

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

**федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Нижегородский государственный университет
им. Н.И. Лобачевского»**

Институт информационных технологий, математики и механики

УТВЕРЖДЕНО

решением президиума Ученого совета ННГУ

протокол № 1 от 16.01.2024 г.

Рабочая программа дисциплины

Введение в анализ данных

Уровень высшего образования

Магистратура

Направление подготовки / специальность

01.04.02 - Прикладная математика и информатика

Направленность образовательной программы

Анализ данных в прикладных областях

Форма обучения

очная

г. Нижний Новгород

2024 год начала подготовки

1. Место дисциплины в структуре ОПОП

Дисциплина Б1.О.06 Введение в анализ данных относится к обязательной части образовательной программы.

2. Планируемые результаты обучения по дисциплине, соотнесенные с планируемыми результатами освоения образовательной программы (компетенциями и индикаторами достижения компетенций)

Формируемые компетенции (код, содержание компетенции)	Планируемые результаты обучения по дисциплине (модулю), в соответствии с индикатором достижения компетенции		Наименование оценочного средства	
	Индикатор достижения компетенции (код, содержание индикатора)	Результаты обучения по дисциплине	Для текущего контроля успеваемости	Для промежуточной аттестации
ОПК-2: Способен совершенствовать и реализовывать новые математические методы решения прикладных задач	ОПК-2.1: Знает современные математические методы решения прикладных задач ОПК-2.2: Умеет совершенствовать математические методы решения прикладных задач ОПК-2.3: Имеет навыки создания новых математических методов решения прикладных задач	ОПК-2.1: Знает требования, необходимые для выполнения при подготовке данных к анализу Знает способы проведения разведывательного анализа данных. ОПК-2.2: Умеет осуществлять автоматизированный сбор данных. Умеет выявлять зависимости в анализируемых данных Умеет находить оценки параметров распределения данных ОПК-2.3: Владеет навыками проверки качества построенных оценок при помощи готовых библиотек	Задачи	Экзамен: Задачи
ПК-3: Способен представлять результаты проведенной работы в области профессиональной деятельности	ПК-3.1: Знает методы подготовки отчетов, статей, докладов, презентаций, публикаций по результатам проведенной работы в области профессиональной деятельности ПК-3.2: Умеет оформлять отчеты, статьи, доклады,	ПК-3.1: Знает приемы и возможности визуализации данных средствами Python. ПК-3.2: Умеет адекватно интерпретировать результаты аналитической деятельности, реализованной	Собеседование	Экзамен: Контрольные вопросы

	презентации по результатам проведенной работы в области профессиональной деятельности ПК-3.3: Имеет опыт подготовки отчетов, докладов, статей, презентаций по результатам проведенной работы в области профессиональной деятельности	средствами Python. ПК-3.3: Владеет навыками проведения полного цикла работ по анализу данных от сбора данных до интерпретации полученных результатов и подготовки соответствующих отчетов.		
--	---	--	--	--

3. Структура и содержание дисциплины

3.1 Трудоемкость дисциплины

	очная
Общая трудоемкость, з.е.	6
Часов по учебному плану	216
в том числе	
аудиторные занятия (контактная работа):	
- занятия лекционного типа	32
- занятия семинарского типа (практические занятия / лабораторные работы)	32
- КСР	2
самостоятельная работа	114
Промежуточная аттестация	36 Экзамен

3.2. Содержание дисциплины

(структурированное по темам (разделам) с указанием отведенного на них количества академических часов и виды учебных занятий)

Наименование разделов и тем дисциплины	Всего (часы)	в том числе			
		Контактная работа (работа во взаимодействии с преподавателем), часы из них			Самостоятельная работа обучающегося, часы
		Занятия лекционного типа	Занятия семинарского типа (практические занятия/лабораторные работы), часы	Всего	
	Ф	Ф	Ф	Ф	Ф
Роль теории вероятностей и математической статистики при анализе данных	4	2	0	2	2
Типы статистических данных (числовые, ординальные, номинальные). Генеральная совокупность, выборка. Репрезентативность выборки.	13	3	2	5	8
Одномерные и многомерные данные. Законы распределения.	27	5	6	11	16

