

МИНОБРНАУКИ РОССИИ
ВЛАДИВОСТОКСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
НАУЧНО-ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР "ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ"

Рабочая программа дисциплины (модуля)
БИОИНФОРМАТИКА

Направление и направленность (профиль)
09.04.03 Прикладная информатика. Интеллектуальный анализ данных

Год набора на ОПОП
2025

Форма обучения
очная

Владивосток 2025

Рабочая программа дисциплины (модуля) «Биоинформатика» составлена в соответствии с требованиями ФГОС ВО по направлению подготовки 09.04.03 Прикладная информатика (утв. приказом Минобрнауки России от 19.09.2017г. №916) и Порядком организации и осуществления образовательной деятельности по образовательным программам высшего образования – программам бакалавриата, программам специалитета, программам магистратуры (утв. приказом Минобрнауки России от 06.04.2021 г. N245).

Составитель(и):

Вишневский А.А., ассистент, Научно-образовательный центр "Искусственный интеллект", Andrey.Vishnevskiy2025@vvsu.ru

Кригер А.Б., кандидат физико-математических наук, доцент, Научно-образовательный центр "Искусственный интеллект", Aleksandra.Kriger@vvsu.ru

Утверждена на заседании научно-образовательный центр "искусственный интеллект" от 05.06.2025 , протокол № 6

СОГЛАСОВАНО:

Заведующий кафедрой (разработчика)

Кригер А.Б.

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ	
Сертификат	1582918206
Номер транзакции	0000000000DC615F
Владелец	Кригер А.Б.

1 Цель, планируемые результаты обучения по дисциплине (модулю)

Цели освоения учебной дисциплины:

Получение базовых знаний и формирование навыков, необходимых для обработки и анализа данных в биологии, медицине и других естественнонаучных дисциплин актуальными методами математической статистики: изучение данных, выявление закономерностей, определение зависимостей и построения моделей с помощью современных технологий.

Задачи освоения дисциплины:

- приобретение студентами знаний о видах статистических данных в биологии, медицине и других естественнонаучных дисциплин, методах их обработки и анализа;
- применение методов математической статистики для анализа экспериментальных данных и клинических наблюдений;
- приобретение практических навыков работы с программой RStudio.

Планируемыми результатами обучения по дисциплине (модулю), являются знания, умения, навыки. Перечень планируемых результатов обучения по дисциплине (модулю), соотнесенных с планируемыми результатами освоения образовательной программы, представлен в таблице 1.

Таблица 1 – Компетенции, формируемые в результате изучения дисциплины (модуля)

Название ОПОП ВО, сокращенное	Код и формулировка компетенции	Код и формулировка индикатора достижения компетенции	Результаты обучения по дисциплине		
			Код результата	Формулировка результата	
09.04.03 «Прикладная информатика» (М-ПИ)	ПКВ-1 : Способен использовать и развивать методы научных исследований и инструментария в области интеллектуального анализа данных и интеллектуальных систем	ПКВ-1.1к : Использует существующие методы и технологии научных исследований, современные методы в области интеллектуального анализа данных	РД1	Знание	о видах статистических данных в биологии, медицине и других естественнонаучных дисциплин, методах их обработки и анализа
			РД1	Умение	анализировать полученные теоретические знания для выбора наиболее соответствующих данным статистических методов
			РД1	Навык	применения основных методов математической статистики для обработки и анализа конкретных данных

В процессе освоения дисциплины решаются задачи воспитания гармонично развитой, патриотичной и социально ответственной личности на основе традиционных российских духовно-нравственных и культурно-исторических ценностей, представленные в таблице 1.2.

