

**Федеральное государственное автономное образовательное
учреждение высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»**

УТВЕРЖДЕНО

**Директор физтех-школы
прикладной математики и
информатики**

А.М. Райгородский

	Рабочая программа дисциплины (модуля)
по дисциплине:	Причинно-следственные выводы в статистике
по направлению:	Информатика и вычислительная техника
профиль подготовки:	Прикладная математика и информатика Физтех-школа Прикладной Математики и Информатики кафедра дискретной математики
курс:	2
квалификация:	магистр

Семестр, формы промежуточной аттестации: 3 (осенний) - Экзамен

Аудиторных часов: 45 всего, в том числе:

лекции: 30 час.

семинары: 15 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 60 час.

Подготовка к экзамену: 30 час.

Всего часов: 135, всего зач. ед.: 3

Количество контрольных работ, заданий: 2

Программу составил: А.М. Ченцов, ассистент

Программа обсуждена на заседании кафедры дискретной математики 20.01.2025

Аннотация

Концепция причинно-следственной связи тесно связана с понятием контрфактуальности, ``что было бы, если?'' Это увеличивает сложность статистического анализа на данных, получаемых с помощью пассивного наблюдения, поскольку в таком случае одно или более извозможных состояний исследуемой системы является ненаблюдаемым. Анализ причинно- следственного влияния длительное время считался периферийной темой в статистике, несмотря на попытки математической формализации этого понятия в работах Неймана и Рубина, а также внимание эконометристов к этой теме. Систематизация исследований в этой области началась в 1990е годы в работах Ангриста, Пишке и др., а общая статистическая теория причинно- следственного анализа возникла в 2010х годах в работах Перла, Спиртеса и др., которые предложили концепцию моделей на направленных ациклических графах (DAG), помогающих упростить и унифицировать разработанные ранее методы (инструментальных переменные, разрывный регрессионный дизайн и др.), и в ряде случаев даже позволяют автоматизировать поиск причинно-следственной структуры в данных (платформы DoWhy, CausalML, causal-learn, Econ-ML, CausalImpact и др.). В последние 15 лет область причинно-следственного анализа в статистике переживает бурный рост, во многом в связи с пониманием её необходимости для развития сильного искусственного интеллекта (Дж. Перл, 2019). Среди новейших работ в этой области центральное место занимают работы В.Черножукова, А.Беллони и др., посвященные анализу причинно-следственного влияния при использовании больших данных (big-p асимптотика, подразумевающая рост числа регрессоров вместе с ростом размера выборки). В этих случаях необходимость регуляризации обучающего алгоритма для оценки мешающего параметра высокой размерности приводит к проблемам в статистических свойствах оценок интересующей переменной. Для решения этих проблем был предложен метод двойного машинного обучения структурной частично-линейной модели, который является центральным предметом изучения в предлагаемом курсе.

1. Цели и задачи

Цель дисциплины

Знакомство слушателей с ключевыми положениями теории статистического причинно-следственного вывода (causal inference), активно развивающегося раздела науки о данных.

Задачи дисциплины

- ознакомление студентов с максимально широким спектром задач и методов теории, включая дисперсионный анализ, корреляционный анализ, дискриминантный анализ, регрессионный анализ, анализ и прогнозирование временных рядов, анализ выживаемости, анализ панельных данных, факторный анализ, кластерный анализ, многомерное шкалирование, выборочный анализ, множественную проверку гипотез;
- приобретение теоретических знаний и практических умений и навыков в области статистического анализа;
- оказание консультаций и помощи студентам в проведении собственных теоретических исследований в области прикладной статистики.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
УК-2 Способен управлять проектом на всех этапах его жизненного цикла	УК-2.2 Способен прогнозировать результат деятельности и планировать последовательность шагов для достижения данного результата. Формирует план-график реализации проекта в целом и план контроля его выполнения
	УК-2.3 Способен организовать и координировать работу участников проекта, обеспечивать работу команды необходимыми ресурсами
	ОПК-2.1 Имеет представление о современном состоянии исследований в рамках тематической области своей профессиональной деятельности

