Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования «Ростовский государственный экономический университет (РИНХ)»

УТВЕРЖДАЮ
Начальник
учебно-методического управления
Платонова Т.К.
«25» июня 2024 г.

Рабочая программа дисциплины
Многомерные статистические методы

Направление 01.03.05 Статистика
Направленность 01.03.05.01 Анализ больших данных

Для набора 2022 года

Квалификация
Бакалавр

КАФЕДРА **Статистики, эконометрики и оценки рисков**

Распределение часов дисциплины по семестрам

Семестр (<Курс>.<Семестр на курсе>)	**7 (4.1)**		Итого	
Недель	16			
Вид занятий	УП	РП	УП	РП
Лекции	10	10	10	10
Лабораторные	10	10	10	10
Практические	10	10	10	10
Итого ауд.	30	30	30	30
Контактная работа	30	30	30	30
Сам. работа	249	249	249	249
Часы на контроль	9	9	9	9
Итого	288	288	288	288

ОСНОВАНИЕ

Учебный план утвержден учёным советом вуза от 25.06.2024 г. протокол № 18.

Программу составил(и): к.э.н., доц., Трегубова А.А.;к.э.н., доц., Кракашова О.А.

Зав. кафедрой: д.э.н., проф. Ниворожкина Л.И.

Методический совет направления: к.э.н., доцент Андреева О.В.

1. ЦЕЛИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

1.1	Цель изучения дисциплины: овладение методологией многомерного статистического анализа, инструментальными средствами обработки данных, навыками использования современного программного обеспечения для построения многомерных моделей.

2. ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ

ОПК-3: Способен осознанно применять методы математической и дескриптивной статистики для анализа количественных данных, в том числе с применением необходимой вычислительной техники и стандартных компьютерных программ, содержательно интерпретировать полученные результаты, готовить статистические материалы для докладов, публикаций и других аналитических материалов

ПК-6: Способен осуществлять поиск статистической информации, ее первичную обработку и подготовку для проведения аналитических исследований, в том числе с использованием технологий больших данных

В результате освоения дисциплины обучающийся должен:

Знать:

методы математической и дескриптивной статистики для анализа данных (соотнесено с индикатором ОПК-3.1); способы снижения размерности исследуемых многомерных признаков и отбора наиболее информативных показателей (соотнесено с индикатором ПК-6.1)

Уметь:

содержательно интерпретировать результаты расчетов и обосновывать полученные при моделировании выводы, использовать современное программное обеспечение для построения многомерных моделей (соотнесено с индикатором ОПК-3.2); критически оценивать полученные при моделировании результаты (соотнесено с индикатором ПК-6.2)

Владеть:

методами многомерного анализа с использованием специализированных программных средств (соотнесено с индикатором ОПК-3.3); инструментальными средствами обработки данных (соотнесено с индикатором ПК-6.3)

3. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

Раздел 1. Регрессионный анализ и классификация.

№	Наименование темы / Вид занятия	Семестр / Курс	Часов	Компетен-ции	Литература
1.1	Тема "Первичная обработка данных". Многомерное признаковое пространство. Многомерное нормальное распределение. Методы шкалирования при обработке качественных признаков. Проблема размерности в многомерных методах исследования. Статистическое оценивание и сравнение многомерных. генеральных совокупностей. Распределение и характеристики многомерной совокупности. Многомерное нормальное распределение. Статистические оценки многомерной генеральной совокупности. / Лек /	7	2	ОПК-3, ПК -6	Л1.1, Л1.2, Л1.3, Л1.4, Л2.1, Л2.2, Л2.3, Л2.4, Л2.5, Л2.6, Л2.7
1.2	Тема "Первичная обработка данных". Многомерное признаковое пространство. Многомерное нормальное распределение. Методы шкалирования при обработке качественных признаков. Проблема размерности в многомерных методах исследования. Статистическое оценивание и сравнение многомерных. генеральных совокупностей. Распределение и характеристики многомерной совокупности. Многомерное нормальное распределение. Статистические оценки многомерной генеральной совокупности. Работа в LibreOffice Calc, RStudio, STADIA. Знакомство с источниками данных: База данных Центрального банка РФ http://cbr.ru/hd_base/ База статистических данных https://rosstat.gov.ru/databases Единая межведомственная информационно-статистическая система https://www.fedstat.ru/ / Пр /	7	4	ОПК-3, ПК -6	Л1.1, Л1.2, Л1.3, Л1.4, Л2.1, Л2.2, Л2.3, Л2.4, Л2.5, Л2.6, Л2.7
1.3	Тема "Первичная обработка данных". Многомерное признаковое пространство. Многомерное нормальное распределение. Методы шкалирования при обработке качественных признаков. Проблема размерности в многомерных методах исследования. Статистическое оценивание и сравнение многомерных.	7	8	ОПК-3, ПК -6	Л1.1, Л1.2, Л1.3, Л1.4, Л2.1, Л2.2, Л2.3, Л2.4, Л2.5, Л2.6, Л2.7

	генеральных совокупностей. Распределение и характеристики многомерной совокупности. Многомерное нормальное распределение. Статистические оценки многомерной генеральной совокупности. / Ср /				
1.4	Тема "Корреляционно-регрессионный анализ". Корреляционный анализ. Построение и интерпретация модели множественной линейной регрессии. Ранговая корреляция. Корреляция категоризованных переменных. Статистический анализ экспертных оценок. / Ср /	7	12	ОПК-3, ПК -6	Л1.1, Л1.2, Л1.3, Л1.4, Л2.1, Л2.2, Л2.3, Л2.4, Л2.5, Л2.6, Л2.7
1.5	Тема "Дискриминантный анализ". Построение и интерпретация модели линейного дискриминантного анализа. Пошаговый дискриминантный анализ. Оценка качества дискриминантной функции. / Ср /	7	14	ОПК-3, ПК -6	Л1.1, Л1.2, Л1.3, Л1.4, Л2.1, Л2.2, Л2.3, Л2.4, Л2.5, Л2.6, Л2.7
1.6	Тема "Дискриминантный анализ". Построение и интерпретация модели линейного дискриминантного анализа. Пошаговый дискриминантный анализ. Оценка качества дискриминантной функции. Работа в LibreOffice Calc, RStudio, STADIA. / Лаб /	7	2	ОПК-3, ПК -6	Л1.1, Л1.2, Л1.3, Л1.4, Л2.1, Л2.2, Л2.3, Л2.4, Л2.5, Л2.6, Л2.7
1.7	Тема "Кластерный анализ". Непараметрический случай классификации без обучения: кластерный анализ. Расстояние между объектами. Меры близости между объектами. Меры близости между кластерами. Иерархические кластер-процедуры. Метод k-средних. Расщепление смесей вероятностных распределений. / Ср /	7	26	ОПК-3, ПК -6	Л1.1, Л1.2, Л1.3, Л1.4, Л2.1, Л2.2, Л2.3, Л2.4, Л2.5, Л2.6, Л2.7
1.8	Тема "Кластерный анализ". Непараметрический случай классификации без обучения: кластерный анализ. Расстояние между объектами. Меры близости между объектами. Меры близости между кластерами. Иерархические кластер-процедуры. Метод k-средних. Расщепление смесей вероятностных распределений. Работа в RStudio, STADIA. / Лаб /	7	2	ОПК-3, ПК -6	Л1.1, Л1.2, Л1.3, Л1.4, Л2.1, Л2.2, Л2.3, Л2.4, Л2.5, Л2.6, Л2.7
1.9	Тема "Первичная обработка данных". Многомерное признаковое пространство. Многомерное нормальное распределение. Методы шкалирования при обработке качественных признаков. Проблема размерности в многомерных методах исследования. Знакомство с источниками данных: База данных Центрального банка РФ http://cbr.ru/hd_base/, База статистических данных https://rosstat.gov.ru/databases, Единая межведомственная информационно-статистическая система https://www.fedstat.ru/ Работа с Консультант + / Ср /	7	28	ОПК-3, ПК -6	Л1.1, Л1.2, Л1.3, Л1.4, Л2.1, Л2.2, Л2.3, Л2.4, Л2.5, Л2.6, Л2.7
1.10	Тема "Дискриминантный анализ". Построение и интерпретация модели линейного дискриминантного анализа. Пошаговый дискриминантный анализ. Оценка качества дискриминантной функции. / Ср /	7	35	ОПК-3, ПК -6	Л1.1, Л1.2, Л1.3, Л1.4, Л2.1, Л2.2, Л2.3, Л2.4, Л2.5, Л2.6, Л2.7
1.11	Тема "Кластерный анализ". Непараметрический случай классификации без обучения: кластерный анализ. Расстояние между объектами. Меры близости между объектами. Меры близости между кластерами. Иерархические кластер-процедуры. Метод k-средних. Расщепление смесей вероятностных распределений. / Ср /	7	20	ОПК-3, ПК -6	Л1.1, Л1.2, Л1.3, Л1.4, Л2.1, Л2.2, Л2.3, Л2.4, Л2.5, Л2.6, Л2.7

Раздел 2. Снижение размерности. Комплексный многомерный анализ.

