Программа учебной дисциплины

«Современные методы анализа данных»

Утверждена

Академическим советом ООП

Протокол № от «__»_____20__ г.

Автор	**к. ф.-м. н. Горяинова Е.Р.,** доцент департамента математики на факультете экономических наук
Число кредитов	3
Контактная работа (час.)	42
Самостоятельная работа (час.)	72
Курс	1
Формат изучения дисциплины	Без использования онлайн курса

I. ЦЕЛЬ, РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ И ПРЕРЕКВИЗИТЫ

Целями дисциплины «Современные методы анализа данных» являются:

- сформировать теоретические знания в области математической статистики;
-обучить студентов применять основные модели и методы математической статистики для обработки реальных социально-экономических данных.

В результате освоения дисциплины студент должен

Знать:
-основные методы первичной обработки статистических данных;
- основные методы проверки однородности экспериментальных данных;
- основные методы дисперсионного анализа;
- принципы сравнения статистических критериев;
- методы оценивание параметров линейных регрессионных моделей;

У*меть:*
– строить математические модели, адекватно описывающие социально-экономические явления;
- использовать статистические критерии для проверки гипотез относительно наблюдаемых случайных данных;

Владеть:
-навыками решения типовых задач математической статистики;
- основными определениями, методами и алгоритмами анализа данных, содержащих случайную составляющую;
- стандартными инструментариями обработки статистической информации.

В результате освоения дисциплины студент осваивает следующие компетенции:

Компетенция	**Код по ФГОС / НИУ**	**Дескрипторы – основные признаки освоения (показатели достижения результата)**	**Формы и методы обучения, способствующие формированию и развитию компетенции**
Способен рефлексировать (оценивать и перерабатывать) освоенные научные методы	СК-М1	Распознаѐт, воспроизводит и использует основные научные понятия, методы и приёмы решения задач статистической обработки данных и применяет их в практической деятельности.	Лекции, семинары, выполнение домашних заданий, групповая работа на семинарах
Способен предлагать концепции, модели, изобретать и апробировать способы и инструменты профессиональной деятельности	СК-М2	Использует доступную информацию для построения математической модели проблемной ситуации, формулирует концепции и методы решения поставленной задачи из своей профессиональной области.	Лекции, семинары, выполнение домашних заданий, групповая работа на семинарах
Способен принимать управленческие решения, оценивать их возможные последствия и нести за них ответственность	СК-М5	Предлагает и обосновывает методы решения поставленных задач в области социальных, экономических и политических процессов, и оценивает их достоинства и недостатки.	Лекции, семинары, выполнение домашних заданий, групповая работа на семинарах
Способен разрешать мировоззренческие, социально и личностно значимые проблемы	СЛК-М6	Способен строить и анализировать математические модели реальных задач в соответствии с направлением подготовки и специализацией.	Лекции, семинары, выполнение домашних заданий, групповая работа на семинарах

ПРЕРЕКВИЗИТЫ

Для изучения курса «Современные методы анализа данных» требуются знания базового курса «Теория вероятностей и математическая статистика».

II. СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

Тема 1. Понятие о робастных и непараметрических методах статистического анализа данных (Л.-4ч., С.-4ч., СРС-14ч.)

Краткий обзор известных методов оценивания параметров и основ проверки статистических гипотез. Демонстрация примеров правильного и неправильного выбора альтернативных гипотез в задачах с дихотомическими данными.

Причины возникновения робастных и непараметрических статистических методов.

Понятие о робастных оценках в терминах кривой чувствительности SC (sensitivity curve) и высокой пороговой точки (high breakdown point). Определение B-робастной оценки, исследование свойства B-робастности для выборочного среднего и выборочной медианы. Вычисление пороговой точки для выборочного среднего, $\alpha-$*урезанного* среднего и выборочной медианы. M-оценки (MAD и MADN) для оценивания параметра масштаба.

Методы сравнения статистических критериев. Функция мощности. Относительная асимптотическая эффективность (ОАЭ) статистических критериев по Питмену.

Литература по разделу.

1. Горяинова Е.Р., Панков А.Р., Платонов Е.Н. Прикладные методы анализа статистических данных. М.: Изд. Дом ВШЭ,2012.- 312 с.
2. Хампель Ф., Рончетти Э., Рауссей П., Штаэль В. Робастность в статистике. Подход на основе функции влияния. – М.: Мир, 1989.
3. Hettmansperger T.P., McKean J.W. Robust nonparametric statistical methods. Boca Raton: CRC Press, 2011.
4. Maronna R.A., Martin D., Yohai V. Robust Statistics: Theory and Methods. Chichester: Wiley, 2006.
5. Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере. – М.: Инфра. – М, 2003.

