., X_n$ опять подсоединяются к полученным кластерам, при этом веса продолжают накапливаться. Новое разбиение сравнивается с предыдущим. Если они совпадают, то работа алгоритма завершается. В противном случае цикл повторяется. Окончательное разбиение имеет центры тяжести, которые не совпадают с эталонами, их можно обозначить через $C_1, C_2,..., C_k$. При этом каждая точка $X_i$ ($i$ = 1, 2, ..., $n$) будет относиться к такому кластеру (классу) $l$, для которого

$$d(x_i, c_l) = \min_{1 \le j \le K} d(x_i, c_j)$$

Возможны две модификации метода $k$-средних. Первая предполагает пересчет центра тяжести кластера после каждого изменения его состава, а вторая – лишь после того, как будет завершен просмотр всех данных. В обоих случаях итеративный алгоритм этого метода минимизирует дисперсию внутри каждого кластера, хотя в явном виде такой критерий оптимизации не используется. Пример работы итеративного алгоритма метода $k$-средних рассмотрен в гл. 9.

Вычислительные процедуры большинства итеративных методов классификации сводятся к выполнению следующих шагов:

*Шаг 1.* Выбирается число кластеров, на которые должна быть разбита совокупность, задается первоначальное разбиение объектов и определяются центры тяжести кластеров.

*Шаг 2.* В соответствии с выбранными мерами сходства определяется новый состав каждого кластера.

*Шаг 3.* После полного просмотра всех объектов и распределения их по кластерам осуществляется пересчет центров тяжести кластеров.

*Шаг 4.* Процедуры 2 и 3 повторяются до тех пор, пока следующая

итерация не даст такой же состав кластеров, что и предыдущая.

*Критерии качества классификации*

При использовании различных методов кластерного анализа для одной и той же совокупности могут быть получены различные варианты разбиения. Существенное влияние на характеристики кластерной структуры оказывают: во-первых, набор признаков, по которым осуществляется классификация, во-вторых, тип выбранного алгоритма. Например, иерархические и итеративные методы приводят к образованию различного числа кластеров. При этом сами кластеры различаются и по составу, и по степени близости объектов. Выбор меры сходства также влияет на результат разбиения. Если используются методы с эталонными алгоритмами, например метод *k*-средних, то задаваемые начальные условия разбиения в значительной степени определяют конечный результат разбиения. После завершения процедур классификации необходимо оценить полученные результаты. Для этой цели используется некоторая мера качества классификации, которую принято называть *функционалом* или *критерием качества.* Наилучшим по выбранному функционалу следует считать такое разбиение, при котором достигается экстремальное (минимальное или максимальное) значение целевой функции – функционала качества.

В большинстве случаев алгоритмы классификации и критерии качества связаны между собой, т.е. определенный алгоритм обеспечивает получение экстремального значения соответствующего функционала качества. Например, использование метода Уорда приводит к получению кластеров с минимальной внутриклассовой дисперсией.

Рассмотрим наиболее распространенные функционалы качества.

1. Сумма квадратов расстояний до центров классов

$$F_1 = \sum_{l=1}^{k} \sum_{i \in S_l} d^2(X_i, \overline{X_l})$$

где $l$ – номер кластера ($l$ = 1,2,..., $k$),

$X_i$ – вектор значений переменных для $i$-го объекта, входящего в $l$-й кластер;

$\overline{X}_l$ – центр $l$-*го* кластера;

$d(X_i, \overline{X}_l\}$ – расстояние между $i$-м объектом и центром $l$-го кластера.

При использовании этого критерия стремятся получить такое разбиение совокупности объектов на $k$ кластеров, при котором значение $F_1$ было бы минимальным.

2. Сумма внутриклассовых расстояний между объектами

$$F_2 = \sum_{l=1,}^{k} \sum_{i,j \in S_l} d_{ij}^2$$

В этом случае наилучшим следует считать такое разбиение, при котором достигается минимальное значение $F_2$, т.е. получены кластеры большой «плотности». Объекты, попавшие в один кластер, близки между собой по значениям тех переменных, которые использовались для классификации.

3. Суммарная внутриклассовая дисперсия

$$F_3 = \sum_{l=1}^{k} \sum_{j=1}^{p} \sigma_{ij}^2$$

где $\sigma^2_{ij}$ – дисперсия $j$-й переменной в кластере $S_l$.

В данном случае разбиение, при котором сумма внутриклассовых (внутригрупповых) дисперсий будет минимальной, следует считать оптимальным.

