Белорусский государственный университет

УТВЕРЖДАЮ
Проректор по учебной работе
и образовательным инновациям
_____ О.Н. Здрок
«02» декабря 2021 г.

Регистрационный № УД – 10489/уч.

СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ В ПРОГРАММНОЙ СРЕДЕ R

Учебная программа учреждения высшего образования по учебной дисциплине для специальности:

1-31 03 07 Прикладная информатика (по направлениям)

Направление специальности:
1-31 03 07-01 Прикладная информатика (программное обеспечение компьютерных систем)

2021 г.

Учебная программа составлена на основе образовательного стандарта высшего образования ОСВО 1-31 03 07-2013, учебных планов G31-167/уч., G31и-194/уч. от 30.05.2013 г.

СОСТАВИТЕЛИ:

Казаченок Виктор Владимирович, заведующий кафедрой компьютерных технологий и систем Белорусского государственного университета, доктор педагогических наук, кандидат физико-математических наук, профессор

РЕЦЕНЗЕНТЫ:

Баровик Дмитрий Валентинович, заместитель начальника управления автоматизации банковских операций ОАО «Центр банковских технологий», кандидат физико-математических наук

РЕКОМЕНДОВАНА К УТВЕРЖДЕНИЮ:

Кафедрой компьютерных технологий и систем Белорусского государственного университета
(протокол № 4 от 26.10.2021 г.)

Методической комиссией факультета прикладной математики и информатики Белорусского государственного университета
(протокол № 3 от 30.11.2021 г.)

Заведующий кафедрой _____ В. В. Казаченок

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Цели и задачи учебной дисциплины

Учебная дисциплина «Статистический анализ данных в программной среде R» входит в разряд дисциплин, изучаемых студентами специальности 1-31 03 07 Прикладная информатика (по направлениям). Для успешного освоения дисциплины студентам понадобятся базовые навыки работы с приложениями Microsoft Office.

Дисциплина «Статистический анализ данных в программной среде R» знакомит студентов с прикладными задачами и способами исследования зависимостей между различными данными и дальнейшем использовании знаний об этих зависимостях в задачах обработки и визуализации. Решение различных проблем зависит, прежде всего, от выявления взаимосвязи между различными явлениями, которые представляют собой результат одновременного воздействия большого числа причин. Поэтому при изучении этих явлений необходимо в первую очередь выявлять главные, основные причины. В то же время для получения достоверных выводов важно правильно оценить другие неучтенные и случайные причины. Решение перечисленных задач проводится, в первую очередь, методами прикладного статистического анализа, чем и обусловлено основное содержание дисциплины.

Цели учебной дисциплины:

1. Изучение теоретических основ предварительного (домодельного) статистического анализа данных (Exploratory Data Analysis).

2. Формирование навыков практического решения задач статистического анализа данных и представления получаемых результатов с использованием языка *R*.

Задачи учебной дисциплины:

Основная задача - освоение студентами теории и практики анализа данных, анализ возможностей их визуализации и обработки.

Место учебной дисциплины в системе подготовки специалиста с высшим образованием.

Учебная дисциплина относится **к циклу** дисциплин специализации компонента учреждения высшего образования.

Связи с другими учебными дисциплинами, включая учебные дисциплины компонента учреждения высшего образования, дисциплины специализации и др.

Учебная дисциплина «Статистический анализ данных в программной среде R» изучается после получения базовых знаний по дисциплине «Программирование».

Требования к компетенциям

Освоение учебной дисциплины «Статистический анализ данных в программной среде R» должно обеспечить формирование следующей компетенции:

ПК-33. Осуществлять поиск, систематизацию и анализ информации по перспективам развития отрасли, инновационным технологиям, проектам и решениям.

В результате изучения дисциплины студент должен:

знать:

– базовые математические модели анализа данных;
– основные математические методы анализа данных;
– различные математические алгоритмы анализа данных;
– основные математические методы визуализации;

уметь:

– использовать основные результаты дисциплины «Статистический анализ данных в программной среде R»;
– использовать теоретические и практические навыки применения анализа данных в практической деятельности;

владеть:

– основными методами анализа данных;
– основными методами визуализации данных.

