Белорусский государственный университет

УТВЕРЖДАЮ
Проректор по учебной работе

_____ А.Л. Толстик

_____

Регистрационный № УД-1592 / уч.

МНОГОМЕРНЫЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ

Учебная программа учреждения высшего образования по учебной дисциплине для специальности:

1-31 03 03 Прикладная математика (по направлениям)

2015 г.

Учебная программа составлена на основе образовательного стандарта высшего образования ОСВО 1-31 03 03-2013 и учебного плана G31-173/уч., 30.05.2013.

Составители:

Е. Н. Орлова, доцент кафедры математического моделирования и анализа данных Белорусского государственного университета, кандидат физико-математических наук, доцент;

Ю.С. Харин, заведующий кафедрой математического моделирования и анализа данных Белорусского государственного университета, доктор физико-математических наук, чл.-корр. НАНБ, профессор.

Рекомендована к утверждению:

Кафедрой математического моделирования и анализа данных Белорусского государственного университета (протокол № 19 от 07 апреля 2015 г.);

Учебно-методической комиссией факультета прикладной математики и информатики Белорусского государственного университета (протокол № 6 от 12 мая 2015 г.).

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Многомерный статистический анализ данных – это раздел математики, который занимается изучением методов сбора и предварительного анализа многомерных статистических данных, их систематизации и обработки с целью выявления характера и структуры взаимосвязей между признаками, присущими исследуемому объекту или системе. Методы многомерного статистического анализа данных активно применяются в технических исследованиях, экономике, теории и практике управления, социологии, медицине и т.д. С результатами наблюдений, измерений, испытаний, опытов, с их анализом имеют дело специалисты во всех отраслях практической деятельности, почти во всех областях теоретических исследований.

В основе многомерного статистического анализа лежит принцип: данные имеют вероятностную (стохастическую) природу. Поэтому для их описания и анализа используются вероятностно-статистические модели и методы.

Развитие вычислительной техники и программного обеспечения способствует широкому внедрению методов многомерного статистического анализа в практику. Пакеты прикладных программ (ППП) с удобным пользовательским интерфейсом, такие как IBM Statistics, STATISTICA, а также современные технологии компьютерного анализа данных с использованием и системы R, снимают трудности в применении указанных методов, заключающиеся в сложности математического аппарата, опирающегося на линейную алгебру, теорию вероятностей и математическую статистику, и громоздкости вычислений.

Учебная дисциплина «Многомерный статистический анализ данных» знакомит студентов с классическими и современными методами решения задач статистического анализа данных, к которым относятся описание данных и их первичная статистическая обработка, оценивание неизвестных характеристик и параметров, исследование статистических свойств полученных оценок, исследование взаимосвязей между показателями, проверка статистических гипотез.

Целью дисциплины «Многомерный статистический анализ данных» является:

– расширение фундаментального математического образования;

– формирование у студентов профессиональных знаний в предметной области, подкрепленных владением математическими методами и пакетами прикладных программ.

Задачами дисциплины «Многомерный статистический анализ данных» являются:

– систематическое изучение теоретических основ многомерного статистического анализа данных;

– приобретение практических навыков для правильного использования арсенала методов статистического анализа данных в конкретных ситуациях.

В первом разделе дисциплины «Многомерный статистический анализ данных» приводятся сведения о математических моделей данных, функциональных и числовых характеристиках многомерных вероятностных моделей данных, а также о методах их оценивания. Формулируется проблема сжатия данных и приводится критерий выбора информативных признаков.

Во втором разделе изучаются свойства многомерного нормального распределения, которое является удобной математической моделью для изложения методов многомерного статистического анализа. Приводятся числовые и функциональные характеристики этого распределения. Даются определения парного, частного, множественного коэффициентов корреляции, а также вводится понятие функции регрессии. Вычисляются оценки максимального правдоподобия вектора математического ожидания и ковариационной матрицы.

Третий раздел посвящен статистическому исследованию зависимостей. Определяются выборочные парный, частный и множественный коэффициенты корреляции. Приводятся критерии для проверки статистических гипотез о значениях парного, частного и множественного коэффициентов корреляции, а также критерий отношения правдоподобия для проверки гипотезы о независимости компонент гауссовского случайного вектора.