Тема 2. Исследование однородности двух выборок (Л.-4ч., С.-4ч., СРС-12ч.)

Понятие об однородности выборок.

Ранги, связки, средние ранги. Непараметрические ранговые критерии.

Выявление неоднородности, связанной со сдвигом (классический критерий Стьюдента, ранговый критерий Вилкоксона, Фишера – Йейтса) или масштабом (классический F-критерий, ранговый критерий Ансари-Брэдли). Проверка однородности против альтернатив общего вида (критерий Колмогорова – Смирнова, критерий омега-квадрат).

Сравнительный анализ ОАЭ изученных критериев для различных распределений выборок.

Литература по разделу.

1. Горяинова Е.Р., Панков А.Р., Платонов Е.Н. Прикладные методы анализа статистических данных. М.: Изд. Дом ВШЭ,2012.- 312 с.
2. Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере. – М.: Инфра. – М, 2003.
3.Холлендер М., Вулф Д. Непараметрические методы статистики. – М.: Финансы и статистика, 1983 (глава 5 с.101-110, глава 10 с.232-242).

Тема 3. Дисперсионный анализ (Л.-2ч., С.-2ч., СРС-10ч.)
Задача однофакторного анализа (классический F-критерий, ранговый критерий Краскела – Уоллиса). Доверительное оценивание контрастов в гауссовской модели. Критерий Джонкхиера для упорядоченных альтернатив. ОАЭ классического критерия и критерия Краскела – Уоллиса.

Литература по разделу.
1. Горяинова Е.Р., Панков А.Р., Платонов Е.Н. Прикладные методы анализа статистических данных. М.: Изд. Дом ВШЭ,2012.- 312 с.
2. Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере. – М.: Инфра. – М, 2003.
3.Холлендер М., Вулф Д. Непараметрические методы статистики. – М.: Финансы и статистика, 1983

Тема 4. Анализ статистической взаимосвязи социально-экономических явлений (Л.-8ч., С.-10ч., СРС-20ч.)
Шкалы измерений (количественная, порядковая, номинальная).
Исследование связи между номинальными переменными (таблица сопряженности признаков, критерий хи-квадрат, меры связи признаков: коэффициенты контингенции, ассоциации, среднеквадратической сопряженности, Пирсона, Крамера). λ –меры прогноза Гутмана.
Исследование связи между порядковыми переменными (ранговый коэффициент корреляции Спирмена, коэффициент согласованности Кендалла, коэффициент конкордации).
Случайные векторы. Ковариационная матрица. Независимость и некоррелированность компонент случайного вектора. Выборочный коэффициент корреляции. Частные коэффициенты корреляции.
Анализ структуры и тесноты связи между количественными переменными. Критерий хи-квадрат.
Измерение тесноты связи при нелинейной зависимости (индекс корреляции и его оценивание по сгруппированным и несгруппированным данным).
Анализ множественных связей (множественный коэффициент корреляции, его вычисление и свойства для общих и нормальных моделей).

Литература по разделу.
1. Горяинова Е.Р., Панков А.Р., Платонов Е.Н. Прикладные методы анализа статистических данных. М.: Изд. Дом ВШЭ,2012.- 312 с.
2. Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере. – М.: Инфра. – М, 2003.
3. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Исследование зависимостей. Справочное издание под ред. Айвазяна С.А. – М.: Финансы и статистика, 1985 (главы 1,2 с.56-124).

Тема 5. Регрессионный анализ. Робастные методы оценивания параметров линейной регрессии. Сравнение свойств оценок, полученных различными методами (Л.-4ч., С.-2ч., СРС-14ч.)

Обзор методов оценивания параметров в линейных регрессионных моделях (ЛРМ): МНК, метод наименьших модулей (МНМ), ранговые методы, монотонные М-оценки Хьюбера и немонотонные М-оценки Тьюки, LMS и LTS -оценки. Асимптотические распределения М, R и L-оценок в ЛРМ. Определение пороговой точки ВР (breakdown point) оценки параметров ЛРМ. Оценки с высокой пороговой точкой (НВР-оценки). Сравнительный анализ свойств оценок параметров ЛРМ, полученных различными методами.

Литература по разделу.

