

**Федеральное государственное автономное образовательное
учреждение высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»**

УТВЕРЖДЕНО

**Директор физтех-школы
прикладной математики и
информатики
А.М. Райгородский**

	Рабочая программа дисциплины (модуля)
по дисциплине:	Прикладная статистика и анализ данных
по направлению:	Прикладная математика и информатика
профиль подготовки:	А1360: Передовые методы искусственного интеллекта Физтех-школа Прикладной Математики и Информатики кафедра дискретной математики
курс:	3
квалификация:	бакалавр

Семестр, формы промежуточной аттестации: 5 (осенний) - Экзамен

Аудиторных часов: 60 всего, в том числе:

лекции: 30 час.

семинары: 30 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 45 час.

Подготовка к экзамену: 30 час.

Всего часов: 135, всего зач. ед.: 3

Программу составил: Н.А. Волков, ассистент

Программа обсуждена на заседании кафедры дискретной математики 12.02.2024

Аннотация

Курс посвящен современным статистическим подходам к анализу данных, которые позволяют получать интерпретируемые результаты. В курсе рассматриваются теоретические основы используемых методов и критериев, изучаются способы их применения в языках Python и R, а также обсуждается применение теории к реальным задачам, в частности, АВ-тестированию. В результате прохождения курса слушатель сможет проводить полноценную аналитику реальных данных.

1. Цели и задачи

Цель дисциплины

- изучение математических и теоретических основ современного статистического анализа, а также подготовка слушателей к дальнейшей самостоятельной работе в области анализа статистических задач прикладной математики, физики и экономики.

Задачи дисциплины

- изучение математических основ математической статистики;
- приобретение слушателями теоретических знаний в области современного статистического анализа.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
УК-1 Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач	УК-1.1 Анализирует задачу, выделяя этапы ее решения, действия по решению задачи
	УК-1.2 Находит, критически анализирует и выбирает информацию, необходимую для решения поставленной задачи
	УК-1.3 Рассматривает различные варианты решения задачи, оценивает их преимущества и недостатки
	УК-1.4 Грамотно, логично, аргументированно формирует собственные суждения и оценки
	УК-1.5 Определяет и оценивает практические последствия возможных вариантов решения задачи
ОПК-1 Способен применять фундаментальные знания, полученные в области физико-математических и (или) естественных наук и использовать их в профессиональной деятельности	ОПК-1.2 Способен строить математические модели, производить количественные расчеты и оценки
	ОПК-1.1 Способен анализировать поставленную задачу, намечать пути ее решения
	ОПК-1.3 Способен определять границы применимости полученных результатов
ОПК-2 Способен использовать современные информационные технологии и программные средства при решении задач профессиональной деятельности, соблюдая требования информационной безопасности	ОПК-2.1 Способен применять современные вычислительную технику и сервисы сети Интернет в области (сфере) профессиональной деятельности
	ОПК-2.2 Знает и умеет применять численные математические методы и прикладное программное обеспечение для решения научных задач в профессиональной области
	ОПК-2.3 Знает основные требования информационной безопасности
ОПК-3 Способен составлять и оформлять научные и (или) технические (технологические, инновационные) отчеты	ОПК-3.1 Знает основные правила оформления научных публикаций и научно-технической документации, в том числе с использованием прикладного программного обеспечения
	ОПК-3.2 Владеет на практике методологией составления научно-технических отчетов (проектов)