В четвертой главе рассматриваются задачи проверки различных статистических гипотез относительно параметров многомерного нормального распределения задачи проверки статистических гипотез относительно параметров многомерной линейной регрессии. Приводятся также сведения о таких разделах многомерного статистического анализа, как дискриминантный и дисперсионный анализ.

Учебная дисциплина «Многомерный статистический анализ данных» относится к циклу дисциплин вузовского компонента и взаимосвязана с учебной дисциплиной «Теория вероятностей и математическая статистика». Методы, излагаемые в дисциплине «Многомерный статистический анализ данных», могут быть использованы при изучении ряда других дисциплин по специальности «Прикладная математика»

В результате изучения дисциплины студент должен

***знать*:**

- методы статистического оценивания многомерных распределений;
- методы предварительного статистического анализа данных;
- методы корреляционного анализа;
- методы регрессионного анализа данных;
- методы статистической проверки гипотез;

уметь:

– осуществлять предварительный статистический анализ данных с целью установления модели данных, выявления кластерной структуры данных и аномальных наблюдений;

– проводить статистический анализ многомерных данных с целью установления статистических зависимостей методами корреляционного, регрессионного и дисперсионного анализа между переменными;

– осуществлять статистический анализ однородности многомерных данных на основе графического анализа и статистических критериев;

– осуществлять классификацию неоднородных данных с помощью методов дискриминантного и кластерного анализа.

владеть:

– методами решения основных задач статистического анализа многомерных данных;

– навыками по подготовке данных и решения типовых задач статистического анализа данных;

– навыками применения современных ППП для решения задач статистического анализа многомерных данных в различных приложениях;

– навыками по подготовке отчетов с результатами статистического анализа данных, включающих содержательную интерпретацию результатов анализа, комментарии, выводы и рекомендации.

В соответствии с образовательным стандартом специальности 1-31 03 03 «Прикладная математика» (по направлениям) учебная программа предусматривает для изучения дисциплины всего 164 часа, из них 68 аудиторных часа, в том числе лекций – 34 часа, лабораторных занятий – 26 часов, УСР – 8 часов (3 курс, 5 семестр).

Форма текущей аттестации по учебной дисциплине – экзамен.

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Раздел I. Разведочный анализ и сжатие данных

Тема 1.1. Актуальность задач многомерного статистического анализа данных (МСАД). Принципы и общая схема многомерного статистического анализа. Типы данных. Шкалы измерения. Классификация задач и методов МСАД.

Тема 1.2. Предварительный (первичный) статистический анализ данных. Математические модели данных. Наблюдения как случайные величины, векторы, функции. Модель данных «случайная выборка». Альтернативные модели данных: неоднородная выборка, выборка с засорениями, зависимые наблюдения. Примеры.

Тема 1.3. Функциональные и числовые характеристики многомерных вероятностных моделей данных. Функциональные характеристики вероятностных моделей данных: распределение вероятностей, функция распределения, плотность распределения вероятностей, характеристическая функция. Числовые характеристики вероятностных моделей данных: характеристики положения, характеристики рассеяния, характеристики формы. Ковариация. Коэффициент корреляции. Ковариационная и корреляционная матрицы.

Тема 1.4. Статистическое оценивание функциональных и числовых характеристик вероятностных моделей данных. Параметрический и непараметрический подходы к статистическому оцениванию характеристик вероятностных моделей данных. Эмпирическая функция распределения. Гистограмма.

Тема 1.5. Проблема сжатия данных. Метод главных компонент. Постановка задачи сжатия данных. Главные компоненты и их вероятностные свойства. Критерий выбора информативных признаков в пространстве главных компонент.

Раздел II. Многомерное нормальное распределение и оценивание его параметров

Тема 2.1. Многомерное нормальное (гауссовское) распределение как модель многомерных данных. Многомерное нормальное распределение и его основные свойства. Функциональные и числовые характеристики, маргинальные распределения, линейные преобразования гауссовских случайных векторов.

Тема 2.2. Условное распределение гауссовского вектора. Функция регрессии и ее оптимальные свойства. Частный и множественный коэффициенты корреляции: определение и свойства. Прогнозирование. Функция регрессии.

Тема 2.3. Статистическое оценивание параметров многомерной гауссовской модели. Оценивание параметров многомерной гауссовской модели по методу максимального правдоподобия. Выборочное среднее и выбо-

рочная ковариационная матрица. Вероятностные свойства выборочного среднего и выборочной ковариационной матрицы. Несмещенная выборочная ковариационная матрица.