Числовые характеристики одномерных случайных величин.	14	2	2	4	10
Числовые характеристики многомерной случайной величины.	16	2	2	4	12
Регрессия, подгонка прямой под облако точек.	30	4	6	10	20
Кластеризация данных, задача и методы решения.	29	5	6	11	18
Оценивание неизвестных параметров распределения.	22	4	4	8	14
Критерий согласия и их применение.	23	5	4	9	14
Аттестация	36				
КСР	2			2	
Итого	216	32	32	66	114

Содержание разделов и тем дисциплины

Роль теории вероятностей и математической статистики при анализе данных, возможности языка Python для анализа.

Типы статистических данных (числовые, ординальные, номинальные). Генеральная совокупность, выборка. Репрезентативность выборки. Способы обеспечения сопоставимости данных. Обзор существующих открытых библиотек данных.

Одномерные и многомерные данные. Понятие одномерной случайной величины, дискретные и непрерывные одномерные случайные величины. Понятие многомерной случайной величины.

Эмпирическая плотность и эмпирическая функция распределения, примеры их построения.

Визуализация эмпирических распределений: построение гистограмм, графиков функций распределения.

Числовые характеристики одномерных случайных величин. Статистические числовые характеристики.

Анализ данных на основе статистических числовых характеристик: характеристик центрального положения, разброса. Оценка функции распределения на основе квантилей.

Числовые характеристики многомерной случайной величины. Ковариация, коэффициент корреляции.

Построение ковариационной матрицы. Корреляционный анализ. Построение диаграмм рассеивания.

Выявление зависимости между величинами на основе ковариационной матрицы.

Регрессия, подгонка прямой под облако точек. Простая линейная регрессия. Многомерная регрессия.

Полиномиальная регрессия. Построение простой и многомерной линейной регрессии.

Кластеризация данных, задача и методы решения (метод к-средних, иерархическая кластеризация).

Определение оптимального количества кластеров.

Оценивание неизвестных параметров распределения. Точечные оценки. Параметрическое и непараметрическое оценивание. Метод максимального правдоподобия и метод моментов.

Простые и сложные гипотезы. Ошибки первого и второго рода. Критерий согласия и их применение.

Обзор готовых средств для проверки гипотез.

4. Учебно-методическое обеспечение самостоятельной работы обучающихся

Самостоятельная работа обучающихся включает в себя подготовку к контрольным вопросам и заданиям для текущего контроля и промежуточной аттестации по итогам освоения дисциплины приведенным в п. 5.

Для обеспечения самостоятельной работы обучающихся используются:

- электронный курс "Теория вероятностей и математическая статистика ДО" (<https://e-learning.unn.ru/course/view.php?id=789>).

Иные учебно-методические материалы: Самостоятельная работа обучающихся включает в себя подготовку к контрольным вопросам и задачам для текущего контроля и промежуточной

аттестации по итогам освоения дисциплины,
приведенным в п. 5.

5. Фонд оценочных средств для текущего контроля успеваемости и промежуточной аттестации по дисциплине (модулю)

5.1 Типовые задания, необходимые для оценки результатов обучения при проведении текущего контроля успеваемости с указанием критериев их оценивания:

5.1.1 Типовые задания (оценочное средство - Задачи) для оценки сформированности компетенции ОПК-2:

Задача 1.

Дайте интерпретацию характера зависимости между случайными величинами ξ_1 и ξ_2 по графику регрессии $R_{\xi_2|\xi_1}(x)$, представленном на рис. 1.