(публикации, проекты)	ОПК-3.3 Владеет методами визуального и графического представления результатов научной (научно-технической, инновационной технологической) деятельности в виде отчетов, научных публикаций
ПК-1 Способен ставить, формализовывать и решать задачи, в том числе разрабатывать и исследовать математические модели изучаемых явлений и процессов, системно анализировать научные проблемы, получать новые научные результаты	ПК-1.2 Способен выдвигать гипотезы, строить математические модели для описания изучаемых явлений и процессов, оценивать качество разработанной модели
	ПК-1.1 Способен находить, анализировать и обобщать информацию об актуальных результатах исследований в рамках тематической области своей профессиональной деятельности
	ПК-1.3 Способен применять теоретические и (или) экспериментальные методы исследований к конкретной научной задаче и интерпретировать полученные результаты
ПК-2 Способен самостоятельно или в качестве члена (руководителя) малого коллектива организовывать и проводить научные исследования и их апробацию	ПК-2.1 Знает принципы построения научной работы, методы сбора и анализа полученного материала, способы аргументации
	ПК-2.2 Способен планировать и проводить научные исследования самостоятельно или в качестве члена (руководителя) малого научного коллектива
	ПК-2.3 Способен проводить апробацию результатов научно-исследовательской работы посредством публикации научных статей и участия в конференциях

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- основные понятия математической статистики;
- основные подходы к сравнению оценок параметров неизвестного распределения;
- асимптотические и неасимптотические свойства оценок параметров неизвестного распределения;
- основные методы построения оценок с хорошими асимптотическими свойствами: метод моментов, метод максимального правдоподобия, метод выборочных квантилей;
- понятие эффективных оценок и неравенство информации Рао-Крамера;
- определение и главные свойства условного математического ожидания случайной величины относительно сигма-алгебры или другой случайной величины;
- определение общей линейной регрессионной модели и метод наименьших квадратов;
- многомерное нормальное распределение и его основные свойства;
- базовые понятия теории проверки статистических гипотез;
- лемму Неймана – Пирсона и теорему о монотонном отношении правдоподобия;
- критерий хи-квадрат Пирсона для проверки простых гипотез в схеме Бернулли.

уметь:

- обосновывать асимптотические свойства оценок с помощью применения предельных теорем теории вероятностей;
- строить оценки с хорошими асимптотическими свойствами для параметров неизвестного распределения по заданной выборке из него;
- находить байесовские оценки по заданному априорному распределению;
- вычислять условные математические ожидания с помощью условных распределений;
- находить оптимальные оценки с помощью полных достаточных статистик;
- строить точные и асимптотические доверительные интервалы, и области для параметров неизвестного распределения;
- находить оптимальные оценки и доверительные области в гауссовской линейной модели;
- строить равномерно наиболее мощные критерии в случае параметрического семейства с монотонным отношением правдоподобия;
- строить F-критерий для проверки линейных гипотез в линейной гауссовской модели.

владеть:

- основными методами математической статистики построения точечных и доверительных оценок: методом моментов, выборочных квантилей, максимального правдоподобия, методом наименьших квадратов, методом центральной статистики.
- навыками асимптотического анализа статистических критериев;
- навыками применения теорем математической статистики в прикладных задачах физики и экономики.

4. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

4.1. Разделы дисциплины (модуля) и трудоемкости по видам учебных занятий

№	Тема (раздел) дисциплины	Трудоемкость по видам учебных занятий, включая самостоятельную работу, час.			
		Лекции	Семинары	Лаборат. работы	Самост. работа
1	Линейная регрессия, свойства метода наименьших квадратов	2	2		3
2	Анализ остатков	2	2		3
3	Обобщенная линейная модель, статистические свойства оценки коэффициентов, построение доверительных интервалов	2	2		3
4	Пропуски в данных - типы пропусков, методы работы	2	2		3
5	Причины избыточности информации в данных, типы методов снижения размерности	2	2		3
6	Теорема об SVD-разложении. Док-во существования SVD-разложения	2	2		3
7	Коэффициенты корреляции Пирсона, Спирмена и Кендалла, их свойства	2	2		3
8	Виды задач дисперсионного анализа, примеры	2	2		3
9	Виды альтернатив в непараметрическом случае	2	2		3
10	Комбинирование критериев для построения более мощных процедур на примере одновременной проверки на нормальность и однородность двух выборок с условием на контроль FWER	2	2		3
11	Однофакторный дисперсионный анализ для случая независимых выборок	2	2		3
12	Практическая аналитика.	2	2		3
13	Ориентированные ациклические графы, терминология	2	2		4
14	Терминология в ориентированных ациклических графах	4	4		5
Итого часов		30	30		45
Подготовка к экзамену		30 час.			