Раздел III. Статистическое исследование зависимостей

Тема 3.1. Исследование парной зависимости. Выборочный парный коэффициент корреляции: свойства и применения. Z-статистика Фишера и проверка гипотез о значении коэффициента корреляции.

Тема 3.2. Выборочные частный и множественный коэффициенты корреляции. Выборочный частный коэффициент корреляции: определение, свойства и применения. Выборочный множественный коэффициент корреляции: определение, свойства и применения. Статистические выводы о значениях частного и множественного коэффициентов корреляции.

Тема 3.3. Проверка общих гипотез о независимости. Обобщенная статистика отношения правдоподобия. Критерий отношения правдоподобия для проверки гипотезы о независимости компонент гауссовского случайного вектора.

Раздел IV. Проверка гипотез и статистическая классификация

Тема 4.1. Статистическая проверка гипотез на основе T^2-статистики Хоттелинга. Проверка гипотез о значении вектора математического ожидания. T^2-статистика Хоттелинга. Сравнение векторов математических ожиданий по двум выборкам. Многомерная проблема Беренса-Фишера.

Тема 4.2. Проверка гипотез относительно параметров многомерного нормального распределения. Проверка гипотез о значении вектора математического ожидания. Проверка гипотез о значении ковариационной матрицы. Проверка гипотез о совпадении многомерного нормального распределения с наперед заданным многомерным нормальным распределением.

Тема 4.3. Проверка гипотез относительно нескольких выборок из многомерных нормальных распределений. Гипотеза о совпадении векторов математических ожиданий при неизвестной одинаковой ковариационной матрице. Гипотеза о равенстве ковариационных матриц. Гипотеза однородности.

Тема 4.4. Регрессионный анализ данных. Регрессионная модель данных. Статистическое оценивание параметров многомерной линейной регрессии. Свойства оценок параметров. Проверка гипотез относительно параметров модели многомерной линейной регрессии.

Тема 4.5. Статистическая классификация многомерных нормальных распределений. Дискриминантный анализ данных при наличии обучающей выборки. Подстановочное байесовское решающее правило. Оптимальная классификация гауссовских случайных векторов. Обзор методов статистической классификации в условиях неполных данных и непараметрической неопределенности (кластер-анализ данных и непараметрические классификаторы).

Тема 4.6. Дисперсионный анализ. Математическая модель и постановка задачи дисперсионного анализа данных. Методы дисперсионного анализа данных. Статистическое оценивание параметров. Проверка гипотез относительно параметров модели. Таблица дисперсионного анализа и ее интерпретация.

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

№п/п	Название раздела, темы	Количество часов				Количество часов УСР	Форма контроля знаний
		Аудиторные					
		Лекции	Практ. и сем. занятия	Лаб. занятия	Иное		
1	**Разведочный анализ и сжатие данных**	**10**		**6**		**2**	
1.1	Актуальность задач статистического анализа данных.	2		2			Опрос на занятии
1.2	Предварительный (первичный) статистический анализ данных	2		2			Опрос на занятии
1.3	Функциональные и числовые характеристики многомерных вероятностных моделей данных	2		2			Опрос на занятии
1.4	Статистическое оценивание функциональных и числовых характеристик вероятностных моделей данных	2				2	Контрольная работа №1
1.5	Проблема сжатия данных. Метод главных компонент	2		2			Опрос на занятии
2	**Многомерное нормальное распределение и оценивание его параметров**	**6**		**4**		**2**	
2.1	Многомерное нормальное (гауссовское) распределение как модель многомерных данных	2				2	Тест
2.2	Условное распределение гауссовского вектора	2		2			Опрос на занятии
2.3	Статистическое оценивание параметров многомерной гауссовской модели	2		2			Опрос на занятии
3	**Статистическое исследование зависимостей**	**6**		**4**		**2**	
3.1	Исследование парной зависимости	2				2	Контрольная работа №2