Перечисленные выше способы оценки качества разбиения предполагают чисто формальный подход и являются для исследователя только вспомогательными средствами. Основная роль принадлежит содержательному анализу результатов классификации. Выбор лучшего варианта разбиения облегчается в значительной мере серьезной подготовительной работой, в частности выбором признаков, характеризующих классифицируемые объекты. В зависимости от количества признаков, их взаимосвязи, выбранного масштаба измерения подбирается наиболее подходящий алгоритм классификации, задаются начальные параметры разбиения. Все это облегчает интерпретацию результатов разбиения и позволяет судить о его качестве с точки зрения поставленной задачи.

# 3 ПРАКТИЧЕСКИЙ РАЗДЕЛ

## 3.1 Перечень практических занятий

1. Предварительный анализ данных.
2. Точечные и интервальные оценки
3. Проверка статистических гипотез.
4. Дисперсионный анализ.
5. Корреляционный анализ.
6. Регрессионный анализ.
7. Кластерный анализ.

## 3.2 Задания для выполнения практических работ

### 1 Предварительный анализ данных.

**Задание.** Имеются статистические данные о добыче нефти по двум скважинам (k – номе слушателя в журнале группы)

Таблица 1 – Исходные данные

| Номер наблюдения | Добыча нефти среднесут, скважина 1 | Добыча нефти среднесут, скважина 2 |
|---|---|---|
| 1 | 34 + k/10 | 41 |
| 2 | 38 | 37 + k/10 |
| 3 | 35 | 36 |
| 4 | 45 | 43 |
| 5 | 52 + /10 | 52 |
| 6 | 49 | 51 |
| 7 | 45 | 48 |
| 8 | 49 | 51 |
| 9 | 51 | 51 |
| 10 | 49 | 51 |
| 11 | 38 | 43 |
| 12 | 45 | 48 |
| 13 | 49 | 38 + k/10 |
| 14 | 38 | 43 |
| 15 | 41 | 41 |
| 16 | 37 + k/10 | 37 |
| 17 | 41 | 51 |
| 18 | 45 | 43 |
| 19 | 40 | 48 |
| 20 | 40 | 41 |

Требуется построить:
– несгруппированный ряд,
– ранжированный ряд,
– вариационный ряд,
– гистограмму,
– ящик с усами.

**2 Точечные и интервальные оценки**

**Задание.** Определить точечные и интервальные оценки о добыче нефти по двум скважинам.

**3. Проверка статистических гипотез.**

**Задание**. Сравнить неизвестные средние и дисперсии о добыче нефти по двум скважинам по критериям Стьюдента.

**4. Дисперсионный анализ.**

**Задание**. Требуется выяснять, как влияет гипсометрическое положение рудных тел (А) на среднее содержание в них золота.

| Наблюдение | Фактор А | | |
|---|---|---|---|
| | Горизон (+800 м) | Горизон (+500 м) | Горизонт (+200 м) |
| 1 | 1 | 2 | 3 |
| 2 | 5 | 5 | 10 |
| 3 | 4 | 4 | 8 |
| 4 | 6 | 6 | 9 |
| 5 | 3 | 7 | 10 |
| 6 | 4 | 4 | 7 |
| 7 | 5 | 8 | 6 |
| 8 | 2 | 5 | 8 |

**5. Корреляционный анализ.**

**Задание 1.** Необходимо выяснить, существует ли зависимость между содержаниями золота и свинца на одном из месторождений, по данным опробования:

Таблица – Содержания золота и свинца в рудном теле

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *Au*, г/т | 2,5 | 1,2 | 3,6 | 1,1 | 4,8 | 12,1 | 8,2 | 4,9 |
| *Pb*, $n \cdot 10^{-3}$% | 8 | 8 | 8 | 10 | 5 | 15 | 5 | 3 |
| *Au*, г/т | 6,8 | 13,2 | 15,1 | 7,8 | 8,8 | 9,1 | 2,6 | 5,5 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *Pb*, $n \cdot 10^{-3}$% | 4 | 8 | 10 | 5 | 8 | 6 | 6 | 3 |

**6. Регрессионный анализ.**

**Задание**. По данным таблицы оценить модель множественной линейной регрессии добычи нефти от водонасыщенности и сопутствующего газа.