Структура учебной дисциплины

Дисциплина изучается в 6 семестре. Всего на изучение учебной дисциплины «Статистический анализ данных в программной среде R» отведено:

– для очной формы получения высшего образования – 54 часа, в том числе 34 аудиторных часа, из них: лабораторные занятия – 30 часов (в том числе ДО – 12 часов), управляемая самостоятельная работа – 4 часа.

Трудоемкость учебной дисциплины составляет 1,5 зачетных единиц.

Форма текущей аттестации по учебной дисциплине – зачет.

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Тема 1. Первичная обработка статистических данных

Введение. Типы статистических данных и способы их первичной обработки. Качественные и количественные шкалы. Механический и случайный отбор при формировании выборки.

Тема 2. Числовые характеристики вариационных рядов

Методы предварительного анализа данных на основе дескриптивных статистик и графических представлений. Вариационные ряды, робастные показатели. Расчет средних значений, дисперсии, моды, медианы. Построение гистограмм, кумулят частот, квартилей, определение квантилей.

Тема 3. Корреляционный анализ

Методы исследования парных статистических зависимостей на основе корреляционного анализа (для количественных, ранговых и номинальных шкал). Построение диаграммы рассеяния, корреляционной таблицы. Вычисление коэффициента корреляции, определение его значимости. Шкала Чеддока.

Тема 4. Регрессионный анализ

Методы исследования парных статистических зависимостей на основе регрессионного анализа. Метод наименьших квадратов. Вычисление неизвестных параметров линейных одномерных регрессионных моделей; графическое представление таких моделей. Матричное представление линейных регрессионных моделей.

Тема 5. Базовые конструкции языка *R*

Возможности и базовые конструкции языка *R*, предназначенные для реализации изучаемых разделов анализа данных (вычисление числовых характеристик вариационных рядов, корреляционный анализ, регрессионный анализ и др.). Форматы данных (векторы, матрицы, факторы, списки, пропущенные данные). Управляющие конструкции и функции. Статистические функции. Использование расширяющих библиотек и графических возможностей.

Тема 6. Ряды динамики

Методы фильтрации и прогнозирования рядов динамики. Составляющие временного ряда: тренд, циклические и случайные колебания. Индивидуальные и обобщающие характеристики. Методы анализа: а) средних, б) с использованием критерия Валлиса и Мура, в) с использованием критерия серий, г) скользящей средней, д) аналитического выравнивания.

Тема 7. Кластерный анализ

Алгоритмы кластерного анализа неоднородных многомерных данных. Стандартизация переменных. Характеристики кластера: центр, радиус, среднеквадратическое отклонение, размер. Агломеративные и дивизимные, иерархические и итеративные методы кластеризации. Алгоритмы кластеризации *K*-means и *G*-means.

Тема 8. Дискриминантный анализ

Методы и стадии интеллектуального анализа данных Data Mining (деревья решений, алгоритмы кластеризации, регрессионный анализ, нейронные сети, временные ряды и др.). Место интеллектуального анализа данных в процессе поддержки принятия решений. Общий вид искусственного нейрона, нейронная сеть Кохонена и ее геометрическая интерпретация.

Тема 9. Обучение нейронных сетей

Процесс Data Mining. Построение и использование модели. Прогнозирование на основании знания зависимостей, трендов и статистики. Место человека-аналитика в системах интеллектуального анализа данных. Обучение нейронных сетей для определения веса синаптических связей. Алгоритмы обучения с учителем и без учителя.