1. Горяинова Е.Р., Панков А.Р., Платонов Е.Н. Прикладные методы анализа статистических данных. М.: Изд. Дом ВШЭ,2012.- 312 с.
2. Дрейпер Н., Смит Г. Прикладной регрессионный анализ: В 2-х книгах, Кн. 1. – М.: Финансы и статистика, 1986. Кн. 2. – М.: Финансы и статистика, 1987(глава 2 с.104-148).
3. Goryainova E. R., Botvinkin E. A. Experimental and Analytic Comparison of the Accuracy of Different Estimates of Parameters in a Linear Regression Model. Пер. с рус. *Automation and Remote Control*. 2017. Vol. 78. No. 10. P. 1819-1836
4. Maronna R.A., Martin D., Yohai V. Robust Statistics: Theory and Methods. Chichester: Wiley, 2006.

По желанию слушателей некоторые разделы могут быть заменены следующими темами:

1. Понятие репрезентативной выборки. Определение объема репрезентативной выборки для однородной и стратифицированной генеральной совокупности. Оптимальный объём выборки при ограниченном финансировании исследований.
2. Задача дискриминантного анализа. Методы классификации объектов (дискриминантное правило Фишера, «наивное» байесовское правило, метод, использующий логистическую регрессию). Применение указанных методов для классификации объектов с номинальными показателями (на примере реальных данных классификации респондентов на классы принимающих и не принимающих участие в благотворительной деятельности).
3. Задача снижения размерности многомерных показателей. Постановка задачи компонентного анализа, метод главных компонент, матрица нагрузок. Постановка задачи факторного анализа. Методы, позволяющие оценить число латентных факторов, методы ортогонального вращения, позволяющие получить простую факторную структуру.
4. Цепи Маркова. Вероятностные характеристики цепей Маркова. Классификация состояний цепи Маркова. Эргодические цепи Маркова. Эргодическая теорема. Предельные вероятности состояний цепи Маркова. Задача о разорении игрока.

III. ОЦЕНИВАНИЕ

Тип контроля	Форма контроля	1 год				Параметры **
		1	2	3	4	
Текущий	Домашнее задание				*	Письменные работы выполняются студентами самостоятельно по окончании пройденной темы. После проверки работы преподавателем проводится защита
Итоговый	Экзамен				э	письменная работа 80 минут

Для прохождения контроля студент должен, как минимум, продемонстрировать знание основных определений и формулировок; умение решать типовые задачи, разобранные на семинарских занятиях и владеть основными методами обработки статистических данных.

Оценки по всем формам контроля выставляются по 10-ти балльной шкале.
Средняя оценка по результатам проверки всех домашних заданий студента формирует накопленную оценку за курс. Способ округления накопленной оценки – целая часть количества набранных баллов
Результирующая оценка за дисциплину рассчитывается следующим образом:

$$O_{результ} = 0.5 * O_{накопл} + 0.5 * \cdot O_{экз.}$$

Способ округления результирующей оценки – арифметический.

Если результирующая оценка составляет менее 4-х баллов, то округлённая результирующая оценка равна целой части набранной студентом оценки.

Первая пересдача – письменная экзаменационная работа. Накопленная оценка за домашнее задание сохраняется.

Вторая пересдача - письменная экзаменационная работа, оценка за домашнее задание не учитывается.

IV. ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

Оценочные средства текущего контроля

Тематика заданий текущего контроля

Вариант домашней работы:

Изученные (в темах №2 и №4) методы проверки однородности выборок и методы исследования зависимости/независимости показателей, измеряемых в различных шкалах, позволяют слушателям курса сформулировать реальные проблемы, которые могут быть решены указанными методами. В качестве домашнего задания предлагается самостоятельно сформулировать на статистическом языке задачу, выбрать соответствующие реальные данные и представить решение задачи, основанное на использовании изученных методов.

Ниже в качестве примера представлены формулировки задач, предложенные студентами магистратуры предыдущих лет.

1.Средняя стоимость лечения одного пациента-льготника с диагнозом «дуоденит» составляет (в рублях на ноябрь 2007 года):

Дальневосточный фед. округ	Приволжский фед. округ
Амурская обл. 245,61	Кировская обл. 196,27
Еврейская АО 101,45	Оренбургская обл. 309,79
Камчатская обл. 202,84	Пензенская обл. 271,76
Корякский АО 327,63	Пермская обл. 329,58
Магаданская обл. 144,5	Башкортостан 233,49
Приморский край 458,81	Марий-Эл 298,24
	Мордовия 311,6
	Татарстан 284,03
	Чувашия 405,5

Одинакова ли средняя стоимость лечения льготников в Дальневосточном и Приволжском федеральных округах?

