Пермский филиал федерального государственного автономного образовательное учреждение высшего образования
"Национальный исследовательский университет
"Высшая школа экономики"

Программа учебной дисциплины
«Технологии анализа данных»

Утверждена
Академическим советом[1]
_____
название академического совета
Протокол № от __ _____ 20__

Академический руководитель ОП
_____ _____
Подпись ФИО

Разработчик	Радионова М.В., к.ф.-м.н., доцент, кафедра высшей математики
Число кредитов	7
Контактная работа (час.)	90
Самостоятельная работа (час.)	176
Образовательная программа, курс	«Информационная аналитика в управлении предприятием» направления подготовки 38.04.05 Бизнес-информатика, уровень магистратура, 1 курс
Формат изучения дисциплины	Без использования онлайн курса/иное

[1] Для *ПУД из общеуниверситетского пула – Руководитель Департамента / заведующий кафедрой.*

1. Цель, результаты освоения дисциплины и пререквизиты

Целями освоения дисциплины «Технологии анализа данных» по направлению подготовки 38.04.05 Бизнес-информатика магистерской программы «Информационная аналитика в управлении предприятием» являются:

- приобретение студентами базовых знаний по основам анализа данных;
- знакомство с прикладными задачами дисциплины;
- анализ алгоритмов обработки статистических и эмпирических данных, используемых в современном программном обеспечении;
- получения практических навыков использования статистических и эмпирических методов анализа в ходе разработки и экспериментального исследования новых программных продуктов;
- развитие умений, позволяющих использовать опыт применения статистических и эмпирических методов обработки данных для решения задач экономического анализа и моделирования.

Настоящая дисциплина относится к обязательным дисциплинам специализации вариативной части цикла дисциплин программы.

Изучение данной дисциплины базируется на следующих дисциплинах:

Средства Business Intelligence и системы поддержки принятия решений.

Для освоения учебной дисциплины студенты должны владеть следующими знаниями и компетенциями:

Знание основ функционирования персональных компьютеров.

Знание основ организации обработки данных с помощью компьютеров.

Базовые навыки работы с персональным компьютером в среде Microsoft Windows.

Базовые знания и навыки работы с офисными приложениями (текстовым процессором и электронными таблицами).

Знание основных понятий математического анализа, теории вероятности и математической статистики.

Знание основ построения баз данных.

В результате освоения дисциплины студент осваивает следующие компетенции:

Код по ОС ВШЭ	Компетенция
ПК-8	Способен согласовывать с заказчиком требования, руководить процессами проектирования архитектуры предприятия, вырабатывать рекомендации по ее реализации
ПК-9	Способен разрабатывать и внедрять компоненты архитектуры предприятия, разрабатывать рекомендации по их внедрению и эксплуатации
ПК-10	Способен проводить исследования и поиск новых моделей и методов совершенствования архитектуры предприятия
ПК-11	Способен проводить аналитические и поисковые исследования в сфере экономики, управления и ИКТ для выявления продуктовых, технологических, организационных, маркетинговых инновации.
ПК-12	Способен проводить научные исследования для выработки стратегических решений в области ИКТ
ПК-13	Способен согласовывать с заказчиком, планировать и выполнять самостоятельную и коллективную научно-исследовательскую работу
ПК-14	Способен готовить демонстрационные материалы, проводить консультации и разрабатывать рекомендации для потенциальных заказчиков по вопросам разработки и совершенствования архитектуры предприятия
ПК-15	проводить исследования в области новых моделей и методов, направленных

Код по ОС ВШЭ	Компетенция
	на совершенствование архитектуры предприятия, разработку и внедрение отдельных ее компонентов
ПК-17	организовать электронное предприятие, используя новейшие тенденции развития электронной коммерции
ПК-18	совершенствовать методы ведения электронного бизнеса, работы подразделений электронного бизнеса несетевых компаний
ПК-19	оценивать эффективность мероприятий относительно целей совершенствования ИТ-инфраструктуры и архитектуры предприятия и бизнес-процессов
ПК-20	определять политику информационной безопасности на предприятии/в организации;
УК-3	Способен к самостоятельному освоению новых методов исследований, изменению научного и производственного профиля своей деятельности
УК-8	Способен вести профессиональную, в том числе научно-исследовательскую деятельность в международной среде

В результате освоения дисциплины студент должен:

- знать основные понятия анализа данных, необходимые для дальнейшего изучения других дисциплин, предусмотренных учебным планом;
- уметь применять методы дисциплины для решения задач, возникающих в других дисциплинах;
- производить статистические расчеты в стандартных постановках, давать содержательную интерпретацию результатов вычислений, обрабатывать эмпирические и экспериментальные данные
- иметь навыки (приобрести опыт) представление о сферах применения и возможностях статистических и эмпирических методов.

