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
(факультет / институт / филиал)

УТВЕРЖДЕНО
решением Ученого совета ННГУ
протокол от 30 ноября 2022 г. № 13

Рабочая программа дисциплины

Введение в анализ данных

(наименование дисциплины (модуля))

Уровень высшего образования
магистратура
(бакалавриат / магистратура / специалитет)

Направление подготовки / специальность
01.04.02 Прикладная математика и информатика
(указывается код и наименование направления подготовки / специальности)

Направленность образовательной программы
Анализ данных в прикладных областях
(указывается профиль / магистерская программа / специализация)

Форма обучения
очная
(очная / очно-заочная / заочная)

Нижний Новгород
2023 год

1. Место дисциплины в структуре ОПОП

Дисциплина относится к обязательной части

Б1.О.06 Введение в математическую статистику

№ варианта	Место дисциплины в учебном плане образовательной программы	Стандартный текст для автоматического заполнения в конструкторе РПД
1	Блок 1. Дисциплины (модули) Обязательная часть	Дисциплина Б1.О.06, «Введение в анализ данных» относится к обязательной части ООП направления подготовки 01.04.02 «Прикладная математика и информатика».

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 обученияпо дисциплине****	
ОПК-2. Способен совершенствовать и реализовывать новые математические методы решения прикладных задач	ОПК-2.1. Знает современные математические методы решения прикладных задач	*Знает требования, необходимые для выполненияпри подготовке данных к анализу* *Знает способы проведения разведывательногоанализа данных.*	*Собеседование*
	ОПК-2.2. Умеет совершенствоват ь математическиеметоды решения прикладных задач	*Умеет осуществлять автоматизированныйсбор данных.* *Умеет выявлять зависимости в анализируемыхданных* *Умеет находить оценки параметров распределенияданных*	*Задачи (практические задания)*
	ОПК-2.3. Имеет навыки создания новых математических методов решения прикладных задач	*Владеет навыками проверки качества построенных оценок при помощи готовыхбиблиотек*	*Задачи (практические задания)*
ПК-3. Способен представлять результаты проведенной работы в области профессиональной деятельности	ПК-3.1. Знаетметоды подготовки отчетов, статей, докладов, презентаций, публикаций по результатам проведенной работы в области профессиональной деятельности.	*Знает приемы и возможности визуализации данных средствами Python.*	*Задачи (практические задания)*
	ПК-3.2. Умеет оформлять отчеты, статьи, доклады, презентации по результатам проведенной работы в области профессиональной деятельности.	*Умеет адекватно интерпретировать результаты аналитической деятельности, реализованной средствами Python.*	*Собеседование*

	ПК-3.3. Имеет опыт подготовки отчетов, докладов,статей, презентаций по результатам проведенной работы в области профессиональной деятельности	*Владеет навыками проведения полного цикла работ по анализу данных от сбора данных до интерпретации полученных результатов и подготовки соответствующих отчетов.*	*Задачи (практические задания)*

3. Структура и содержание дисциплины

3.1. Трудоемкость дисциплины

	Очная форма обучения
Общая трудоемкость	**7 ЗЕТ**
Часов по учебному плану	**252**
в том числе	
аудиторные занятия (контактная работа): - занятия лекционного типа - занятия семинарского типа - занятия лабораторного типа - текущий контроль (КСР)	**66** 32 32 0 2
самостоятельная работа	**150**
Промежуточная аттестация – экзамен	**36**

3.2. Содержание дисциплины

Наименование и краткое содержание разделов и тем дисциплины	**Всего (часы)**	В том числе				
		Контактная работа (работа во взаимодействии с преподавателем), часы. Из них				**Самостоятельная**
		Занятия лекционноготипа Очная	**Занятия семинарскоготипа** Очная	**Занятия лабораторноготипа** Очная	**Всего** Очная	
Роль теории вероятностей и математической статистики при анализе данных, возможности языка Python для сбора и анализа данных.	8	2	1			5
Типы статистических данных (числовые, ординальные, номинальные). Генеральная совокупность, выборка. Репрезентативность выборки. Способы обеспечения сопоставимости данных. Обзор существующих открытых библиотек данных.	10	4	2			4
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.	28	4	4			20