Таблица 1.2 – Целевые ориентиры воспитания

Воспитательные задачи	Формирование ценностей	Целевые ориентиры
Формирование научного мировоззрения и культуры мышления		

Формирование осознания ценности научного мировоззрения и критического мышления	Гуманизм	Системное мышление
Формирование коммуникативных навыков и культуры общения		
Формирование навыков публичного выступления и презентации своих идей	Взаимопомощь и взаимоуважение	Умение работать в команде и взаимопомощь

2 Место дисциплины (модуля) в структуре ОПОП

Дисциплина "Биоинформатика" относится к дисциплинам формируемым участниками образовательных отношений, Блоку 1 "Дисциплины" (модули)

3. Объем дисциплины (модуля)

Объем дисциплины (модуля) в зачетных единицах с указанием количества академических часов, выделенных на контактную работу с обучающимися (по видам учебных занятий) и на самостоятельную работу, приведен в таблице 2.

Таблица 2 – Общая трудоемкость дисциплины

Название ОПОП ВО	Форма обучения	Часть УП	Семестр (ОФО) или курс (ЗФО, ОЗФО)	Трудоемкость (З.Е.)	Объем контактной работы (час)						СРС	Форма аттестации			
					Всего	Аудиторная			Внеаудиторная						
						лек.	прак.	лаб.	ПА	КСР					
09.04.03 Прикладная информатика	ОФО	M01.B	4	3	33	8	24	0	1	0	75	Э			

4 Структура и содержание дисциплины (модуля)

4.1 Структура дисциплины (модуля) для ОФО

Тематический план, отражающий содержание дисциплины (перечень разделов и тем), структурированное по видам учебных занятий с указанием их объемов в соответствии с учебным планом, приведен в таблице 3.1

Таблица 3.1 – Разделы дисциплины (модуля), виды учебной деятельности и формы текущего контроля для ОФО

№	Название темы	Код результата обучения	Кол-во часов, отведенное на				Форма текущего контроля
			Лек	Практ	Лаб	СРС	
1	Виды статистических данных в биомедицине и других естественнонаучных дисциплин	РД1, РД1, РД2, РД2	1	4	0	15	Выполнение практических работ
2	Статистические методы исследования и анализа данных	РД1, РД1, РД2, РД2, РД2	3	8	0	30	Выполнение практических работ
3	Исследование зависимостей	РД1, РД2, РД2, РД2	4	12	0	30	Выполнение практических работ

Итого по таблице	8	24	0	75
------------------	---	----	---	----

4.2 Содержание разделов и тем дисциплины (модуля) для ОФО

Тема 1 Виды статистических данных в биомедицине и других естественнонаучных дисциплин.

Содержание темы: 1.1 Особенности статистических данных в биомедицине и других естественнонаучных дисциплин. Шкалы измерений: количественные (числовые непрерывные или дискретные) и качественные данные (числовые непрерывные или дискретные), абсолютные и относительные величины (экстенсивные, интенсивные, соотношения и наглядности). Типичные ошибки при использовании относительных величин. 1.2 Генеральная совокупность и выборка. Основные способы организации выборки. Общие рекомендации по численности выборки. Репрезентативность выборки. Зависимые и независимые выборки. .

Формы и методы проведения занятий по теме, применяемые образовательные технологии: Лекционные и практические занятия.

Виды самостоятельной подготовки студентов по теме: Чтение предлагаемой литературы.

Тема 2 Статистические методы исследования и анализа данных.

Содержание темы: 2.1 Первичный и вторичный анализы данных. Показатели описательной статистики: показатели положения, показатели разброса, показатели асимметрии, графические способы представления результатов. Свойства описательных статистик. 2.2 Обработка пропущенных значений: удаление, восстановление. 2.3 Виды распределения и их параметры. Индуктивная статистика. Критерии для проверки гипотезы о законе распределения: Шапиро-Уилка, Колмогорова-Смирнова и др. Уровень статистической достоверности. 2.4 Преобразование данных: нормировка данных (центрирование, нормировка по максимальному значению, нормировка по минимальному значению, нормировка по среднему значению), преобразование признаков к ранговому представлению (операция дискретизации), преобразование признаков бинарному представлению, логарифмическое преобразование, преобразование к нормальному распределению, преобразование к равномерному или произвольному распределению. 2.5 Выявление выбросов. Удаление выбросов из анализа. .