ОПК-2 Имеет представление об актуальных проблемах науки и техники в области информатики и вычислительной техники, способен на научном языке формулировать профессиональные задачи	ОПК-2.2 Способен оценивать актуальность исследований в области информатики и вычислительной техники и их практическую значимость
	ОПК-2.3 Владеет профессиональной терминологией, используемой в современной научно-технической литературе, обладает навыками устного и письменного изложения результатов научной деятельности в рамках профессиональной коммуникации
ПК-2 Понимает и способен применить в научно-исследовательской и прикладной деятельности основные законы естествознания, современный математический аппарат и алгоритмы, современные информационно-коммуникационные технологии	ПК-2.1 Знает основы научно-исследовательской деятельности в области информационных технологий, владеет знанием основ философии и методологии науки; знанием методов научных исследований и навыками их проведения
	ПК-2.2 Умеет применять полученные знания в области фундаментальных научных основ теории информации и решать стандартные задачи в собственной научно-исследовательской деятельности
	ПК-2.3 Имеет практический опыт научно-исследовательской деятельности в области информационно-коммуникационных технологий
	ПК-2.4 Владеет методами и алгоритмами решения задач цифровой обработки сигналов, использования сети Интернет, аннотирования, реферирования, библиографического поиска, опыт работы с научными источниками

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- основы базовой статистики и регрессионного анализа,

уметь:

- понять поставленную задачу;
- оценивать корректность постановок задач;
- строго доказывать или опровергать утверждение;
- самостоятельно находить алгоритмы решения, в том числе и нестандартных, и проводить их анализ;
- самостоятельно видеть следствия полученных результатов;
- давать экспертную оценку финальным результатам решения.

владеть:

- навыками освоения большого объема информации и решения задач;
- навыками самостоятельной работы и освоения новых дисциплин;
- культурой постановки, анализа и решения математических и прикладных задач, требующих для своего решения использования математических подходов;
- предметным языком прикладной статистики и навыками грамотного описания решения задач и представления полученных результатов.
- навыками компьютерной обработки информации.

4. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

4.1. Разделы дисциплины (модуля) и трудоемкости по видам учебных занятий

№	Тема (раздел) дисциплины	Трудоемкость по видам учебных занятий, включая самостоятельную работу, час.			
		Лекции	Семинары	Лаборат. работы	Самост. работа

1	Стандартные модели причинно-следственного анализа	6	3		12
2	Причинно-следственная идентификация в статистике: модель потенциальных исходов Рубина и предположение о стабильности эффекта влияния (SUTVA).	6	3		12
3	Модели на направленных ациклических графах (DAG).	6	3		12
4	Инструментальные переменные, разрывный регрессионный дизайн и синтетические контрольные переменные в моделях DAG.	6	3		12
5	Бутстрап	6	3		12
Итого часов		30	15		60
Подготовка к экзамену		30 час.			
Общая трудоёмкость		135 час., 3 зач.ед.			

4.2. Содержание дисциплины (модуля), структурированное по темам (разделам)

Семестр: 3 (Осенний)

1. Стандартные модели причинно-следственного анализа

Контрфактуальность. Причинно-следственный анализ и предсказание как задачи анализа данных.

2. Причинно-следственная идентификация в статистике: модель потенциальных исходов Рубина и предположение о стабильности эффекта влияния (SUTVA).

Структурные частично-линейные модели (ЧЛМ).

3. Модели на направленных ациклических графах (DAG).

Критерий обходного пути. Причинно-следственный поиск.

4. Инструментальные переменные, разрывный регрессионный дизайн и синтетические контрольные переменные в моделях DAG.

Причинно-следственный анализ с использованием данных высокой размерности.

5. Бутстрап

Асимптотика big-n. Итеративная оценка ЧЛМ. Ядерная плотность, разложение в ряды.

5. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)

Стандартная учебная аудитория. Вычислительное устройство с доступом в Интернет (компьютер, ноутбук, планшет и т.д.) для самостоятельной работы.

