

Белорусский государственный университет

УТВЕРЖДАЮ

**Проректор по учебной работе и
образовательным инновациям**



О.Г. Прохоренко

«08» июля 2022 г.

Регистрационный № УД – _11791/уч.

СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ

**Учебная программа учреждения высшего образования
по учебной дисциплине для специальности:**

1-98 01 01 Компьютерная безопасность (по направлениям)

Направление специальности:

**1-98 01 01-01 Компьютерная безопасность (математические методы
и программные системы)**

2022 г.

Учебная программа составлена на основе образовательного стандарта высшего образования ОСВО 1-98 01 01-2013, учебного плана: № Р98-138/уч. от 30.05.2013 г.

СОСТАВИТЕЛИ:

Ю.С. Харин, профессор кафедры математического моделирования и анализа данных Белорусского государственного университета, доктор физико-математических наук, академик НАН Беларуси, профессор;

И.А. Бодягин, доцент кафедры математического моделирования и анализа данных Белорусского государственного университета, кандидат физико-математических наук, доцент.

РЕЦЕНЗЕНТ:

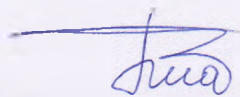
Л.А. Сошникова, профессор кафедры статистики УО «Белорусский государственный экономический университет», доктор экономических наук, профессор

РЕКОМЕНДОВАНА К УТВЕРЖДЕНИЮ:

Кафедрой математического моделирования и анализа данных факультета прикладной математики и информатики БГУ (протокол № 10 от 26 апреля 2022 г.)

Учебно-методической комиссией факультета прикладной математики и информатики БГУ (протокол № 7 от 31.05.2022).

Заведующий кафедрой



И.А. Бодягин

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Статистический анализ данных (САД) – это самостоятельная научная дисциплина, изучающая методы сбора и предварительного анализа статистических данных, их систематизации и обработки с целью выявления характера и структуры взаимосвязей между признаками, присущими исследуемому объекту или системе. Методы 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. Важную роль методы статистического анализа данных играют при решении задач защиты информации.

В основе 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 статистического анализа в практику. Пакеты прикладных программ (ППП) с удобным пользовательским интерфейсом и также современные технологии компьютерного анализа данных с использованием языков программирования Python и 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 «Статистический анализ данных» для специальности 1-98 01 01 «Компьютерная безопасность (по направлениям)» 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.

Теоретический курс поддерживается лабораторным компьютерным практикумом, предполагающим использование статистических пакетов и языков программирования R или Python.

Цели и задачи учебной дисциплины

Цели учебной дисциплины «Статистический анализ данных»:

- освоение студентами теоретических основ статистического анализа данных;
- расширение фундаментального математического образования;
- формирование у студентов профессиональных знаний в предметной области, подкрепленных владением математическими методами и пакетами прикладных программ.

Задачи учебной дисциплины «Статистический анализ данных»:

- систематическое изучение теоретических основ статистического анализа данных;
- приобретение практических навыков для правильного использования арсенала методов статистического анализа данных в конкретных ситуациях;
- овладение инструментальными средствами, включая статистические пакеты с пользовательским интерфейсом, а также языки R и Python.

Место учебной дисциплины в системе подготовки специалиста с высшим образованием.

Учебная дисциплина «Статистический анализ данных» относится к циклу дисциплин специализаций компонента учреждения образования.

Связи с другими учебными дисциплинами

Учебная дисциплина «Статистический анализ данных», основываясь на теоретических положениях, моделях и методах учебной дисциплины «Теория вероятностей и математическая статистика», обобщает и развивает их применительно к задачам статистического анализа данных.

Знания, полученные в рамках данной дисциплины, будут использованы при изучении других дисциплин специальности 1-98 01 01 Компьютерная безопасность (по направлениям), требующих углубленной подготовки в области анализа данных, включая дисциплины: «Компьютерный статистический анализ дискретных данных», «Тестирование генераторов случайных и псевдослучайных последовательностей».

Требования к компетенциям

Освоение учебной дисциплины «Статистический анализ данных» должно обеспечить формирование следующих академических, социально-личностных и профессиональных компетенций:

Академические компетенции:

АК-1. Уметь применять базовые научно-теоретические знания для решения теоретических и практических задач.