Формы и методы проведения занятий по теме, применяемые образовательные технологии: Лекционные и практические занятия.

Виды самостоятельной подготовки студентов по теме: Чтение предлагаемой литературы.

Тема 3 Исследование зависимостей.

Содержание темы: 3.1 Параметрические методы сравнения данных: критерий Колмогорова-Смирнова, критерий согласия хи-квадрат Пирсона, t-тест Стьюдента. 3.2 Непараметрические методы сравнения для зависимых и независимых выборок: U-тест Манна-Уитни, критерий Уилкоксона, критерий Краскела-Уоллиса, критерий Фридмана. 3.3 Методы сравнения номинальных данных: критерий ?2 (хи-квадрат) для проверки статистической гипотезы о наличии связи между двумя качественными признаками; тест Кохрана-Мантеля-Хензеля для проверки нулевой гипотезы о том, что две номинальные переменные условно независимы при каждом значении третьей переменной. 3.4 Коэффициент корреляции Пирсона, ранговые коэффициенты корреляции Спирмена и Кендалла. Частная корреляция. Автокорреляция. 3.5 Анализ выживаемости. Таблица времени жизни. Оценка шансов (OR, odds ratio). Относительный риск. Индекс риска. Лог-ранк тест. Оценка Каплана Мейера кривых выживаемости. Обобщенный (Геханом) критерий Вилкоксона, F-критерий Кокса, логарифмический ранговый критерий,

обобщенный Пето (Peto R.и Peto J.) критерий Вилкоксона. Модель Кокса. Пропорциональный риск. Частичное правдоподобие. .

Формы и методы проведения занятий по теме, применяемые образовательные технологии: Лекционные и практические занятия.

Виды самостоятельной подготовки студентов по теме: Чтение предлагаемой литературы.

5 Методические указания для обучающихся по изучению и реализации дисциплины (модуля)

5.1 Методические рекомендации обучающимся по изучению дисциплины и по обеспечению самостоятельной работы

Программой данной дисциплины предусмотрено чтение лекций, проведение практических занятий и самостоятельная работа студентов для формирования и развития профессиональных навыков обучающихся.

5.2 Особенности организации обучения для лиц с ограниченными возможностями здоровья и инвалидов

При необходимости обучающимся из числа лиц с ограниченными возможностями здоровья и инвалидов (по заявлению обучающегося) предоставляется учебная информация в доступных формах с учетом их индивидуальных психофизических особенностей:

- для лиц с нарушениями зрения: в печатной форме увеличенным шрифтом; в форме электронного документа; индивидуальные консультации с привлечением тифлосурдопереводчика; индивидуальные задания, консультации и др.

- для лиц с нарушениями слуха: в печатной форме; в форме электронного документа; индивидуальные консультации с привлечением сурдопереводчика; индивидуальные задания, консультации и др.

- для лиц с нарушениями опорно-двигательного аппарата: в печатной форме; в форме электронного документа; индивидуальные задания, консультации и др.

6 Фонд оценочных средств для проведения текущего контроля и промежуточной аттестации обучающихся по дисциплине (модулю)

В соответствии с требованиями ФГОС ВО для аттестации обучающихся на соответствие их персональных достижений планируемым результатам обучения по дисциплине (модулю) созданы фонды оценочных средств. Типовые контрольные задания, методические материалы, определяющие процедуры оценивания знаний, умений и навыков, а также критерии и показатели, необходимые для оценки знаний, умений, навыков и характеризующие этапы формирования компетенций в процессе освоения образовательной программы, представлены в Приложении 1.

