МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное учреждение
высшего профессионального образования
"Казанский (Приволжский) федеральный университет"
Институт вычислительной математики и информационных технологий

УТВЕРЖДАЮ
Проректор
по образовательной деятельности КФУ
Проф. Минзарипов Р.Г.
_____
"___"_____20___ г.

Программа дисциплины

Программное обеспечение анализа данных М1.В.3

Направление подготовки: 010400.68 - Прикладная математика и информатика
Профиль подготовки: Анализ данных и его приложения
Квалификация выпускника: магистр
Форма обучения: очное
Язык обучения: русский
Автор(ы):
Григорьева И.С.
Рецензент(ы):
Миссаров М.Д.

СОГЛАСОВАНО:
Заведующий(ая) кафедрой: Турилова Е. А.
Протокол заседания кафедры No ___ от "____" _____ 201__г
Учебно-методическая комиссия Института вычислительной математики и информационных технологий:
Протокол заседания УМК No ____ от "____" _____ 201__г

Регистрационный No

Казань
2014

Содержание

Программу дисциплины разработал(а)(и) доцент, к.н. (доцент) Григорьева И.С. кафедра математической статистики отделение прикладной математики и информатики , Irina.Grigorieva@kpfu.ru

1. Цели освоения дисциплины

Современный анализ данных невозможен без привлечения программных средств об-работки. Курс призван познакомить с некоторыми комплексами таких программ, а также с задачами, которые они решают. Курс содержит как теоретическую, так и практическую часть, состоящую в решении конкретных задач обработки данных с помощью статистиче-ского языка программирования R.

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

Данная учебная дисциплина включена в раздел " М1.В.3 Общенаучный" основной образовательной программы 010400.68 Прикладная математика и информатика и относится к вариативной части. Осваивается на 1 курсе, 1 семестр.

Современный анализ данных невозможен без привлечения программных средств обработки. Курс призван познакомить с некоторыми комплексами таких программ, а также с задачами, которые они решают. Курс содержит как теоретическую, так и практическую часть, состоящую в решении конкретных задач обработки данных с помощью статистического языка программирования R.

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

В результате освоения дисциплины формируются следующие компетенции:

Шифр компетенции	Расшифровка приобретаемой компетенции
ОК-1 (общекультурные компетенции)	Способность понмать философские концепции естествознания, владеть основами методологии научного познания при изучении различных уровней организации материи, пространства и времени
ОК-2 (общекультурные компетенции)	Способность иметь представление о современном состоянии и проблемах прикладной математики информатики, истории и методологии развития
ОК-3 (общекультурные компетенции)	Способность использовать углубленные теоретические и практические знания в области прикладной математики информатики
ОК-4 (общекультурные компетенции)	Способность самостоятельно приобретать с помощью информационных технологий и использовать в практической деятельности, расширять и углублять свое научное мировоззрение.
ОК-5 (общекультурные компетенции)	Способность порождать новые идеи и демонстрировать навыки самостоятельной научно-исследовательской работы и работы в научном коллективе
ПК-1 (профессиональные компетенции)	Способность проводить научные исследования и получать новые научные и прикладные результаты
ПК-2 (профессиональные компетенции)	Способность разрабатывать концептуальные и теоретические модели решаемых проблем

Шифр компетенции	Расшифровка приобретаемой компетенции
ПК-3 (профессиональные компетенции)	Способность углубленного анализа проблем, постановки и обоснования задач научной и проектно-технологической деятельности

В результате освоения дисциплины студент:

1. должен знать:

- перечень основных задач обработки и анализа данных;
- основные методы обработки и анализа;
- проблемы, возникающие при подборе метода;
- возможности, предоставляемые пакетами прикладных программ для анализа данных;

2. должен уметь:

- формулировать (ставить) задачу обработки и анализа данных;
- подбирать методы обработки в соответствии с поставленной задачей;
- составлять программы для обработки данных на языке R;
- анализировать результаты компьютерной обработки данных.