№	Наименование темы / Вид занятия	Семестр / Курс	Часов	Компетен-ции	Литература
2.1	Тема "Снижение размерности исследуемых многомерных признаков". Метод главных компонент. Собственные векторы и собственные значения и их использование для получения матрицы весовых коэффициентов. Построение и интерпретация модели главных компонент. / Ср /	7	20	ОПК-3, ПК -6	Л1.1, Л1.2, Л1.3, Л1.4, Л2.1, Л2.2, Л2.3, Л2.4, Л2.5, Л2.6, Л2.7
2.2	Тема "Снижение размерности исследуемых многомерных признаков". Метод главных компонент. Собственные векторы и собственные значения и их использование для получения матрицы весовых коэффициентов. Построение и интерпретация модели главных компонент. / Пр /	7	2	ОПК-3, ПК -6	Л1.1, Л1.2, Л1.3, Л1.4, Л2.1, Л2.2, Л2.3, Л2.4, Л2.5, Л2.6, Л2.7

2.3	Тема "Факторный анализ". Модель ортогональных факторов. Определение факторных нагрузок методом главных факторов. Вращение пространства общих факторов. Статистическая оценка надежности решений методом факторного анализа. Построение сводного (интегрального) показателя качества сложной системы. / Ср /	7	22	ОПК-3, ПК -6	Л1.1, Л1.2, Л1.3, Л1.4, Л2.1, Л2.2, Л2.3, Л2.4, Л2.5, Л2.6, Л2.7
2.4	Тема "Факторный анализ". Модель ортогональных факторов. Определение факторных нагрузок методом главных факторов. Вращение пространства общих факторов. Статистическая оценка надежности решений методом факторного анализа. Построение сводного (интегрального) показателя качества сложной системы. Работа в RStudio, STADIA. / Лаб /	7	2	ОПК-3, ПК -6	Л1.1, Л1.2, Л1.3, Л1.4, Л2.1, Л2.2, Л2.3, Л2.4, Л2.5, Л2.6, Л2.7
2.5	Тема "Факторный анализ". Модель ортогональных факторов. Определение факторных нагрузок методом главных факторов. Вращение пространства общих факторов. Статистическая оценка надежности решений методом факторного анализа. Построение сводного (интегрального) показателя качества сложной системы. / Пр /	7	2	ОПК-3, ПК -6	Л1.1, Л1.2, Л1.3, Л1.4, Л2.1, Л2.2, Л2.3, Л2.4, Л2.5, Л2.6, Л2.7
2.6	Тема "Многомерное шкалирование". Многомерное шкалирование: алгоритм и примеры. / Лек /	7	2	ОПК-3, ПК -6	Л1.1, Л1.2, Л1.3, Л1.4, Л2.1, Л2.2, Л2.3, Л2.4, Л2.5, Л2.6, Л2.7
2.7	Тема "Комплексный многомерный анализ". Регрессия на главные компоненты/общие факторы. Кластерный анализ на главных компонентах/общих факторах. / Лек /	7	2	ОПК-3, ПК -6	Л1.1, Л1.2, Л1.3, Л1.4, Л2.1, Л2.2, Л2.3, Л2.4, Л2.5, Л2.6, Л2.7
2.8	Тема "Прикладной многомерный анализ". Решение практических задач с помощью инструментов многомерного статистического анализа (первичная обработка данных, корреляционно-регрессионный анализ, методы снижения размерности и классификации). / Пр /	7	2	ОПК-3, ПК -6	Л1.1, Л1.2, Л1.3, Л1.4, Л2.1, Л2.2, Л2.3, Л2.4, Л2.5, Л2.6, Л2.7
2.9	Тема "Прикладной многомерный анализ". Пример решения практической задачи с помощью инструментов многомерного статистического анализа (первичная обработка данных, корреляционно-регрессионный анализ, методы снижения размерности и классификации). / Лек /	7	4	ОПК-3, ПК -6	Л1.1, Л1.2, Л1.3, Л1.4, Л2.1, Л2.2, Л2.3, Л2.4, Л2.5, Л2.6, Л2.7
2.10	Тема "Снижение размерности исследуемых многомерных признаков". Метод главных компонент. Собственные векторы и собственные значения и их использование для получения матрицы весовых коэффициентов. Построение и интерпретация модели главных компонент. Работа в RStudio, STADIA. / Лаб /	7	4	ОПК-3, ПК -6	Л1.1, Л1.2, Л1.3, Л1.4, Л2.1, Л2.2, Л2.3, Л2.4, Л2.5, Л2.6, Л2.7
2.11	Тема "Факторный анализ". Модель ортогональных факторов. Определение факторных нагрузок методом главных факторов. Вращение пространства общих факторов. Статистическая оценка надежности решений методом факторного анализа. Построение сводного (интегрального) показателя качества сложной системы. / Ср /	7	64	ОПК-3, ПК -6	Л1.1, Л1.2, Л1.3, Л1.4, Л2.1, Л2.2, Л2.3, Л2.4, Л2.5, Л2.6, Л2.7
2.12	/ Экзамен /	7	9	ОПК-3, ПК -6	Л1.1, Л1.2, Л1.3, Л1.4, Л2.1, Л2.2, Л2.3, Л2.4, Л2.5, Л2.6, Л2.7

4. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

Структура и содержание фонда оценочных средств для проведения текущей и промежуточной аттестации представлены в Приложении 1 к рабочей программе дисциплины.

5. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

5.1. Основная литература

	Авторы,	Заглавие	Издательство, год	Колич-во
Л1.1	Ниворожкина Л. И.	Статистические методы анализа данных: учеб.	М.: РИО□, 2016	108
Л1.2	Ниворожкина Л. И., Арженовский С. В.	Многомерные статистические методы в экономике: учеб. для вузов	М.: Дашков и К, 2008	196

	Авторы,	Заглавие	Издательство, год	Колич-во
Л1.3	Афанасьев В. Н., Леушина Т. В., Лебедева Т., Цыпин А. П., Афанасьев В. Н.	Эконометрика: учебник	Оренбург: Оренбургский государственный университет, 2012	https://biblioclub.ru/index.php?page=book&id=260747 неограниченный доступ для зарегистрированных пользователей
Л1.4	Александровская, Ю. П.	Многомерный статистический анализ в экономике: учебное пособие	Казань: Казанский национальный исследовательский технологический университет, 2017	https://www.iprbookshop.ru/79330.html неограниченный доступ для зарегистрированных пользователей

5.2. Дополнительная литература

	Авторы,	Заглавие	Издательство, год	Колич-во
Л2.1	Арженовский С. В.	Применение многомерных методов анализа в оценке рисков с использованием ППП: метод. указания к лаборатор. занятиям	Ростов н/Д: Изд-во РГЭУ (РИНХ), 2015	95
Л2.2	Елисеева И. И.	Эконометрика: учеб. для бакалавриата и магистратуры	М.: Юрайт, 2016	59
Л2.3	Симчера В. М.	Методы многомерного анализа статистических данных: учеб. пособие для студентов, обучающихся по спец. "Финансы и кредит", "Бухгалт. учет, анализ и аудит", "Мировая экономика", "Налоги и налогообложение"	М.: Финансы и статистика, 2008	50
Л2.4	Кремер Н. Ш.	Теория вероятностей и математическая статистика: Учеб. для вузов	М.: ЮНИТИ-ДАНА, 2000	87
Л2.5		Журнал "Вопросы статистики"	,	1
Л2.6	Зехин В. А., Мхитарян В. С., Айвазян С. А.	Практикум по многомерным статистическим методам: учебное пособие	Москва: Московский государственный университет экономики, статистики и информатики, 2003	https://biblioclub.ru/index.php?page=book&id=90409 неограниченный доступ для зарегистрированных пользователей
Л2.7	Шорохова, И. С., Кисляк, И. В., Мариев, О. С.	Статистические методы анализа: учебное пособие	Екатеринбург: Уральский федеральный университет, ЭБС АСВ, 2015	https://www.iprbookshop.ru/65987.html неограниченный доступ для зарегистрированных пользователей

5.3 Профессиональные базы данных и информационные справочные системы

База данных Центрального банка РФ http://cbr.ru/hd_base/
База статистических данных https://rosstat.gov.ru/databases
Единая межведомственная информационно-статистическая система https://www.fedstat.ru/
Консультант +

5.4. Перечень программного обеспечения

Операционная система РЕД ОС
Libre Office
RStudio
Универсальный статистический пакет STADIA

5.5. Учебно-методические материалы для студентов с ограниченными возможностями здоровья

При необходимости по заявлению обучающегося с ограниченными возможностями здоровья учебно-методические материалы предоставляются в формах, адаптированных к ограничениям здоровья и восприятия информации. Для лиц с нарушениями зрения: в форме аудиофайла; в печатной форме увеличенным шрифтом. Для лиц с нарушениями слуха: в форме электронного документа; в печатной форме. Для лиц с нарушениями опорно-двигательного аппарата: в форме электронного документа; в печатной форме.

6. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Помещения для всех видов работ, предусмотренных учебным планом, укомплектованы необходимой специализированной учебной мебелью и техническими средствами обучения:
- столы, стулья;
- персональный компьютер / ноутбук (переносной);
- проектор;

- экран / интерактивная доска.
Лабораторные занятия проводятся в компьютерных классах, рабочие места в которых оборудованы необходимыми лицензионными и/или свободно распространяемыми программными средствами и выходом в Интернет.

7. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ

Методические указания по освоению дисциплины представлены в Приложении 2 к рабочей программе дисциплины.

Приложение 1

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

1. Описание показателей и критериев оценивания компетенций на различных этапах их формирования, описание шкал оценивания

1.1 Показатели и критерии оценивания компетенций:

ЗУН, составляющие компетенцию	Показатели оценивания	Критерии оценивания	Средства оценивания
ОПК-3: способность осознанно применять методы математической и дескриптивной статистики для анализа количественных данных, в том числе с применением необходимой вычислительной техники и стандартных компьютерных программ, содержательно интерпретировать полученные результаты, готовить статистические материалы для докладов, публикаций и других аналитических материалов			
Знания: методов математической и дескриптивной статистики для анализа данных	Формулирует ответы на поставленные вопросы коллоквиума и тестов в части методов регрессионного анализа и классификации	Полнота и содержательность ответа; умение приводить примеры. Верные ответы на коллоквиум и тестовые вопросы.	Коллоквиум (вопросы 1-23) Тестовые задания (1-15) Экзаменационные билеты (1-10)
Умения: содержательно интерпретировать результаты расчетов и обосновывать полученные при моделировании выводы; использовать современное программное обеспечение для построения многомерных моделей	Выполняет лабораторные задания, анализирует и интерпретирует полученные результаты. Формирует отчет по заданию к лабораторной работе с использованием стандартных пакетов прикладных программ	Полнота и правильность решений; обоснованность обращения к базам данных; содержательность выводов и интерпретации полученных результатов. Правильность использования методов обработки данных, корреляционно-регрессионного анализа при выполнении задания к лабораторной работе, качество анализа и интерпретации полученных результатов, обоснованность выводов; качество оформления.	Лабораторное задание (задания 1-4) Экзаменационные билеты (1-10)
Навыки: владения методами многомерного анализа с использованием специализированных программных средств	Выполняет лабораторные задания, анализирует и интерпретирует полученные результаты. Формирует отчет по заданию к лабораторной работе в части методов многомерного анализа с использованием стандартных пакетов прикладных программ	Полнота и правильность решений; обоснованность обращения к базам данных; содержательность выводов и интерпретации полученных результатов. Правильность использования методов обработки данных, стандартных прикладных программ, методов многомерного анализа при выполнении задания к лабораторной работе, качество анализа и интерпретации полученных результатов, обоснованность выводов; качество оформления	Лабораторное задание (задания 3-7) Экзаменационные билеты (1-10)
ПК-6: способность осуществлять поиск статистической информации, ее первичную обработку и подготовку для проведения аналитических исследований, в том числе с использованием технологий больших данных			
Знания: способов снижения размерности исследуемых многомерных признаков и отбора наиболее информативных показателей	Формулирует ответы на поставленные вопросы коллоквиума в части методов снижения размерности многомерных признаков, комплексного многомерного анализа. Формирует отчет по заданию к лабораторной работе с	Полнота и содержательность ответа; умение приводить примеры. Правильность использования методов обработки данных, снижения размерности многомерных признаков, комплексного многомерного анализа при выполнении задания к лабораторной работе, качество анализа и интерпретации полученных результатов, обоснованность выводов; качество оформления. Верные ответы на тестовые вопросы.	Коллоквиум (вопросы 24-38) Тестовые задания (16-20) Лабораторное задание (задания 5-7) Экзаменационные билеты (1-10)

	использованием стандартных пакетов прикладных программ		
Умения: критически оценивать полученные при моделировании результаты	Выполняет лабораторные задания, анализирует и интерпретирует полученные результаты.	Полнота и правильность решений; обоснованность обращения к базам данных; содержательность выводов и интерпретации полученных результатов. Правильность использования методов обработки данных, корреляционно-регрессионного анализа, методов снижения размерности многомерных признаков, комплексного многомерного анализа при выполнении задания к лабораторной работе, качество анализа и интерпретации полученных результатов, обоснованность выводов; качество оформления.	Лабораторное задание (задания 1-7) Экзаменационные билеты (1-10)
Навыки: владения инструментальными средствами обработки данных	Выполняет лабораторные задания, анализирует и интерпретирует полученные результаты. Формирует отчет по заданию к лабораторной работе с использованием стандартных пакетов прикладных программ	Полнота и правильность решений; обоснованность обращения к базам данных; содержательность выводов и интерпретации полученных результатов. Правильность использования методов обработки данных, корреляционно-регрессионного анализа, методов снижения размерности многомерных признаков, комплексного многомерного анализа при выполнении задания к лабораторной работе, качество анализа и интерпретации полученных результатов, обоснованность выводов; качество оформления.	Лабораторное задание (задания 1-7) Экзаменационные билеты (1-10)

1.2 Шкалы оценивания:

Текущий контроль успеваемости и промежуточная аттестация осуществляется в рамках накопительной балльно-рейтинговой системы в 100-балльной шкале:

84-100 баллов (оценка «отлично»)

67-83 баллов (оценка «хорошо»)

50-66 баллов (оценка «удовлетворительно»)

0-49 баллов (оценка «неудовлетворительно»)

2. Типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций в процессе освоения образовательной программы

Экзаменационные билеты

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 1

1. Содержание и основные этапы многомерного статистического анализа.
2. Параметрический случай классификации без обучения: расщепление смесей вероятностных распределений.
3. Задача.

Фирма изучает спрос на автомобили. Сформирована по опросам покупателей выборка по переменным: цена автомобиля, объем двигателя, расход бензина, безопасность, фирма производитель. Предложите многомерный статистический метод для сегментирования рынка автомобилей. Обоснуйте свое решение.

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 2

1. Постановка задачи корреляционного анализа многомерной генеральной совокупности.
2. Непараметрический случай классификации без обучения: кластерный анализ. Постановка задачи автоматической классификации.
3. Задача.

Получите дискриминантную функцию Фишера для следующих исходных данных:

$$X_1 = \begin{pmatrix} 3 & 7 \\ 2 & 4 \\ 4 & 7 \end{pmatrix}, X_2 = \begin{pmatrix} 6 & 9 \\ 5 & 7 \\ 4 & 8 \end{pmatrix}, \overline{x}_1 = \begin{pmatrix} 3 \\ 6 \end{pmatrix}, \overline{x}_2 = \begin{pmatrix} 5 \\ 8 \end{pmatrix}$$

и общая ковариационная матрица

$$\Sigma = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}.$$

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 3

1. Корреляционный анализ количественных признаков: множественные и частные коэффициенты корреляции.
2. Кластерный анализ: расстояние между объектами и меры близости объектов друг к другу.
3. Задача.

Получите дискриминантную функцию Фишера для следующих исходных данных:

$$X_1 = \begin{pmatrix} 3 & 7 \\ 2 & 5 \\ 4 & 7 \end{pmatrix}, X_2 = \begin{pmatrix} 6 & 9 \\ 4 & 7 \\ 3 & 8 \end{pmatrix}, \overline{x}_1 = \begin{pmatrix} 2 \\ 6 \end{pmatrix}, \overline{x}_2 = \begin{pmatrix} 5 \\ 7 \end{pmatrix}$$

и общая ковариационная матрица

$$\Sigma = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}.$$

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 4

1. Проверка статистических гипотез о параметрах многомерной нормально распределенной генеральной совокупности
2. Кластерный анализ: расстояние между классами объектов.
3. Задача.

Получены следующие результаты дискриминантного анализа

Discriminant Function Analysis Summary
No. of vars in model: 3; Grouping: Var4 (2 grps)
Wilks' Lambda: ,14454 approx. F (3,5)=9,8640 p< ,0153

N=9	**Wilks'**	**Partial**	**F-remove**	**p-level**	**Toler.**	**1-Toler.**
Производительность труда	0,181529	0,796246	1,279471	0,309315	0,687015	0,312985
Удельный вес потерь от брака	0,153153	0,943771	0,297895	0,608664	0,877639	0,122361
Фондоотдача	0,163325	0,884994	0,649754	0,456810	0,627656	0,372345

Проинтерпретируйте результаты и сделайте соответствующие выводы.