| | Y--добыча нефти | X-водонасыщенность | Z--сопутствующий газ |
|---|---|---|---|
| 1 | 32 | 32 | 74 |
| 2 | 38 | 20 | 73 |
| 3 | 36 | 24 | 91 |
| 4 | 47 | 18 | 115 |
| 5 | 52 | 19 | 136 |
| 6 | 54 | 15 | 137 |
| 7 | 50 | 20 | 129 |
| 8 | 51 | 25 | 133 |
| 9 | 51 | 20 | 129 |
| 10 | 50 | 22 | 129 |
| 11 | 44 | 24 | 110 |
| 12 | 45 | 26 | 113 |
| 13 | 44 | 29 | 106 |
| 14 | 44 | 26 | 93 |
| 15 | 41 | 25 | 106 |
| 16 | 37 | 29 | 92 |
| 17 | 41 | 25 | 107 |
| 18 | 42 | 23 | 99 |
| 19 | 40 | 25 | 95 |
| 20 | 41 | 25 | 96 |

**7. Кластерный анализ.**

**Задание**. Провести классификацию скважин.

| Номер скважины | Y- добыча нефти | X - водонасыщенность |
|---|---|---|
| 1 | 32 | 23 |
| 2 | 38 | 20 |
| 3 | 36 | 24 |
| 4 | 47 | 18 |
| 5 | 52 | 19 |
| 6 | 54 | 15 |
| 7 | 50 | 20 |

# 4 РАЗДЕЛ КОНТРОЛЯ ЗНАНИЙ

## 4.1 Вопросы к зачету учебной дисциплине «Статистические методы в геологии»

1 Предварительный анализ данных: ряд данных, несгруппированный ряд, ранжированный ряд, вариативный ряд, интервальный ряд, полигон частот, гистограмма.

2 Точечные оценки параметров генеральной совокупности. Доверительные интервалы.

3 Проверка статистических гипотез.

4 Гипотезы о параметрах многомерной нормально распределенной генеральной совокупности.

5 Сравнение двух генеральных совокупностей: параметрические критерии.

6 Сравнение двух генеральных совокупностей: непараметрические критерии.

7 Однофакторный дисперсионный анализ, критерий Фишера. Коэффициент детерминации.

8 Непараметрический однофакторный анализ. Критерий Краскала-Уоллиса.

9 Взаимосвязь факторов. Парные коэффициенты корреляции. Частные коэффициенты корреляции. Проверка значимости параметров связи.

10 Множественный коэффициент детерминации и его свойства.

11 Критерий независимости на основе рангового коэффициента корреляции Спирмена.

12 Модель множественной линейной регрессии. Значимость коэффициентов. Значимость коэффициента детерминации.

13 Иерархические кластер-процедуры.

## 4.2 Критерии оценок результатов учебной деятельности студентов по учебной дисциплине «Статистические методы в геологии» (на основании письма Министерства образования Республики Беларусь от 28.05.2013 г. № 09- 10/53-ПО)

Десятибалльная шкала в зависимости от величины балла и отметки включает следующие критерии:

**10 (десять) баллов, зачтено** выставляется за систематизированные, глубокие и полные знания по всем разделам учебной программы учреждения высшего образования по учебной дисциплине, а также по основным вопросам, выходящим за её пределы; за точное использование научной терминологии (в том числе на иностранном языке), грамотное, логически правильное изложение ответа на вопросы; за безупречное владение инструментарием учебной дисциплины, умение его эффективно использовать в постановке и решении научных и профессиональных задач; за выраженную способность самостоятельно и творчески решать сложные проблемы в нестандартной ситуации; за полное и глубокое усвоение основной, дополнительной литературы, по изучаемой учебной дисциплине; за умение свободно ориентироваться в теориях, концепциях и направлениях по изучаемой учебной дисциплине и давать им аналитическую оценку, использовать научные достижения других дисциплин; за творческую самостоятельную работу на практических, лабораторных занятиях, активное творческое участие в групповых обсуждениях, высокий уровень культуры исполнения заданий.

**9 (девять) баллов, зачтено** выставляется за систематизированные, глубокие и полные знания по всем разделам учебной программы учреждения высшего образования по учебной дисциплине; за точное использование научной терминологии (в том числе на иностранном языке), грамотное, логически правильное изложение ответа на вопросы; за владение инструментарием учебной дисциплины, умение его эффективно использовать в постановке и решении научных и профессиональных задач; за способность самостоятельно и творчески решать сложные проблемы в нестандартной ситуации в рамках учебной программы учреждения высшего образования по учебной дисциплине; за полное усвоение основной и дополнительной литературы, рекомендованной учебной программой учреждения высшего образования по учебной дисциплине; за умение ориентироваться в теориях, концепциях и направлениях по изучаемой учебной дисциплине и давать им аналитическую оценку; за систематическую, активную самостоятельную работу на практических, лабораторных занятиях, творческое участие в групповых обсуждениях, высокий уровень культуры исполнения заданий.