7 Учебно-методическое и информационное обеспечение дисциплины (модуля)

7.1 Основная литература

1. Айвазян, С. А. Методы эконометрики : учебник / С. А. Айвазян. — Москва : Магистр : ИНФРА-М, 2022. — 512 с. - ISBN 978-5-9776-0153-5. - Текст : электронный. - URL: <https://znanium.ru/catalog/product/1840468> (Дата обращения -18.06.2025)
2. Барский, А. Б., Искусственный интеллект и интеллектуальные системы управления : монография / А. Б. Барский. — Москва : Русайнс, 2022. — 185 с. — ISBN 978-5-4365-8166-8. — URL: <https://book.ru/book/943706> (дата обращения: 18.06.2025). — Текст : электронный.
3. Миркин, Б. Г. Базовые методы анализа данных : учебник и практикум для вузов / Б. Г. Миркин. — 3-е изд., перераб. и доп. — Москва : Издательство Юрайт, 2024. — 297 с. — (Высшее образование). — ISBN 978-5-534-19709-9. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: <https://urait.ru/bcode/556941> (дата обращения: 12.03.2025).

7.2 Дополнительная литература

1. Криволапов, С. Я., Математика на Python : учебник / С. Я. Криволапов, М. Б. Хрипунова. — Москва : КноРус, 2024. — 455 с. — ISBN 978-5-406-12069-9. — URL: <https://book.ru/book/950432> (дата обращения: 18.06.2025). — Текст : электронный.
2. Чернышев С. А. ОСНОВЫ ПРОГРАММИРОВАНИЯ НА PYTHON. Учебное пособие для вузов [Электронный ресурс] : Москва : Издательство Юрайт , 2022 - 286 - Режим доступа: <https://urait.ru/book/osnovy-programmirovaniya-na-python-496893>

7.3 Ресурсы информационно-телекоммуникационной сети "Интернет", включая профессиональные базы данных и информационно-справочные системы (при необходимости):

1. Образовательная платформа "ЮРАЙТ"
2. Электронно-библиотечная система "BOOK.ru"
3. Электронно-библиотечная система "ZNANIUM.COM"
4. Электронно-библиотечная система ЮРАЙТ - Режим доступа: <https://urait.ru/>
5. Open Academic Journals Index (OAJI). Профессиональная база данных - Режим доступа: <http://oaji.net/>
6. Президентская библиотека им. Б.Н.Ельцина (база данных различных профессиональных областей) - Режим доступа: <https://www.prlib.ru/>
7. Информационно-справочная система "Консультант Плюс" - Режим доступа: <http://www.consultant.ru/>

8 Материально-техническое обеспечение дисциплины (модуля) и перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень программного обеспечения

Основное оборудование:

- Монитор облачный 23" LG23CAV42K/мышь Genius Optical Wheel проводная/клавиатура Genius KB110 проводная
- Мультимедийный проектор CASIO (Япония)
- Облачный монитор LG Electronics черный +клавиатура+мышь
- Уст-во бесп.пит.SmartUPS 3000

Программное обеспечение:

- Microsoft Office Professional Plus 2013 Russian
- Python

МИНОБРНАУКИ РОССИИ
ВЛАДИВОСТОКСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
НАУЧНО-ОБРАЗОВАТЕЛЬНЫЙ ЦЕНТР "ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ"

Фонд оценочных средств
для проведения текущего контроля
и промежуточной аттестации по дисциплине (модулю)

БИОИНФОРМАТИКА

Направление и направленность (профиль)
09.04.03 Прикладная информатика. Интеллектуальный анализ данных

Год набора на ОПОП
2025

Форма обучения
очная

Владивосток 2025

1 Перечень формируемых компетенций

Название ОПОП ВО, сокращенное	Код и формулировка компетенции	Код и формулировка индикатора достижения компетенции
09.04.03 «Прикладная информатика» (М-ПИ)	ПКВ-1 : Способен использовать и развивать методы научных исследований и инструментария в области интеллектуального анализа данных и интеллектуальных систем	ПКВ-1.1к : Использует существующие методы и технологии научных исследований, современные методы в области интеллектуального анализа данных

Компетенция считается сформированной на данном этапе в случае, если полученные результаты обучения по дисциплине оценены положительно (диапазон критериев оценивания результатов обучения «зачтено», «удовлетворительно», «хорошо», «отлично»). В случае отсутствия положительной оценки компетенция на данном этапе считается несформированной.