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 5

1. Ранговая корреляция: по Спирмену, Кендаллу.
2. Кластерный анализ: оценка качества разбиения объектов на классы.
3. Задача. Получены следующие результаты дискриминантного анализа

Classification Functions; grouping: Var4

	G_1:1 P=0.44	**G_2:2 P=0.56**
Производительность труда	9,1345	6,4151
Удельный вес потерь от брака	6,3900	11,0694
Фондоотдача	10,5584	3,3542
Constant	-54,8703	-25,1824

Posterior Probabilities (Lab 3)
Incorrect classifications are marked with *

Case	Observed	G_1:1	G_2:2	Case	Observed	G_1:1	G_2:2
1	G_1:1	0,999868	0,000132	**7**	G_2:2	0,009825	0,990175
2	G_1:1	0,999568	0,000432	**8**	G_2:2	0,010251	0,989749
3	G_1:1	0,999740	0,000260	**9**	G_2:2	0,000000	1,000000
4	G_1:1	0,999989	0,000011	**10**	---	0,000492	0,999508
5	G_2:2	0,000478	0,999522	**11**	---	0,000935	0,999065
6	G_2:2	0,000000	1,000000	**12**	---	1,000000	0,000000

Проинтерпретируйте результаты и сделайте соответствующие выводы.

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 6

1. Корреляция категоризованных переменных: таблицы сопряженности и меры степени тесноты статистической связи.
2. Кластерный анализ: принцип построения агломеративных иерархических процедур классификации.
3. Задача.

Получены следующие результаты факторного анализа

X_6 - удельный вес покупных изделий; X_{11} - среднегодовая численность ППП; X_{12} - среднегодовая стоимость ОПФ; X_{14} - фондовооруженность труда; X_{15} - оборачиваемость нормируемых оборотных средств; X_{17} - непроизводственные расходы.

Factor Loadings (Varimax normalized)
Extraction: Principal components
(Marked loadings are > ,700000)

Variables	Factor 1	Factor 2	Factor 3
X6	0,088887	0,002317	0,902880
X11	0,755472	-0,010037	0,395243
X12	0,957894	-0,086188	0,164689
X14	0,760477	-0,176215	-0,350881

X15	-0,219036	0,858033	-0,165881
X17	0,023075	0,888763	0,168818
Expl.Var	2,123035	1,564707	1,177664
Prp.Totl	0,353839	0,260785	0,196277

Проинтерпретируйте результаты и сделайте соответствующие выводы.

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 7

1. Многомерная классификация: постановка задачи, основные определения. Классификации с обучением и без обучения.
2. Кластерный анализ: последовательные кластер-процедуры, метод *k*-средних.
3. Задача.

По данным о 5 домохозяйствах провести компонентный анализ на основе показателей удельного веса доходов, не связанных с основной работой, в общей сумме доходов x_1 и удельного веса расходов на питание x_2, если ковариационная матрица $\Sigma = \begin{bmatrix} 1 & -1 \\ -1 & 5 \end{bmatrix}$.

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 8

1. Многомерная классификация: оптимальная (байесовская) процедура классификации.
2. Снижение размерности многомерных признаков: метод главных компонент.
3. Задача.

Для данных по 138 индивидам выполните корреляционный анализ и сделайте выводы.

Доход	Удовлетворение работой	
	Полное неудовлетворение	Полное удовлетворение
< $250	42	27
> $500	7	62

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 9

1. Параметрический дискриминантный анализ в случае нормальных классов. Линейная дискриминантная функция Фишера.
2. Алгоритм вычисления главных компонент.
3. Задача.

Пусть $\boldsymbol{X}$ имеет многомерное нормальное распределение $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$,

где $\boldsymbol{\mu}^T = [1, -1, 2]$ и $\Sigma = \begin{bmatrix} 4 & 0 & -1 \\ 0 & 5 & 0 \\ -1 & 0 & 2 \end{bmatrix}$.

Какие из следующих случайных величин являются независимыми? Объясните.

а). X_1 и X_2.

б). X_1 и X_3.

в). (X_1, X_3) и X_2.

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 10

1. Алгоритм дискриминантного анализа в случае двух нормальных классов.
2. Главные компоненты многомерной нормально распределенной совокупности. Главные компоненты стандартизованных переменных.
3. Задача.

Пусть переменные x_1 и x_2 измерены на четырех объектах A, B, C и D:

Объекты	A	B	C	D
x_1	5	1	-1	3
x_2	4	-2	1	1

Необходимо классифицировать объекты на две группы методом k-средних.

Критерии оценивания:

Экзаменационный билет оценивается максимально в 100 баллов:

- 84-100 баллов (оценка «отлично»)
- 67-83 баллов (оценка «хорошо»)
- 50-66 баллов (оценка «удовлетворительно»)
- 0-49 баллов (оценка «неудовлетворительно»)

Задача оценивается максимально в 50 баллов. Критерии оценивания задачи:

- 42-50 баллов. Задача решена в полном объеме, выбраны верные инструментальные методы и приемы решения, проведены верные расчеты, сделан полный, содержательный вывод по результатам проведенных расчетов.
- 34-41 балла. Задача решена в полном объеме с небольшими погрешностями, выбраны верные инструментальные методы и приемы решения, проведены верные расчеты, сделан полный, содержательный вывод по результатам проведенных расчетов, в расчетах и выводах содержатся незначительные ошибки.
- 25-33 балла. Задача решена частично, частично выбраны верные инструментальные методы и приемы решения, проведены частичные расчеты, сделан вывод по результатам проведенных расчетов с отдельными, незначительными погрешностями.
- 0-24 балла. Задача не решена или решена частично, частично выбраны необходимые инструментальные методы и приемы решения, расчеты не проведены или проведены частично, вывод по результатам проведенных расчетов не сделан или ошибочен.

Каждый вопрос оценивается отдельно, максимально в 25 баллов. Максимальная общая сумма – 50 баллов. Критерии оценивания отдельного вопроса:

- 21-25 баллов. Ответ на вопрос верный; продемонстрировано наличие глубоких исчерпывающих / твердых и достаточно полных знаний, грамотное и логически стройное изложение материала при ответе.
- 17-20 балла. Ответ на вопрос верный, но с отдельными погрешностями и ошибками, уверенно исправленными после дополнительных вопросов; продемонстрировано наличие глубоких исчерпывающих / твердых и достаточно полных знаний, грамотное и логически стройное изложение материала при ответе.
- 13-16 балла. Ответ на вопрос частично верен, продемонстрирована некоторая неточность ответов на дополнительные и наводящие вопросы.
- 0-12 балла. Ответ на вопрос не верен, продемонстрирована неуверенность и неточность ответов на дополнительные и наводящие вопросы.

Вопросы для коллоквиума

Раздел 1 «Регрессионный анализ и классификация»

1. В чем особенности МСА?
2. Основные этапы МСА.
3. Формы представления данных, используемых в МСА.
4. Понятие признакового пространства. Приведите примеры.
5. Виды зависимостей исследуемых многомерными статистическими методами.
6. Кратко поясните логическую схему построения статистического критерия для проверки однородности нормальной выборочной совокупности.
7. Каковы основные характеристики многомерной случайной величины?
8. Кратко поясните особенности множественного коэффициента корреляции, частного коэффициента корреляции.
9. В чем особенности измерения степени тесноты статистической связи между категоризованными переменными?
10. В чем особенности дискриминантного анализа?
11. Как определяется качество дискриминантных функций?
12. В чем суть непараметрического дискриминантного анализа?
13. Приведите пример (графически), когда дискриминантная функция будет нелинейной.
14. Как определяется количество дискриминантных функций?
15. Суть оптимального байесовского правила классификации.
16. Какие задачи решаются с помощью кластерного анализа?
17. Какие меры сходства используются при проведении кластерного анализа?
18. Особенности параметрической классификации без обучения.
19. Какие меры расстояний между объектами используются в кластерном анализе?
20. Как оценивается качество полученного разбиения на классы?
21. Принцип "работы" иерархических процедур классификации.
22. Особенности метода Уорда.
23. Алгоритм метода *k*-средних.

Раздел 2 «Снижение размерности. Комплексный многомерный анализ»

24. Какие задачи решаются с помощью компонентного анализа?
25. Как находятся главные компоненты?
26. Как интерпретируются результаты компонентного анализа?
27. В чем суть факторного анализа? Какие виды факторного анализа используются на практике?
28. Как определить достаточное число факторов для характеристики изучаемого явления или процесса?
29. Модель ортогональных факторов.
30. Метод главных факторов.
31. Вращение системы факторов.
32. В чем отличие факторного анализа от компонентного?
33. Как проверить надежность результатов факторного анализа?
34. В чем суть задачи многомерного шкалирования?
35. Как решается задача метрического шкалирования по Торгерсону?
36. В чем отличие метрического шкалирования от неметрического?
37. Как строится матрица различий объектов?
38. Каков алгоритм решения задачи неметрического шаклирования?

Критерии оценивания:
Каждый вопрос оценивается отдельно, максимально в 2 балла. Максимальное количество вопросов за семестр – 16 вопросов. Максимальная общая сумма – 32 балла. Критерии оценивания отдельного вопроса:

- 1,7-2 балла. Ответ на вопрос верный; продемонстрировано наличие глубоких исчерпывающих / твердых и достаточно полных знаний, грамотное и логически стройное изложение материала при ответе.
- 1,4-1,6 балла. Ответ на вопрос верный, но с отдельными погрешностями и ошибками, уверенно исправленными после дополнительных вопросов; продемонстрировано наличие глубоких исчерпывающих / твердых и достаточно полных знаний, грамотное и логически стройное изложение материала при ответе.
- 1-1,3 балла. Ответ на вопрос частично верен, продемонстрирована некоторая неточность ответов на дополнительные и наводящие вопросы.
- 0-0,9 балла. Ответ на вопрос не верен, продемонстрирована неуверенность и неточность ответов на дополнительные и наводящие вопросы.