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.	24	3	4			17
Числовые характеристики многомерной случайной величины. Ковариация, коэффициент корреляции. Построение ковариационной матрицы. Корреляционный анализ. Построение диаграмм рассеивания. Выявление зависимости между величинами на основе ковариационной матрицы.	29	3	4			22
Регрессия, подгонка прямой под облако точек. Простая линейная регрессия. Многомерная регрессия. Полиномиальная регрессия. Построение простой и многомерной линейной регрессии.	25	5	5			15
Кластеризация данных, задача и методы решения (метод к-средних, иерархическая кластеризация). Определение оптимального количества кластеров.	27	3	4			20
Оценивание неизвестных параметров распределения. Точечные оценки. Параметрическое и непараметрическое оценивание. Метод максимального правдоподобия и метод моментов.	33	4	4			25
Простые и сложные гипотезы. Ошибки первого и второго рода. Критерий согласия и их применение. Обзор готовых средств для проверки гипотез.	30	4	4			22
Текущий контроль (КСР)	2				2	
Промежуточная аттестация – экзамен	36					
Итого	**252**	**32**	**32**		**66**	**150**

Текущий контроль успеваемости реализуется в формах опросов на занятиях семинарского типа и проверке отчетов по практическим заданиям. Промежуточная аттестация проходит в традиционной форме (экзамен).

4. Учебно-методическое обеспечение самостоятельной работы обучающихся

Самостоятельная работа студента включает выполнение практических заданий под контролем преподавателя, самостоятельного изучения конспектов лекций и и подготовке к промежуточной аттестации.

Контрольные вопросы и задания для проведения текущего контроля и промежуточной аттестации по итогам освоения дисциплины приведены в п. 5.2.

5. Фонд оценочных средств для промежуточной аттестации по дисциплине (модулю), включающий:

5.1. Описание шкал оценивания результатов обучения по дисциплине

Уровень сформированности компетенций (индикатора	**Шкала оценивания сформированности компетенций**						
	плохо	**неудовлетворительно**	**удовлетворительно**	**хорошо**	**очень хорошо**	**отлично**	**превосходно**

достижения компетенций)	Не зачтено		Зачтено				
Знания	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 ошибки.	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, без ошибок.	Уровень знаний в объеме, превышающем программу подготовки.
Умения	Отсутствие минимальных умений. Невозможность оценить наличие умений вследствие отказа обучающегося от ответа	При решении стандартных задач не продемонстрированы основные умения. Имели место грубые ошибки.	Продемонстрированы основные умения. Решены типовые задачи с негрубыми ошибками. Выполнены все задания, но не в полном объеме.	Продемонстрированы все основные умения. Решены все основные задачи с негрубыми ошибками. Выполнены все задания, в полном объеме, но некоторые с недочетами.	Продемонстрированы все основные умения. Решены все основные задачи. Выполнены все задания, в полном объеме, но некоторые с недочетами.	Продемонстрированы все основные умения, решены все основные задачи с отдельными несущественным недочетами, выполнены все задания в полном объеме.	Продемонстрированы все основные умения, решены все основные задачи. Выполнены все задания, в полном объеме без недочетов
Навыки	Отсутствие владения материалом. 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.	Продемонстрированы базовые навыки при решении стандартных задач с некоторыми недочетами	Продемонстрированы базовые навыки при решении стандартных задач без ошибок и недочетов.	Продемонстрированы навыки при решении нестандартных задач без ошибок и недочетов.	Продемонстрирован творческий подход к решению нестандартных задач.