АК-2. Владеть системным и сравнительным анализом.

АК-3. Владеть исследовательскими навыками.

АК-4. Уметь работать самостоятельно.

АК-5. Быть способным вырабатывать новые идеи (креативность).

АК-6. Владеть междисциплинарным подходом при решении проблем.

АК-7. Иметь навыки, связанные с использованием технических устройств, управлением информацией и работой с компьютером.

АК-8. Иметь лингвистические навыки (устная и письменная коммуникация).

АК-9. Уметь учиться, повышать свою квалификацию в течение всей жизни.

Социально-личностные компетенции:

СЛК-6. Уметь работать в команде.

Профессиональные компетенции:

ПК-1. Работать с научной, нормативно-справочной и специальной литературой с целью получения последних сведений о новых методах защиты информации, о стойкости существующих систем защиты информации.

ПК-2. Формулировать задачи, возникающие при организации защиты информации.

ПК-3. Разрабатывать модели явлений, процессов или систем при организации защиты информации.

ПК-4. Выбирать необходимые методы исследования, модифицировать существующие, разрабатывать новые методы и применять их для решения поставленных задач при организации защиты информации.

ПК-5. Выполнять оценку эффективности методов защиты информации.

В результате изучения дисциплины студент должен

знать:

– методы статистического оценивания параметров распределения вероятностей;

– методы предварительного статистического анализа данных;

– методы корреляционного анализа;

– методы регрессионного анализа данных;

– методы дискриминантного и кластерного анализа;

– методы статистической проверки гипотез;

– методы статистического анализа случайных последовательностей;

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 применения современных ППП для решения задач статистического анализа временных рядов в различных приложениях;
- навыками по подготовке отчетов с результатами статистического анализа данных, включающих содержательную интерпретацию результатов анализа, комментарии, выводы и рекомендации.

Структура учебной дисциплины

Дисциплина изучается в 5 семестре. Всего на изучение учебной дисциплины «Статистический анализ данных» отведено:

- для очной формы получения высшего образования – 152 часа, в том числе 68 аудиторных часов, из них: лекции – 34 часа, лабораторные занятия – 30 часов (в том числе 12 ч/ДО), управляемая самостоятельная работа – 4 ч (ДО).

Трудоемкость учебной дисциплины составляет 4 зачетные единицы.

Форма текущей аттестации – экзамен.

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Раздел 1. Статистический анализ многомерных данных

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

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. Методы предварительного анализа данных. Примеры.

Тема 1.3. Статистическое оценивание функциональных и числовых характеристик вероятностных моделей данных. Функциональные и числовые характеристики вероятностных моделей данных. Параметрический и непараметрический подходы к статистическому оцениванию характеристик вероятностных моделей данных.

Тема 1.4. Многомерное нормальное (гауссовское) распределение как модель многомерных данных. Многомерное нормальное распределение как модель многомерных данных. Функциональные и числовые характеристики, маргинальные распределения, линейные преобразования гауссовских случайных векторов. Условное распределение гауссовского вектора. Функция регрессии и ее оптимальные свойства. Частный и множественный коэффициенты корреляции: определение и свойства.

Тема 1.5. Статистическое оценивание параметров многомерной гауссовской модели. Оценивание параметров многомерной гауссовской модели по методу максимального правдоподобия. Свойства и распределение оценок.

Тема 1.6. Корреляционный анализ. Выборочный коэффициент корреляции: свойства и применения. Выборочный частный коэффициент корреляции: свойства и применения. Выборочный множественный коэффициент корреляции: свойства и применения. Статистические выводы о значениях коэффициентов корреляции.

Тема 1.7. Статистическая проверка гипотез на основе T^2 -статистики Хоттелинга. Многомерный T^2 -критерий Стюдента. Свойства статистики Хоттелинга. Применение статистики Хоттелинга в анализе данных. Проверка гипотезы о значениях вектора математического ожидания. Построение доверительной области. Проверка гипотезы о равенстве средних в двух выборках. Проблема Беренса-Фишера.

Тема 1.8. Регрессионный анализ данных. Регрессионная модель данных. Статистическое оценивание параметров многомерной линейной регрессии. Свойства оценок параметров. Проверка гипотез относительно параметров модели многомерной линейной регрессии.