**8 (восемь) баллов, зачтено** выставляется за систематизированные, глубокие и полные знания по всем разделам учебной программы учреждения высшего образования по учебной дисциплине в объеме учебной программы учреждения высшего образования по учебной дисциплине; за использование

научной терминологии (в том числе на иностранном языке), грамотное, логически правильное изложение ответа на вопросы, умение делать обоснованные выводы и обобщения; за владение инструментарием учебной дисциплины (методами комплексного анализа, техникой информационных технологий), умение его использовать в постановке и решении научных и профессиональных задач; за способность самостоятельно решать сложные проблемы в рамках учебной программы учреждения высшего образования по учебной дисциплине; за усвоение основной и дополнительной литературы, рекомендованной учебной программой учреждения высшего образования по учебной дисциплине; за умение ориентироваться в теориях, концепциях и направлениях по изучаемой учебной дисциплине и давать им аналитическую оценку; за активную самостоятельную работу на практических, лабораторных занятиях, систематическое участие в групповых обсуждениях, высокий уровень культуры исполнения заданий.

**7 (семь) баллов, зачтено** выставляется за систематизированные, глубокие и полные знания по всем разделам учебной программы учреждения высшего образования по учебной дисциплине; за использование научной терминологии (в том числе на иностранном языке), грамотное, логически правильное изложение ответа на вопросы, умение делать обоснованные выводы и обобщения; за владение инструментарием учебной дисциплины, умение его использовать в постановке и решении научных и профессиональных задач; за свободное владение типовыми решениями в рамках учебной программы учреждения высшего образования по учебной дисциплине; за усвоение основной и дополнительной литературы, рекомендованной учебной программой учреждения высшего образования по учебной дисциплине; за умение ориентироваться в основных теориях, концепциях и направлениях по изучаемой учебной дисциплине и давать им аналитическую оценку; за самостоятельную работу на практических, лабораторных занятиях, участие в групповых обсуждениях, высокий уровень культуры исполнения заданий.

**6 (шесть) баллов, зачтено** выставляется за достаточно полные и систематизированные знания в объеме учебной программы учреждения высшего образования по учебной дисциплине; за использование необходимой научной терминологии, грамотное, логически правильное изложение ответа на вопросы, умение делать обобщения и обоснованные выводы; за владение инструментарием учебной дисциплины, умение его использовать в решении учебных и профессиональных задач; за способность самостоятельно применять типовые решения в рамках, учебной программы учреждения высшего образования по учебной дисциплине; за усвоение основной литературы, рекомендованной учебной программой учреждения высшего образования по учебной дисциплине; за умение ориентироваться в базовых теориях, концепциях и направлениях по изучаемой дисциплине и давать им сравнительную оценку; за активную самостоятельную работу на практических, лабораторных занятиях, периодическое участие в групповых

обсуждениях, высокий уровень культуры исполнения заданий.

**5 (пять) баллов, зачтено** выставляется за достаточные знания в объеме учебной программы учреждения высшего образования по учебной дисциплине; за использование научной терминологии, грамотное, логически правильное изложение ответа на вопросы, умение делать выводы; за владение инструментарием учебной дисциплины, умение его использовать в решении учебных и профессиональных задач; за способность самостоятельно применять типовые решения в рамках учебной программы учреждения высшего образования по учебной дисциплине; за усвоение основной литературы, рекомендованной учебной программой учреждения высшего образования по учебной дисциплине; за умение ориентироваться в базовых теориях, концепциях и направлениях по изучаемой учебной дисциплине и давать им сравнительную оценку; за самостоятельную работу на практических, лабораторных занятиях, фрагментарное участие в групповых обсуждениях, достаточный уровень культуры исполнения заданий.