3.2	Выборочные частный и множественный коэффициенты корреляции	2		2			Опрос на занятии
3.3	Проверка общих гипотез о независимости	2		2			Опрос на занятии
4	**Проверка гипотез и статистическая классификация**	**12**		**10**		**2**	
4.1	Статистическая проверка гипотез на основе T^2-статистики Хоттелинга	2				2	Тест
4.2	Проверка гипотез относительно параметров многомерного нормального распределения	2		2			Опрос на занятии
4.3	Проверка гипотез относительно нескольких выборок из многомерных нормальных распределений	2		2			Опрос на занятии
4.4	Регрессионный анализ данных	2		2			Опрос на занятии
4.5	Статистическая классификация многомерных нормальных распределений	2		2			Опрос на занятии
4.6	Дисперсионный анализ	2		2			Опрос на занятии
ИТОГО		**34**		**26**		**8**	

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Рекомендуемая литература

Основная

1. Андерсон Т.В. Введение в многомерный статистический анализ. М., Физматгиз,1963.
2. Андерсон Т.В. Статистический анализ временных рядов. М.: Мир, 1976.
3. Кендалл М.Дж., Стьюарт А. Многомерный статистический анализ и временные ряды. М.: Мир, 1976.
4. Харин Ю.С., Жук Е.Е. Математическая и прикладная статистика. Мн.: БГУ, 2005.
5. Бендат Дж., Пирсол А. Прикладной анализ случайных данных. Москва: Мир 1989.
6. Бриллинджер Д. Временные ряды. М.: Мир, 1990.
7. Харин Ю.С., Малюгин В.И., Абрамович М.С. Математические и компьютерные основы статистического моделирования и анализа данных. Мн.: БГУ, 2008.
8. Харин Ю.С., Абрамович М.С., Малюгин В.И. Компьютерный учебник по статистике. Мн.: БГУ, 2001.
9. Ли Ц., Джадж Д., Зельнер А. Оценивание параметров марковских моделей по агрегированным временным рядам. М.: Статистика, 1977.
10. Харин Ю.С., Степанова М.Д. Практикум на ЭВМ по математической статистике. Мн.: Университетское, 1987.

Дополнительная

1. Болч Б., Хуань К.Дж. Многомерные статистические методы для экономики. М.: Мир, 1979.
2. Афифи С.А., Эйзен. Статистический анализ. Подход с использованием ЭВМ. М.: Мир, 1982.
3. Айвазян С.А. и др. Прикладная статистика. Основы моделирования и первичная обработка данных. М.: Финансы и статистика, 1983.
4. Айвазян С.А. и др. Прикладная статистика. Исследование зависимостей. М.: Финансы и статистика, 1985.
5. Айвазян С.А. и др. Прикладная статистика. Классификация и снижение размерности. М.: Финансы и статистика, 1989.

6. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики, т. 1. Теория вероятностей и прикладная статистика, М.: ЮНИТИ, 2001.
7. Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы. М.: Финансы и статистика, 1998.
8. Харин Ю.С., Степанова М.Д. Практикум на ЭВМ по математической статистике. - Мн.: Университетское, 1987.
9. Харин Ю.С. и др. Математические основы криптологии. – Мн.: БГУ, 1999
10. Харин Ю.С., Агиевич С.В. Компьютерный практикум по математическим методам защиты информации. – Мн.: БГУ, 2001.
11. Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере. М.: ИНФРА-М, 1995.
12. Боровиков В.П. STATISTICA.Статистический анализ и обработка данных в среде Windows.М.: Филинъ, 1997.
13. Многомерный статистический анализ в экономике / Под ред. В.Н. Тамашевича. М.: ЮНИТИ, 1999.
14. Кнут Д. Искусство программирования на ЭВМ. Т.2: Получисленные алгоритмы. - М.: Мир, 1977.
15. Ивченко Г.И., Медведев Ю.И. Введение в математическую статистику. – М.:URSS, 2009.
16. Ивченко Г.И., Глибоченко А.Ф., Иванов В.А., Медведев Ю.И. Статистический анализ дискретных последовательностей. – М.: МИЭМ, 1994.

Организация управляемой самостоятельной работы студентов

Управляемая самостоятельная работа (УСР) студентов – это самостоятельная работа, выполняемая по заданию и при методическом руководстве преподавателя, а также контролируемая преподавателем на определенном этапе обучения. Целью УСР является целенаправленное обучение студентов основным навыкам и умению индивидуальной самостоятельной работы.