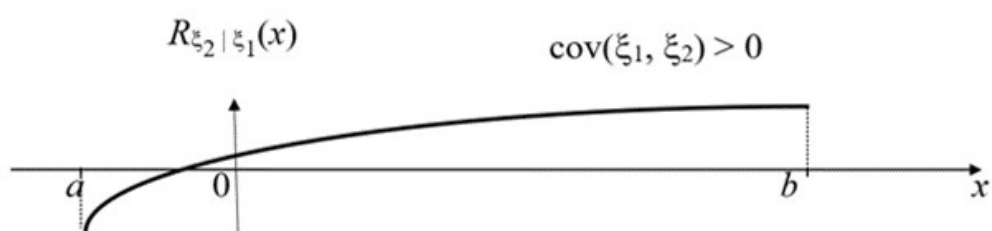


Рис. 1

Задача 2.

К задаче прилагаются два файла с данными о числе организаций, осуществляющих образовательную деятельность по субъектам РФ. В документе **01_Образование_организации_01.xlsx** содержатся данные за 2016, 2017 года, а в документе **01_Образование_организации_02.xlsx** – за 2015, 2018 года.

1. Для данных документов необходимо:

- Загрузить данные из документов для работы в Python средствами библиотеки `xlrd`.
- Составить словарь (dictionary), ключом в котором является название субъекта РФ, а значением – список из четырех элементов: число образовательных организаций в 2015, 2016, 2017, 2018 годах.

Примечания: 1) исключить из рассмотрения сводную информацию по федеральным округам и РФ; 2) в списках к каждому субъекту соблюсти хронологический порядок данных: начиная с 2015го и заканчивая 2018м годом.

Критерии оценивания (оценочное средство - Задачи)

Оценка	Критерии оценивания
превосходно	Продемонстрированы все основные умения, решены все основные задачи в

Оценка	Критерии оценивания
	полном объеме без недочетов
отлично	Продemonстрированы все основные умения, решены все основные задачи в полном объеме, с одним недочетом.
очень хорошо	Продemonстрированы все основные умения. Решены все основные задачи, но некоторые с недочетами.
хорошо	Продemonстрированы все основные умения. Решены все основные задачи с негрубыми ошибками и некоторые с недочетами.
удовлетворительно	Продemonстрированы основные умения. Решены типовые задачи с негрубыми ошибками. Решены все предложенные задачи но не в полном объеме.
неудовлетворительно	При решении стандартных задач не продemonстрированы основные умения. Имели место грубые ошибки.
плохо	Отсутствие минимальных умений. Невозможность оценить наличие умений вследствие отказа обучающегося от ответа.

5.1.2 Типовые задания (оценочное средство - Собеседование) для оценки сформированности компетенции ПК-3:

1. Перечислить способы представления выборочных значений.
2. Назовите основные библиотеки Python, используемые для сбора, анализа данных.
3. Перечислить типы шкал измерений и соответствующие типы статистических данных.
4. Дать определение одномерных и многомерных данных.

Критерии оценивания (оценочное средство - Собеседование)

Оценка	Критерии оценивания
зачтено	Полный ответ на вопрос задания
не зачтено	Неполный ответ на вопрос задания или отсутствие ответа

5.2. Описание шкал оценивания результатов обучения по дисциплине при промежуточной аттестации