Тестовые задания

(один верный ответ)

1. Мультиколлинеарность факторных переменных − это:

А. Отсутствие связи между факторными переменными
Б. Тесная связь между факторными переменными
В. Многомерная связь между факторными переменными
Г. Множественная регрессионная модель.

2. Дискриминантный анализ – совокупность статистических методов многомерной классификации объектов при наличии:

А. Средних значений
Б. «Обучающих» выборок
В. «Обычных» выборок
Г. Коэффициентов корреляции

3. При использовании метода *k*-средних для классификации многомерных объектов в состав кластера включаются новые объекты таким образом, чтобы внутриклассовая дисперсия:

А. Значение дисперсии может быть любым
Б. Стремилась к максимуму
В. Оставалась постоянной
Г. Стремилась к минимуму

4. Какая матрица в факторном анализе используется для определения матрицы значений факторов?

А. Корреляционная матрица исходных наблюдений
Б. Единичная матрица
В. Обратная матрица исходных наблюдений
Г. Необходимы дополнительные сведения

5. Методом *k*-средних необходимо множество объектов разделить на три группы. Сколько кластеров будет образовано по завершении кластеризации?

А. 3
Б. 1
В. 2
Г. 4

6. При кластерном анализе были пронормированы значения признаков. Укажите значения математического ожидания и дисперсии этих признаков

А. Математическое ожидание равно 1, дисперсия равна 1
Б. Математическое ожидание равно 1, дисперсия равна 0
В. Математическое ожидание равно 0, дисперсия равна 1
Г. Математическое ожидание равно 0, дисперсия равна 0

7. В каком соотношении должны быть число пар дискриминантных переменных p и количество наблюдений n?

А. $p > n$
Б. $p < n$
В. $p = n$
Г. В любом соотношении

8. Что представляет собой доверительная область для математического ожидания k-мерного нормально распределенного вектора?

А. k-мерный эллипсоид
Б. k-мерный вектор
В. k-мерный параллелепипед
Г. Необходимы дополнительные сведения

9. В методе главных компонент первая главная компонента соответствует направлению, вдоль которого дисперсия векторов исходного набора

А. Минимальна
Б. Постоянна
В. Максимальна
Г. Любая

10. Метод главных компонент предназначен

А. Для определения значимых коэффициентов регрессии
Б. Для уменьшения размерности модели
В. Для увеличения размерности модели
Г. Для фиксирования размерности модели

11. Метод Варда предназначен для …

А. Снижения размерности признакового пространства
Б. Многомерного шкалирования
В. Классификации без обучения
Г. Нет такого метода.

12. К какой группе методов относится решение задачи расщепления смеси распределений?

А. Параметрический случай классификации без обучения
Б. Непараметрический случай классификации без обучения
В. Параметрический случай классификации с обучением
Г. Непараметрический случай классификации с обучением

13. Метод N ближайших соседей относится к …

А. Параметрический случай классификации без обучения
Б. Непараметрический случай классификации без обучения
В. Параметрический случай классификации с обучением
Г. Непараметрический случай классификации с обучением

14. К методу многомерного шкалирования относится …

А. Подход Торгенсона
Б. Определение сходства объектов
В. Подход Фишера

Г. Метод Краскала.

15. Для изучения тесноты связи между несколькими (больше двух) ранжировками используется коэффициент:

А. Спирмена

Б. Пирсона

В. Фишера

Г. Кендалла (конкордации)

16. Методом *k*-средних необходимо множество объектов разделить на три группы. Сколько кластеров будет образовано по завершении кластеризации? _____.

17. Имеются данные о 150 абитуриентах, сдававших вступительный экзамен в магистратуру некоторого экономического факультета:

Y – количество баллов за вступительный экзамен по экономической теории.

D – фиктивная переменная равная единице, если соответствующий абитуриент посещал подготовительные курсы для поступающих, и равная нулю в противном случае.

EF – фиктивная переменная равная единице, если соответствующий абитуриент является выпускником бакалавриата данного экономического факультета, и равная нулю в противном случае.

Используя эти данные, исследователь оценил параметры линейной регрессионной модели:

$$\widehat{Y_i} = \underset{(0,1)}{20} + \underset{(4,5)}{30}\, EF_i - \underset{(1,3)}{10}\, D_i + \underset{(1,4)}{15}\, D_i \cdot EF_i$$

В соответствии с полученными результатами, определите, какое количество баллов в среднем получает абитуриент, который заканчивал бакалавриат данного экономического факультета и не посещал курсы? _____ (*округлите до целых, если требуется*)**.**

18. Исследуется зависимость среднедушевого потребления алкоголя по странам мира от различных факторов.

Модель 1:

$$ALCO_i=\beta_1+\beta_2 GDP_i+\beta_3 MUSL_i+\beta_4 BUDD_i+\beta_5 HINDU_i+\varepsilon_i,$$

где $ALCO_i$ – среднедушевое потребление чистого спирта на человека (л), GDP_i – ВВП на душу населения (долларов США), $MUSL_i$, $BUDD_i$, $HINDU_i$ – доли населения исповедующего, соответственно, мусульманство, буддизм и индуизм (в % от общей численности населения). В ходе МНК-оценивания модели на основе данных о 50 странах получены следующие результаты: сумма квадратов остатков $RSS=200$, объясненная сумма квадратов $ESS=300$.

Также для проверки гипотезы о том, что религия не оказывает существенного влияния на потребление алкоголя, были оценены параметры второй модели:

Модель №2:

$$ALCO_i=\beta_1+\beta_2 GDP_i+\varepsilon_i.$$

Во второй модели, по сравнению с первой, значение *ESS* увеличился на 100.

Сколько составит скорректированный R^2 во второй модели? _____ (*округлите до 4-х знаков после запятой*).

19. Моделируется прибыль фирм в некоторой отрасли экономики России. y_i - прибыль *i*-ой фирмы (млн. руб.), d_i - фиктивная переменная, которая принимает значение 1, если *i*-ая располагается в Москве и значение 0 в противном случае, $x_i^{(1)}$, $x_i^{(2)}$, $x_i^{(3)}$ - некоторые количественные переменные. По 30 наблюдениям было оценено следующее уравнение регрессии (в скобках указаны стандартные отклонения оценок коэффициентов):

$$\widehat{\ln(y_i)} = \underset{(24,0)}{5,0} - \underset{(0,3)}{0,8} \cdot x_i^{(1)} + \underset{(0,02)}{0,07} \cdot x_i^{(2)} + \underset{(0,01)}{0,03} \cdot x_i^{(3)} - \underset{(0,5)}{1,0} \cdot d_i,\ R^2 = 0,8$$

Проверьте на 5% уровне значимости гипотезу о значимости влияния местоположения на прибыль фирмы. В ответе укажите значение тестовой статистики (*округлить до целых*) и результат проверки гипотезы H_0 *(вывод)*.

20. По 100 наблюдениям исследовательница Глафира оценила модель зависимости заработной платы wagei (в тысячах рублей) от длительности обучения educ.yearsi (годы) и дамми на тип учебного заведения (высшее образование или нет, educ.typei). Оценённая модель имеет вид:

$$\widehat{wage}_i = 15 + 20 \cdot educ.years_i + 30 \cdot educ.type_i$$

Также у Глафиры есть данные ESS=110, TSS=250. Глафира решила добавить в модель образование родителей fath.educi и moth.educi (годы), после чего ESSnew=205.

На уровне значимости 5% проверяя гипотезу о влиянии длительности обучения отца и матери на заработную плату их ребёнка, определите, чему равно наблюдаемое значение F-статистики _____ (*ответ округлите с точностью до целых*).

Критерии оценивания:
Каждое тестовое задание оценивается отдельно, максимально в 2 балла. Максимальная общая сумма – 40 баллов. Критерии оценивания отдельного задания:

- 2 балла. Ответ верен.
- 0 баллов. Ответ не верен.

Лабораторные задания

Раздел 1 «Регрессионный анализ и классификация»

Лабораторное задание 1. «Первичная обработка данных»

Данные об издержках на транспортировку продуктов питания 10-ти фирм, занимающихся снабжением, представлены в таблице. Необходимо построить 95% доверительные интервалы для средних значений трех имеющихся признаков в предположении, что они имеют нормальное распределение, а также доверительную область для первых двух признаков.