2.Содержание учебной дисциплины

Темы, объем часов и планируемые результаты обучения представлены в таблице

Разделы / темы дисциплины	Объем в часах				Планируемые результаты обучения (ПРО), подлежащие контролю	Формы контроля
	лк	см	ср	onl		
Раздел 1. Предварительный анализ данных и визуализация	6	8	30		Решает задачи на проверку данных на выбросы, нормальное распределение.	
Раздел 2. Корреляционный анализ данных	6	12	30		Решает задачи теории статистического оценивания и проверки гипотез. Может оценить взаимосвязь между несколькими переменными	письменная работа 60 минут
Раздел 3. Методы классификации многомерных наблюдений	10	12	30		Решает задачи классификации и кластеризации пространства наблюдений	

Раздел 4. Регрессионный анализ	16	10	30		Решает задачи построения и проверки на качество регрессионных моделей	Постановка задачи (желательно по интересной для студента тематике) и её решение методами, обсуждаемыми на курсе. Постановка задачи сдаётся в конце третьего модуля, решение задачи – в конце четвертого модуля.Работа сдается в бумажной и электронной форме. Срок выполн
Часов по видам учебных занятий:	40	50	176			

Формы учебных занятий:
лк – лекции в аудитории;
см - семинары/ практические занятия/ лабораторные работы в аудитории;
onl – лекции или иные виды работы студента с помощью онлайн-курса;
ср – самостоятельная работа студента.

Содержание разделов дисциплины:

Раздел 1. Предварительный анализ данных и визуализция

Статистические методы обработки экспериментальных данных. Основные понятия и задачи математической статистики. Генеральная совокупность, выборка, результаты наблюдений, статистика, статистическая оценка, требования к оценкам. Классификация признаков по шкалам измерений. Описательная статистика: среднее значение, математическое ожидание, медиана, мода, дисперсия, среднее квадратичное отклонение, коэффициент вариации, показатель точности опыта, минимум, максимум, размах выборки, моменты распределения. Вариационная статистика: параметры классовых интервалов, группировка, функции эмпирического распределения. Ранжирование. Проверка случайности выборки из нормальной совокупности. Статистические гипотезы и статистические критерии. Характеристики критериев. Тест Вальда. Тест Стьюдента. Лемма Неймана-Пирсона. Критерий отношения правдоподобия. Введение в A/B-тестирование. Критерий последовательного отношения правдоподобия. Непараметрические критерии.

Формы и методы проведения занятий по разделу, применяемые учебные технологии: лекционные занятия, решение задач на семинарах, самостоятельная работа, проверка усвоенного материала микроконтролем.

Раздел 2. Корреляционный анализ данных

Парный коэффициент корреляции. Проверка гипотезы о значимости коэффициента корреляции. Доверительный интервал для парного коэффициента корреляции. Частный коэффициент корреляции. Проверка гипотезы о значимости частного коэффициента корреляции. Доверительный интервал для частного коэффициента корреляции. Множественный коэффициент корреляции. Проверка гипотезы о значимости множественного коэффициента корреляции. Корреляционный анализ качественных данных. Исследование связи между номинальными переменными (таблица сопряженности признаков, критерий хи-квадрат, меры связи признаков: коэффициенты контингенции, ассоциации, среднеквадратической сопряженности, Пирсона, Крамера). $\lambda-$меры прогноза

Гутмана. Исследование связи между порядковыми переменными (ранговый коэффициент корреляции Спирмена, коэффициент согласованности Кендалла, коэффициент конкордации).

Формы и методы проведения занятий по разделу, применяемые учебные технологии: лекционные занятия, решение задач на семинарах, самостоятельная работа, проверка усвоенного материала микроконтролем.

Раздел 3. Методы классификации многомерных наблюдений

Классификация многомерных наблюдений без обучения (непараметрический случай), кластерный анализ.

Общая постановка задачи автоматической классификации, классификации без обучения, понятия кластерного анализа. Метрики расстояния и близости между объектами, расстояния между кластерами. Функционалы качества разбиения. Основные типы задач и алгоритмов кластерного анализа. Иерархические, параллельные и последовательные процедуры кластерного анализа. Метод к-средних. Зависимость выбора алгоритма классификации от цели статистического исследования.

Классификация многомерных наблюдений при наличии обучающих выборок, дискриминантный анализ.

Методы классификации с обучением, основные понятия дискриминантного анализа. Обучающие выборки. Линейный дискриминантый анализ при известных параметрах многомерного нормального закона распределения (случай двух классов и общий случай). Вероятность ошибочной классификации с помощью дискриминантной функции. Оценка качества дискриминантной функции и информативности отдельных признаков.

Классификация многомерных наблюдений без обучения (параметрический случай), расщепление смесей вероятностных распределений. Канонические корреляции.

Понятие и задача расщепления смеси вероятностных распределений. Алгоритм решения задачи автоматической классификации в рамках модели смеси распределений, приводящий к схеме дискриминантного анализа

Канонические корреляции и канонические величины генеральной совокупности, их оценивание, интерпретация и использование в экономических исследованиях.

