МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ

«Национальный исследовательский ядерный университет «МИФИ»

Обнинский институт атомной энергетики –

филиал федерального государственного автономного образовательного учреждения высшего профессионального образования «Национальный исследовательский ядерный университет «МИФИ»

(ИАТЭ НИЯУ МИФИ)

Отделение интеллектуальных кибернетических систем

Одобрено на заседании УМС
ИАТЭ НИЯУ МИФИ Протокол
от 30.08.2022 № 2-8/2022

МЕТОДИЕСКИЕ МАТЕРИАЛЫ

учебной дисциплины

ОБРАБОТКА И СТАТИСТИЧЕСКИЙ АНАЛИЗ БОЛЬШИХ ДАННЫХ

для студентов направления подготовки

09.04.01 – Информатика и вычислительная техника

программа
Большие данные и машинное обучение в атомной энергетике

Форма обучения: очная

г. Обнинск 2022г.

Методические указания для обучающихся по освоению дисциплины

Вид учебного занятия	Организация деятельности студента
Лекция	Написание конспекта лекций: кратко, схематично, последовательно фиксировать основные положения, выводы, формулировки, обобщения; помечать важные мысли, выделять ключевые слова, термины. Проверка терминов, понятий с помощью энциклопедий, словарей, справочников с выписыванием толкований в тетрадь. Обозначить вопросы, термины, материал, который вызывает трудности, пометить и попытаться найти ответ в рекомендуемой литературе. Если самостоятельно не удается разобраться в материале, необходимо сформулировать вопрос и задать преподавателю на консультации, на практическом занятии. Уделить внимание следующим понятиям: эмпирическая функция распределения, параметр сглаживания, метрика Колмогорова, Мизеса, Андерсона-Дарлинга, ядерная оценка плотности распределения, проекционная оценка плотности распределения, систематическая и случайная ошибки оценивания, оптимальный параметр сглаживания, базисные функции, метод локальной аппроксимации, простая и сложная гипотеза, нулевая гипотеза, альтернатива, ошибка первого и второго рода, тест, критерий, критическая область, ранг и инверсия, ранговая корреляция, параметры положения, сдвига и масштаба, фактор
Практические и лабораторные занятия	Проработка рабочей программы, уделяя особое внимание целям и задачам, структуре и содержанию дисциплины. Работа с конспектом лекций, просмотр рекомендуемой литературы. Изучение выбранной предметной области, включая задачи семинарских занятий и домашних работ.
Курсовая работа	Не предусмотрена
Контрольная работа	Ознакомиться с основной и дополнительной литературой, включая справочные издания, зарубежные источники, основополагающие термины. Попрактиковаться в решении задач по всем темам двух контрольных работ
Подготовка к зачету	При подготовке к экзамену в письменной форме ориентироваться на методические пособия, конспект лекций, рекомендуемую литературу и др.

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

1. Язык для статистической обработки данных и работы с графикой R для создания и демонстрации графиков законов распределения случайной величины и выполнения расчетов ИДЗ.
2. MS Excel для выполнения расчетов ИДЗ
3. Издательская система LaTeX для подготовки докладов, презентаций и учебного материала
4. Материалы открытой энциклопедии Wikipedia // Корневая URL: http://ru.wikipedia.org/wiki/Математическая статистика
5. Сайт проекта R: http://www.r-project.org
6. Язык **программирования** R/**Введение — Викиучебник** https://ru.wikibooks.org/wiki/Язык_программирования_R/Введение
7. R — объектно-ориентированная статистическая среда: http://ashipunov.info/shipunov/software/r/r-ru.htm
8. Графическая галерея R — примеры графики, генерируемой R: http://web.archive.org/web/20130113002105/http://gallery.r-enthusiasts.com/.

2. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине

Видеопроектор, компьютер, язык и среда R для статистической обработки данных и работы с графикой для создания и демонстрации графиков законов распределения случайной величины,

выполнения контрольных работ и экзамена, MS Office Excel, издательская система LaTeX для подготовки докладов, презентаций и учебного материала.

.

3. Иные сведения и (или) материалы

3.1. Перечень образовательных технологий, используемых при осуществлении образовательного процесса по дисциплине

Часов в интерактивной форме – 26.

Лекционный материал основан на большом количестве задач непараметрической математической статистики. Некоторые лекции сопровождаются презентациями, к примеру, разъясняющими статистические методы и модели. С привлечением языка и среды разработки R поясняются необходимые методы обработки статистической информации. Вариация входных параметров закона распределения, объема выборки и графическое сопровождение позволяет студенту лучше понять смысл получаемых оценок. Также в среде разработки R выполняется контрольные работы, готовится реферат и экзамен.