**4 (четыре) балла, зачтено** выставляется за достаточный объем знаний в рамках образовательного стандарта высшего образования; за усвоение основной литературы, рекомендованной учебной программой учреждения высшего образования по учебной дисциплине; за использование научной терминологии, логическое изложение ответа на вопросы, умение делать выводы без существенных ошибок; за владение инструментарием учебной дисциплины, умение его использовать в решении стандартных (типовых) задач; за умение под руководством преподавателя решать стандартные (типовые) задачи; за умение ориентироваться в основных теориях, концепциях и направлениях по изучаемой учебной дисциплине и давать им оценку; за работу под руководством преподавателя на практических, лабораторных занятиях, допустимый уровень культуры исполнения заданий.

**3 (три) балла, не зачтено** выставляется за недостаточно полный объем знаний в рамках образовательного стандарта высшего образования; за знание части основной литературы, рекомендованной учебной программой учреждения высшего образования по учебной дисциплине; за использование научной терминологии, изложение ответа на вопросы с существенными, логическими ошибками; за слабое владение инструментарием учебной дисциплины, некомпетентность в решении стандартных (типовых) задач; за неумение ориентироваться в основных теориях, концепциях и направлениях изучаемой учебной дисциплины; за пассивность на практических и лабораторных занятиях, низкий уровень культуры исполнения заданий.

**2 (два) балла, не зачтено** выставляется за фрагментарные знания в рамках образовательного стандарта высшего образования; за знания отдельных литературных источников, рекомендованных учебной программой учреждения высшего образования по учебной дисциплине; за неумение использовать научную терминологию учебной дисциплины, наличие в ответе грубых, логических ошибок; за пассивность на практических и лабораторных занятиях, низкий уровень культуры исполнения заданий.

**1 (один) балл, не зачтено** выставляется за отсутствие знаний и (компетенций) в рамках образовательного, стандарта высшего образования, отказ от ответа, неявка на аттестацию без уважительной причины.

Экзаменуемые имеют право пользоваться во время проведения зачета учебной программой курса. Результаты зачета объявляются, как правило, в день его проведения и заносятся в зачетную книжку экзаменуемого и экзаменационную ведомость.

Студенты, не выполнившие в полном объеме требования по изучению данной дисциплины, не допускаются кафедрой к сдаче экзамена.

## 4.3 Самостоятельная работа по дисциплине «Статистические методы в геологии»

В ходе освоения студентам рекомендуется использовать в качестве информационно-методической поддержки электронные тексты лекционных занятий, задания для самостоятельной работы, контрольные вопросы. Наиболее эффективными формами и методами организации самостоятельной работы студентов являются: выполнение практических заданий.

# 5 ВСПОМОГАТЕЛЬНЫЙ РАЗДЕЛ

## 5.1 Учебная программа по учебной дисциплине «Статистические методы в геологии»

Министерство образования Республики Беларусь

Учреждение образования
«Гомельский государственный университет имени Франциска Скорины»

Институт дополнительного образования

УТВЕРЖДАЮ
Ректор ГГУ имени Ф.Скорины
С.А.Хахомов

# УЧЕБНАЯ ПРОГРАММА ПО УЧЕБНОЙ ДИСЦИПЛИНЕ

«Статистические методы в геологии»

1-51 02 71 Разработка и эксплуатация нефтяных и газовых месторождений
(квалификация: горный инженер)

в соответствии с типовым учебным планом переподготовки, утвержденным 01.06.2021, регистрационный № № 110

2022

Разработчик программы:

Л. Н. Марченко, заведующий кафедрой фундаментальной и прикладной математики

Рекомендована к утверждению:
кафедрой переподготовки и повышения квалификации
Протокол заседания от 01.09.2022 № 1;

научно-методическим советом учреждения образования «Гомельский государственный университет имени Франциска Скорины»
Протокол заседания от 13.09.2022 № 1

## Пояснительная записка

Учебная программа по учебной дисциплине «Статистические методы в геологии» специальности переподготовки 1-51 02 71 Разработка и эксплуатация нефтяных и газовых месторождений (квалификация: горный инженер) разработана с целью реализации образовательной программы переподготовки руководящих работников и специалистов, имеющих высшее образование.

Учебная дисциплина «Статистические методы в геологии» специальности переподготовки (далее – дисциплина) относится к компоненту «Общенаучные и общепрофессиональные дисциплины» учебного плана переподготовки.

