

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение
высшего образования
"Казанский (Приволжский) федеральный университет"
Институт математики и механики им. Н.И. Лобачевского



УТВЕРЖДАЮ

Проректор
по образовательной деятельности КФУ
Проф. Таюрский Д.А.

_____ 20__ г.

Программа дисциплины

Статистический анализ данных с использованием компьютерных пакетов Б1.В.ДВ.8

Направление подготовки: 02.03.01 - Математика и компьютерные науки

Профиль подготовки: Математическое и компьютерное моделирование

Квалификация выпускника: бакалавр

Форма обучения: очное

Язык обучения: русский

Автор(ы):

Новиков А.А.

Рецензент(ы):

Насыров С.Р.

СОГЛАСОВАНО:

Заведующий(ая) кафедрой: Насыров С. Р.

Протокол заседания кафедры No ____ от " ____ " _____ 201__ г

Учебно-методическая комиссия Института математики и механики им. Н.И. Лобачевского :

Протокол заседания УМК No ____ от " ____ " _____ 201__ г

Регистрационный No

Казань
2018

Содержание

1. Цели освоения дисциплины
2. Место дисциплины в структуре основной образовательной программы
3. Компетенции обучающегося, формируемые в результате освоения дисциплины /модуля
4. Структура и содержание дисциплины/ модуля
5. Образовательные технологии, включая интерактивные формы обучения
6. Оценочные средства для текущего контроля успеваемости, промежуточной аттестации по итогам освоения дисциплины и учебно-методическое обеспечение самостоятельной работы студентов
7. Литература
8. Интернет-ресурсы
9. Материально-техническое обеспечение дисциплины/модуля согласно утвержденному учебному плану

Программу дисциплины разработал(а)(и) ассистент, к.н. Новиков А.А. Кафедра математического анализа отделение математики , A.Nobukob@gmail.com

1. Цели освоения дисциплины

Основные понятия, терминологию анализа данных.

Основные методы исследования и примеры использования анализа данных.

Обучающийся должен знать теоретические основы анализа данных.

Уметь решать типовые задачи и формулировать прикладные задачи в терминах анализа данных.

Уметь строить модели естественных явлений, исследуемых методами анализа данных.

Владеть основными методами исследования, использующими анализ данных.

Владеть методами статистического анализа данных.

Владеть навыками интуитивного подхода к решению нестандартных задач с привлечением анализа данных.

Демонстрировать способность и готовность решать задачи, использующие методы анализа данных.

2. Место дисциплины в структуре основной образовательной программы высшего профессионального образования

Данная учебная дисциплина включена в раздел "Б1.В.ДВ.8 Дисциплины (модули)" основной образовательной программы 02.03.01 Математика и компьютерные науки и относится к дисциплинам по выбору. Осваивается на 4 курсе, 7 семестр.

Курс базируется на курсах, охватывающих теорию вероятностей, математическую статистику и теорию случайных процессов, а также на курсе компьютерных технологий и численных методов.

3. Компетенции обучающегося, формируемые в результате освоения дисциплины /модуля

В результате освоения дисциплины формируются следующие компетенции:

Шифр компетенции	Расшифровка приобретаемой компетенции
ОК-5 (общекультурные компетенции)	способность к коммуникации в устной и письменной формах на русском и иностранном языках для решения задач межличностного и межкультурного взаимодействия
ПК-4 (профессиональные компетенции)	способность публично представлять собственные и известные научные результаты
ПК-5 (профессиональные компетенции)	способностью использовать методы математического и алгоритмического моделирования при решении теоретических и прикладных задач
ПК-7 (профессиональные компетенции)	способность использовать методы математического и алгоритмического моделирования при анализе управленческих задач в научно-технической сфере, в экономике, бизнесе и гуманитарных областях знаний
ПК-8 (профессиональные компетенции)	способность представлять и адаптировать знания с учетом уровня аудитории
ПК-9 (профессиональные компетенции)	способность к организации учебной деятельности в конкретной предметной области (математика, физика, информатика)

Шифр компетенции	Расшифровка приобретаемой компетенции
ОПК-4 (профессиональные компетенции)	способность находить, анализировать, реализовывать программно и использовать на практике математические алгоритмы, в том числе с применением современных вычислительных систем
ПК-3 (профессиональные компетенции)	способность строго доказать утверждение, сформулировать результат, увидеть следствия полученного результата
ПК-6 (профессиональные компетенции)	способность передавать результаты проведенных физико-математических и прикладных исследований в виде конкретных рекомендаций, выраженных в терминах предметной области изучаемого явления

В результате освоения дисциплины студент:

1. должен знать:

Основные понятия, терминологию анализа данных.