Общая трудоёмкость	135 час., 3 зач.ед.
--------------------	---------------------

4.2. Содержание дисциплины (модуля), структурированное по темам (разделам)

Семестр: 5 (Осенний)

1. Линейная регрессия, свойства метода наименьших квадратов

Коэффициент детерминации (R^2), информационные критерии (AIC, BIC), метрики (MSE, MAE, MAPE). Гауссовская линейная модель - доверительные интервалы для коэффициентов модели и для отклика, гипотезы о незначимости коэффициента и группы коэффициентов, общая линейная гипотеза, сравнение моделей.

2. Анализ остатков

Дисперсия остатков линейной модели в условиях гетероскедастичности, визуальный анализ. Критерии проверки на гомоскедастичность: Бройша-Пагана и Голдфелда-Квандта. Преобразование Бокса-Кокса. Устойчивые оценки дисперсии Уайта, асимптотическая нормальность.

3. Обобщенная линейная модель, статистические свойства оценки коэффициентов, построение доверительных интервалов

Частные случаи - пуассоновская регрессии. (лог. регрессия в МЛ). Байесовский классификатор. Построение оптимального байесовского классификатора с док-вом теоремы об оптимальности. Квадратичный и линейный дискриминантный анализ. Оценки параметров, вид разделяющей поверхности. Наивный байесовский классификатор.

4. Пропуски в данных - типы пропусков, методы работы

Робастная регрессия. Методы на основе ближайшего соседа - kNN, взвешенный kNN, их свойства. Непараметрическая регрессия, локальное усреднение, оценка Надарая-Ватсона. Условия сходимости оценки Надарая-Ватсона, выбор ширины ядра, доверительная лента. Локальная линейная регрессионная модель. Регрессионное дерево, метод построения, свойства. Случайный лес и его свойства.

5. Причины избыточности информации в данных, типы методов снижения размерности

Метод главных компонент (PCA) как выбор направлений с максимальной дисперсией, формулы перехода в сжатое пространство и обратно. Дисперсии образа, выбор размерности сжатого пространства на основе доли необъясненной дисперсии

6. Теорема об SVD-разложении. Док-во существования SVD-разложения

Методы SNE и t-SNE: первоначальный вариант SNE, симметричный SNE, проблема скученности, метод t-SNE как решение проблемы. Метод UMAP. Постановка задачи: графы, функционал качества (KL). Общие слова о том, какая “метрика” используется, и почему в этом случае нет проблемы проклятия размерности

7. Коэффициенты корреляции Пирсона, Спирмена и Кендалла, их свойства

Таблицы сопряженности 2x2, точный тест Фишера, меры взаимосвязи, определение количества наблюдений. Общий случай таблиц сопряженности, типы вероятностных моделей, критерий хи-квадрат. Влияние признаков на целевую переменную: корреляции, подход с помощью решающих деревьев – важность признаков на основе Mean Decrease Impurity, Permutation feature importance, Drop Column feature importance.

8. Виды задач дисперсионного анализа, примеры

Критерии проверки однородности для бернуллиевских выборок, доверительные интервалы для разности (простые и Уилсона). Проверка на равенство средних нормальных выборок (t-test, 3 сл.), проверка равенства дисперсий, проверка однородности нормальных выборок. АВ-тестирование. Принципы разбиения, особенности. АА-тесты. Разбиение на тестовые группы, сроки теста, проверка нескольких гипотез. Пример, в котором события, соответствующие одному пользователю, зависимы. Бакетное семплирование как способ решения проблемы

9. Виды альтернатив в непараметрическом случае

Критерии Смирнова и Розенблатта. Критерий Уилкоксона-Манна-Уитни, его свойства, связанная с ним оценка параметра сдвига. Связные выборки, предположения модели, пример, когда предположения не выполняются. Критерий знаков, его свойства, связанная с ним оценка параметра сдвига. Критерий ранговых сумм Уилкоксона, его свойства, связанная с ним оценка параметра сдвига. Проверка симметрии