2 Показатели оценивания планируемых результатов обучения

Компетенция ПКВ-1 «Способен использовать и развивать методы научных исследований и инструментария в области интеллектуального анализа данных и интеллектуальных систем»

Таблица 2.1 – Критерии оценки индикаторов достижения компетенции

Код и формулировка индикатора достижения компетенции	Результаты обучения по дисциплине			Критерии оценивания результатов обучения
	Код ре- з- та	Тип ре- з- та	Результат	
ПКВ-1.1к : Использует существующие методы и технологии научных исследований, современные методы в области интеллектуального анализа данных	РД 1	Знание	о видах статистических данных в биологии, медицине и других естественнонаучных дисциплинах, методах их обработки и анализа	сформировавшиеся знания о видах статистических данных в биологии, медицине и других естественнонаучных дисциплинах, методах их обработки и анализа
	РД 1	умение	анализировать полученные теоретические знания для выбора наиболее соответствующих данным статистических методов	сформировавшиеся умения анализировать полученные теоретические знания для выбора наиболее соответствующих данным статистических методов
	РД 1	навык	применения основных методов математической статистики для обработки и анализа конкретных данных	сформировавшиеся навыки применения основных методов математической статистики для обработки и анализа конкретных данных

Таблица заполняется в соответствии с разделом 1 Рабочей программы дисциплины (модуля).

3 Перечень оценочных средств

Таблица 3 – Перечень оценочных средств по дисциплине (модулю)

Контролируемые планируемые результаты обучения	Контролируемые темы дисциплины	Наименование оценочного средства и представление его в ФОС	
		Текущий контроль	Промежуточная аттестация
Очная форма обучения			
РД1	Знание : о видах статистических данных в биологии, медицине и других естественнонаучных дисциплин, методах их обработки и анализа	1.1. Виды статистических данных в биомедицине и других естественнонаучных дисциплин	Практическая работа Список вопросов
РД1	Умение : анализировать полученные теоретические знания для выбора наиболее соответствующих данным статистических методов	1.1. Виды статистических данных в биомедицине и других естественнонаучных дисциплин	Практическая работа Список вопросов
		1.2. Статистические методы исследования и анализа данных	Практическая работа Список вопросов
РД1	Навык : применения основных методов математической статистики для обработки и анализа конкретных данных	1.2. Статистические методы исследования и анализа данных	Практическая работа Список вопросов
		1.3. Исследование зависимостей	Практическая работа Список вопросов
РД2	Знание : современных научных принципов и методов исследований в биомедицине	1.1. Виды статистических данных в биомедицине и других естественнонаучных дисциплин	Практическая работа Список вопросов
		1.2. Статистические методы исследования и анализа данных	Практическая работа Список вопросов
		1.3. Исследование зависимостей	Практическая работа Список вопросов
РД2	Умение : применять на практике теоретические знания	1.1. Виды статистических данных в биомедицине и других естественнонаучных дисциплин	Практическая работа Список вопросов
		1.2. Статистические методы исследования и анализа данных	Практическая работа Список вопросов
		1.3. Исследование зависимостей	Практическая работа Список вопросов
РД2	Навык : практического и спользования современных научных принципов и методов исследования в программе RStudio	1.2. Статистические методы исследования и анализа данных	Практическая работа Список вопросов
		1.3. Исследование зависимостей	Практическая работа Список вопросов

4 Описание процедуры оценивания

Качество сформированности компетенций на данном этапе оценивается по результатам текущих и промежуточных аттестаций при помощи количественной оценки,

выраженной в баллах. Максимальная сумма баллов по дисциплине (модулю) равна 100 баллам.