Формы и методы проведения занятий по разделу, применяемые учебные технологии: лекционные занятия, решение задач на семинарах, самостоятельная работа, проверка усвоенного материала микроконтролем.

Раздел 4. Регрессионный анализ

Анализ однофакторной регрессионной модели

Простейшая линейная регрессионная модель (ПЛРМ). Природа случайной ошибки. Корреляционное поле наблюдений и его применение к выбору формы регрессии. Оценки методом наименьших квадратов коэффициентов ПЛРМ. Интерпретация коэффициентов ПЛРМ. Коэффициент детерминации и его свойства. Теорема Гаусса–Маркова. Доверительные интервалы для коэффициентов регрессии и проверка гипотез об их значимости (t – тест). Проверка значимости всей регрессии на основе критерия Фишера. Прогнозирование значения зависимой переменной по ПЛРМ, точность прогноза. Функциональные преобразования в линейной регрессионной модели. Линеаризация нелинейной регрессионной модели.

Анализ общей линейной модели наблюдений при классических предположениях

Множественный регрессионный анализ: особенности спецификации модели, отбор факторов при построении множественной регрессии. Классическая нормальная линейная модель множественной регрессии, оценка параметров методом МНК, ковариационная матрица и ее выборочная оценка. Оценка дисперсии возмущений. Определение доверительных интервалов для коэффициентов и функции регрессии. Оценка значимости множественной регрессии.

Анализ линейной модели наблюдений при отклонениях от классических предположений

Возможные отклонения от предположений классической общей линейной модели наблюдений (ОЛМН): закон распределения, отличный от нормального; автокорреляция, ее суть, причины, последствия, обнаружение и методы устранения; гетероскедастичность, ее суть, последствия, обнаружение и методы смягчения проблемы гетероскедастичности; Исследовательские методы проверки отсутствия гомоскедастичности: тесты Спирмена, Голдфелда–Квандта, Уайта. Мультиколлинеарность, ее суть, последствия, определение и методы устранения. Взвешенный МНК как частный случай обобщенного МНК; содержательный смысл этого подхода. Метод максимального правдоподобия. Реализация этого метода для модели с двумя группами однородных наблюдений.

Формы и методы проведения занятий по разделу, применяемые учебные технологии: лекционные занятия, решение задач на семинарах, самостоятельная работа, проверка усвоенного материала микроконтролем.

3. Оценивание

Текущий контроль по дисциплине «Технологии анализа данных» включает в себя следующие элементы:

1. Понятия генеральной совокупности, выборки и результатов наблюдений, статистика, статистическая оценка, требования к оценкам.
2. Описательная статистика: среднее значение, математическое ожидание, медиана, мода, дисперсия, среднее квадратичное отклонение, коэффициент вариации, показатель точности опыта, минимум, максимум, размах выборки, моменты распределения.
3. Вариационная статистика: параметры классовых интервалов, группировка, функции эмпирического распределения. Ранжирование.
4. Проверка случайности выборки из нормальной совокупности.
5. Статистические гипотезы и статистические критерии. Характеристики критериев.
6. Парный коэффициент корреляции.
7. Проверка гипотезы о значимости коэффициента корреляции.
8. Доверительный интервал для парного коэффициента корреляции.
9. Частный коэффициент корреляции.
10. Проверка гипотезы о значимости частного коэффициента корреляции.
11. Доверительный интервал для частного коэффициента корреляции.
12. Множественный коэффициент корреляции.
13. Проверка гипотезы о значимости множественного коэффициента корреляции.
14. Корреляционный анализ качественных данных. **Непараметрические показатели связи**. Таблицы сопряженности.
15. Общая постановка задачи автоматической классификации, классификации без обучения, понятия кластерного анализа.
16. Метрики расстояния и близости между объектами, расстояния между кластерами. Функционалы качества разбиения.
17. Иерархические, параллельные и последовательные процедуры кластерного анализа.