10. Комбинирование критериев для построения более мощных процедур на примере одновременной проверки на нормальность и однородность двух выборок с условием на контроль FWER

Сравнение интенсивностей двух экспоненциальных выборок. Сравнение интенсивностей пуассоновских процессов. Перестановочные критерии - идея, примеры для гипотез о среднем, а также для гипотез о равенстве средних двух выборок. Множественная проверка гипотез с помощью перестановок: версия max-T, обобщенный вариант

11. Однофакторный дисперсионный анализ для случая независимых выборок

F-критерий и критерий Бартлетта, их применимость. Критерий Краскела-Уоллиса и Джонкхиера. Post-hoc анализ: LSD Фишера, HSD Тьюки, критерии Неменья и Данна, оценка контраста. Однофакторный дисперсионный анализ для случая связанных выборок. F-критерий, критерии Фридмана и Пейджа. Post-hoc анализ. Двухфакторный дисперсионный анализ, случай дополнительной контрольной группы.

12. Практическая аналитика.

Какие особенности в данных могут присутствовать? Воронка. Парадокс Симпсона, примеры и выводы. Контрафактивная модель, причинно-следственный эффект, статистическая связь, утверждение о том, что связь не есть причинность. Равенство величины причинно-следственного эффекта и статистической связи при случайном назначении воздействия. Контрафактивная модель на примере парадокса Симпсона

13. Ориентированные ациклические графы, терминология

Марковское распределение на графе, примеры. Условная независимость и ее свойства. Оценка распределений в графе методом максимального правдоподобия. Интервенция, средний условный эффект как способ оценки причинно-следственного эффекта по графу. Примеры. Связь оценки причинно-следственного эффекта методом интервенции с контрафактивной моделью.

14. Терминология в ориентированных ациклических графах

Марковское свойство, примеры. Свойства d-разделимости и d-связности, теорема об условной независимости на множестве вершин. Построение причинно-следственных графов по данным: метод индуктивной причинности. Оценка условной независимости: частная корреляция, причинность по Грейнджеру.

5. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)

Стандартная учебная аудитория.

6.Перечень рекомендуемой литературы

Основная литература

000000593, Введение в математическую статистику [Текст] : [учебник для вузов] / Г. И. Ивченко, Ю. И. Медведев .— М. : ЛКИ, 2010, 2014, 2015 .— 600 с.

000002458, Наглядная математическая статистика [Текст] : учеб. пособие для вузов / М. Б. Лагутин .— 2-е изд., испр. — М. : Бином. Лаб. знаний, 2009 .— 472 с.

Дополнительная литература

Прикладная математическая статистика [Текст] : для инженеров и научных работников / А. И. Кобзарь .— 2-е изд., испр. / [Научное изд.] .— М. : Физматлит, 2012 .— 816 с. — (Современные методы в математике). - Библиогр.: с. 737-759. - Предм. указ.: с. 806-810. - Имен. указ.: с. 811-813. - 500 экз. - ISBN 978-5-9221-1375-5 (в пер.) .— Полный текст (Доступ из сети МФТИ / Удаленный доступ).

7. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля)

Не используются

8. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень необходимого программного обеспечения и информационных справочных систем (при необходимости)

На занятиях используются мультимедийные технологии, включая демонстрацию презентаций.

9. Методические указания для обучающихся по освоению дисциплины (модуля)

Студент, изучающий дисциплину, должен с одной стороны, овладеть общим понятийным аппаратом, а с другой стороны, должен научиться применять теоретические знания на практике.

Успешное освоение дисциплины требует:

- посещения студентом всех видов аудиторных занятий;
- качественной самостоятельной подготовки к практическим занятиям, активной работы на них;
- активной самостоятельной и аудиторной работы студента;
- своевременной сдачи преподавателю заданий по аудиторным видам работ.