Распределение баллов по видам учебной деятельности

Вид учебной деятельности	Оценочное средство		
	Вопросы к экзамену	Практические работы	Итого
Практические занятия		60	60
Самостоятельная работа	10	10	20
Промежуточная аттестация	20		20
Итого	30	70	100

Сумма баллов, набранных студентом по всем видам учебной деятельности в рамках дисциплины, переводится в оценку в соответствии с таблицей.

Сумма баллов по дисциплине	Оценка по промежуточной аттестации	Характеристика качества сформированности компетенции
от 91 до 100	«зачтено» / «отлично»	Студент демонстрирует сформированность дисциплинарных компетенций, обнаживает всестороннее, систематическое и глубокое знание учебного материала, усвоил основную литературу и знаком с дополнительной литературой, рекомендованной программой, умеет свободно выполнять практические задания, предусмотренные программой, свободно оперирует приобретенными знаниями, умеет применять их в ситуациях повышенной сложности.
от 76 до 90	«зачтено» / «хорошо»	Студент демонстрирует сформированность дисциплинарных компетенций: основные знания, умения освоены, но допускаются незначительные ошибки, неточности, затруднения при аналитических операциях, переносе знаний и умений на новые, нестандартные ситуации.
от 61 до 75	«зачтено» / «удовлетворительно»	Студент демонстрирует сформированность дисциплинарных компетенций: в ходе контрольных мероприятий допускаются значительные ошибки, проявляется отсутствие отдельных знаний, умений, навыков по некоторым дисциплинарным компетенциям, студент испытывает значительные затруднения при оперировании знаниями и умениями при их переносе на новые ситуации.
от 41 до 60	«не зачтено» / «неудовлетворительно»	У студента не сформированы дисциплинарные компетенции, проявляется недостаточность знаний, умений, навыков.
от 0 до 40	«не зачтено» / «неудовлетворительно»	Дисциплинарные компетенции не сформированы. Проявляется полное или практически полное отсутствие знаний, умений, навыков.

5 Примерные оценочные средства

5.1 Экзаменационные вопросы

Вопросы к экзамену

1. Виды медико-биологических данных, их особенности.
2. Описательные статистики и их свойства.
3. Основные способы организации выборки.
4. Методы обработки пропущенных значений.
5. Статистическая достоверность.
6. Критерии для проверки на нормальность распределения, условия их применения.
7. Способы преобразования данных.
8. Критерии определения выбросов.
9. Параметрическая статистика. Тесты для зависимых и независимых выборок.

10. Непараметрическая статистика. Тесты для непрерывных и дискретных выборок.

11. Преимущества и недостатки непараметрических методов.
12. Таблицы сопряженности признаков.
13. Корреляционный анализ. Коэффициенты корреляции, их значимость.
14. Автокорреляция.
15. Регрессионный анализ.
16. Анализ выживаемости.
17. Модель пропорциональных интенсивностей Кокса.
18. Многомерный анализ данных.
19. Взаимодействие между переменными (эффект модификации).

Краткие методические указания

Для самостоятельной подготовки к экзамену рекомендуется использовать предложенную в РПД литературу и информационные ресурсы сети Интернет.

Шкала оценки

Критерии оценки

№	Баллы	Описание
5	16-20	Студент демонстрирует глубокие и всесторонние знания учебного материала по выбранной теме, свободно оперирует приобретенными знаниями при ответе на дополнительные вопросы.
4	11-15	Студент демонстрирует знания материала по выбранной теме на среднем уровне: основные знания освоены, но допускаются незначительные ошибки, неточности, затруднения при ответе на некоторые вопросы.
3	6-10	Студент демонстрирует освоение выбранной темы на базовом уровне: допускаются значительные ошибки, проявляется отсутствие отдельных знаний, студент испытывает значительные затруднения при ответе на дополнительные вопросы.
2	3-5	Студент демонстрирует освоение выбранной темы на уровне ниже базового, проявляется недостаточность знаний, навыков.
1	0-2	Студент не смог подготовить доклад, тем самым проявил полное или практически полное отсутствие знаний по выбранной теме.