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

Дневная форма получения образования с применением электронных средств обучения (ДО)

Номер раздела, темы	Название раздела, темы	Количество аудиторных часов					Количество часов УСР	Форма контроля знаний
		Лекции	Практические и семинарские занятия	Лабораторные занятия (ауд.)	Лабораторные занятия (ДО)	Иное		
1	2	3	4	5		6	7	8
1	Первичная обработка статистических данных			2	2 (ДО)			Отчет по лабораторной работе № 1
2	Числовые характеристики вариационных рядов			2	2 (ДО)			
3	Корреляционный анализ			2	2 (ДО)			Электронный тест
4	Регрессионный анализ			2	2 (ДО)			
5	Базовые конструкции языка *R*			2			2	Отчет по лабораторной работе № 2
6	Ряды динамики			2	2 (ДО)			
7	Кластерный анализ			2	2 (ДО)			
8	Дискриминантный анализ			2			1	Электронный тест
9	Обучение нейронных сетей			2			1	
И т о г о				**18**	**12**		**4**	

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Перечень основной литературы

№№ п/п	Список литературы	Год издания
1.	*Буяльская Ю.В.* Введение в компьютерный и интеллектуальный анализ данных/ Метод. указания / Ю.В. Буяльская, В.В. Казаченок – Минск: БГУ, 2016. – 46 с.	2016
2.	*Шипунов А.Б. и др.* Наглядная статистика. Используем R! – М.: ДМК Пресс, 2020. – 298 с.	2020
3.	Шипунов А.Б., Балдин Е.М. Анализ данных с R. – М., 2016. – 224 с.	2016
4.	*Ефимова М.Р., Ганченко О.И., Петрова Е.В.* Практикум по общей теории статистики. 3-е издание. – М: Финансы и статистика, 2011. – 368 с.	2011
5.	Теория статистики: Учебник/ Под ред. Р.А. Шмойловой. 5-е издание, доп. и перераб. – М.: Финансы и статистика, 2014. – 656 с.	2014

Перечень дополнительной литературы

№№ п/п	Список литературы	Год издания
1.	*Казаченок В.В.* Применение нейронных сетей в обучении // Информатика и образование. – 2020. – № 2. – С. 41-47.	2020
2.	Информатизация образования – 2014: материалы междунар. научн. конференции, Минск, 22–25 окт. 2014 г. / Белорусский гос. ун-т. – Минск: БГУ, 2014. – 464 с.	2014

Перечень рекомендуемых средств диагностики и методика формирования итоговой оценки

Для диагностики компетенций студентов используются следующие формы**:**

1. Устно-письменная форма.
2. Техническая форма.

К устно-письменной форме диагностики компетенций относятся:

- отчет по лабораторной работе.

К технической форме диагностики компетенций относятся:

- электронные тесты.

Перечень используемых средств диагностики результатов управляемой самостоятельной работы студентов:

- электронные тесты;
- отчет по лабораторной работе.

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.

Формой текущей аттестации по дисциплине «Статистический анализ данных в программной среде R» учебным планом предусмотрен зачет.

Оценка текущей успеваемости рассчитывается на основе модульно-рейтинговой системы, основанной на Положении о рейтинговой системе оценки знаний обучающихся по учебной дисциплине в БГУ (приказ ректора БГУ от 31.03.2020 № 189-ОД).

1. Текущая аттестация предусматривает проведение зачёта (в том числе с учётом результатов промежуточного и итогового тестирования) и проводится согласно Правилам проведения аттестации студентов, курсантов, слушателей при освоении содержания образовательных программ высшего образования (Постановление Министерства образования Республики Беларусь от 29 мая 2012 г. N 53).

Весовые коэффициенты, определяющие вклад текущего контроля знаний и текущей аттестации в итоговую оценку:

Формирование оценки за текущую успеваемость:

- отчет по лабораторной работе – 60 %;
- выполнение электронного теста – 40 %.

Итоговая оценка по дисциплине рассчитывается на основе оценки текущей успеваемости и оценки на зачете с учетом их весовых коэффициентов. Вес оценки по текущей успеваемости составляет 40 %, оценка на зачете – 60 %.