Тема 1.9. Дискриминантный анализ данных. Задача статистической классификации. Дискриминантный анализ данных при наличии обучающей выборки. Подстановочное байесовское решающее правило. Оптимальная классификация гауссовских случайных векторов. Непараметрический дискриминантный анализ.

Тема 1.10. Кластерный анализ данных. Кластерный анализ данных: общая характеристика алгоритмов классификации. Алгоритмы К-средних и иерархического кластерного анализа. Оценка качества кластеризации. EM-алгоритм расщепления смесей распределений многомерных данных в режиме самообучения.

Раздел 2. Статистический анализ случайных последовательностей

Тема 2.1. Случайная функция как модель статистических наблюдений в динамике. Временные ряды и их основные характеристики. Отсчет, реализация и ансамбль реализаций временного ряда. Совместное распределение вероятностей отсчетов, математическое ожидание, дисперсия, ковариационная функция, корреляционная функция временного ряда.

Тема 2.2. Стационарные временные ряды. Стационарные временные ряды в узком и широком смысле. Связь между стационарностью в узком и широком смысле. Спектральная плотность стационарного в широком смысле временного ряда и ее связь с ковариационной функцией.

Тема 2.3. Статистическое оценивание спектральной плотности и выявление скрытых периодичностей. Выборочная спектральная плотность: ее связь с выборочной ковариационной функцией, статистические свойства и сглаживание. Периодограмма и её свойства. Сглаженная периодограмма. Применение периодограммных статистик для выявления скрытых периодичностей.

Тема 2.4. Статистическое обнаружение скрытых рекуррентных последовательностей. Понятие рекуррентных последовательностей. Обзор методов статистического обнаружения скрытых рекуррентных последовательностей.

Тема 2.5. Цепи Маркова s -го порядка. Малопараметрические модели. Условия эргодичности.

Тема 2.6. Оценки максимального правдоподобия параметров однородных цепей Маркова. Метод максимального правдоподобия для оценивания параметров однородных цепей Маркова. Оценивание матрицы вероятностей одношаговых переходов и стационарного распределения вероятностей.

Тема 2.7. Проверка гипотез об однородных цепях Маркова. Проверка гипотезы о независимости отсчетов. Проверка гипотезы о совпадении матрицы вероятностей одношаговых переходов с наперед заданной матрицей.

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

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

№	Название раздела, темы	Количество аудиторных часов					Количество часов УСП	Форма контроля знаний
		Лекции	Практические занятия	Семинарские занятия	Лабораторные занятия	Иное		
1	Статистический анализ многомерных данных	20			18		2	
1.1	Актуальность задач статистического анализа данных (САД).	2			2			Устный опрос
1.2	Предварительный (первичный) статистический анализ данных	2			2			Устный опрос
1.3	Статистическое оценивание функциональных и числовых характеристик вероятностных моделей данных	2			2 (ДО)			Устный опрос
1.4	Многомерное нормальное (гауссовское) распределение как модель многомерных данных	2			2			Контрольная работа №1
1.5	Статистическое оценивание параметров многомерной гауссовской модели.	2			2 (ДО)			Устный опрос

1.6	Корреляционный анализ	2			2			Контрольная работа №2
1.7	Статистическая проверка гипотез на основе T2-статистики Хотеллинга	2					2 (ДО)	Устный опрос
1.8	Регрессионный анализ данных	2			2			Устный опрос
1.9	Дискриминантный анализ данных.	2			2			Тест
1.10	Кластерный анализ данных	2			2 (ДО)			Устный опрос
2	Статистический анализ случайных последовательностей	14			12		2	
2.1	Случайная функция как модель статистических наблюдений в динамике	2			2 (ДО)			Тест
2.2	Стационарные временные ряды	2			2			Устный опрос
2.3	Статистическое оценивание спектральной плотности и выявление скрытых периодичностей	2					2 (ДО)	Устный опрос
2.4	Статистическое обнаружение скрытых рекуррентных последовательностей	2			2 (ДО)			Устный опрос

2.5	Цепи Маркова s -го порядка	2			2			Устный опрос
2.6	Оценки максимального правдоподобия параметров однородных цепей Маркова	2			2 (ДО)			Устный опрос
2.7	Проверка гипотез об однородных цепях Маркова	2			2			Устный опрос
ИТОГО		34			30		4	