3. должен владеть:

- навыками использования языка R, векторного подхода, командной строки;
- навыками поиска в интернет обновлений и расширений языка, подключения их к работе.

4. должен демонстрировать способность и готовность:

- выбирать методы анализа данных и подходящее программное обеспечение
- следить за развитием существующих средств анализа данных, а также за появлением новых.

4. Структура и содержание дисциплины/ модуля

Общая трудоемкость дисциплины составляет зачетных(ые) единиц(ы) 108 часа(ов).

Форма промежуточного контроля дисциплины зачет в 1 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. Введение.	1	1	2	0	0	устный опрос
2.	Тема 2. Особенности статистического языка R.	1	2-4	2	4	0	домашнее задание
3.	Тема 3. Первичная обработка данных.	1	4-6	2	4	0	домашнее задание
4.	Тема 4. Поиск взаимосвязей. Линейные модели.	1	7-9	2	4	0	устный опрос
5.	Тема 5. Проверка статистических гипотез. Библиотека проверки нормальности.	1	9-12	2	6	0	контрольная работа
6.	Тема 6. Построение отчетов по результатам статистической обработки.	1	12-14	2	4	0	творческое задание
7.	Тема 7. Кластерный анализ и задачи классификации.	1	15-17	2	6	0	устный опрос
.	Тема . Итоговая форма контроля	1		0	0	0	зачет
	Итого			14	28	0	

4.2 Содержание дисциплины

Тема 1. Введение.

лекционное занятие (2 часа(ов)):

Класс задач обработки данных. ППП обработки и анализа данных. Проблемы, возникающие при постановке задач и подборе методов обработки. Повторение основ статистики

Тема 2. Особенности статистического языка R.

лекционное занятие (2 часа(ов)):

Язык командной строки. Принципы векторных вычислений. Типы данных и объекты языка R, Векторы, факторы, таблицы, списки. Обращение к их компонентам. Ввод и вывод данных и результатов исследования

практическое занятие (4 часа(ов)):

Освоение консоли и командной строки. Самостоятельное создание объектов языка R, работа с ними. Работа со встроенными данными. Ввод данных из внешних источников (Excel-таблицы с расширением .csv, текстовые таблицы). Вывод результатов расчетов на экран и в файлы.

Тема 3. Первичная обработка данных.

лекционное занятие (2 часа(ов)):

Проверка полноты и достоверности данных (выбросы, ошибки при наборе). Вычисление основных статистических характеристик. Простейшие способы визуализации данных.

практическое занятие (4 часа(ов)):

Получение навыка первичной обработки и поиска ошибок в данных (объектах) средствами языка R. Первичная визуализация (скаттерплоты, боксплоты, барплоты для данных разного типа) Вычисление основных статистических параметров для объектов разного типа (применение операторов группы apply). Построение гистограмм. Шкалирование (scale) и группировка (cut) данных.

Тема 4. Поиск взаимосвязей. Линейные модели.

лекционное занятие (2 часа(ов)):

Корреляция и регрессия. Линейная регрессия. Метод главных компонент. Визуализация данных.

практическое занятие (4 часа(ов)):

Операторы plot, lines, points, text и т.п., их варианты для разных типов данных. Оператор lm() построения линейных моделей. Построение линий регрессии. Применение метода главных компонент для визуализации данных большой размерности.

Тема 5. Проверка статистических гипотез. Библиотека проверки нормальности.

лекционное занятие (2 часа(ов)):

Команды проверки гипотез, их варианты. Гипотезы значимости, однородности и согласия. Использование критического уровня значимости (p-value). R как язык программирования.

практическое занятие (6 часа(ов)):

Команды типа .test (t.test, var.test, wilcox.test, oneway.test, cor.test и другие). Состав выводимой информации. Подключение библиотеки проверки нормальности. Операторы циклов и условные операторы. Создание и использование пользовательских функций. Использование линейных моделей (продолжение изучения команды lm), а также дисперсионного анализа и других методов, использующих полученные модели.