Примерный перечень заданий для управляемой самостоятельной работы студентов

Тема 5. Базовые конструкции языка *R* (2 ч)
Использование расширяющих библиотек и графических возможностей. Сравнение возможностей языка *R* и стандартных средств Microsoft Office.
Форма контроля **–** отчет по лабораторной работе.

Тема 8. Дискриминантный анализ (1 ч)
Общий вид искусственного нейрона, нейронная сеть Кохонена и ее геометрическая интерпретация.
Форма контроля – электронный тест.

Тема 9. Обучение нейронных сетей (1 ч)
Обучение нейронных сетей для определения веса синаптических связей. Понятие алгоритмов обучения с учителем и без учителя.
Форма контроля – электронный тест.

Примерная тематика лабораторных занятий

Тема 1. Первичная обработка статистических данных.

– Качественные и количественные шкалы измерений.

Тема 2. Числовые характеристики вариационных рядов.

– Методы предварительного анализа данных на основе дескриптивных статистик и графических представлений. Вариационные ряды, робастные показатели.

– Расчет средних значений, дисперсии, моды, медианы. Построение гистограмм, кумулят частот.

Тема 3. Корреляционный анализ.

– Методы исследования парных статистических зависимостей на основе корреляционного анализа (для количественных, ранговых и номинальных шкал). Построение диаграммы рассеяния, корреляционной таблицы. Вычисление коэффициента корреляции, определение его значимости. Шкала Чеддока.

Тема 4. Регрессионный анализ.

– Методы исследования парных статистических зависимостей на основе регрессионного анализа. Понятие метода наименьших квадратов. Программные средства вычисления неизвестных параметров линейных одномерных регрессионных моделей; графическое представление таких моделей.

Тема 5. Базовые конструкции языка *R*.

– Возможности и базовые конструкции языка *R*, предназначенные для реализации изучаемых разделов анализа данных (вычисление числовых характеристик вариационных рядов, корреляционный анализ, регрессионный анализ и др.).

– Управляющие конструкции и функции. Статистические функции. Импорт и экспорт данных в языке *R*. Использование расширяющих библиотек и графических возможностей. Сравнение возможностей языка *R* и стандартных средств Microsoft Office.

Тема 6. Ряды динамики.

– Методы фильтрации и прогнозирования рядов динамики. Составляющие временного ряда: тренд, циклические и случайные колебания. Индивидуальные и обобщающие характеристики. Методы анализа: а) средних, б) скользящей средней, в) аналитического выравнивания.

Тема 7. Кластерный анализ.

– Алгоритмы кластерного анализа неоднородных многомерных данных. Стандартизация переменных. Характеристики кластера: центр, радиус, среднеквадратическое отклонение, размер. Агломеративные и дивизимные методы кластеризации. Алгоритм кластеризации *K*-means.

Тема 8. Дискриминантный анализ.

– Методы и стадии интеллектуального анализа данных Data Mining (деревья решений, алгоритмы кластеризации, регрессионный анализ, нейронные сети, временные ряды и др.). Место интеллектуального анализа данных в процессе поддержки принятия решений.

– Общий вид искусственного нейрона, нейронная сеть Кохонена и ее геометрическая интерпретация.

Тема 9. Обучение нейронных сетей.

– Процесс Data Mining. Приложения Data Mining (торговля, банковское дело, телекоммуникации и др.). Прогнозирование на основании знания зависимостей, трендов и статистики. Место человека-аналитика в системах интеллектуального анализа данных.

– Обучение нейронных сетей для определения веса синаптических связей. Понятие алгоритмов обучения с учителем и без учителя.

Описание инновационных подходов и методов к преподаванию учебной дисциплины

При организации образовательного процесса используется ***практико-ориентированный подход,*** который предполагает***:***

- освоение содержание образования через решения практических задач;
- приобретение навыков эффективного выполнения разных видов профессиональной деятельности;
- ориентацию на генерирование идей, реализацию групповых студенческих проектов, развитие предпринимательской культуры;
- использованию процедур, способов оценивания, фиксирующих сформированность профессиональных компетенций.