1. Виды медико-биологических данных, их особенности.

Медико-биологические данные включают клинические, генетические, биохимические, физиологические и эпидемиологические данные. Клинические данные содержат информацию о пациентах: диагнозы, симптомы, результаты осмотров. Генетические данные представляют последовательности ДНК, РНК, SNP и другие молекулярные маркеры. Биохимические данные отражают концентрации метаболитов, белков, гормонов. Физиологические данные включают показатели работы органов (ЭКГ, ЭЭГ, давление). Эпидемиологические данные охватывают заболеваемость, распространенность патологий в популяции. Особенности таких данных: неоднородность, большой объем, наличие шумов и пропусков, необходимость стандартизации и интерпретации в контексте биологических процессов.

2. Описательные статистики и их свойства.

Описательные статистики используются для первичного анализа данных и включают меры центральной тенденции (среднее, медиана, мода), меры изменчивости (дисперсия, стандартное отклонение, размах) и меры формы распределения (асимметрия, эксцесс). Среднее чувствительно к выбросам, медиана устойчива к ним. Дисперсия показывает разброс данных вокруг среднего, а стандартное отклонение — его масштаб. Асимметрия указывает на скошенность распределения, эксцесс — на остроту пика. Эти статистики помогают понять структуру данных перед углубленным анализом.

3. Основные способы организации выборки.

Выборка может быть организована как простая случайная, стратифицированная, кластерная или систематическая. Простая случайная выборка предполагает равную вероятность отбора каждого элемента. Стратифицированная делит популяцию на однородные группы (страты), из которых делается выбор. Кластерная выборка разбивает данные на кластеры, затем случайно отбирает некоторые из них. Систематическая выборка использует фиксированный интервал (например, каждый 10-й элемент). Выбор метода зависит от структуры данных и целей исследования.

4. Методы обработки пропущенных значений.

Пропущенные значения можно удалить (если их мало), заменить

средним/медианой (для числовых данных) или модой (для категориальных). Более сложные методы включают множественную импутацию (предсказание пропусков на основе других переменных) и использование алгоритмов машинного обучения (KNN, регрессия). Если пропуски неслучайны (MNAR), требуется специальный анализ. Важно учитывать влияние обработки пропусков на итоговые результаты.

5. Статистическая достоверность.

Статистическая достоверность отражает надежность выводов, обычно выражается через p -уровень (вероятность ошибки при отклонении нулевой гипотезы). Если $p < 0.05$, результат считается статистически значимым. Доверительные интервалы показывают диапазон, в котором с заданной вероятностью находится истинный параметр. Мощность теста — вероятность обнаружить эффект, если он есть. Достоверность зависит от объема выборки, размера эффекта и вариабельности данных.

6. Критерии для проверки на нормальность распределения, условия их применения.

Для проверки нормальности используют графические методы (Q-Q plot, гистограммы) и статистические тесты: Шапиро-Уилка (для малых выборок), Колмогорова-Смирнова (для больших выборок), Андерсона-Дарлинга (чувствителен к хвостам распределения). Тесты применяются перед использованием параметрических методов, требующих нормальности данных. Если распределение ненормальное, применяют непараметрические тесты или преобразования данных.

7. Способы преобразования данных.

Для нормализации данных используют логарифмирование (для правосторонней асимметрии), квадратный корень (для умеренной асимметрии), преобразование Бокса-Кокса (оптимизирует нормальность). Для категориальных данных применяют бинаризацию или one-hot encoding. Стандартизация (z-оценка) и масштабирование (Min-Max) используют для приведения данных к единому диапазону перед анализом.

8. Критерии определения выбросов.