Тема 6. Построение отчетов по результатам статистической обработки.

лекционное занятие (2 часа(ов)):

Вывод графики и текстовой информации в файлы разного типа. Анализ результатов. Оптимизация программных кодов.

практическое занятие (4 часа(ов)):

Операторы управления потоками текстового и графического вывода: - sink("Файл"), sink() - pdf, jpg, png, dev.off() и другими Операторы вывода текста (print, cat, write, ...), управление графическими окнами. Создание отчетов средствами языка R.

Тема 7. Кластерный анализ и задачи классификации.

лекционное занятие (2 часа(ов)):

Постановка задачи разбиения. Различные матрицы расстояний. Методы кластеризации (k-means, иерархическая кластеризация и т.п.).

практическое занятие (6 часа(ов)):

Оператор построения матрицы расстояний dist(). Метод k-means (оператор kmeans). Исследование устойчивости результатов кластеризации. Визуализация результатов. Иерархическая кластеризация (hclust), ее визуализация с помощью дендрограммы. Выделение классов на основе иерархической кластеризации (оператор cutree). Исследование поведения метода в зависимост от способа объекдинения кластеров (метод ближайшего соседа, полной связи, Варда и т.п.)

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. Введение.	1	1	подготовка к устному опросу	8	устный опрос

N	Раздел Дисциплины	Семестр	Неделя семестра	Виды самостоятельной работы студентов	Трудоемкость (в часах)	Формы контроля самостоятельной работы
2.	Тема 2. Особенности статистического языка R.	1	2-4	подготовка домашнего задания	8	домашнее задание
3.	Тема 3. Первичная обработка данных.	1	4-6	подготовка домашнего задания	8	домашнее задание
4.	Тема 4. Поиск взаимосвязей. Линейные модели.	1	7-9	подготовка к устному опросу	8	устный опрос
5.	Тема 5. Проверка статистических гипотез. Библиотека проверки нормальности.	1	9-12	Подбор данных для анализа	4	устный опрос
				подготовка к контрольной работе	8	контрольная работа
6.	Тема 6. Построение отчетов по результатам статистической обработки.	1	12-14	Подбор данных для творческого анализа	4	устный опрос
				подготовка к творческому заданию	8	творческое задание
7.	Тема 7. Кластерный анализ и задачи классификации.	1	15-17	подготовка к устному опросу	10	устный опрос
	Итого				66	

5. Образовательные технологии, включая интерактивные формы обучения

Дисциплина представляет собой цикл лекционных и лабораторных занятий. Лабора-торные занятия посвящены выработке базовых навыков создания и использования про-грамм на языке программирования R для решения различных задач обработки данных. Практические занятия проходят в компьютерных классах. Практические занятия проходят в интерактивной форме обсуждения решения различных задач или в активной форме са-мостоятельного решения задач студентами. Контроль за выполнением самостоятельной работы проявляется в функциональном тестировании выполненных студентами заданий на примерах, предложенных преподавателем.

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

Тема 1. Введение.

устный опрос , примерные вопросы:

Описать задачи анализа данных, реализованные в ППП. Подобрать примеры из других дисциплин курса.

Тема 2. Особенности статистического языка R.

домашнее задание , примерные вопросы:

Изучить особенности языка R, интерфейс командной строки, векторизованные вычисления. Научиться выполнять простейшие действия на R. Записать результаты работы в виде скриптов (сценариев).

Тема 3. Первичная обработка данных.

домашнее задание , примерные вопросы:

Повторить понятия из математической стаистики. Случайные величины, законы распределения, параметры законов распределения. Выборочные оценки параметров. Написать скрипты для вычисления параметров.

Тема 4. Поиск взаимосвязей. Линейные модели.

устный опрос , примерные вопросы:

Знать понятия "корреляция, линия регрессии". Идею метода главных компонент.

Тема 5. Проверка статистических гипотез. Библиотека проверки нормальности.