При геологических исследованиях накапливается большое количество геологической информации: результаты геологической документации буровых скважин, спектральных и химических анализов руд, пород и минералов, данные геофизических и геохимических измерений и др. Одно из важнейших направлений в геологии состоит в широком внедрении автоматизированных методов накопления, хранения, обработки и передачи геологической информации с целью повышения эффективности геологических исследований. Практическое использование статистических методов и моделей на ПК необходимо при решении реальных геологических задач и основано на знаниях стандартных пакетов прикладных программ. Поэтому дисциплина «Статистические методы в геологии» актуальна для слушателей специальности 1-25 02 71 «Разработка и эксплуатация нефтяных и газовых месторождений»

***Целью*** дисциплины «Статистические методы в геологии» является овладение слушателями основами статистического анализа данных на ПК.

***Задачами*** являются:

– усвоение основных технологий использования пакетов прикладных программ в статистическом анализе;

– овладение практическими навыками использования пакетов прикладных программ в экономическом анализе;

– анализ реальных геологических объектов с помощью пакетов прикладных программ.

В результате изучения дисциплины необходимо:

*знать:*

– основные статистические методы для анализа геологических данных;

– основы построения статистических моделей в MS Excel;

*уметь:*

– использовать MS Excel при проведении корреляционно-регрессионного анализа,

– проводить дисперсионный анализ MS Excel;

– анализировать временные ряды в MS Excel.

*владеть:*

– основами статистического моделирования в MS Excel;

– навыками использования ПК при анализе реальных геологических объектов.

Изучение учебной дисциплины компонента «Статистические методы в геологии» направлено на повышение качества воспитательного процесса посредством формирования системного и логического мышления, ответственности за результаты учебы, трудолюбия, обеспечивает развитие математических способностей и склонностей обучающихся, а также их личностный рост и гражданское становление.

Изучение дисциплины рассчитано на **30 часов.**

**Распределение учебных часов по видам занятий:**

в **заочной** форме получения образования: лекции – 10 часов, практические занятия – 8 часов, самостоятельная работа – 10 часов;

в **очной (вечерней)** форме получения образования: лекции – 14 часов, практические занятия – 14 часов, самостоятельная работа – 2 часов;

**Формы текущей аттестации** слушателей по дисциплине – зачет.

## СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Тема 1 Математическое моделирование геологических объектов

Геологические объекты и их свойства. Свойства геологических объектов. Основные виды математических моделей, применяемых в геологии. Основные этапы и модели статистического анализа.

Тема 2 Предварительный анализ данных

Ранжированный ряд. Вариационный ряд. Частота и относительная частота. Полигон частот. Кумулянта. Интервальный ряд. Гистограмма. Ящик с усами.

Тема Точечные и интервальные оценки

Меры положения. Медиана. Мода. Среднее значение. Меры разброса. Дисперсия. Среднеквадратическое отклонение. Меры формы. Моменты случайной величины. Асимметрия. Эксцесс.

Тема 4 Проверка статистических гипотез

Общая схема проверки гипотез. Возможные ошибки при проверке гипотез. Сравнение математических ожиданий в двух выборках. Сравнение дисперсий в двух выборках. Проверка гипотезы о виде законов распределения вероятности генеральной совокупности. Критерии согласия.

Тема 5 Дисперсионный анализ

Уровни фактора. Результативный фактор. Однофакторный дисперсионный анализ. Критерий Фишера. Выборочный коэффициент детерминации. Двухфакторный дисперсионный анализ без повторений и с повторениями. Общая вариация. Критерий Фишера. Непараметрический дисперсионный анализ.

Тема 6 Корреляционный анализ

Основы корреляционного анализа. Выборочный парный коэффициент корреляции. Шкала Чеддока. Корреляционная матрица. Значимость парного коэффициента корреляции. Выборочный частный коэффициент корреляции и его значимость. Выборочный множественный коэффициент корреляции и его значимость.

Тема 7 Регрессионный анализ

Эндогенные и экзогенные переменные. Основные предпосылки модели множественной линейной регрессии. Метод наименьших квадратов (МНК). Анализ адекватности линейной модели множественной регрессии. Коэффициент детерминации модели. Значимость коэффициентов регрессии. Проверка гипотезы об адекватности модели в целом. $F$-критерий Фишера

Тема 8 Кластерный анализ

Измерение близости объектов. Характеристики близости: линейное расстояние, евклидово расстояние, расстояние Чебышева, Манхэттенское расстояние. Расстояние между кластерами. Дендрограмма.

## Требования к проверке результатов самостоятельной работы

Самостоятельная работа слушателей по дисциплине «Статистические методы в геологии» представляет собой важный компонент познавательно-практической деятельности слушателей.