18. Метод к-средних.
19. Методы классификации с обучением, основные понятия дискриминантного анализа. Обучающие выборки.
20. Линейный дискриминантый анализ при известных параметрах многомерного нормального закона распределения (случай двух классов и общий случай).
21. Вероятность ошибочной классификации с помощью дискриминантной функции.
22. Оценка качества дискриминантной функции и информативности отдельных признаков.
23. Понятие и задача расщепления смеси вероятностных распределений. Алгоритм решения задачи автоматической классификации в рамках модели смеси распределений, приводящий к схеме дискриминантного анализа
24. Сущность и типологизация задач снижения размерности.
25. Математическая модель, ее обоснование и алгоритм метода главных компонент.
26. Собственные векторы и собственные значения корреляционной матрицы, их использование для получения матрицы факторных нагрузок.
27. Основные числовые характеристики главных компонент.
28. Корреляционное поле наблюдений и его применение к выбору формы регрессии.
29. Общая линейная модель наблюдений (ОЛМН) с классическими предположениями (запись в скалярной и матричной формах).
30. Метод наименьших квадратов (МНК) и его геометрическая интерпретация в случае ОЛМН
31. Теорема Гаусса-Маркова для ОЛМН.
32. Анализ качества множественной линейной регрессионной модели с использованием коэффициента детерминации и скорректированного коэффициента детерминации.
33. Формулировка общей линейной гипотезы. Содержательные примеры линейных гипотез: о значимости коэффициентов; о значимости регрессионной модели в целом, для проверки свойств функции Кобба-Дугласа и др. F – статистика для проверки линейной гипотезы. Ее запись в матричном виде, а также с использованием остаточной суммы квадратов или коэффициента детерминации.
34. Фиктивные переменные и их применение в множественных регрессионных моделях для анализа сезонности; для описания структурных изменений; к исследованию влияния неколичественной переменной.
35. Запись множественной линейной регрессионной модели в центрированных и нормированных переменных. Представление оценки МНК параметров ОЛМН и коэффициента детерминации через элементы выборочной корреляционной матрицы исходных переменных. Интерпретация бета-коэффициентов.
36. Возможные отклонения от предположений классической ОЛМН: автокорреляция, гетероскедастичность различных наблюдений; закон распределения отличный от нормального. Неформальные методы обнаружения их обнаружения, возможные экономические причины возникновения.
37. Природа проблемы гетероскедастичности. Виды гетероскедастичности. Последствия гетероскедастичности. Способы выявления гетероскедастичности. Методы преодоления гетероскедастичности.
38. Нелинейные модели. Классы нелинейных регрессий. Методы оценивания нелинейных регрессий. Показатели качества нелинейных моделей.
39.
40. Оценка наименьших квадратов коэффициентов ПЛРМ. Интерпретация коэффициентов ПЛРМ.
41. Остаточная сумма квадратов. Разложение выборочной дисперсии зависимой переменной в виде суммы дисперсии эмпирической регрессии и дисперсии остатков.

42. Коэффициент детерминации и его свойства. Оценка дисперсии ошибки модели и ее свойства.
43. Предположения классической ПЛРМ. Теорема Гаусса-Маркова.
44. Основные показатели качества парной линейной регрессионной модели.
45. Статистические свойства оценок наименьших квадратов коэффициентов ПЛРМ.
46. Доверительные интервалы для коэффициентов регрессии и проверка гипотез об их значимости (t – тест).

Примеры тестовых и контрольных заданий приведены в приложении (см. Фонды оценочных средств по дисциплине)

Промежуточная аттестация по дисциплине проводиться в форме экзамена. Экзамен проводится в письменном виде (в виде теста).

Оценка по дисциплине ($O_{дисциплине}$) определяется, как взвешенная сумма оценок по всем видам контроля и рассчитывается по следующей формуле:

$$O_{дисциплине} = 0,2*O_{ЭК1} + 0,2*O_{ЭК2} + 0,1*O_{ЭК3} + 0,1*O_{ЭК4} + 0,4*O_{экзамен}$$

где $O_{ЭК1}$ – оценка за контрольную работу;
$O_{ЭК2}$ – домашнюю работу;
$O_{ЭК3}$ – оценка за самостоятельную работу;
$O_{ЭК4}$ – оценка за аудиторную работу;
$O_{экзамен}$ – оценка за экзамен;
Способ округления – арифметический.

Критерии оценивания

Стандартные критерии оценивания контрольной (экзаменационной) тестовой работы:

Характеристика решения (первичные баллы)	**Оценка**
Верно выполнены от 25 до 30 заданий	10
Верно выполнены 23 или 24 задания	9
Верно выполнены 21 или 22 задания	8
Верно выполнены 19 или 20 заданий	7
Верно выполнены 17 или 18 заданий	6
Верно выполнены 15 или 16 заданий	5
Верно выполнены 13 или 14 заданий	4
Верно выполнены 11 или 12 заданий	3
Верно выполнены 7 или 10 заданий	2
Верно выполнены 0 или 6 заданий	1

Стандартные критерии оценивания домашней работы:

высшая оценка в 9 баллов (10 баллов проставляется в исключительных случаях) проставляются при отличном выполнении заданий: полных (с детальными или многочисленными примерами и возможными обобщениями) ответах на вопросы, правильном решении задачи и четком и исчерпывающем ее представлении,

почти отличная оценка в 8 баллов проставляется при полностью правильных ответах и решении задач, но при отсутствии какого-либо из выше перечисленных отличительных

признаков, как, например: детальных примеров или обобщений, четкого и исчерпывающего представления решаемой задачи,

оценка в 7 баллов проставляется при правильных ответах на вопросы и правильном решении задачи, но при отсутствии пояснений, примеров, обобщений, без представления алгоритма или последовательности решения задач,