Основные методы исследования и примеры использования анализа данных.

Обучающийся должен знать теоретические основы анализа данных.

2. должен уметь:

Уметь решать типовые задачи и формулировать прикладные задачи в терминах анализа данных.

Уметь строить модели естественных явлений, исследуемых методами анализа данных.

3. должен владеть:

Владеть основными методами исследования, использующими анализ данных.

Владеть методами статистического анализа данных.

Владеть навыками интуитивного подхода к решению нестандартных задач с привлечением анализа данных.

4. должен демонстрировать способность и готовность:

Демонстрировать способность и готовность решать задачи, использующие методы анализа данных.

4. Структура и содержание дисциплины/ модуля

Общая трудоемкость дисциплины составляет 4 зачетных(ые) единиц(ы) 144 часа(ов).

Форма промежуточного контроля дисциплины: зачет в 7 семестре.

Суммарно по дисциплине можно получить 100 баллов, из них текущая работа оценивается в 50 баллов, итоговая форма контроля - в 50 баллов. Минимальное количество для допуска к зачету 28 баллов.

86 баллов и более - "отлично" (отл.);

71-85 баллов - "хорошо" (хор.);

55-70 баллов - "удовлетворительно" (удов.);

54 балла и менее - "неудовлетворительно" (неуд.).

4.1 Структура и содержание аудиторной работы по дисциплине/ модулю

Тематический план дисциплины/модуля

N	Раздел Дисциплины/ Модуля	Семестр	Неделя семестра	Виды и часы аудиторной работы, их трудоемкость (в часах)			Текущие формы контроля
				Лекции	Практические занятия	Лабораторные работы	
N	Раздел Дисциплины/ Модуля	Семестр	Неделя семестра	Виды и часы аудиторной работы, их трудоемкость (в часах)			Текущие формы контроля
				Лекции	Практические занятия	Лабораторные работы	
1.	Тема 1. первичная обработка данных	7		5	2	0	
2.	Тема 2. оценка параметров	7		5	5	0	
3.	Тема 3. проверка гипотез	7		5	6	0	
4.	Тема 4. регрессионный анализ	7		5	5	0	
5.	Тема 5. разбиение на кластеры	7		5	5	0	
6.	Тема 6. дисперсионный анализ	7		5	5	0	
7.	Тема 7. метод главных компонент	7		4	6	0	
4.2 Содержание дисциплины							
	Тема 1. первичная обработка данных			0	0	0	Зачет
лекционное занятие (5 часа(ов)):							
выборка, сгруппированные данные, вариационный ряд, эмпирическая функция распределения, гистограмма, полигон							
практическое занятие (2 часа(ов)):							
выборка, сгруппированные данные, вариационный ряд, эмпирическая функция распределения, гистограмма, полигон							
Тема 2. оценка параметров							
лекционное занятие (5 часа(ов)):							
выборочное среднее, выборочная дисперсия, выборочные моменты, выборочные квантили, выборочная медиана, выборочная корреляция, выборочный коэффициент корреляции, интервальные оценки,							
практическое занятие (5 часа(ов)):							
выборочное среднее, выборочная дисперсия, выборочные моменты, выборочные квантили, выборочная медиана, выборочная корреляция, выборочный коэффициент корреляции, интервальные оценки,							
Тема 3. проверка гипотез							
лекционное занятие (5 часа(ов)):							
одновыборочный критерий Стьюдента, двухвыборочный критерий Стьюдента, критерий Фишера, критерий Пирсона, критерий Кендалла, критерий Спирмена, критерий Вилкоксона, критерий хи-квадрат, критерий совпадения наблюдаемой частоты с гипотетической вероятностью, критерий совпадения параметров биномиальных распределений.							
практическое занятие (6 часа(ов)):							

одновыборочный критерий Стьюдента, двухвыборочный критерий Стьюдента, критерий Фишера, критерий Пирсона, критерий Кендалла, критерий Спирмена, критерий Вилкоксона, критерий хи-квадрат, критерий совпадения наблюдаемой частоты с гипотетической вероятностью, критерий совпадения параметров биномиальных распределений.