Выбросы определяют с помощью межквартильного размаха (IQR: значения за пределами $Q1 - 1.5 \cdot IQR$ или $Q3 + 1.5 \cdot IQR$), z-оценок (значения > 3 или < -3),

методов машинного обучения (Isolation Forest, DBSCAN). Визуально их можно обнаружить на boxplot или scatter plot. Выбросы могут быть артефактами или важными наблюдениями, поэтому их анализ требует осторожности.

9. Параметрическая статистика. Тесты для зависимых и независимых выборок.

Параметрические тесты предполагают нормальность распределения и однородность дисперсий. Для независимых выборок: t-тест Стьюдента (2 группы), ANOVA (>2 группы). Для зависимых выборок: парный t-тест, повторные измерения ANOVA. Эти тесты более мощные, чем непараметрические, но требуют строгих допущений.

10. Непараметрическая статистика. Тесты для непрерывных и дискретных выборок.

Непараметрические тесты не требуют нормальности. Для непрерывных данных: Манна-Уитни (независимые выборки), Вилкоксона (зависимые выборки), Краскела-Уоллиса (>2 группы). Для дискретных данных: хи-квадрат, точный тест Фишера. Эти тесты устойчивы к выбросам, но менее мощные.

11. Преимущества и недостатки непараметрических методов.

Преимущества: не требуют нормальности, устойчивы к выбросам, применимы к порядковым данным. Недостатки: менее мощные при выполнении условий для параметрических тестов, могут терять информацию при ранжировании данных. Используются при малых выборках или нарушении допущений параметрических тестов.

12. Таблицы сопряженности признаков.

Таблицы сопряженности показывают совместное распределение категориальных переменных. Анализируются с помощью теста хи-квадрат (для больших выборок) или точного теста Фишера (для малых). Используются для изучения взаимосвязей между признаками, например, генетическими маркерами и заболеваниями.

13. Корреляционный анализ. Коэффициенты корреляции, их значимость.

Корреляция измеряет силу и направление связи между переменными. Коэффициент Пирсона (для линейной связи), Спирмена (для монотонной), Кендалла (для ранговых данных). Значимость проверяется через р-уровень.

Корреляция не означает причинность, возможны ложные связи из-за confounding-факторов.

14. Автокорреляция.

Автокорреляция — связь между значениями одного временного ряда с лагом. Измеряется коэффициентом автокорреляции, используется в анализе динамических процессов (ЭКГ, экспрессия генов). Может указывать на периодичность или тренды. Учитывается в моделях ARIMA.

15. Регрессионный анализ.

Регрессия моделирует зависимость между предикторами и откликом. Линейная регрессия — для непрерывного отклика, логистическая — для бинарного. Множественная регрессия включает несколько предикторов. Важные аспекты: проверка мультиколлинеарности, остатков, интерпретация коэффициентов.

16. Анализ выживаемости.

Анализирует время до наступления события (смерть, рецидив). Методы: Каплана-Майера (кривые выживаемости), лог-ранк тест (сравнение групп). Учитывает цензурированные данные (наблюдения, где событие не произошло). Применяется в клинических исследованиях.

17. Модель пропорциональных интенсивностей Кокса.

Регрессионная модель для анализа выживаемости с учетом ковариат. Предполагает пропорциональность рисков. Коэффициенты показывают влияние предикторов на hazard ratio. Используется для многомерного анализа факторов риска.

18. Многомерный анализ данных.

Включает методы снижения размерности (PCA, t-SNE), кластеризации (k-means, иерархическая), классификации (SVM, случайный лес). Позволяет выявлять паттерны в высокоразмерных данных (например, транскриптомных).

19. Взаимодействие между переменными (эффект модификации).

Взаимодействие означает, что эффект одной переменной зависит от значения другой. Анализируется через включение произведения предикторов в регрессию. Пример: влияние гена на болезнь может зависеть от пола. Важно для персонализированной медицины.