оценка в 6 баллов проставляется при наличии отдельных неточностей в ответах на вопросы (включая грамматические ошибки) или неточностях в решении задачи непринципиального характера (описки и случайные ошибки арифметического характера),

оценка в 5 баллов проставляется в случаях, когда в ответах и в решении задач имеются неточности и ошибки, свидетельствующие о недостаточном понимании вопросов и требующие дополнительного обращения к тематическим материалам,

оценка в 4 балла проставляется при наличии серьезных ошибок и пробелов в знании по контролируемой тематике,

оценка в 3 балла проставляется при наличии лишь отдельных положительных моментов в ответах на вопросы и в решении задач, говорящих о потенциальной возможности в последующем более успешно выполнить задания; оценка в 3 балла, как правило, ведет к повторному написанию ответов на вопросы или решению дополнительной задачи,

оценка в 2 балла проставляется при полном отсутствии положительных моментов в ответах на вопросы и решении задач и, как правило, ведет к повторному написанию контрольной работы в целом,

оценка в 1 балл проставляется, когда неправильные ответы и решения, кроме того, сопровождаются какими-либо демонстративными проявлениями безграмотности или неэтичного отношения к изучаемой теме.

1. Примеры оценочных средств

Пример типовой контрольной работы:

Решите следующие задания:

Задача 1. Известны следующие результаты наблюдений:

$$n=10,\ \sum_{i=1}^{n} X_i = 30,\ \sum_{i=1}^{n} Y_i = 80,\ \sum_{i=1}^{n} X_i Y_i = 270,\ \sum_{i=1}^{n} X_i^2 = 100,\ \sum_{i=1}^{n} Y_i^2 = 1000\cdot$$

Определить выборочный коэффициент корреляции. Построить 98%-й доверительный интервал для генерального коэффициента корреляции. На уровне значимости 0,05 проверить гипотезу о значимости коэффициента корреляции.

Задача 2. Результаты обследования успеваемости студентов и их работы по специальности характеризуются следующими данными:

Студенты-заочники	Число студентов	Из них	
		получившие положительные оценки	получившие неудовлетворительные оценки
Работающие по специальности	150	115	35
Не работающие по специальности	100	80	20
Итого	250	195	55

Рассчитать коэффициент ассоциации и коэффициент контингенции. Проверить гипотезу о независимости признаков. Сделать выводы.

Задача 3. Экспертами оценивались вложения в инвестиционные проекты. Суммарные оценки представлены в таблице

Проект	Первый эксперт	Второй эксперт
1	10	10
2	14	13
3	17	17
4	11	14
5	13	12
6	13	13
7	18	16
8	14	11
9	19	12
10	18	14

Определить, согласуются ли оценки экспертов, с помощью рангового коэффициента корреляции Спирмена и Кенделла. Проверить гипотезу о значимости коэффициента корреляции Спирмена. Сделать выводы.

Домашняя работа

Статистический анализ данных

Цель работы – проведение самостоятельного исследования с применением изученных в рамках курса эконометрического моделирования. Необходимо выбрать данные для исследования. Данные могут быть кросс-секционными или временными рядами.

Что должно быть предъявлено в работе

1. Постановка исследовательской задачи. Актуальность темы. Постановка исследовательского вопроса. Формулировка интересных содержательных гипотез. (Например, H1: размер банка отрицательно влияет на вероятность банкротства банка). Гипотез возможно 4-5.
2. **Данные.** Описание имеющихся данных (как рассчитываются и в чем измеряются), обозначения, источники данных (со ссылками), предварительный анализ данных с выявлением и исключением имеющихся выбросов, анализ описательных статистик и графический анализ переменных и интерпретация. (Проанализируйте исходную выборку на наличие статистических выбросов, используя различные способы: здравый смысл, анализ описательных статистик и гистограммы, анализ ящичковых диаграмм. Рассчитайте и проинтерпретируйте описательные статистики по каждой переменной, включая фиктивную переменную. Проверьте однородность данных с помощью коэффициента вариации по каждой переменной, и нормальность распределения переменных с помощью: гистограмм, графиков Квантиль-Квантиль, коэффициентов асимметрии и эксцесса, теста Жарке-Бера (Jarque-Bera)). Если были обнаружены выбросы, то от них можно избавиться и снова проверить данные на однородность, нормальность, наличие выбросов.
3. **Корреляционный анализ**, проверка на значимость коэффициентов корреляции. Выводы о характере распределения переменных. Сделайте предварительные предположения о наличии мультиколлинеарности. Если требуется, то нахождение автокорреляционных функций, построение графиков временных рядов и также выводы и предположения.
4. **Кластерный анализ данных.** Провести классификацию данных наблюдений (на 3-4 класса), обосновать выбор классифицирующей функции. Сделать выводы о качестве классификации.
5. **Эконометрическая модель.** Выбор и обоснование спецификации модели (необходимо использование как линейной, так и нелинейной функциональной формы). Проверка предпосылок теоремы Гаусса-Маркова и гипотезы о нормальности случайной ошибки. Проверка линейного/нелинейного ограничения с помощью теста Вальда и структурного сдвига/изменения с помощью теста Чоу с обоснованием их природы и интерпретацией полученного результата. Проверка гипотезы о лишних/пропущенных

переменных. Тестирование функциональной формы модели с помощью теста Рамсея. Если данные временные ряды, то проверка на стационарность. Проверка на коинтеграцию данных. Соответствующие выводы.