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению: Прикладная математика и информатика
профиль подготовки: АІ360: Передовые методы искусственного интеллекта
Физтех-школа Прикладной Математики и Информатики
кафедра дискретной математики
курс: 3
квалификация: бакалавр

Семестр, формы промежуточной аттестации: 5 (осенний) - Экзамен

Разработчик: Н.А. Волков, ассистент

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
УК-1 Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач	УК-1.1 Анализирует задачу, выделяя этапы ее решения, действия по решению задачи
	УК-1.2 Находит, критически анализирует и выбирает информацию, необходимую для решения поставленной задачи
	УК-1.3 Рассматривает различные варианты решения задачи, оценивает их преимущества и недостатки
	УК-1.4 Грамотно, логично, аргументированно формирует собственные суждения и оценки
	УК-1.5 Определяет и оценивает практические последствия возможных вариантов решения задачи
ОПК-1 Способен применять фундаментальные знания, полученные в области физико-математических и (или) естественных наук и использовать их в профессиональной деятельности	ОПК-1.2 Способен строить математические модели, производить количественные расчеты и оценки
	ОПК-1.1 Способен анализировать поставленную задачу, намечать пути ее решения
	ОПК-1.3 Способен определять границы применимости полученных результатов
ОПК-2 Способен использовать современные информационные технологии и программные средства при решении задач профессиональной деятельности, соблюдая требования информационной безопасности	ОПК-2.1 Способен применять современные вычислительную технику и сервисы сети Интернет в области (сфере) профессиональной деятельности
	ОПК-2.2 Знает и умеет применять численные математические методы и прикладное программное обеспечение для решения научных задач в профессиональной области
	ОПК-2.3 Знает основные требования информационной безопасности
ОПК-3 Способен составлять и оформлять научные и (или) технические (технологические, инновационные) отчеты (публикации, проекты)	ОПК-3.1 Знает основные правила оформления научных публикаций и научно-технической документации, в том числе с использованием прикладного программного обеспечения
	ОПК-3.2 Владеет на практике методологией составления научно-технических отчетов (проектов)
	ОПК-3.3 Владеет методами визуального и графического представления результатов научной (научно-технической, инновационной технологической) деятельности в виде отчетов, научных публикаций
ПК-1 Способен ставить, формализовывать и решать задачи, в том числе разрабатывать и исследовать математические модели изучаемых явлений и процессов, системно анализировать научные проблемы, получать новые научные результаты	ПК-1.2 Способен выдвигать гипотезы, строить математические модели для описания изучаемых явлений и процессов, оценивать качество разработанной модели
	ПК-1.1 Способен находить, анализировать и обобщать информацию об актуальных результатах исследований в рамках тематической области своей профессиональной деятельности
	ПК-1.3 Способен применять теоретические и (или) экспериментальные методы исследований к конкретной научной задаче и интерпретировать полученные результаты
ПК-2 Способен самостоятельно или в качестве члена (руководителя) малого коллектива организовывать и проводить научные исследования и их апробацию	ПК-2.1 Знает принципы построения научной работы, методы сбора и анализа полученного материала, способы аргументации
	ПК-2.2 Способен планировать и проводить научные исследования самостоятельно или в качестве члена (руководителя) малого научного коллектива

2. Показатели оценивания компетенций

В результате изучения дисциплины «Прикладная статистика и анализ данных» обучающийся должен:

знать:

- основные понятия математической статистики;
- основные подходы к сравнению оценок параметров неизвестного распределения;
- асимптотические и неасимптотические свойства оценок параметров неизвестного распределения;
- основные методы построения оценок с хорошими асимптотическими свойствами: метод моментов, метод максимального правдоподобия, метод выборочных квантилей;
- понятие эффективных оценок и неравенство информации Рао-Крамера;
- определение и главные свойства условного математического ожидания случайной величины относительно сигма-алгебры или другой случайной величины;
- определение общей линейной регрессионной модели и метод наименьших квадратов;
- многомерное нормальное распределение и его основные свойства;
- базовые понятия теории проверки статистических гипотез;
- лемму Неймана – Пирсона и теорему о монотонном отношении правдоподобия;
- критерий хи-квадрат Пирсона для проверки простых гипотез в схеме Бернулли.