Результативность самостоятельной работы студентов во многом определяется наличием активных методов ее контроля:

– входной контроль знаний и умений студентов в начале изучения очередной дисциплины;

– текущий контроль, то есть регулярное отслеживание уровня усвоения материала на лекциях, практических и лабораторных занятиях;

– промежуточный контроль по окончании изучения раздела или модуля курса;

– самоконтроль, осуществляемый студентом в процессе изучения дисциплины при подготовке к контрольным мероприятиям;

– итоговый контроль по дисциплине в виде зачета или экзамена;

– контроль остаточных знаний и умений спустя определенное время после завершения изучения дисциплины.

На освоение учебного материала в рамках УСР для дисциплины «Многомерный статистический анализ данных» отводится 8 аудиторных часов.

Для самостоятельного изучения в рамках УСР дисциплины «Многомерный статистический анализ данных» выносятся следующие темы:

1. Многомерное нормальное (гауссовское) распределение как модель многомерных данных [1, 4, 5];
2. Статистическое оценивание параметров многомерной гауссовской модели [1, 4, 5];
3. Регрессионный анализ данных [5, 7, 8];
4. Случайная функция как модель статистических наблюдений в динамике [2, 3, 6].

По первой и третьей темам из приведенных выше тем проводится контрольная работа, по второй и четвертой темам контроль осуществляется в виде тестирования.

Рекомендации по контролю качества усвоения знаний и проведению аттестации

На лекционных занятиях по учебной дисциплине «Многомерный статистический анализ данных» рекомендуется использовать элементы проблемного обучения: проблемное изложение некоторых аспектов, использование частично-поискового метода.

Для аттестации обучающихся на соответствие их персональных достижений поэтапным и конечным требованиям образовательной программы создаются фонды оценочных средств, включающие типовые задания, контрольные работы и тесты. Оценочными средствами предусматривается оценка способности обучающихся к творческой деятельности, их готовность вести поиск решения новых задач, связанных с недостаточностью конкретных специальных знаний и отсутствием общепринятых алгоритмов.

Для диагностики компетенций в рамках учебной дисциплины рекомендуется использовать следующие формы:

– устная форма: собеседование, устный промежуточный зачет, итоговый зачет;

– письменная форма: тест, контрольный опрос, контрольная работа;

– устно-письменная форма: отчет по домашним практическим упражнениям с их устной защитой.

Контрольные мероприятия проводятся в соответствии с учебно-методической картой дисциплины. В случае неявки на контрольное мероприятие по уважительной причине студент вправе по согласованию с преподавателем выполнить его в дополнительное время. Для студентов, получивших неудовлетворительные оценки за контрольные мероприятия, либо не явившихся по неуважительной причине, по согласованию с преподавателем и с разрешения заведующего кафедрой мероприятие может быть проведено повторно.

Оценка текущей успеваемости рассчитывается как среднее оценок за каждую из письменных контрольных работ, оценки за отчеты по домашним практическим упражнениям и оценки за итоговый тест.

Итоговая аттестация предусматривает проведение зачета. При этом рекомендуется использовать оценивание успеваемости на основе модульно-рейтинговой системы.

ПРОТОКОЛ СОГЛАСОВАНИЯ УЧЕБНОЙ ПРОГРАММЫ

Название учебной дисциплины, с которой требуется согласование	Название кафедры	Предложения об изменениях в содержании учебной программы учреждения высшего образования по учебной дисциплине	Решение, принятое кафедрой, разработавшей учебную программу (с указанием даты и номера протокола)
Теория вероятностей и математическая статистика	Кафедра математического моделирования и анализа данных	нет	Оставить содержание учебной дисциплины без изменения, протокол № 19 от 07 апреля 2015 г.

ДОПОЛНЕНИЯ И ИЗМЕНЕНИЯ К УЧЕБНОЙ ПРОГРАММЕ

на _____/_____ учебный год

№№ Пп	Дополнения и изменения	Основание

Учебная программа пересмотрена и одобрена на заседании кафедры дискретной математики и алгоритмики (протокол № ____ от _____ 201_ г.)

Заведующий кафедрой

_____ _____ _____

(ученая степень, звание) (подпись) (И.О. Фамилия)

УТВЕРЖДАЮ
Декан факультета

_____ _____ _____

(ученая степень, звание) (подпись) (И.О.Фамилия)