6. **Эмпирические результаты**. Выбор наилучшей модели с интерпретацией статистической и экономической значимости найденных коэффициентов и модели в целом. Все коэффициенты должны быть проинтерпретированы в соответствии с их значимостью.

7. **Заключение.** Обсуждение основных результатов, обсуждение ограничений достигнутого решения. Выводы о том, какие гипотезы были выполнены, а какие нет. Обоснование, почему некоторые гипотезы не были приняты. Каким образом можно улучшить построенную модель?

1. Метрики расстояния и близости между объектами, расстояния между кластерами. Функционалы качества разбиения.
2. Иерархические, параллельные и последовательные процедуры кластерного анализа.
3. Метод к-средних.
4. Методы классификации с обучением, основные понятия дискриминантного анализа. Обучающие выборки.
5. Линейный дискриминантый анализ при известных параметрах многомерного нормального закона распределения (случай двух классов и общий случай).
6. Вероятность ошибочной классификации с помощью дискриминантной функции.
7. Оценка качества дискриминантной функции и информативности отдельных признаков.
8. Понятие и задача расщепления смеси вероятностных распределений. Алгоритм решения задачи автоматической классификации в рамках модели смеси распределений, приводящий к схеме дискриминантного анализа
9. Сущность и типологизация задач снижения размерности.
10. Общая линейная модель наблюдений (ОЛМН) с классическими предположениями (запись в скалярной и матричной формах).
11. Метод наименьших квадратов (МНК) и его геометрическая интерпретация в случае ОЛМН
12. Теорема Гаусса-Маркова для ОЛМН.
13. Анализ качества множественной линейной регрессионной модели с использованием коэффициента детерминации и скорректированного коэффициента детерминации.
14. Формулировка общей линейной гипотезы. Содержательные примеры линейных гипотез: о значимости коэффициентов; о значимости регрессионной модели в целом, для проверки свойств функции Кобба-Дугласа и др. F – статистика для проверки линейной гипотезы. Ее запись в матричном виде, а также с использованием остаточной суммы квадратов или коэффициента детерминации.
15. Фиктивные переменные и их применение в множественных регрессионных моделях для анализа сезонности; для описания структурных изменений; к исследованию влияния неколичественной переменной.
16. Запись множественной линейной регрессионной модели в центрированных и нормированных переменных. Представление оценки МНК параметров ОЛМН и коэффициента детерминации через элементы выборочной корреляционной матрицы исходных переменных. Интерпретация бета-коэффициентов.
17. Возможные отклонения от предположений классической ОЛМН: автокорреляция, гетероскедастичность различных наблюдений; закон распределения отличный от нормального. Неформальные методы обнаружения их обнаружения, возможные экономические причины возникновения.

18. Природа проблемы гетероскедастичности. Виды гетероскедастичности. Последствия гетероскедастичности. Способы выявления гетероскедастичности. Методы преодоления гетероскедастичности.
19. Нелинейные модели. Классы нелинейных регрессий. Методы оценивания нелинейных регрессий. Показатели качества нелинейных моделей.

Тестовая экзаменационная работы состоит из 30 заданий, в полной мере охватывающих учебный материал дисциплины «Статистические и эмпирические методы компьютинга».

Пример типовой экзаменационной работы:

1. Известны следующие результаты наблюдений: $n=10,\ \sum_{i=1}^{n} X_i = 30,\ \sum_{i=1}^{n} Y_i = 80,\ \sum_{i=1}^{n} X_i Y_i = 270,\ \sum_{i=1}^{n} X_i^2 = 100,\ \sum_{i=1}^{n} Y_i^2 = 1000$. Определить выборочный коэффициент корреляции

1. 0,8	**2.** 0,2	**3.** 0,6	**4.** 0, 3	**5.** среди приведенных нет правильного ответа

2. Экспертами оценивались вложения в инвестиционные проекты. Суммарные оценки представлены в таблице

Проект	Первый эксперт	Второй эксперт
1	11	10
2	14	13
3	17	17
4	15	14
5	9	12
6	13	15
7	18	16

Определить, согласуются ли оценки экспертов, с помощью рангового коэффициента корреляции Спирмена.