Учебный материал, представленный в разделе самостоятельной работы, направлен на закрепление и углубление знаний по основным разделам дисциплины, глубокое самостоятельное овладение теоретическим материалом, активизацию познавательной деятельности слушателей и развитие профессионально значимых умений и навыков.

Контроль самостоятельной работы слушателей осуществляется преподавателем во время проведения практических занятий, а также текущей аттестации слушателей.

Примерный перечень вопросов для самостоятельного изучения включается в материалы для текущей аттестации слушателей.

| **Тема занятия** | **Вид задания** | **№ литературного источника в соответствии со списком рекомендуемой литературы (при необходимости - объем)** | **Форма контроля** |
|---|---|---|---|
| Предварительный анализ данных | Проанализировать литературные источники и составить конспект. Выполнить индивидуальное задание. | Осн. [1 -11] | Конспект |
| Числовые характеристики случайной величины | Проанализировать литературные источники и составить конспект. Выполнить индивидуальное задание. | Осн. [1 -11] | Конспект |
| Проверка статистических гипотез | Проанализировать литературные источники и составить конспект. Выполнить индивидуальное задание. | Осн. [1-11] | Конспект |
| Дисперсионный анализ | Изучить методы использования однофакторного дисперсионного анализа в разработке месторождений нефти и газа. Выполнить индивидуальное задание. | Осн. [1-11] | Конспект |
| Корреляционный анализ | Исследовать взаимосвязь показателей в разработке нефтяных месторождений. Выполнить индивидуальное задание. | Осн. [1-11] | Конспект, |
| Регрессионный анализ | Построить регрессионные модели в разработке нефтяных месторождений. | Осн. [1-11] | Конспект, презентация |

| | | | |
|---|---|---|---|
| | Выполнить индивидуальное задание. | | |
| Кластерный анализ | Проанализировать литературные источники и составить конспект. Выполнить индивидуальное задание. | Осн. [1-11] | Конспект, презентация |

## Время, отведенное на изучение отдельных тем

**Заочная** форма получения образования

| Наименование разделов и тем | Количество часов | | | | | | | Формы контроля знаний |
|---|---|---|---|---|---|---|---|---|
| | Всего | Лекции | Практические | Семинарские | Кругл. стол | Конференция | Сам. работа | |
| | **30** | **10** | **8** | | | | **12** | |
| **Математическое моделирование геологических объектов** | 2 | 2 | – | | | | | Устный опрос |
| Предварительный анализ данных | 4 | 2 | – | | | | | Решение практических задач |
| Числовые характеристики случайной величины | 4 | 2 | – | | | | | Решение практических задач |
| Проверка статистических гипотез | 4 | 2 | – | | | | 2 | Решение практических задач |
| Дисперсионный анализ | 4 | 2 | 2 | | | | 2 | Решение практических задач |
| Корреляционный анализ | 4 | – | 2 | | | | 2 | Решение практических задач |
| Регрессионный анализ | 4 | – | 2 | | | | 2 | Решение практических задач |
| Кластерный анализ | 4 | – | 2 | | | | | Решение практических задач |
| **Итого** | **30** | **10** | **8** | | | | **12** | |
| Формы текущей аттестации | **Зачет** | | | | | | | |

## Очная (вечерняя) форма получения образования

| Наименование разделов и тем | Количество часов | | | | | | | Формы контроля знаний |
|---|---|---|---|---|---|---|---|---|
| | Всего | Лекции | Практические | Семинарские | Кругл. стол | Конференция | Сам. работа | |
| | **30** | **10** | **8** | | | | **12** | |
| **Математическое моделирование геологических объектов** | 2 | 2 | – | | | | | Устный опрос |
| Предварительный анализ данных | 4 | 2 | 2 | | | | | Решение практических задач |
| Числовые характеристики случайной величины | 4 | 2 | 2 | | | | | Решение практических задач |
| Проверка статистических гипотез | 4 | 2 | 2 | | | | | Решение практических задач |
| Дисперсионный анализ | 4 | 2 | 2 | | | | | Решение практических задач |
| Корреляционный анализ | 4 | 2 | 2 | | | | | Решение практических задач |
| Регрессионный анализ | 4 | 2 | 2 | | | | | Решение практических задач |
| Кластерный анализ | 4 | – | 2 | | | | 2 | Решение практических задач |
| **Итого** | **30** | **14** | **14** | | | | **2** | |
| Формы текущей аттестации | Зачет | | | | | | | |

## МАТЕРИАЛЫ ДЛЯ ТЕКУЩЕЙ И ИТОГОВОЙ АТТЕСТАЦИИ СЛУШАТЕЛЕЙ

### Вопросы для проведения зачета

14 Предварительный анализ данных: ряд данных, несгруппированный ряд, ранжированный ряд, вариативный ряд, интервальный ряд, полигон частот, гистограмма.