№ п.п.	Затраты топлива, л, X_1	Затраты на ремонт, у.е., X_2	Капитал фирмы, тыс. у.е., X_3
1	16,44	12,43	11,23
2	7,19	2,70	3,92
3	9,92	1,35	9,75
4	4,24	5,78	7,78
5	11,20	5,05	10,67
6	14,25	5,78	9,88
7	13,50	10,98	10,60
8	13,32	14,27	9,45
9	29,11	15,09	3,28
10	12,68	7,61	10,23

Лабораторное задание 2. «Корреляционно-регрессионный анализ»

Загрузите данные из файла *Lab1.csv* (*https://disk.yandex.ru/d/QnfUkoAkWtD1JQ*). Файл содержит подвыборку 4794 наблюдений из массива msm.sta по индивидам. Описание переменных: *lw* – логарифм заработной платы, *edu* – число лет образования, *expr* – опыт работы, *expr2* – квадрат переменной *expr*. Просмотрите дескриптивные статистики переменных в выборке. Сделайте выводы об эмпирическом распределении каждой

переменной. Постройте корреляционную матрицу переменных. Постройте уравнение линейной множественной регрессии переменной заработной платы *lw* от переменных *edu*, *expr* и *expr2*. Сделайте выводы по результатам всех расчетов. Постройте доверительные интервалы для коэффициентов регрессии.

Лабораторное задание 3. «Дискриминантный анализ»

В файле *firm.csv (https://disk.yandex.ru/d/QnfUkoAkWtD1JQ)* имеются данные по 12 предприятиям, характеризуемых тремя экономическими показателями: labor – производительность труда, defect – удельный вес потерь от брака (%) и fund – фондоотдача активной части основных производственных фондов. Из этих предприятий выделены две обучающие выборки (переменная firm), первая из которых включает 4 предприятия группы A, а вторая 5 предприятий группы B. Требуется классифицировать в одну из групп A или B оставшиеся три предприятия.

Лабораторное задание 4. «Кластерный анализ»

Загрузите данные из файла *Lab3.csv (https://disk.yandex.ru/d/QnfUkoAkWtD1JQ)*. Файл содержит подвыборку из массива msm.sta по индивидам. Описание переменных: lw – логарифм заработной платы, edu – число лет образования, nhh – число членов домохозяйства, dd – доля доходов главы домохозяйства в семейном бюджете, age – возраст, pm – процент заработков, который дает основная работа. Необходимо классифицировать наблюдения.

Раздел 2 «Снижение размерности. Комплексный многомерный анализ»

Лабораторное задание 5. «Снижение размерности исследуемых многомерных признаков»

Загрузите данные из файла *Lab3.csv (https://disk.yandex.ru/d/QnfUkoAkWtD1JQ)*. Файл содержит подвыборку из массива msm.sta по индивидам. Описание переменных: lw – логарифм заработной платы, edu – число лет образования, nhh – число членов домохозяйства, dd – доля доходов в семейном бюджете, age – возраст, pm – процент заработков, который дает основная работа. Необходимо провести компонентный анализ данных, а затем классифицировать наблюдения.

Лабораторное задание 6. «Факторный анализ»

Загрузите данные из файла *Lab3.csv (https://disk.yandex.ru/d/QnfUkoAkWtD1JQ)*. Выполните факторный анализ по имеющимся выборочным данным.

Лабораторное задание 7. «Прикладной многомерный анализ»

Деятельность предприятий характеризуется следующими показателями

№ п.п.	Трудоемкость единицы продукции, x_1	Удельный вес покупных изделий, x_2	Коэффициент сменности оборудования, x_3	Индекс снижения себестоимости продукции, q
1	0,51	0,20	1,47	21,9
2	0,36	0,64	1,27	48,4
3	0,23	0,42	1,51	173,5
4	0,26	0,27	1,46	74,1
5	0,27	0,37	1,27	68,6
6	0,29	0,38	1,43	60,8
7	0,01	0,35	1,50	355,6
8	0,02	0,42	1,35	264,8

9	0,18	0,32	1,41	526,6
10	0,25	0,33	1,47	118,6

Приняв за результативный признак q, построить уравнение регрессии на главные компоненты, наиболее тесно связанные с q. Дать экономическую интерпретацию результатов.

Критерии оценивания:

Каждое лабораторное задание оценивается отдельно, максимально в 4 балла. Максимальная общая оценка – 28 баллов. Критерии оценивания:

- 2,6-3 баллов. Задание решено в полном объеме, выбраны верные инструментальные методы и приемы решения, проведены верные расчеты, сделан полный, содержательный вывод по результатам проведенных расчетов.
- 2,1-2,5 балла. Задание решено в полном объеме с небольшими погрешностями, выбраны верные инструментальные методы и приемы решения, проведены верные расчеты, сделан полный, содержательный вывод по результатам проведенных расчетов, в расчетах и выводах содержатся незначительные ошибки.
- 1,5-2 балла. Задание решено частично, частично выбраны верные инструментальные методы и приемы решения, проведены частичные расчеты, сделан вывод по результатам проведенных расчетов с отдельными, незначительными погрешностями.
- 0-1,4 балла. Задание не решено или решено частично, частично выбраны необходимые инструментальные методы и приемы решения, расчеты не проведены или проведены частично, вывод по результатам проведенных расчетов не сделан или ошибочен.

3. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций

Процедуры оценивания включают в себя текущий контроль и промежуточную аттестацию.

Текущий контроль успеваемости проводится с использованием оценочных средств, представленных в п. 2 данного приложения. Результаты текущего контроля доводятся до сведения студентов до промежуточной аттестации.

Промежуточная аттестация проводится в форме экзамена. Экзамен проводится по расписанию промежуточной аттестации в письменном виде. Количество теоретических вопросов в задании – 2, количество задач – 1. Проверка ответов и объявление результатов производится в день экзамена. Результаты аттестации заносятся в ведомость и зачетную книжку студента. Студенты, не прошедшие промежуточную аттестацию по графику сессии, должны ликвидировать задолженность в установленном порядке.

Приложение 2

МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ

Учебным планом предусмотрены следующие виды занятий:

- лекции;
- практические занятия;
- лабораторные занятия.

В ходе лекционных занятий рассматриваются основные теоретические положения и понятия, методы многомерного анализа данных, даются рекомендации для самостоятельной работы и подготовке к практическим и лабораторным занятиям.

В ходе практических и лабораторных занятий углубляются и закрепляются знания студентов по ряду рассмотренных на лекциях вопросов, развиваются навыки многомерного статистического анализа больших массивов данных в пакетах прикладных программ, а также самостоятельной работы и работы в коллективе.

При подготовке к практическим и лабораторным занятиям каждый студент должен:

1) изучить рекомендованную учебную литературу;
2) изучить конспекты лекций;
3) подготовить ответы на все вопросы по изучаемой теме.

В процессе подготовки к практическим и лабораторным занятиям студенты могут воспользоваться консультациями преподавателя.

Вопросы, не рассмотренные на лекциях, практических и лабораторных занятиях, должны быть изучены студентами в ходе самостоятельной работы. Контроль самостоятельной работы студентов над учебной программой курса осуществляется в ходе занятий методом коллоквиума. В ходе самостоятельной работы каждый студент обязан прочитать основную и, по возможности, дополнительную литературу по изучаемой теме, дополнить конспекты лекций недостающим материалом, выписками из рекомендованных первоисточников. Выделить непонятные термины, найти их значение в энциклопедических словарях.

Для подготовки к занятиям, текущему контролю и промежуточной аттестации студенты могут воспользоваться электронно-библиотечными системами. Также обучающиеся могут взять на дом необходимую литературу на абонементе университетской библиотеки или воспользоваться читальными залами.

Методические рекомендации по выполнению лабораторных заданий

Лабораторное задание 1. «Первичная обработка данных»

Данные об издержках на транспортировку продуктов питания 10-ти фирм, занимающихся снабжением, представлены в таблице. Необходимо построить 95% доверительные интервалы для средних значений трех имеющихся признаков в предположении, что они имеют нормальное распределение, а также доверительную область для первых двух признаков.

Вычислим вектор средних значений и ковариационную матрицу для трехмерного нормального распределения.

№ п.п.	Затраты топлива, л, X_1	Затраты на ремонт, у.е., X_2	Капитал фирмы, тыс. у.е., X_3
1	16,44	12,43	11,23
2	7,19	2,70	3,92
3	9,92	1,35	9,75
4	4,24	5,78	7,78
5	11,20	5,05	10,67
6	14,25	5,78	9,88
7	13,50	10,98	10,60
8	13,32	14,27	9,45
9	29,11	15,09	3,28
10	12,68	7,61	10,23

Применяя формулу $\bar{X}_l = \frac{1}{n}\sum_{i=1}^{n} X_{il}, l = 1,2,3,$ получим для, например, X_1: $\bar{X}_1$ =(16,44+7,19+9,92+4,24+…+12,68)/10=13,185. Аналогично получаем $\bar{X}_2$=8,104, $\bar{X}_3$=8,679 и вектор средних имеет вид: $\bar{\mathbf{X}}$ = [13,185, 8,104, 8,679].