2. Проведен социологический опрос 655 человек. Каждый из опрошенных отвечал на два вопроса. Вопрос А: «Удовлетворены ли Вы своим образом жизни?» (варианты ответов: да, нет). Вопрос В: «Каково Ваше материальное положение?» (варианты ответов: плохое, ниже среднего, среднее, выше среднего, хорошее. Результаты опроса сведены в следующую таблицу:

B A	плохое	ниже среднего	среднее	выше среднего	хорошее
Нет	92	64	48	23	3
Да	22	46	136	148	72

Имеется ли зависимость между материальным положением (признак В) и удовлетворенностью образом жизни (признак А)?

Прокомментируйте характер связи между А и В с помощью коэффициентов Пирсона, Крамера, среднеквадратической сопряженности, мер прогноза Гутмана, мер прогноза Краскела-Гудмана.

3. В таблице представлены данные за 1997 год показателей X (индекс человеческого развития) и Y (суточная калорийность питания населения, ккал на душу)

для следующих стран: Австрия, Аргентина, Великобритания, Германия, Египет, Норвегия, Украина, Республика Корея, ЮАР, США.

X	0.904	0.827	0.918	0.906	0.616	0.927	0.721	0.852	0.695	0.927
Y	3343	3136	3237	3330	3289	3350	2753	3336	2933	3642

Являются ли показатели X и Y зависимыми?

Вопросы для оценки качества освоения дисциплины

Тема 1.

1. Что такое статистическая гипотеза?
2. В чем состоят ошибки I и II рода?
3. Дайте определение функции мощности статистического критерия.
4. Дайте определение квантили. Чему равна 0,05-квантиль стандартного гауссовского распределения, если 0,95-квантиль этого распределения равна 1,65?
5. Каков порядок проверки параметрических статистических гипотез?
6. Опишите задачу, которая решается с помощью биномиального критерия.
7. Что такое относительная эффективность по Питмену?
8. Что такое B-робастная оценка?
9. Приведите пример оценки, которая является B-робастной.
10. Приведите пример оценки, которая не является B-робастной.

Тема 2.

1. Какие выборки называют однородными?
2. Назовите основные типы неоднородности выборок.
3. Опишите условия применимости классических и ранговых критериев для проверки гипотезы об однородности.
4. Какие преимущества и какие недостатки имеют ранговые критерии по сравнению с классическими?
5. Какие критерии применяют для проверки гипотезы об однородности двух выборок?
6. Чему равна АОЭ по Питмену критерия Стьюдента по отношению к критерию Вилкоксона, если наблюдения имеют гауссовское распределение?

Тема 3.

1. Назовите основные термины дисперсионного анализа.
2. В чем состоит задача однофакторного дисперсионного анализа?
3. В чем состоит задача двухфакторного дисперсионного анализа?
4. Опишите условия применимости классических и ранговых критериев в задачах однофакторного дисперсионного анализа.
5. Как построить доверительный интервал контраста в задаче однофакторного анализа для гауссовских наблюдений?
6. В каких ситуациях следует применять критерий Джонкхиера и критерий Пейджа?

Тема 4.

1. Опишите основные типы шкал измерений и допустимые преобразования в этих шкалах.
2. Что такое таблица сопряженности признаков?
3. Дайте определение независимости признаков, измеряемых в номинальной шкале.
4. Как проверить гипотезу о независимости признаков в номинальной шкале?
5. Как проверить гипотезу о независимости признаков в порядковой шкале?
6. Назовите основные коэффициенты, измеряющие связь признаков в номинальной шкале.
7. Что такое коэффициент корреляции? Каковы его основные свойства?
8. Как проверить гипотезу о некоррелированности признаков?
9. В каком случае проверка некоррелированности наблюдений эквивалентна проверке независимости?
10. Как измерить тесноту связи двух нелинейно зависимых переменных?
11. Что такое множественный коэффициент корреляции? Каковы его свойства?

Тема 5.
1. Какие методы оценивания параметров линейной регрессии вы знаете?
2. Как определяется пороговая точка оценки параметров регрессионной модели?
3. Какие оценки параметров регрессии с высокой пороговой точкой вы знаете?
4. Какие преимущества и недостатки имеют HBP-оценки?
5. Какими преимуществами обладают MHM, M и R-оценки по сравнению с МНК-оценками?