15 Точечные оценки параметров генеральной совокупности. Доверительные интервалы.

16 Проверка статистических гипотез.

17 Гипотезы о параметрах многомерной нормально распределенной генеральной совокупности.

18 Сравнение двух генеральных совокупностей: параметрические критерии.

19 Сравнение двух генеральных совокупностей: непараметрические критерии.

20 Однофакторный дисперсионный анализ, критерий Фишера. Коэффициент детерминации.

21 Непараметрический однофакторный анализ. Критерий Краскала-Уоллиса.

22 Взаимосвязь факторов. Парные коэффициенты корреляции. Частные коэффициенты корреляции. Проверка значимости параметров связи.

23 Множественный коэффициент детерминации и его свойства.

24 Критерий независимости на основе рангового коэффициента корреляции Спирмена.

25 Модель множественной линейной регрессии. Значимость коэффициентов. Значимость коэффициента детерминации.

26 Иерархические кластер-процедуры.

## Информационно-методическая часть

### *Перечень практических занятий*

1. Предварительный анализ данных.
2. Числовые характеристики случайной величины.
3. Проверка статистических гипотез.
4. Дисперсионный анализ.
5. Корреляционный анализ.
6. Регрессионный анализ.
7. Кластерный анализ.

### *Формы контроля знаний*

1. Защита практических работ.
2. Устный опрос.

## РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

*Основная*

1 Айвазян, С. А. Прикладная статистика. Основы эконометрики: учебное пособие : в 2 т. / С. А. Айвазян, В. С. Мхитарян. – М.: ЮНИТИ, 2002.

2 Елисеева, И. И. Эконометрика: учебное пособие / И. И. Елисеева. – М.: Финансы и статистика, 2004.

3 Магнус, Я. Р. Эконометрика : учебное пособие / Я. Р. Магнус [и др.]. – М.: Дело, 2004.

4 Кобзарь, А. И. Прикладная математическая статистика : для инженеров и научных работников / А. И. Кобзарь. – Москва: ФИЗМАТЛИТ, 2006. – 816 с.

5 Марченко Л.Н. Эконометрика : модель множественной линейной регрессии : практическое руководство / Л.Н.Марченко, Ю.Е Дудовская, Ю.В.Синюгина. – Изд-во: ГГУ им. Ф. Скорины. – 2016. – 48 с.

6 Практикум по эконометрике (+CD) : учебное пособие / И.И. Елисеева и [и др.], под ред. И.И. Елисеевой. – М.: Финансы и статистика, 2008.

7 Терещенко, О. В. Многомерный статистический анализ данных в социальных науках : учеб.пособие для студентов учреждений высшего образования по социально-гуманитарным специальностям / О. В.Терещенко. – Минск : БГУ, 2012. – 355.

8 Эконометрика и экономико-математические методы и модели : учебное пособие. – Изд-во: БГЭУ. – 2018.

9 Дэвис, Дж. С. Статистический анализ данных в геологии : в 2-х кн. Кн.1: пер.с англ / Дж. С. Дэвис. – Москва : Недра, 1990.

10 Харин, Ю.С. Математическая и прикладная статистика : учебное издание : учеб. пособие для студентов вузов по математич. и экономич. специальностям / Ю. С. Харин. – Минск : БГУ, 2005. – 17. – 355с.

11 SPSS/PC+Guide for use with applied statistics for business and economics. : Методические указания как работать с книгой Уэбстэра Аллана Прикладная статистика в бизнесе и экономике. – USA, 1992. – 27. – 353 с.

*Дополнительная*

12 Замков, О. О. Эконометрические методы в макроэкономическом анализе: курс лекций / О. О. Замков. – М. : ГУ ВШЭ, 2001.

13 Малюгин, В. И. Система эконометрического моделирования и прогнозирования СЭМП 1.1: руководство пользователя / В. И. Малюгин [и др.]. – Мн. : БГУ, 2000.

## Источники информации, рекомендуемые слушателям, с указанием объемов для самостоятельного изучения

Указаны в разделе «Содержание программы»: «Требования к проверке результатов самостоятельной работы».