1. 0,8	**2.** 0,2	**3.** 0,6	**4.** 0, 3	**5.** среди приведенных нет правильного ответа

3. Результаты обследования успеваемости студентов и их работы по специальности характеризуются следующими данными:

Студенты-заочники	Из них	
	получившие положительные оценки	получившие неудовлетворительные оценки
Работающие по специальности	115	35
Не работающие по специальности	80	20
Итого	195	55

Проверить гипотезу о независимости признаков. Вычисленное значение критерия будет равно

1. 1,5	**2.** 2	**3.** 1	**4.** 3	**5.** среди приведенных нет правильного ответа

4. Данные о четырех фирмах, деятельность которых характеризуется 2 показателями, представлены в таблице.

Номер	1	2	3	4
Первый показатель	7	8	10	9
Второй	9	12	7	8

показатель				

Провести классификацию фирм, используя метод дальнего соседа с обычным Евклидовым расстоянием. Найти расстояние между 1 и 4 фирмами.

1. 1,5	**2.** 2	**3.** 1	**4.** 3	**5.** среди приведенных нет правильного ответа

5. Расстояния между пятью объектами (n = 5) характеризуется матрицей расстояний:

$$\mathbf{D} = \begin{pmatrix} 0 & 2,2 & 3,0 & 5, \\ 2,2 & 0 & 1,4 & 5, \\ 3,0 & 1,4 & 0 & 6, \\ 5,1 & 5,0 & 6,4 & 0 \\ 5,8 & 6,4 & 7,8 & 2, \end{pmatrix}$$

Чему равно расстояние между кластерами S12 и S345 , в которые входят соответственно объекты (1,2) и (3,4,5), если исходить из принципа «ближайшего соседа».

1. 1,5	**2.** 2	**3.** 1	**4.** 3	**5.** среди приведенных нет правильного ответа

6. Пусть Y и ε – случайные величины, X – неслучайная переменная, c *и* d – неизвестные параметры модели. Какая из регрессионных моделей после некоторых преобразований переменных допускает описание с помощью парной линейной регрессионной модели?

1. $Y = \exp(-c + dX) + \varepsilon$; **2.** $Y = \dfrac{1}{1 + c\exp(-dX + \varepsilon)}$;

3. $Y = cX^d + \varepsilon$; **4.** $Y = \dfrac{1}{c + dX} + \varepsilon$;

1. 1	**2.** 2**	**3.** 3	**4.** 4	**5.** Нет ответа

7. Какое из уравнений регрессий является нелинейным по параметрам

1. $Y_X = ab^X$; **2.** $Y_X = a + bX + cX^2$;

3. $Y_X = a + \dfrac{b}{X} + \varepsilon$; **4.** $Y_X = aX_1 + b\ln X_2$.

1. 1 **	**2.** 2	**3.** 3	**4.** 4	**5.** Нет ответа

8. В результате анализа данных найдены $\bar{X} = 1$, $\bar{Y} = 4$, $\overline{XY} = 1$, $S_X^2 = 4$, $S_Y^2 = 1$. В этом случае точечная оценка функции регрессии $Y_X = a + bX$ при $X = 1$ равна:

1. 2,5	**2.** 4,5	**3.** 3,0	**4.** 4,0**	**5.** Нет ответа

9. Какая из систем нормальных уравнений

1. $\begin{cases} \sum_i x_i^2 - bn = \sum_i y_i x_i \\ \sum_i y_i = -an + b\sum_i \frac{1}{x_i^2} \end{cases}$ 2. $\begin{cases} a\sum_i \frac{1}{x_i^2} - b\sum_i \frac{1}{x_i} = \sum_i \frac{y_i}{x_i} \\ -a\sum_i \frac{1}{x_i} + bn = -\sum_i y_i \end{cases}$

3. $\begin{cases} a\sum_i x_i^2 - b\sum_i \frac{1}{x_i} = -\sum_i x_i y_i \\ -a\sum_i \frac{1}{x_i} + b\sum_i \frac{1}{x_i^4} = \sum_i \frac{y_i}{x_i^2} \end{cases}$ 4. $\begin{cases} a\sum_i \frac{1}{x_i^2} - b\sum_i \frac{1}{x_i} = \sum_i \frac{y_i}{x_i} \\ -a\sum_i \frac{1}{x_i} + bn = \sum_i y_i \end{cases}$

соответствует регрессионной модели $y_i = -ax_i + \frac{b}{x_i^2} + \varepsilon_i,\ i = 1\text{K}\ n.$

1.	1	**2.**	2	**3.**	3**	**4.**	4	**5.**	Нет ответа

10. По результатам 10 наблюдений исследовалась зависимость Y от X. Найдены значения следующих статистик: $\bar{X} = 1,\ \bar{Y} = 4,\ \overline{XY} = 1,\ S_X^2 = 4,\ S_Y^2 = 10,\ RSS = 4.$ В этом случае с вероятностью 0,8 разница между верхней и нижней границами интервальной оценки функции регрессии $Y_X = a + bX$ при $X = 3$ равна:

1.	0,89**	**2.**	1,18	**3.**	1,48	**4.**	1,78	**5.**	Нет ответа

11. Остаточная сумма квадратов *RSS* равна 1. Выборочная дисперсия зависимой переменной по 20 наблюдениям равна 2. Тогда коэффициент детерминации R^2 равен:

1.	0,975**	**2.**	0,5	**3.**	0,95	**4.**	0,8	**5.**	0,025

12. По данным 10 фирм получено уравнение регрессии для объема реализации товарной продукции *Y* в зависимости от затрат на рекламу *X*: $\hat{Y}_X = 6 + 3X$ в предположении, что результаты этих наблюдений связаны моделью $y_i = a + bx_i + \varepsilon_i$. Были найдены также стандартные ошибки оценок параметров регрессии: $S_{\hat{a}} = 2,\ S_{\hat{b}} = 3.$ Значение *t*-статистики при проверке гипотезы о значимости (существенности) коэффициента b оказалось равным:

1.	1,5	**2.**	2	**3.**	1**	**4.**	3	**5.**	6

2. Ресурсы

5.1. Рекомендуемая основная литература

№п/п	Наименование
1.	Анализ данных: учебник для академического бакалавриата / В. С. Мхитарян [и др.] ; под ред. В. С. Мхитаряна. — М. : Издательство Юрайт, 2016. — 490 с. — (Серия : Бакалавр. Академический курс). — ISBN 978-5-9916-5591-0. — Режим доступа : www.biblio-online.ru/book/AF1D197F-1759-422E-9593-8B43E2D1093B.
2.	Миркин, Б. Г. Введение в анализ данных : учебник и практикум / Б. Г. Миркин. — М. : Издательство Юрайт, 2017. — 174 с. — (Серия : Авторский учебник). — ISBN 978-5-9916-5009-0. — Режим доступа : www.biblio-online.ru/book/A5995FCA-A5B5-4402-BBCB-3CA6B8BA2A5B.
3.	Эконометрика : учебник для бакалавриата и магистратуры / С. В. Курышева, Елисеева И.И. [и др.]. — М. : Издательство Юрайт, 2016. — 449 с. — (Бакалавр и магистр. Академический курс). — http://www.biblio-online.ru/viewer/695328A6-B66E-4F13-BE2A-7C1B9BB084BA#page/1

5.2. Рекомендуемая дополнительная литература

№п/п	Наименование
1.	Эконометрика. Практикум: Учебное пособие / С.А. Бородич. - М.: НИЦ ИНФРА-М; Мн.: Нов. знание, 2014. - 329 с.: ил.; 60x90 1/16. - (Высшее образование: Бакалавриат). -http://znanium.com/catalog.php?bookinfo=440758
2.	Демидова О.А., Малахов Д.И. Эконометрика. Учебник и практикум для прикладного бакалавриата. М.: Юрайт, 2016. (доступно для чтения через электронные ресурсы НИУ ВШЭ https://proxylibrary.hse.ru:2180/viewer/ekonometrika-432950#page/1).

5.3. Программное обеспечение

№п/п	Наименование	Условия доступа
1.	MS Office 2010	Из внутренней сети НИУ ВШЭ - Пермь (договор)

5.4. Профессиональные базы данных, информационные справочные системы, интернет-ресурсы (электронные образовательные ресурсы)

№п/п	Наименование	Условия доступа
1.	Электронно-библиотечные ресурсы	По подписке НИУ ВШЭ

5.5. Материально-техническое обеспечение дисциплины

Для проведения лекций и семинаров по дисциплине необходимо наличие ноутбука (компьютера) с установленным пакетом Microsoft® PowerPoint, мультимедийного проектора и аудиооборудования. Для выполнения самостоятельной работы необходим компьютер с подключением к сети Интеренет.

6. Особенности организации обучения для лиц с ограниченными возможностями здоровья и инвалидов

В случае необходимости, обучающимся из числа лиц с ограниченными возможностями здоровья (по заявлению обучающегося), а для инвалидов также в соответствии с индивидуальной программой реабилитации инвалида, могут предлагаться следующие варианты восприятия учебной информации с учетом их индивидуальных психофизических особенностей, в том числе с применением электронного обучения и дистанционных технологий:

6.1.1. *для лиц с нарушениями зрения:* в печатной форме увеличенным шрифтом; в форме электронного документа; в форме аудиофайла (перевод учебных материалов в аудиоформат); индивидуальные консультации с привлечением тифлосурдопереводчика; индивидуальные задания и консультации.

6.1.2. *для лиц с нарушениями слуха*: в печатной форме; в форме электронного документа; видеоматериалы с субтитрами; индивидуальные консультации с привлечением сурдопереводчика; индивидуальные задания и консультации.

6.1.3. *для лиц с нарушениями опорно-двигательного аппарата*: в печатной форме; в форме электронного документа; в форме аудиофайла; индивидуальные задания и консультации.

7. Дополнительные сведения

Особенности самостоятельной работы по курсу отражены в Приложении 1.