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Перечень основной литературы

1. Теория вероятностей, математическая и прикладная статистика : учебник для студ. учреждений высшего образования по спец. "Прикладная математика", "Информатика", "Экономическая кибернетика", "Актуарная математика", "Компьютерная безопасность", "Статистика" / Ю. С. Харин, Н. М. Зуев, Е. Е. Жук. - Минск : БГУ, 2011. – 463 с.
2. Харин Ю.С., Малюгин В.И., Абрамович М.С. Математические и компьютерные основы статистического моделирования анализа данных. Минск: БГУ, 2008. – 455 с.
3. Труш, Н.Н. Введение в компьютерный и интеллектуальный анализ данных : учеб. материалы / Н. Н. Труш. – Минск : БГУ, 2022. – 69 с.

Перечень дополнительной литературы

1. Айвазян С.А. и др. Прикладная статистика. Исследование зависимостей. М.: Финансы и статистика, 1985.
2. Айвазян С.А. и др. Прикладная статистика. Классификация и снижение размерности. М.: Финансы и статистика, 1989.
3. Андерсон Т.В. Введение в многомерный статистический анализ. М., Физматгиз, 1963.
4. Андерсон Т.В. Статистический анализ временных рядов. М.: Мир, 1976.
5. Бриллинджер Д. Временные ряды. М.: Мир, 1990.
6. Харин Ю.С., Малюгин В.И., Абрамович М.С. Математические и компьютерные основы статистического моделирования и анализа данных. Мн.: БГУ, 2008.
7. Харин Ю.С., Абрамович М.С., Малюгин В.И. Компьютерный учебник по статистике. Мн.: БГУ, 2001.
8. Математические и компьютерные основы криптологии : Учеб. пособие для студ. матем. и инженерно-техн. спец. вузов / Ю. С. Харин, В. И. Берник, Г. В. Матвеев, С. В. Агиевич. - Минск : Новое знание, 2003. – 381с.
9. Кнут Д. Искусство программирования на ЭВМ. Т.2: Получисленные алгоритмы. - М.: Мир, 1977.
10. Ивченко Г.И., Медведев Ю.И. Введение в математическую статистику. – М.: URSS, 2009.

Перечень рекомендуемых средств диагностики и методика формирования итоговой отметки

На лекционных занятиях по дисциплине «Статистический анализ данных» рекомендуется особое внимание обращать на установлении связей между теоретическими темами курса и использованием, изучаемых методов и алгоритмов для решения практических задач анализа данных.

Контрольные мероприятия проводятся в соответствии с учебно-методической картой дисциплины.

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

- устная форма: устные опросы по текущим темам;
- письменная форма: контрольная работа, тест.

Отчеты загружаются для проверки в специально организованный онлайн-курс на портале <https://edufpmi.bsu.by/course/view.php?id=308>

Формой текущей аттестации по дисциплине «Статистический анализ данных» учебным планом предусмотрен **экзамен**.

При формировании итоговой отметки используется рейтинговая система оценки знаний студента, дающая возможность проследить и оценить динамику процесса достижения целей обучения. Рейтинговая система предусматривает использование весовых коэффициентов для текущего контроля знаний и текущей аттестации студентов по дисциплине.

Весовые коэффициенты, определяющие вклад текущего контроля знаний в итоговую отметку:

формирование отметки за текущую успеваемость:

отметка текущей успеваемости рассчитывается на основе отметок по 4 контрольным точкам (две контрольные работы и два тестовых задания):

- контрольные работы – 60%;
- тесты – 40%.

Итоговая отметка по дисциплине рассчитывается на основе отметки текущей успеваемости (рейтинговой системы оценки знаний) и экзаменационной отметки с учетом их весовых коэффициентов. Вес отметки по текущей успеваемости составляет 40 %, экзаменационной отметки – 60 %.

Примерный перечень заданий для 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.

На освоение учебного материала в рамках УСР для дисциплины «Статистический анализ данных» отводится 4 аудиторных часа по двум следующим темам в соответствии с учебно-методической картой дисциплины.

Тема № 1.7 Статистическая проверка гипотез на основе T^2 -статистики Хоттелинга (2 ч/ДО).

Перечень вопросов для углубленного самостоятельного изучения:

- проверка гипотезы о значении вектора математического ожидания;
- распределение T^2 -статистики Хоттелинга.