Шкала оценивания сформированности компетенций

Уровень сформированности компетенций (индикатора достижения компетенций)	плохо	неудовлетворительно	удовлетворительно	хорошо	очень хорошо	отлично	превосходно
	не зачтено		зачтено				
<u>Знания</u>	Отсутствие знаний теоретического материала. Невозможность оценить полноту знаний вследствие отказа обучающегося от ответа	Уровень знаний ниже минимальных требований. Имели место грубые ошибки	Минимально допустимый уровень знаний. Допущено много негрубых ошибок	Уровень знаний в объеме, соответствующем программе подготовки. Допущено несколько негрубых ошибок	Уровень знаний в объеме, соответствующем программе подготовки. Допущено несколько несущественных ошибок	Уровень знаний в объеме, соответствующем программе подготовки. Ошибок нет.	Уровень знаний в объеме, превышающем программу подготовки.
<u>Умения</u>	Отсутствие минимальных умений. Невозможность оценить наличие умений вследствие отказа обучающегося от ответа	При решении стандартных задач не продемонстрированы основные умения. Имели место грубые ошибки	Продemonстрированы основные умения. Решены типовые задачи с негрубыми ошибками. Выполнены все задания, но не в полном объеме	Продemonстрированы все основные умения. Решены все основные задачи с негрубыми ошибками. Выполнены все задания в полном объеме, но некоторые с недочетами	Продemonстрированы все основные умения. Решены все основные задачи. Выполнены все задания в полном объеме, но некоторые с недочетами.	Продemonстрированы все основные умения. Решены все основные задачи с отдельными и несущественными недочетами, выполнены все задания в полном объеме	Продemonстрированы все основные умения. Решены все основные задачи. Выполнены все задания, в полном объеме без недочетов
<u>Навыки</u>	Отсутствие базовых навыков. Невозможность оценить наличие навыков вследствие отказа обучающегося от ответа	При решении стандартных задач не продемонстрированы базовые навыки. Имели место грубые ошибки	Имеется минимальный набор навыков для решения стандартных задач с некоторыми недочетами	Продemonстрированы базовые навыки при решении стандартных задач с некоторыми недочетами	Продemonстрированы базовые навыки при решении стандартных задач без ошибок и недочетов	Продemonстрированы навыки при решении нестандартных задач без ошибок и недочетов	Продemonстрирован творческий подход к решению нестандартных задач

Шкала оценивания при промежуточной аттестации

Оценка		Уровень подготовки
зачтено	превосходно	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «превосходно», продемонстрированы знания, умения, владения по соответствующим компетенциям на уровне выше предусмотренного программой

	отлично	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «отлично».
	очень хорошо	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «очень хорошо»
	хорошо	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «хорошо».
	удовлетворительно	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «удовлетворительно», при этом хотя бы одна компетенция сформирована на уровне «удовлетворительно»
не зачтено	неудовлетворительно	Хотя бы одна компетенция сформирована на уровне «неудовлетворительно».
	плохо	Хотя бы одна компетенция сформирована на уровне «плохо»

5.3 Типовые контрольные задания или иные материалы, необходимые для оценки результатов обучения на промежуточной аттестации с указанием критериев их оценивания:

5.3.1 Типовые задания (оценочное средство - Задачи) для оценки сформированности компетенции ОПК-2

Задание 1.

К заданию прилагается два файла с данными о числе организаций, осуществляющих образовательную деятельность по субъектам РФ. В документе `01_Образование_организации_01.xlsx` содержатся данные за 2016, 2017 года, а в документе `01_Образование_организации_02.xlsx` – за 2015, 2018 года.

1. Для данных документов необходимо:

- Загрузить данные из документов для работы в Python средствами библиотеки `xlsx`.
- Составить словарь (dictionary), ключом в котором является название субъекта РФ, а значением – список из четырех элементов: число образовательных организаций в 2015, 2016, 2017, 2018 годах.

Примечания: 1) исключить из рассмотрения сводную информацию по федеральным округам и РФ; 2) в списках к каждому субъекту соблюсти хронологический порядок данных: начиная с 2015го и заканчивая 2018м годом. Пример итоговых записей в словаре:

```
'Белгородская область': [193, 906, 916, 885],
'Брянская область': [163, 607, 715, 727]}
```

2. Для данных документов необходимо:

- В полученном словаре для каждого субъекта РФ добавить в список 2 дополнительных элемента: среднее количество образовательных организаций за 4 года и год, в котором было достигнуто максимальное количество образовательных организаций.