Шкала оценки при промежуточной аттестации

Оценка	**Уровень подготовки**
Превосходно	Все компетенции (части компетенций), на формирование которыхнаправлена дисциплина, сформированы на уровне не ниже «превосходно»
Отлично	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «отлично», при этом хотя бы одна компетенция сформирована на уровне «отлично»
Очень хорошо	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «очень хорошо», при этом хотя бы одна компетенция сформирована на уровне «очень хорошо»
Хорошо	Все компетенции (части компетенций), на формирование которых направлена дисциплина, сформированы на уровне не ниже «хорошо», при этом хотя бы одна компетенция сформирована на уровне «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»
Неудовлетворительно	Хотя бы одна компетенция сформирована на уровне «неудовлетворительно», ни одна из компетенций не сформирована науровне «плохо»
Плохо	Хотя бы одна компетенция сформирована на уровне «плохо»

5.2. Типовые контрольные задания или иные материалы, необходимые для оценки результатов обучения

5.2.1 Контрольные вопросы

Вопросы	*Код формируемой компетенции*
1. Перечислить способы представления выборочныхзначений.	ОПК-2
2. Назовите основные библиотеки Python**,** используемыедля сбора, анализа данных**.**	ПК-3
3. Перечислить типы шкал измерений и соответствующиетипы статистических данных.	ОПК-2
4. Дать определение одномерных и многомерных данных.	ОПК-2
5. Дать определение статистической (выборочной, эмпирической) функции распределения.	ОПК-2
6. С помощью каких графиков в Python визуализируются многомерные данные?	ПК-3
7. Дать определение статистической (выборочной, эмпирической) плотностью вероятностей и гистограммы.	ОПК-2
8. Определить, основные статистические числовые характеристик (выборочное среднее, дисперсию и стандарт).	ОПК-2
9. Привести формулы для вычисления выборочныйначального и центрального моментов k-го порядка, статистической медианы.	ОПК-2
10. Привести способы вычисления статистической ковариации и выборочного коэффициента корреляции, определить понятие несмещенной выборочной ковариации.	ОПК-2
11. Какая библиотека в Python содержит функции длявычисления ковариации и коэффициента корреляции.	ПК-3
12. Перечислить способы выявления статистическаязависимости двух случайных величин.	ОПК-2
13. Дать определения понятию регрессии двух случайныхвеличин.	ОПК-2
14. Определить простую линейную регрессию. Перечислить Ключевые различия между корреляцией илинейной регрессией.	ОПК-2
15. Описать метод использования регрессии для прогнозирования. Определить, что такое подогнанныезначения и остатки.	ОПК-2
16. Дать определение множественной линейнойрегрессии.	ОПК-2
17. Какими средствами в Python строится линейнаярегрессия?	ПК-3
18. Определить понятие нелинейной регрессии, указать,на какие классы она подразделяется.	ОПК-2

19. Перечислить способы диагностики качества регрессионной модели.	ОПК-2
20. Какими средствами в Python проверяется качествопостроенной регрессионной модели?	ОПК-2
21. Дать определения понятия точечного оцениваниянеизвестных параметров распределения. Перечислить критерии качества статистических оценок.	ОПК-2
22. Описать Метод моментов.	ОПК-2
23. Описать Метод максимального правдоподобия.	ОПК-2
24. Какими средствами реализуются в Python реализуетсяметод моментов?	ПК-3
25. Описать метод интервального оценивания неизвестных параметров распределения. Определить понятиедоверительного интервала.	ОПК-2
26. Описать методологию построения доверительныхинтервалов для математического ожидания.	ОПК-2
27. Описать методологию построения доверительныхинтервалов для неизвестной дисперсии.	ОПК-2
28. Описать методологию построения доверительныхинтервалов для неизвестной вероятности события.	ОПК-2
29. Дать определения понятию статистической гипотезы.	ОПК-2
30. Перечислить основные принципы построениякритериев согласия.	ОПК-2
31. Описать метод проверки простых гипотез с помощьюкритерия согласия Колмогорова.	ОПК-2
32. Описать метод проверки простых гипотез о виде распределения с помощью критерия согласия хи-квадратПирсона.	ОПК-2
33. Какими средствами в Python проверяются простыегипотезы о виде распределения?	ПК-3

5.2.2. Типовые задания/задачи для оценки сформированности компетенции «ОПК-2»

Задание 1.