Тема 4. регрессионный анализ

лекционное занятие (5 часа(ов)):

МНК-метод, доверительные интервалы для коэффициентов, статистика Стьюдента для коэффициентов, статистика Фишера, тест Чоу, метод апостериори, метод пошаговой селекции, исследование случайных ошибок, ОМНК-метод, ДОМНК-метод.

практическое занятие (5 часа(ов)):

МНК-метод, доверительные интервалы для коэффициентов, статистика Стьюдента для коэффициентов, статистика Фишера, тест Чоу, метод апостериори, метод пошаговой селекции, исследование случайных ошибок, ОМНК-метод, ДОМНК-метод.

Тема 5. разбиение на кластеры

лекционное занятие (5 часа(ов)):

нормировка, расстояние, кластеры, эвристические методы, иерархические методы, быстрые алгоритмы, функционалы качества разбиения, неизвестное число кластеров, сравнение методов, представление результатов.

практическое занятие (5 часа(ов)):

нормировка, расстояние, кластеры, эвристические методы, иерархические методы, быстрые алгоритмы, функционалы качества разбиения, неизвестное число кластеров, сравнение методов, представление результатов.

Тема 6. дисперсионный анализ

лекционное занятие (5 часа(ов)):

однофакторный дисперсионный анализ, двухфакторный дисперсионный анализ, многофакторный дисперсионный анализ, модели со случайными факторами.

практическое занятие (5 часа(ов)):

однофакторный дисперсионный анализ, двухфакторный дисперсионный анализ, многофакторный дисперсионный анализ, модели со случайными факторами.

Тема 7. метод главных компонент

лекционное занятие (4 часа(ов)):

геометрия главных компонент, эллипсоид рассеяния, вычисление главных компонент, шкалирование, ранговая корреляция, множественная и частная корреляция, таблицы сопряженности.

практическое занятие (6 часа(ов)):

геометрия главных компонент, эллипсоид рассеяния, вычисление главных компонент, шкалирование, ранговая корреляция, множественная и частная корреляция, таблицы сопряженности.

4.3 Структура и содержание самостоятельной работы дисциплины (модуля)

N	Раздел Дисциплины	Семестр	Неделя семестра	Виды самостоятельной работы студентов	Трудоемкость (в часах)	Формы контроля самостоятельной работы
1.	Тема 1. первичная обработка данных	7		написание программы для ЭВМ	10	проверка письменного домашнего задания
2.	Тема 2. оценка параметров	7		написание программы для ЭВМ	16	проверка письменного домашнего задания

N	Раздел Дисциплины	Семестр	Неделя семестра	Виды самостоятельной работы студентов	Трудоемкость (в часах)	Формы контроля самостоятельной работы
3.	Тема 3. проверка гипотез	7		написание программы для ЭВМ	9	проверка письменного домашнего задания
4.	Тема 4. регрессионный анализ	7		написание программы для ЭВМ	9	проверка письменного домашнего задания
5.	Тема 5. разбиение на кластеры	7		написание программы для ЭВМ	10	проверка письменного домашнего задания
6.	Тема 6. дисперсионный анализ	7		написание программы для ЭВМ	12	проверка письменного домашнего задания
7.	Тема 7. метод главных компонент	7		написание программы для ЭВМ	10	проверка письменного домашнего задания
	Итого				76	

5. Образовательные технологии, включая интерактивные формы обучения

предусмотрена встреча с представителями компании SAP и 1С.

6. Оценочные средства для текущего контроля успеваемости, промежуточной аттестации по итогам освоения дисциплины и учебно-методическое обеспечение самостоятельной работы студентов

Тема 1. первичная обработка данных

проверка письменного домашнего задания, примерные вопросы:

Написать программу на Python с использованием библиотеки pandas для объединения множественных файлов в один.

Тема 2. оценка параметров

проверка письменного домашнего задания, примерные вопросы:

Написать программу на Python с использованием библиотеки pandas и sklearn для оценки различных параметров данных, как-то: наиболее частотных слов, средних значений в столбцах, столбцах с наибольшим количеством пропущенных значений.

Тема 3. проверка гипотез

проверка письменного домашнего задания, примерные вопросы:

Написать программу на Python с использованием библиотеки pandas и sklearn для проверки гипотез о характере распределения различных данных.

Тема 4. регрессионный анализ

проверка письменного домашнего задания, примерные вопросы:

Написать программу на Python с использованием библиотеки pandas и sklearn для построения многомерной полиномиальной регрессии по данным.