Форма контроля – устный опрос.

Тема 2.3 Статистическое оценивание спектральной плотности и выявление скрытых периодичностей (2 ч/ДО).

Перечень вопросов для углубленного самостоятельного изучения:

- спектральное представление ковариационной функции случайного процесса;
- теорема Бохнера.

Форма контроля – устный опрос.

Примерная тематика лабораторных занятий

1. Первичный статистический анализ данных.
2. Многомерный нормальный закон распределения вероятностей
3. Регрессионный анализ данных.
4. Корреляционный анализ данных.
5. Статистическая классификация данных.

Описание инновационных подходов и методов к преподаванию учебной дисциплины

При организации образовательного процесса используется **эвристический и практико-ориентированный подходы.**

Эвристический подход предполагает:

- демонстрацию многообразия решений задач анализа данных в условиях выполнения и невыполнения модельных предположений;
- творческую самореализацию обучающихся в процессе решения задач;
- индивидуализацию обучения через возможность самостоятельно ставить цели и задачи анализа и проводить исследования.

Практико-ориентированный подход предполагает:

- освоение содержания 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

Студенты самостоятельно выполняют следующую работу:

- осуществляют углубленное изучение моделей и методов анализа данных по избранным темам с использованием рекомендуемой литературы;
- выполняют задания лабораторного практикума в полном объеме с использованием различных статистических пакетов;
- готовят отчет с результатами проведенных исследований в соответствии с установленными требованиями;
- работают над устранением указанных при проверке отчетов недостатков.

Условия для самостоятельной работы студентов, в частности, для развития навыков самоконтроля, способствующих интенсификации учебного процесса, обеспечиваются наличием и полной доступностью электронных (и бумажных) курсов лекций, учебно-методических материалов по основным темам дисциплины, компьютерных тестовых заданий и данных для их выполнения на портале edufpmi.bsu.by.

Примерный перечень вопросов к экзамену

1. Предварительный (первичный) статистический анализ данных.
2. Статистическое оценивание функциональных и числовых характеристик вероятностных моделей данных.
3. Критерии согласия.
4. Статистическое оценивание плотности распределения вероятностей.
5. Теорема о свойствах гистограммной оценки плотности.
6. Многомерное нормальное распределение и его свойства.
7. Характеристическая функция многомерного нормального распределения.
8. Теорема о линейном преобразовании многомерного нормального вектора.
9. Статистическое оценивание параметров многомерного нормального распределения.
10. Коэффициент корреляции и парная зависимость признаков.
11. Условное распределение многомерного нормального вектора.

12. Частный коэффициент корреляции.
13. Критерий согласия Пирсона.
14. Риск как мера оптимальных решающих правил и его свойства.
15. Задача статистической классификации, рандомизированные и нерандомизированные решающие правила.
16. Оптимальное (байесовское) решающее правило (БРП).
17. Статистическая классификация многомерных нормальных наблюдений.
18. Статистическая классификация многомерных нормальных наблюдений в случае двух классов.
19. Квадратичное и линейное БРП.
20. Основные понятия дискриминантного анализа.
21. Метод k ближайших соседей.
22. Простая линейная регрессия.
23. Множественная линейная регрессия.
24. Временные ряды и их основные характеристики.
25. Временные ряды авторегрессии и скользящего среднего.
26. Цепь Маркова и ее свойства.

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 вероятностей и математическая статистика	Кафедра математического моделирования и анализа данных	нет	Оставить содержание учебной дисциплины без изменения (протокол № 10 от 26 апреля 2022 г.)

ДОПОЛНЕНИЯ И ИЗМЕНЕНИЯ К УЧЕБНОЙ ПРОГРАММЕ

на ____/____ учебный год

№№ Пп	Дополнения и изменения	Основание

Учебная программа пересмотрена и одобрена на заседании кафедры математического моделирования и анализа данных (протокол № ____ от ____).

Заведующий кафедрой

канд. физ.-мат. наук, доцент

(ученая степень, звание)

(подпись)

И.А. Бодягин

(И.О. Фамилия)

УТВЕРЖДАЮ

Декан факультета

доктор техн. наук, доцент

(ученая степень, звание)

(подпись)

А.М. Недзьведь

(И.О. Фамилия)