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

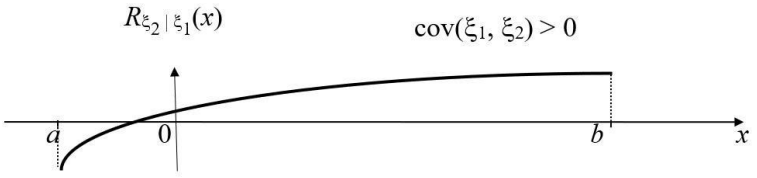

Рис. 1

Задание 2.

Определить типы данных для столбцов B-G (рис.2):

	A	B	C	D	E	F	G
1	event_time	event_type	product_id	category_id	brand	price	user_id
2	2019-11-01 00:00:02 UTC	view	5802432	1487580009286598681		0.32	562076640
3	2019-11-01 00:00:09 UTC	cart	5844397	1487580006317032337		2.38	553329724
4	2019-11-01 00:00:10 UTC	view	5837166	1783999064103190764	pnb	22.22	556138645
5	2019-11-01 00:00:11 UTC	cart	5876812	1487580010100293687	jessnail	3.16	564506666
6	2019-11-01 00:00:24 UTC	remove_from_c	5826182	1487580007483048900		3.33	553329724
7	2019-11-01 00:00:24 UTC	remove_from_c	5826182	1487580007483048900		3.33	553329724
8	2019-11-01 00:00:25 UTC	view	5856189	1487580009026551821	runail	15.71	562076640
9	2019-11-01 00:00:32 UTC	view	5837835	1933472286753424063		3.49	514649199

Рис.2

5.2.3. Типовые задачи для оценки сформированности компетенции «ПК-3»*:*

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:

   a. Загрузить данные из документов для работы в Python средствами библиотеки xlrd.

   b. 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:

   c. 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:

   a. 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.

   b. 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.

6. Учебно-методическое и информационное обеспечение дисциплины

а) основная литература:

1. Федоткин М.А. Основы прикладной теории вероятностей и статистики. — М.: Высшая школа. 2006. - 168 с.
2. Теория вероятностей и математическая статистика. Авторы: Федоткин М.А., Пройдакова Е.В.: Электронный управляемый курс. – Нижний Новгород: Нижегородский госуниверситет, 2014.

   (Идентификационный номер в электронном каталоге фонда электронных образовательных ресурсов ННГУ: 789Е.14.08)

б) дополнительная литература:

1. Лагутин М. Б. Наглядная математическая статистика: учебное пособие. — 2-е изд., испр. — М. : БИНОМ. Лаборатория знаний, 2009. — 472 с.
2. Практическая статистика для специалистов Data Science: Пер. с англ./ П. Брюс, Э. Брюс. — СПб.: БХВ-Петербург, 2018. — 304 с.

в) программное обеспечение и Интернет-ресурсы .

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

Фонд образовательных электронных ресурсов ННГУ им. Лобачевского
http://www.unn.ru/books/resources.html

7.Материально-техническое обеспечение дисциплины

Помещения представляют собой учебные аудитории для проведения учебных занятий, предусмотренных программой, обязательно оснащенные оборудованием и техническими средствами обучения: компьютерный класс, проектор, экран.

Используемое лицензионные программное обеспечение: операционные системы семейства Microsoft Windows, лицензия по подписке Microsoft Imagine.

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 01.04.02 Прикладная математика и информатика.

Автор: к.ф-м.н., доцент_____Е.В. Пройдакова

Рецензент (ы) _____

Заведующий кафедрой_____А.В. Зорин

Программа одобрена на заседании методической комиссии института информационных технологий, математики и механики от 30 ноября 2022 года, протокол № 3.