Тема 5. разбиение на кластеры

проверка письменного домашнего задания, примерные вопросы:

Написать программу на Python с использованием библиотеки pandas и sklearn для построения кластеризации методами k-mean, kNN и ForEI

Тема 6. дисперсионный анализ

проверка письменного домашнего задания, примерные вопросы:

Написать программу на Python с использованием библиотеки pandas и sklearn для построения многомерных дисперсионных характеристик

Тема 7. метод главных компонент

проверка письменного домашнего задания, примерные вопросы:

Написать программу на Python с использованием библиотеки pandas и sklearn для выделения главных компонент многомерных выборок.

Итоговая форма контроля

зачет (в 7 семестре)

Примерные вопросы к зачету:

1. Опишите процесс работы и докажите сходимость алгоритма kNN
2. Представьте вывод решения задачи многомерной регрессии
3. Докажите существование SVD-разложения

7.1. Основная литература:

Теория вероятностей, Ч. 2. Случайные величины, , 2013г.

Теория вероятностей, Ч. 1. Элементарная теория вероятностей, , 2013г.

Технология Data Mining: Интеллектуальный анализ данных, Степанов, Роман Григорьевич, 2009г.

7.2. Дополнительная литература:

Линейная алгебра и геометрия, Кострикин, Алексей Иванович;Манин, Ю.И., 2005г.

Вычислительная линейная алгебра, Вержбицкий, Валентин Михайлович, 2009г.

Аналитическая геометрия и линейная алгебра, Гусак, Алексей Адамович, 2011г.

Теория вероятностей и математическая статистика, Гмурман, Владимир Ефимович, 2004г.

7.3. Интернет-ресурсы:

краткий справочник по математике - <http://mathembook.ru>

платформа MOOK - stepik.org

портал математических интернет-ресурсов - <http://almath.com>

портал математических интернет-ресурсов - <http://math.ru>

сайт учебных материалов по математическим наукам - <http://www.exponenta.ru>

8. Материально-техническое обеспечение дисциплины(модуля)

Освоение дисциплины "Статистический анализ данных с использованием компьютерных пакетов" предполагает использование следующего материально-технического обеспечения:

Мультимедийная аудитория, вместимостью более 60 человек. Мультимедийная аудитория состоит из интегрированных инженерных систем с единой системой управления, оснащенная современными средствами воспроизведения и визуализации любой видео и аудио информации, получения и передачи электронных документов. Типовая комплектация мультимедийной аудитории состоит из: мультимедийного проектора, автоматизированного проекционного экрана, акустической системы, а также интерактивной трибуны преподавателя, включающей тач-скрин монитор с диагональю не менее 22 дюймов, персональный компьютер (с техническими характеристиками не ниже Intel Core i3-2100, DDR3 4096Mb, 500Gb), конференц-микрофон, беспроводной микрофон, блок управления оборудованием, интерфейсы подключения: USB, audio, HDMI. Интерактивная трибуна преподавателя является ключевым элементом управления, объединяющим все устройства в единую систему, и служит полноценным рабочим местом преподавателя. Преподаватель имеет возможность легко управлять всей системой, не отходя от трибуны, что позволяет проводить лекции, практические занятия, презентации, вебинары, конференции и другие виды аудиторной нагрузки обучающихся в удобной и доступной для них форме с применением современных интерактивных средств обучения, в том числе с использованием в процессе обучения всех корпоративных ресурсов. Мультимедийная аудитория также оснащена широкополосным доступом в сеть интернет. Компьютерное оборудование имеет соответствующее лицензионное программное обеспечение.

Компьютерный класс, представляющий собой рабочее место преподавателя и не менее 15 рабочих мест студентов, включающих компьютерный стол, стул, персональный компьютер, лицензионное программное обеспечение. Каждый компьютер имеет широкополосный доступ в сеть Интернет. Все компьютеры подключены к корпоративной компьютерной сети КФУ и находятся в едином домене.

Необходимо, чтобы на компьютерах был установлен интерпретатор Python и R.

Программа составлена в соответствии с требованиями ФГОС ВПО и учебным планом по направлению 02.03.01 "Математика и компьютерные науки" и профилю подготовки Математическое и компьютерное моделирование .

Автор(ы):

Новиков А.А. _____

"__" _____ 201__ г.

Рецензент(ы):

Насыров С.Р. _____

"__" _____ 201__ г.