Пример итоговых записей в словаре:

```
'Белгородская область': [193, 906, 916, 885, 725, 2017],
'Брянская область': [163, 607, 715, 727, 553, 2018]}
```


3. Для данных документов необходимо:

- а. С использованием библиотечного средства Counter составить словарь, в котором ключом является год из диапазона 2015-2018, а значением – количество субъектов РФ, в которых именно в этом году количество образовательных организаций было максимальным.
- б. Отсортировать субъекты РФ по возрастанию среднего числа образовательных организаций за 4 года.

Примечание: для решения указанной задачи написать функцию, возвращающую пятый элемент списка, использовать встроенную функцию сортировки sorted.

Критерии оценивания (оценочное средство - Задачи)

Оценка	Критерии оценивания
превосходно	Продemonстрированы все основные умения, решены все основные задачи в полном объеме без недочетов
отлично	Продemonстрированы все основные умения, решены все основные задачи в полном объеме, с одним недочетом.
очень хорошо	Продemonстрированы все основные умения. Решены все основные задачи, но некоторые с недочетами.
хорошо	Продemonстрированы все основные умения. Решены все основные задачи с негрубыми ошибками и некоторые с недочетами.
удовлетворительно	Продemonстрированы основные умения. Решены типовые задачи с негрубыми ошибками. Решены все предложенные задачи но не в полном объеме.
неудовлетворительно	При решении стандартных задач не продemonстрированы основные умения. Имели место грубые ошибки.
плохо	Отсутствие минимальных умений. Невозможность оценить наличие умений вследствие отказа обучающегося от ответа

5.3.2 Типовые задания (оценочное средство - Контрольные вопросы) для оценки сформированности компетенции ПК-3

1. Перечислить способы представления выборочных значений.
2. Назовите основные библиотеки Python, используемые для сбора, анализа данных.
3. Перечислить типы шкал измерений и соответствующие типы статистических данных.
4. Дать определение одномерных и многомерных данных.
5. Дать определение статистической (выборочной, эмпирической) функции распределения.

6. С помощью каких графиков в Python визуализируются многомерные данные?
7. Дать определение статистической (выборочной, эмпирической) плотностью вероятностей и гистограммы.
8. Определить, основные статистические числовые характеристик (выборочное среднее, дисперсию и стандарт).
9. Привести формулы для вычисления выборочный начального и центрального моментов k-го порядка, статистическую медиану.
10. Привести способы вычисления статистической ковариации и выборочного коэффициента корреляции, определить понятие несмещенной выборочной ковариации.
11. Какая библиотека в Python содержит функции для вычисления ковариации и коэффициента корреляции.
12. Перечислить способы выявления статистической зависимости двух случайных величин.
13. Дать определения понятию регрессии двух случайных величин.
14. Определить простую линейную регрессию. Перечислить Ключевые различия между корреляцией и линейной регрессией.
15. Описать метод использования регрессии для прогнозирования. Определить, что такое подогнанные значения и остатки.
16. Дать определение множественной линейной регрессии.
17. Какими средствами в Python строится линейная регрессия?
18. Определить понятие нелинейной регрессии, указать, на какие классы она подразделяется.
19. Перечислить способы диагностики качества регрессионной модели.
20. Какими средствами в Python проверяется качество построенной регрессионной модели?
21. Дать определения понятия точечного оценивания неизвестных параметров распределения. Перечислить критерии качества статистических оценок.
22. Описать Метод моментов.
23. Описать Метод максимального правдоподобия.
24. Какими средствами реализуются в Python реализуется метод моментов?
25. Описать метод интервального оценивания неизвестных параметров распределения. Определить понятие доверительного интервала.

26. Описать методологию построения доверительных интервалов для математического ожидания.
27. Описать методологию построения доверительных интервалов для неизвестной дисперсии.
28. Описать методологию построения доверительных интервалов для неизвестной вероятности события.
29. Дать определения понятию статистической гипотезы.
30. Перечислить основные принципы построения критериев согласия.
31. Описать метод проверки простых гипотез с помощью критерия согласия Колмогорова.
32. Описать метод проверки простых гипотез о виде распределения с помощью критерия согласия хи-квадрат Пирсона.
33. Какими средствами в Python проверяются простые гипотезы о виде распределения?