контрольная работа , примерные вопросы:

Создание скриптов для ввода и первичной обработки данных, функций для вызова различных тестов (команд проверки гипотез)

устный опрос , примерные вопросы:

Подобрать в интернете данные, пригодные для последующего статистического анализа.

Тема 6. Построение отчетов по результатам статистической обработки.

творческое задание , примерные вопросы:

Провести статистическую обработку учебных файлов данных. Каждый файл содержит сведения об объектах: численные показатели, категориальные показатели. Результат представить в виде скрипта (сценария) и отчета по результатам обработки.

устный опрос , примерные вопросы:

Подобрать в интернете данные, пригодные для последующего кластерного анализа.

Тема 7. Кластерный анализ и задачи классификации.

устный опрос , примерные вопросы:

Провести кластеризацию данных учебных файлов разными методами. Визуализировать полученные результаты. Результат представить в виде скрипта (сценария) и отчета по результатам обработки.

Тема . Итоговая форма контроля

Примерные вопросы к зачету:

Применение скриптов, созданных студентом в течение семестра, для тестовых файлов данных.

7.1. Основная литература:

Основы статистической обработки, Салимов, Фарид Ибрагимович, 2010г.

Статистический анализ данных в экологии и природопользовании с использованием программы STATGRAPHICS Plus, Мальцев, Кирилл Александрович;Мухарамова, Светлана Саясовна, 2011г.

Наглядная статистика. Используем R!, Шипунов, Алексей Борисович;Балдин, Евгений Михайлович;Волкова, Полина Андреевна, 2012г.

Многомерный анализ данных методами прикладной статистики, Барковский, Станислав Станиславович;Захаров, Вячеслав Михайлович;Лукашов, Андрей Михайлович, 2010г.

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

Обработка и анализ данных социологических исследований в пакете SPSS 17.0, Фарахутдинов, Шамиль Фаритович;Бушуев, Алексей Сергеевич, 2011г.

Геостатистический анализ данных в экологии и природопользовании (с применением пакета R), Савельев, Анатолий Александрович, 2012г.

7.2. Дополнительная литература:

Многомерный анализ данных в нефтяной геохимии и химии окружающей среды, Туров, Ю. П.;Гузняева, М. Ю., 2010г.

Экспертный анализ данных в молекулярной фармакологии, Торшин, Иван Юрьевич;Громова, Ольга Алексеевна, 2012г.

Интеллектуальный анализ данных для поддержки принятия решений , Ризаев, Ильдус Султанович;Рахал, Ясер, 2011г.

7.3. Интернет-ресурсы:

Д.Мертц, Б.Хантинг. Статистическое программирование на R: Часть 1. Купаемся в изобилии статистических возможностей - http://www.ibm.com/developerworks/ru/library/l-r1/

Д.Мертц, Б.Хантинг. Статистическое программирование на R. Часть 2. Функциональное програм-мирование и анализ данных. - http://www.ibm.com/developerworks/ru/library/l-r2/

Д.Мертц, Б.Хантинг. Статистическое программирование на R. Часть 3. Повторное использование кода и объектное программирование. - http://www.ibm.com/developerworks/ru/library/l-r3/

Сайт Евгения Балдина - http://www.inp.nsk.su/~baldin/DataAnalysis/index.html

Сайт проекта R - http://www.r-project.org/

8. Материально-техническое обеспечение дисциплины(модуля)

Освоение дисциплины "Программное обеспечение анализа данных" предполагает использование следующего материально-технического обеспечения:

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.

Для проведения занятий по курсу требуется компьютерный класс с выходом в интернет.

Необходимо также иметь возможность подключения пакетов программ (права администратора)

Программа составлена в соответствии с требованиями ФГОС ВПО и учебным планом по направлению 010400.68 "Прикладная математика и информатика" и магистерской программе Анализ данных и его приложения .

Автор(ы):
Григорьева И.С. _____
"__" _____ 201 __ г.

Рецензент(ы):
Миссаров М.Д. _____
"__" _____ 201 __ г.