Методические рекомендации по организации самостоятельной работы обучающихся

Для организации самостоятельной работы студентов по учебной дисциплине «Статистический анализ данных в программной среде R» используются современные информационные ресурсы:

на образовательном портале EDUFPMI размещены:

- учебно-методические материалы,
- учебное издание для теоретического изучения дисциплины,
- методические указания к лабораторным занятиям,
- материалы текущего контроля и текущей аттестации,
- вопросы для подготовки к зачету,
- тесты, вопросы для самоконтроля,
- список рекомендуемой литературы и информационных ресурсов.

Примерный перечень вопросов к зачету

1. Введение. Типы статистических данных и способы их первичной обработки. Качественные и количественные шкалы. Механический и случайный отбор при формировании выборки.
2. Методы предварительного анализа данных на основе дескриптивных статистик и графических представлений. Вариационные ряды, робастные показатели.
3. Расчет средних значений, дисперсии, моды, медианы. Построение гистограмм, кумулят частот.
4. Методы исследования парных статистических зависимостей на основе корреляционного анализа (для количественных, ранговых и

номинальных шкал). Построение диаграммы рассеяния, корреляционной таблицы.

5. Вычисление коэффициента корреляции, определение его значимости. Шкала Чеддока.
6. Методы исследования парных статистических зависимостей на основе регрессионного анализа. Понятие метода наименьших квадратов. Программные средства вычисления неизвестных параметров линейных одномерных регрессионных моделей; графическое представление таких моделей.
7. Возможности и базовые конструкции языка *R*, предназначенные для реализации изучаемых разделов анализа данных (вычисление числовых характеристик вариационных рядов, корреляционный анализ, регрессионный анализ и др.). Управляющие конструкции и функции.
8. Статистические функции. Импорт и экспорт данных в языке *R*. Использование расширяющих библиотек и графических возможностей. Сравнение возможностей языка *R* и стандартных средств Microsoft Office.
9. Ряды динамики. Методы фильтрации и прогнозирования рядов динамики. Составляющие временного ряда: тренд, циклические и случайные колебания. Индивидуальные и обобщающие характеристики. Методы анализа: а) средних, б) скользящей средней, в) аналитического выравнивания.
10. Алгоритмы кластерного анализа неоднородных многомерных данных. Стандартизация переменных. Характеристики кластера: центр, радиус, среднеквадратическое отклонение, размер. Агломеративные и дивизимные методы кластеризации. Алгоритм кластеризации K-means.
11. Методы и стадии интеллектуального анализа данных Data Mining (деревья решений, алгоритмы кластеризации, регрессионный анализ, нейронные сети, временные ряды и др.). Место интеллектуального анализа данных в процессе поддержки принятия решений.
12. Общий вид искусственного нейрона, нейронная сеть Кохонена и ее геометрическая интерпретация.
13. Процесс Data Mining. Приложения Data Mining (торговля, банковское дело, телекоммуникации и др.). Прогнозирование на основании знания зависимостей, трендов и статистики. Место человека-аналитика в системах интеллектуального анализа данных.
14. Обучение нейронных сетей для определения веса синаптических связей. Понятие алгоритмов обучения с учителем и без учителя.

ПРОТОКОЛ СОГЛАСОВАНИЯ УЧЕБНОЙ ПРОГРАММЫ УВО

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 информации	Информационных систем управления	Нет	Изменений не требуется (протокол № 4 от 26.10.2021 г.)

ДОПОЛНЕНИЯ И ИЗМЕНЕНИЯ К УЧЕБНОЙ ПРОГРАММЕ

на _20___/_20___ учебный год

№№ Пп	Дополнения и изменения	Основание

Учебная программа пересмотрена и одобрена на заседании кафедры компьютерных технологий и систем (протокол № от 20__ г.)

Заведующий кафедрой
д.пед. наук _____ В. В. Казаченок
профессор (подпись)

УТВЕРЖДАЮ
Декан факультета
д.т. наук _____ А. М. Недзьведь
(подпись)