Критерии оценивания (оценочное средство - Контрольные вопросы)

Оценка	Критерии оценивания
превосходно	Уровень знаний в объеме, превышающем программу подготовки продемонстрирован при ответах на вопросы.
отлично	Уровень знаний в объеме, соответствующем программе подготовки, отсутствие ошибок при ответах на вопросы.
очень хорошо	Уровень знаний в объеме, соответствующем программе подготовки. Допущено несколько несущественных ошибок при ответе на вопросы.
хорошо	Уровень знаний в объеме, соответствующем программе подготовки. Допущено несколько негрубых ошибок при ответе на вопросы.
удовлетворительно	Минимально допустимый уровень знаний. Допущено много негрубых ошибок при ответе на вопросы.
неудовлетворительно	Уровень знаний ниже минимальных требований. Имели место грубые ошибки в ответах на вопросы.
плохо	Отсутствие знаний теоретического материала. Невозможность оценить полноту знаний вследствие отказа обучающегося от ответа.

6. Учебно-методическое и информационное обеспечение дисциплины (модуля)

Основная литература:

1. Федоткин Михаил Андреевич. Основы прикладной теории вероятностей и статистики : учеб. для студентов вузов, обучающихся по специальности "Прикладная математика и информатика" и

по направлению "Прикладная математика и информатика". - М. : Высшая школа, 2006. - 368 с. : ил. - ISBN 5-06-005328-8 : 215.60., 183 экз.

2. Федоткин Михаил Андреевич. Модели в теории вероятностей : учебник. - М. : Физматлит : ННГУ, 2012. - 608 с. - (Библиотека Нижегородского государственного университета им. Н. И. Лобачевского). - ISBN 978-5-9221-1384-7 : 600.00., 200 экз.

Дополнительная литература:

1. Белько Иван Васильевич. Высшая математика для экономистов : 3 семестр : Теория вероятностей и математическая статистика. - М. : Новое знание, 2002. - 144 с. - (Экспресс-курс). - Лит.: с. 140. - ISBN 5-94735-015-7 : 62-00., 2 экз.

Программное обеспечение и Интернет-ресурсы (в соответствии с содержанием дисциплины):

Учебные аудитории для проведения учебных занятий, предусмотренных образовательной программой, оснащены мультимедийным оборудованием (проектор, экран), техническими средствами обучения (компьютерами)

На компьютерах должны быть установлены:

1. Операционная система Microsoft Windows.
2. Open-source среда Spyder.
3. Веб-интерактивная вычислительная среда Jupyter Notebook (для поддержки языка Python).

Помещения для самостоятельной работы обучающихся оснащены компьютерной техникой с таким же наполнением ПО и с

возможностью подключения к сети "Интернет" и обеспечены доступом в электронную информационно-образовательную среду.

7. Материально-техническое обеспечение дисциплины (модуля)

Учебные аудитории для проведения учебных занятий, предусмотренных образовательной программой, оснащены мультимедийным оборудованием (проектор, экран), техническими средствами обучения, компьютерами.

Помещения для самостоятельной работы обучающихся оснащены компьютерной техникой с возможностью подключения к сети "Интернет" и обеспечены доступом в электронную информационно-образовательную среду.

Программа составлена в соответствии с требованиями ОС ННГУ по направлению подготовки/специальности 01.04.02 - Прикладная математика и информатика.

Автор(ы): Пройдакова Екатерина Вадимовна, кандидат физико-математических наук, доцент.

Заведующий кафедрой: Зорин Андрей Владимирович, доктор физико-математических наук.

Программа одобрена на заседании методической комиссии от 13.12.2023, протокол № 3.