Примеры заданий итогового контроля

Примерный вариант экзаменационного билета

1.В результате проведенного исследования было установлено, что у 309 светлоглазых мужчин жены также имеют светлые газа, а у 214 светлоглазых мужчин жены темноглазые. У 119 темноглазых мужчин жены также темноглазые, а у 132 темноглазых мужчин жены светлоглазые. Имеется ли зависимость между цветом глаз мужей и их жен? Исследуйте силу связи между этими показателями с помощью коэффициентов контингенции и ассоциации.

2. По 20 территориям России были изучены следующие данные: X – среднедушевой доход (в тыс. руб.), Y – доля занятых тяжелым физическим трудом в общей численности занятых (%), Z – доля экономически активного населения в численности всего населения (%). По результатам наблюдений были вычислены выборочные коэффициенты корреляции. Для показателей X и Y выборочный коэффициент корреляции равен 0.746, для X и Z равен 0.507, для Y и Z равен 0.432. Вычислите частный коэффициент корреляции показателей X и Y при условии, что показатель Z зафиксирован.

3.Три квалифицированных эксперта (A,B и C) оценивают в порядке предпочтения 10 бизнес-проектов. Результаты представлены в таблице:

A	1	4	2	5	3	7	6	9	8	10

B	2	1	3	4	5	8	9	10	6	7
C	2	1	4	5	3	7	9	8	6	10

Является ли эта экспертная группа согласованной?

4. В таблице представлены данные за 1995 год показателей X (ВВП в паритетах покупательной способности) и Y (коэффициент детской смертности в %) для следующих стран: Бурунди, Чад, Индия, Египет, Мексика, Бразилия, Республика Корея, Канада, США, Швейцария.

X	2.3	2.6	5.2	12.2	23.7	20	42.4	78.3	100	95.9
Y	98	117	68	16	33	44	10	6	8	6

Считая , что наблюдения имеют гауссовское распределение, выясните являются ли признаки X и Y зависимыми и постройте приближенный доверительный интервал уровня надежности 0.95 для коэффициента корреляции X и Y.

5.Уровень гистамина в мокроте у 7 курильщиков, склонных к аллергии, составил (в микрограммах): 102,4; 100,0; 67,6; 65,9; 64,7; 39,6; 31,2, а у курильщиков, несклонных к аллергии: 48,1; 45,5; 41,7; 35,4; 29,1; 18,9; 58,3; 66,8; 71,3; 94,3. Верно ли предположение о том, что уровень гистамина у курильщиков, подверженных аллергии, выше, чем у неаллергенов? Принять уровень значимости равным 0,05.

6. Какие преимущества и какие недостатки имеют ранговые критерии по сравнению с классическими ?

V. РЕСУРСЫ

1. Основная литература

1. Горяинова Е.Р., Панков А.Р., Платонов Е.Н. Прикладные методы анализа статистических данных. М.: Изд. Дом ВШЭ,2012.- 312 с.

2. Дополнительная литература

1. Тюрин Ю.Н., Макаров А.А. Анализ данных на компьютере. – М.: Инфра, 2003.
2. .Холлендер М., Вулф Д. Непараметрические методы статистики. – М.: Финансы и статистика, 1983
3. Хампель Ф., Рончетти Э., Рауссей П., Штаэль В. Робастность в статистике. Подход на основе функции влияния. – М.: Мир, 1989.
4. Hettmansperger T.P., McKean J.W. Robust nonparametric statistical methods. Boca Raton: CRC Press, 2011.
5. Maronna R.A., Martin D., Yohai V. Robust Statistics: Theory and Methods. Chichester: Wiley, 2006.
6. Goryainova E. R., Botvinkin E. A. Experimental and Analytic Comparison of the Accuracy of Different Estimates of Parameters in a Linear Regression Model. / Пер. с рус. // *Automation and Remote Control*. 2017. Vol. 78. No. 10. P. 1819-1836
7. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Исследование зависимостей. Справочное издание под ред. Айвазяна С.А. – М.: Финансы и статистика, 1985.

8. Горяинова Е.Р., Слепнёва Т.И. Методы бинарной классификации объектов с номинальными показателями. Журнал Новой экономической ассоциации. 2012. № 2. С. 27-49.

3. Программное обеспечение

Не требуется

4. Материально-техническое обеспечение дисциплины

Учебные аудитории для лекционных занятий по дисциплине обеспечивают использование и демонстрацию тематических иллюстраций, соответствующих программе дисциплины в составе:

- мультимедийный проектор с дистанционным управлением.