уметь:

- обосновывать асимптотические свойства оценок с помощью применения предельных теорем теории вероятностей;
- строить оценки с хорошими асимптотическими свойствами для параметров неизвестного распределения по заданной выборке из него;
- находить байесовские оценки по заданному априорному распределению;
- вычислять условные математические ожидания с помощью условных распределений;
- находить оптимальные оценки с помощью полных достаточных статистик;
- строить точные и асимптотические доверительные интервалы, и области для параметров неизвестного распределения;
- находить оптимальные оценки и доверительные области в гауссовской линейной модели;
- строить равномерно наиболее мощные критерии в случае параметрического семейства с монотонным отношением правдоподобия;
- строить F-критерий для проверки линейных гипотез в линейной гауссовской модели.

владеть:

- основными методами математической статистики построения точечных и доверительных оценок: методом моментов, выборочных квантилей, максимального правдоподобия, методом наименьших квадратов, методом центральной статистики.
- навыками асимптотического анализа статистических критериев;
- навыками применения теорем математической статистики в прикладных задачах физики и экономики.

3. Перечень типовых (примерных) вопросов, заданий, тем для подготовки к текущему контролю

Примеры заданий для контрольных работ:

1. Во взвешенном методе наименьших квадратов каждому наблюдению задается некоторый известный вес w_i . Задача имеет вид $\sum_{i=1}^n w_i \left(Y_i - x_i^T \theta \right)^2 \rightarrow \min_{\theta}$. Найдите решение задачи в матричном виде.
2. Проведите эксперимент по определению реального уровня значимости критерия для проверки гипотезы о незначимости коэффициента в гауссовской линейной модели, если на самом деле в данных присутствует гетероскедастичность.
3. Для этого смоделируйте некоторым образом двумерные данные x и посчитайте по ним ожидаемый отклик
4. $y(x) = \theta_0 + \theta_1 x^{(1)} + \theta_2 x^{(2)}$, где коэффициенты выберите по своему усмотрению, причем $\theta_2 = 0$.

5. Зашумите набор значений $y(x_i)$ некоторым шумом, дисперсия которого зависит от x или от номера наблюдения.
6. По таким данным обучите линейную модель и проверьте гипотезу $H_0: \theta_2 = 0$.
7. Повторите эксперимент несколько раз и посчитайте долю случаев, в которых гипотеза отвергается. Распределение шума должно быть одинаковым в каждом эксперименте.
8. Пусть $\mathcal{X} = \mathbb{R}^2$ --- пространство признаков, $\mathcal{Y} = \{0, 1\}$ --- множество классов. Рассматривается квадратичный дискриминантный анализ. Условное распределение X при условии $Y=k$ равно $\mathcal{N}(a_k, \Sigma_k)$. Приведите примеры таких a_k, Σ_k и вероятностей $\text{Prob}(Y = k)$, при которых разделяющая поверхность является
 9. гиперболой; параболой; двумя параллельными прямыми; двумя пересекающимися прямыми.
10. Пусть X_1, \dots, X_n --- выборка в пространстве \mathbb{R}^D , а Y_1, \dots, Y_n --- ее проекция на линейное подпространство размерности $d < D$. Докажите, что величина $\sum_{i=1}^n (X_i - Y_i)^2$ минимальна, если Y_1, \dots, Y_n --- проекция на линейное подпространство, образованное первыми d главными компонентами. Чему она равна?
11. Медицинская лаборатория проводит испытания нового препарата для лечения некоторого заболевания. Для исследований были отобраны 2500 больных. Некоторые из них принимали новый препарат, а другие --- плацебо. В первой группе значимое улучшение состояния наблюдается среди 853 пациентов из 1719 пациентов, принимавших новый препарат. Во второй группе значимое улучшение наблюдается среди 369 пациентов из 781 пациентов, принимавших плацебо. Влияет ли новый препарат на улучшение состояния у пациентов?
12. Рассмотрим критерий Уилкоксона-Манна-Уитни. Докажите, что при отсутствии совпадений среди X_i и Y_j для статистики Манна-Уитни справедливо $U = V - \frac{m(m+1)}{2}$. Найдите $\text{Exp} U$ при справедливости гипотезы H_0 об однородности выборок.