Элементы ковариационной матрицы вычисляем по формуле $s_{lj} = \frac{1}{10}\sum_{i=1}^{10}\left(X_{il} - \bar{X}_l\right)\left(X_{ij} - \bar{X}_j\right), l, j = 1,2,3.$ Например, для i=j=1: s_{11} = [(16,44–13,185)^2 + (7,19–13,185)^2 +…+ (12,68–13,185)^2]/10 = 39,626 и далее, для i=1, j=2: s_{12} = [(16,44–13,185)(12,43–8,104) +(7,19–13,185)(2,7–8,104) + …+ (12,68–13,185)(7,61–8,104] / 10 = 20,615 и т.д. Ковариационная матрица имеет вид: $S = \begin{bmatrix} 39,626 & 20,615 & -4,735 \\ 20,615 & 20,9 & -0,547 \\ -4,735 & -0,547 & 7,236 \end{bmatrix}$ или несмещенная оценка ковариационной матрицы: $\hat{S} = \frac{10}{10-1}\begin{bmatrix} 39,626 & 20,615 & -4,735 \\ 20,615 & 20,9 & -0,547 \\ -4,735 & -0,547 & 7,236 \end{bmatrix} = \begin{bmatrix} 44,029 & 22,905 & -5,261 \\ 22,905 & 23,222 & -0,608 \\ -5,261 & -0,608 & 8,039 \end{bmatrix}$.

Вычислим $T^2_{\alpha;(p,n-p)} = \frac{p(n-1)}{n-p}F_{\alpha;(p,n-p)} = \frac{3(10-1)}{10-3}F_{0,05;(3,7)} = 3,857 \cdot 4,347 = 16,767.$

Тогда получим доверительные интервалы для каждой из средних:

$13,185 - \sqrt{16,767} * \sqrt{\frac{44,029}{10}} \le \mu_1 \le 13,185 + \sqrt{16,767} * \sqrt{\frac{44,029}{10}}$ и 4,593≤μ_1 ≤21,777,

$8,104 - \sqrt{16,767} * \sqrt{\frac{23,222}{10}} \le \mu_2 \le 8,104 + \sqrt{16,767} * \sqrt{\frac{23,222}{10}}$ и 1,864≤μ_2 ≤14,344,

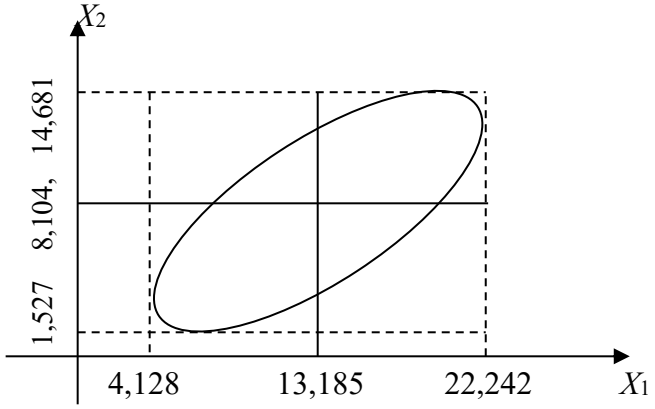
$8,679 - \sqrt{16,767} * \sqrt{\frac{8,933}{10}} \le \mu_3 \le 8,679 + \sqrt{16,767} * \sqrt{\frac{8,933}{10}}$ и 5,007≤μ_3 ≤12,351.

Для первых двух признаков доверительная область строится с учетом формулы $n(\bar{\mathbf{X}}-\boldsymbol{\mu}_0)^T\mathbf{S}^{-1}(\bar{\mathbf{X}}-\boldsymbol{\mu}_0)=\dfrac{p(n-1)}{n-p}F_{\alpha;(p,n-p)}$. Имеем:

$$10\left[13{,}185-\mu_1,\ 8{,}104-\mu_2\right]^T\begin{bmatrix}48{,}921 & 22{,}905\\ 22{,}905 & 25{,}803\end{bmatrix}^{-1}\begin{bmatrix}13{,}185-\mu_1\\ 8{,}104-\mu_2\end{bmatrix}=$$

$$=0{,}35(13{,}185-\mu_1)^2+0{,}66(8{,}104-\mu_2)^2-2\cdot 0{,}31(13{,}185-\mu_1)(8{,}104-\mu_2)\le 16{,}767.$$

Получившаяся доверительная область показана на рис. в виде эллипса.



С целью оценки воздействия состояния окружающей среды на здоровье населения собраны данные* по двум Федеральным округам. Необходимо проверить при $\alpha = 0{,}05$ существенность различий двух округов по выбранным двум показателям.

№ региона	Северо-Западный федеральн. округ		№ региона	Центральный федеральн. округ	
	Число умерших на 1000 чел. населения	Заболеваемость на 1000 чел. населения новообразованиями		Число умерших на 1000 чел. населения	Заболеваемость на 1000 чел. населения новообразованиями
1	16,6	9,6	1	16,1	12,3
2	12,5	7,6	2	17,6	8,5
3	15,3	8,1	3	19,2	10
4	17,1	7,6	4	18,2	8
5	16,3	7,3	5	20,2	10,9
6	20,1	6,4	6	18,1	7,5
7	11,6	10,3	7	19,2	8,2
8	20,9	8,7	8	18,2	6,8
9	22,5	7,4	9	17	10
10	16,4	8,6	10	18,1	9,3
11			11	17,7	10,9
12			12	19,7	9,8
			13	19,9	6,7
			14	18,2	8,9
			15	21,9	8
			16	21,5	9,2
			17	19,5	11,3
			18	15,6	10,3
$\bar{X}$	16,93	8,16	$\bar{X}$	18,66	9,26

1. Определим векторы средних и ковариационные матрицы.

$\bar{X}_1$=[16,93, 8,16]; $\bar{X}_2$=[18,66, 9,26];

$$\mathbf{S}_1=\begin{bmatrix}13{,}38 & -1{,}86\\ -1{,}86 & 1{,}50\end{bmatrix},\ \mathbf{S}_2=\begin{bmatrix}2{,}93 & -0{,}76\\ -0{,}76 & 2{,}62\end{bmatrix}.$$

Объединенная ковариационная матрица рассчитывается так:

* Данные за 2001 г. по: Регионы России. М.: Госкомстат, 2002.

$$\mathbf{S}_{12}=\frac{1}{10+18-2}((10-1)\mathbf{S}_1+(18-1)\mathbf{S}_2)=\begin{bmatrix}6,54 & -1,14\\ -1,14 & 2,24\end{bmatrix}.$$

Найдем обратную матрицу $\mathbf{S}_{12}^{-1}=\begin{bmatrix}0,17 & 0,09\\ 0,09 & 0,49\end{bmatrix}$.

2. Рассчитаем фактическое значение критерия Хоттелинга:

$$T^2=\frac{n_1n_2}{n_1+n_2}\left(\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2\right)^T\mathbf{S}_{12}^{-1}\left(\bar{\mathbf{X}}_1-\bar{\mathbf{X}}_2\right)=$$

$$=\frac{10\cdot18}{10+18}[16,93-18,66,\quad 8,16-9,26]^T\begin{bmatrix}0,17 & 0,09\\ 0,09 & 0,49\end{bmatrix}[16,93-18,66,\quad 8,16-9,26]=1,42.$$

3. Найдем критическое значение критерия Хоттелинга и сравним с фактическим

$$T_{кр}^2=\frac{p(n_X+n_Y-2)}{n_X+n_Y-p-1}F_{\alpha;(p;n_X+n_Y-p-1)}=\frac{2(10+18-2)}{10+18-2-1}F_{0,05;(2,25)}=7,04.$$

Поскольку критическое значение больше фактического, то гипотеза о равенстве векторов средних значений признаков для двух округов не может быть отвергнута.

4. Проверим с помощью критерия Бартлетта равенство ковариационных матриц двух выборок.

Рассчитаем $b=1-\left(\frac{1}{n_X-1}+\frac{1}{n_Y-1}-\frac{1}{n_X+n_Y-2}\right)\frac{2p^2+3p-1}{6(p+1)}$=0,905,

$a=(n_X+n_Y-2)\ln\det\mathbf{S}_{XY}-\left[(n_X-1)\ln\det\mathbf{S}_X+(n_Y-1)\ln\det\mathbf{S}_Y\right]$=8,835.

W = *ab*=7,99. По таблице находим $\chi^2_{0,05;3}=7,81$. Таким образом, фактическое значение критерия Бартлетта больше табличного и гипотеза о равенстве ковариационных матриц и, следовательно, однородности двух федеральных округов по выбранным двум признакам, отвергается с надежностью 95%.

Проверьте полученные результаты при помощи пакета RStudio.

Лабораторное задание 2. «Корреляционно-регрессионный анализ»

1. Загрузите данные из файла Lab1.csv (https://disk.yandex.ru/d/QnfUkoAkWtD1JQ). Файл содержит подвыборку 4794 наблюдений из массива msm.sta по индивидам. Описание переменных: lw – логарифм заработной платы, edu – число лет образования, expr – опыт работы, expr2 – квадрат переменной expr.

2. Просмотрите дескриптивные статистики переменных в выборке: среднее значение, медиану, моду, дисперсию, эксцесс и асимметрию. Сделайте выводы.