В ходе практических занятий происходит публичное обсуждение каждой решаемой задачи и статистического метода, его достоинства и недостатки. При этом студенты высказывают свои мнения и дополняют построенную статистическую модель или предлагают свои модели, в той ситуации, когда задача этого требует.

После проведения контрольных работ на консультациях проводится разбор допущенных студентами ошибок.

3.2. Формы организации самостоятельной работы обучающихся (темы, выносимые для самостоятельного изучения; вопросы для самоконтроля; типовые задания для самопроверки

Самостоятельно изучается некоторые темы математической статистики при подготовке реферата. Для изучения используется приведенная в списке основная и дополнительная литература. Контроль освоения материала осуществляется в ходе экзамена.

№	Тема и часть, изучаемая (осваиваемая) самостоятельно
1.1	**Оценка функции распределения.** Статистики Колмогорова, Мизеса, Андерсона–Дарлинга. Теорема Гливенко-Кантелли. Построение эмпирической ф.р. в среде R имеющимися средствами.
1.2	**Оценка плотности распределения.** Построение гистограмм и ядерных оценок в среде R имеющимися средствами.
1.3	**Порядковые статистики.** Непараметрические доверительные интервалы.
2.1	Ранги и ранговые критерии в среде R. Некоторые свойства рангов, инверсий и статистик от них.
2.2	**Непараметрический регрессионный анализ.** Прикладные задачи идентификации. Метод локальной аппроксимации.
3.1	**Задача о положении.** Критерий Фищера, Вилкоксона, знаковых рангов Вилкоксона, Гупта и оценки, связанные с ними.
3.2	**Задача о рассеянии.** Критерий Ансари-Бредли, Мозеса, Миллера и оценки, связанные с ними.
3.3	**Дисперсионный анализ.** Критерий Краскела-Уоллеса,Джонкхиера-Терпстры, Фридмана и др., и оценки, связанные с ними.
3.4	**Другие критерии и задачи.** Биномиальный критерий, критерии Холлендера-Прошана, критерии нормальности и др. критерии и оценки, связанные с ними.

4.Вопросы и задания для самоконтроля по всем темам:

1. Смоделировать в среде R выборку известного распределения заданного объема.

2. Построить графики эмпирической функции распределения, гистограммы, полигона частот, ядерной оценки плотности распределения.
3. Найти числовые выборочные характеристики.
4. Вариацией входных параметров встроенных функций определить их возможности.
5. Сравнить с методикой выполнения тех же задач в других средах.
6. Разобраться с возможностями МНК в среде R.
7. Решить задачу линейного регрессионного анализа с предложенными преподавателем базисными функциями в среде R.
8. Как построить д.и. для квантили?
9. Как найти м.о. числа инверсий, суммы рангов?
10. Является ли сумма рангов с.в.? А если есть связи?
11. Чем критерий Вилкоксона отличается от критерия знаковых рангов Вилкоксона?
12. В каких задачах необходимо применять критерий Краскела-Уоллеса, а в каких – Фридмана?
13. Что такое фактор в дисперсионном анализе?
14. Означает ли равенство двух параметров положения (масштаба) однородность данных?

12.3. Краткий терминологический словарь

Выборка	Массив независимых одинаково распределенных с.в.
Объем выборки	Размер массива
Статистика	Функция от выборки
Оценка параметра	Статистика, приближенно оценивающая неизвестный параметр
Несмещенная оценка	Оценка, условное м.о. которой равно оцениваемому параметру
Состоятельная оценка	Оценка, в том или ином вероятностном смысле приближающаяся к неизвестному при росте объема выборки
Эффективная оценка	Несмещенная оценка с минимально возможной дисперсией
Асимптотически нормальная оценка	Оценка, распределение которой приближается к нормальному закону при росте объема выборки
Доверительный интервал	Интервал со случайными границами, накрывающий оцениваемый параметр с вероятностью не ниже заданной
Доверительная вероятность (надежность)	Уровень вероятности, определяющий длину д.и. Обычно берется равным – 90, 95, 100%
Статистическая гипотеза	Утверждение о законе распределения с.в., о виде зависимости и т.д.
Нулевая гипотеза	Основная гипотеза
Альтернатива	Гипотеза, хотя бы в чем- то противоречащая нулевой гипотезе