4. Перечень типовых (примерных) вопросов и тем для проведения промежуточной аттестации обучающихся

1. Как проводить проверку линейных гипотез в линейной регрессии при наличии гетероскедастичности?
2. Какова цель снижения размерности в данных?
3. Какие методы снижения размерности в данных можно применять при 1000 признаков?
4. В чем заключается проклятие размерности?
5. Как проверить независимости вещественного и категориального признака?
6. Как строится критерий Уилкоксона-Манна-Уитни?
7. Как проводить АВ-тестирование? Как проводить разбиение пользователей?
8. В чем преимущество перестановочных критериев?
9. В чем отличие статистической связи от величины причинно-следственного эффекта?
10. Как оценить величину причинно-следственного эффекта с помощью интервенции в графе?
11. Какие методы статистического анализа применяются для описательной статистики и в каких случаях они используются?
12. Какие основные понятия и методы корреляционного анализа вы знаете? Как они применяются на практике?
13. Что такое регрессионный анализ? Какие типы регрессии существуют и какие задачи они решают?
14. Как проводится анализ дисперсии (ANOVA) и в каких областях он применяется?
15. Что такое метод максимального правдоподобия и как он используется для оценки параметров статистических моделей?
16. Какие методы проверки статистических гипотез вы знаете? В чем заключается процедура проверки гипотез?
17. Какие методы машинного обучения могут быть использованы для анализа данных? В чем отличие подхода машинного обучения от классической статистики?
18. Какие методы анализа временных рядов существуют? Какие задачи решаются при анализе временных рядов?

19. Как проводится кластерный анализ? В каких областях он может быть применен для анализа данных?

20. Какие программные средства и инструменты вы используете для проведения статистического анализа данных? Какие преимущества и недостатки у этих инструментов?

Критерии оценивания

Оценка "Отлично" (10) - полностью и вовремя решены все задачи без ошибок. Продemonстрирован грамотный подход к решению задач, реализованы оптимальные алгоритмы, код оформлен в едином удобочитаемом стиле.

Оценка "Отлично" (9) - полностью и вовремя решены все задачи без ошибок. Продemonстрирован грамотный подход к решению задач, реализованы оптимальные алгоритмы.

Оценка "Отлично" (8) - полностью и вовремя решены все задачи без ошибок. Продemonстрирован грамотный подход к решению задач.

Оценка "Хорошо" (7) - полностью решены все задачи. Допущены несущественные ошибки.

Оценка "Хорошо" (6) - полностью решено большинство задач. В некоторых задачах допущены и не исправлены ошибки, либо некоторые задачи решены частично.

Оценка "Хорошо" (5) - полностью решено две трети задач. В некоторых задачах допущены и не исправлены ошибки, либо некоторые задачи решены частично.

Оценка "Удовлетворительно" (4) - полностью решено более половины задач. В остальных задачах допущены и не исправлены ошибки, либо некоторые задачи решены частично.

Оценка "Удовлетворительно" (3) - полностью решено более половины задач.

Оценка "Неудовлетворительно" (2) - решено менее половины задач.

Оценка "Неудовлетворительно" (1) - не решено ни одной задачи.

5. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности

Экзамен может проводиться по итогам текущей успеваемости и сдачи заданий и других видов работ, предусмотренных программой дисциплины и (или) путем организации специального опроса, проводимого в устной и (или) письменной форме.

При проведении устного экзамена обучающемуся предоставляется 30 минут на подготовку. Опрос обучающегося не должен превышать одного астрономического часа.

Во время проведения экзамена обучающиеся могут пользоваться программой дисциплины, а также справочной литературой, конспектами лекций или другими материалами.