Постройте диаграммы распределения для каждой из переменных, а также сравните их с кривой нормального распределения. Рассчитайте значение статистики Колмогорова-Смирнова для проверки соответствия эмпирического распределения нормальному закону распределения. Сделайте выводы об эмпирическом распределении каждой переменной.

3. Постройте корреляционную матрицу переменных. Оцените значимость коэффициентов корреляции. Прокомментируйте результаты.

Рассчитайте частные коэффициенты корреляции. Сравните их с парными коэффициентами корреляции и сделайте выводы.

4. Постройте уравнение линейной множественной регрессии переменной заработной платы *lw* от переменных *edu*, *expr* и *expr2*.

Проанализируйте остатки регрессии, выполнив необходимые тесты на автокорреляцию и гетероскедастичность.

Сделайте выводы по результатам всех расчетов, выполненных в этом пункте задания.

5. Постройте доверительные интервалы для коэффициентов регрессии.

6. Сохраните скрипт и прикрепите его к отчету.

Лабораторное задание 3. «Дискриминантный анализ»

1. В файле *firm.csv* (https://disk.yandex.ru/d/QnfUkoAkWtD1JQ) имеются данные по 12 предприятиям, характеризуемым тремя экономическими показателями: labor – производительность труда, defect – удельный вес потерь от брака (%) и fund – фондоотдача активной части основных

производственных фондов. Из этих предприятий выделены две обучающие выборки (переменная firm), первая из которых включает 4 предприятия группы A, а вторая 5 предприятий группы B. Требуется классифицировать в одну из групп A или B оставшиеся три предприятия.

2. Перед выполнением дискриминантного анализа необходимо убедиться в том, что переменные характеризующие предприятия являются нормально распределенными и дисперсии и ковариации этих переменных внутри групп однородны. Для этого используется дисперсионный анализ (ANOVA). Необходимые опции реализованы в функции anov.

Затем, выполните один из тестов на в однородность дисперсий и ковариаций внутри двух групп (например, t-тест или M-тест Бокса).

Для проверки на нормальность распределения воспользуйтесь, например, графиками поля рассеяния или соответствующим тестом.

3. Выполните дискриминантный анализ имеющихся 9 предприятий. Получите результаты дискриминантного анализа по каждой переменной, в частности, лямбды Уилкса как для всей дискриминации, так и отдельно для каждой переменной и значимость переменных для классификации.

Получите значения коэффициентов дискриминантных функций с после опытными вероятностями попадания предприятия в одну из групп. Должны получиться следующие дискриминантные функции:

$$f_A = -54{,}87 + 9{,}13labor + 6{,}39defect + 10{,}56\,fund,$$

$$f_B = -25{,}18 + 6{,}42labor + 11{,}07defect + 3{,}35\,fund.$$

Получите матрицу, по строкам которой фактическая классификация, а по столбцам – полученная по модели. В идеальном случае они должны совпадать и матрица должна иметь диагональный вид при проценте корректных наблюдений 100%.

Получите соответственно классификацию по наблюдениям и вероятности отнесения каждого наблюдения к каждой из двух групп (A или B). Причем классифицированы будут и последние 3 наблюдения, для которых мы не имели первоначально информации о том, к какой из групп они относятся (наблюдения 10 и 11 – к группе B, а 12 – к группе A). Также можно получить квадрат расстояния Махаланобиса от центра каждой из групп.

4. Выполните пошаговый анализ методом последовательного включения переменных и методом исключения. Обратите внимание на возможности изменения значений F критерия для включения/исключения переменной и вида отображения – конечного результата или результатов по шагам и метод пошагового анализа: включения или исключения.

5. Получите результаты в п. 2–4. Сделайте содержательные выводы по результатам всех выполненных расчетов.

Лабораторное задание 4. «Кластерный анализ»

1. Загрузите данные из файла Lab3.csv (https://disk.yandex.ru/d/QnfUkoAkWtD1JQ). Файл содержит подвыборку из массива msm.sta по индивидам. Описание переменных: lw – логарифм заработной платы, edu – число лет образования, nhh – число членов домохозяйства, dd – доля доходов главы домохозяйства в семейном бюджете, age – возраст, pm – процент заработков, который дает основная работа. Необходимо классифицировать наблюдения.

2. Выполните расчет описательных статистик по переменным выборки. Сделайте выводы. Почему нельзя использовать данные в натуральном виде? В файле Lab3a.csv (https://disk.yandex.ru/d/QnfUkoAkWtD1JQ) содержатся стандартизованные переменные.

3. Выполните кластерный анализ имеющихся индивидов методом классификации k-средних, указав переменные классификации (все имеющиеся переменные).

Сформируйте таблицу со средними по переменным для каждого кластера и расстояниями между кластерами.

Выполните дисперсионный анализ для проверки значимости для классификации каждой из использованных переменных.

Получить описательные статистики (математическое ожидание, стандартное отклонение и дисперсию) для каждого кластера, а также объекты (наблюдения) каждого класса и расстояние от объектов до центра кластера, которому принадлежит этот объект. Сохранить результаты классификации.

Попробуйте изменить количество кластеров и состав переменных, по которым строится классификация. Как это влияет на результаты разбиения?

4. Организуйте случайную подвыборку наблюдений и выполните кластерный анализ имеющихся данных несколькими методами агломеративной классификации (иерархического объединения кластеров:

одиночной связи (ближайшего соседа), полной связи (дальнего соседа), невзвешенный метод средней связи, взвешенный метод средней связи, невзвешенный центроидный метод, взвешенный центроидный метод (медианной связи), метод Уорда).

Построить иерархическое дерево результатов агломеративной классификации.

Рекомендуется выбирая различные правила объединения кластеров и меры расстояний между объектами, сравнить результаты полученных классификаций, выбрать оптимальный по вашему мнению вариант и сделать содержательные экономические выводы по его результатам.

Лабораторное задание 5. «Снижение размерности исследуемых многомерных признаков»

1. Загрузите данные из файла Lab3.csv (https://disk.yandex.ru/d/QnfUkoAkWtD1JQ). Файл содержит подвыборку из массива msm.sta по индивидам. Описание переменных: lw – логарифм заработной платы, edu – число лет образования, nhh – число членов домохозяйства, dd – доля доходов в семейном бюджете, age – возраст, pm – процент заработков, который дает основная работа. Необходимо провести компонентный анализ данных, а затем классифицировать наблюдения.

2. Выполните расчет описательных статистик по переменным выборки. Сделайте выводы. Рассчитайте корреляционную матрицу переменных. Необходимо ли стандартизировать переменные?

3. Выполните анализ главных компонент.

Выберите число факторов и получите соответствующий процент объясненной этими факторами вариации. Просмотрите коэффициенты факторных нагрузок. Далее можно получить график собственных чисел и сами собственные числа, соответствующие им собственные векторы. Дайте интерпретацию факторам по значениям полученных собственных векторов.

Рекомендуется далее самостоятельно ознакомиться с другими опциями в представлении результатов компонентного анализа.

4. Для имеющихся данных определите и обоснуйте оптимальное число факторов и дайте их интерпретацию. Какой процент вариации они объясняют?

5. Выполните кластерный анализ по значениям выбранных в п. 4 главных компонент для выборки индивидов. Сделайте выводы.

6. Постройте уравнение множественной регрессии lw на выделенные в п. 4 главные компоненты. Дайте интерпретацию полученного результата.

Лабораторное задание 6. «Факторный анализ»

1. Загрузите данные из файла Lab3.csv (https://disk.yandex.ru/d/QnfUkoAkWtD1JQ).

2. Выполните факторный анализ по имеющимся выборочным данным.

Поиска латентных факторов осуществите методом главных факторов (рекомендуется сравнить результаты, полученные разными методами оценивания), укажите максимальное количество факторов, например 2, и минимальное собственное число фактора для включения его в анализ – 0.

Получите факторные нагрузки. Укажите какие коэффициенты по модулю больше 0,7. Также можно получить значения собственных чисел и долю объясненной факторами вариации. Получите значения общностей для каждого из фактора.

3. Осуществите вращение факторов. Для этого выберите один из методов вращения: варимакс, биквартимакс, квартимакс, эквимакс, например, Varimax normalized. Посмотрите как изменились факторные нагрузки и другие результаты факторного анализа. Дайте интерпретацию факторов.

Поэкспериментируйте, выбирая различные значения числа факторов на начальной стадии анализа и различные методы вращения факторов. Выберите оптимальный с вашей точки зрения результат.

4. Выполните кластерный анализ наблюдений (по объектам) по факторам, полученным после варимакс вращения в предыдущем пункте.

Получите таблицу значений факторов для каждого наблюдения. Скопируйте два столбца значений таблицы в окно исходных данных на место добавленных новых переменных. Далее методом классификации k-средних (см. лабораторную работу по кластерному анализу) выполните кластеризацию для вновь полученных переменных.

5. Постройте уравнение множественной регрессии lw на выделенные факторы.

6. Постройте необходимые графики и рассчитайте соответствующие статистики. Сделайте содержательные экономические выводы по результатам статистического анализа.