6.Перечень рекомендуемой литературы

Основная литература

1. Статистика для всех / С. Бослаф. — Москва, ДМК Пресс, 2015.— URL: <https://e.lanbook.com/book/66475> (дата обращения: 26.01.2021). - Полный текст (Режим доступа : из сети МФТИ / Удаленный доступ)

Дополнительная литература

1. Вероятность и статистика , Электрон. версия печ. публикации / В. Б. Монсик, А. А. Скрынников. — Москва, Лаборатория знаний, 2020

7. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля)

Не используются

8. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень необходимого программного обеспечения и информационных справочных систем (при необходимости)

На лекционных занятиях используются мультимедийные технологии, включая демонстрацию презентаций.

Для контроля и коррекции знаний, обучающиеся могут использовать компьютерное тестирование.

9. Методические указания для обучающихся по освоению дисциплины (модуля)

Успешное освоение курса требует напряжённой самостоятельной работы студента. В программе курса приведено минимально необходимое время для работы студента над темой. Самостоятельная работа включает в себя:

- чтение и конспектирование рекомендованной литературы,
- проработку учебного материала (по конспектам лекций, учебной и научной литературе), подготовку ответов на вопросы, предназначенных для самостоятельного изучения, доказательство отдельных утверждений, свойств;
- решение задач, предлагаемых студентам на практических занятиях и в качестве курсового задания,
- подготовку к экзамену.

Руководство и контроль за самостоятельной работой студента осуществляется в форме индивидуальных консультаций.

Показателем владения материалом служит умение решать задачи. Для формирования умения применять теоретические знания на практике студенту необходимо решать как можно больше задач. При решении задач каждое действие необходимо аргументировать, ссылаясь на известные теоретические сведения.

Важно добиться понимания изучаемого материала, а не механического его запоминания. При затруднении изучения отдельных тем, вопросов, следует обращаться за консультациями к лектору или преподавателю, ведущему практические занятия.

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению: Информатика и вычислительная техника
профиль подготовки: Прикладная математика и информатика
Физтех-школа Прикладной Математики и Информатики
кафедра дискретной математики
курс: 2
квалификация: магистр

Семестр, формы промежуточной аттестации: 3 (осенний) - Экзамен

Разработчик: А.М. Ченцов, ассистент

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
УК-2 Способен управлять проектом на всех этапах его жизненного цикла	УК-2.2 Способен прогнозировать результат деятельности и планировать последовательность шагов для достижения данного результата. Формирует план-график реализации проекта в целом и план контроля его выполнения
	УК-2.3 Способен организовать и координировать работу участников проекта, обеспечивать работу команды необходимыми ресурсами
ОПК-2 Имеет представление об актуальных проблемах науки и техники в области информатики и вычислительной техники, способен на научном языке формулировать профессиональные задачи	ОПК-2.1 Имеет представление о современном состоянии исследований в рамках тематической области своей профессиональной деятельности
	ОПК-2.2 Способен оценивать актуальность исследований в области информатики и вычислительной техники и их практическую значимость
	ОПК-2.3 Владеет профессиональной терминологией, используемой в современной научно-технической литературе, обладает навыками устного и письменного изложения результатов научной деятельности в рамках профессиональной коммуникации
ПК-2 Понимает и способен применить в научно-исследовательской и прикладной деятельности основные законы естествознания, современный математический аппарат и алгоритмы, современные информационно-коммуникационные технологии	ПК-2.1 Знает основы научно-исследовательской деятельности в области информационных технологий, владеет знанием основ философии и методологии науки; знанием методов научных исследований и навыками их проведения
	ПК-2.2 Умеет применять полученные знания в области фундаментальных научных основ теории информации и решать стандартные задачи в собственной научно-исследовательской деятельности
	ПК-2.3 Имеет практический опыт научно-исследовательской деятельности в области информационно-коммуникационных технологий
	ПК-2.4 Владеет методами и алгоритмами решения задач цифровой обработки сигналов, использования сети Интернет, аннотирования, реферирования, библиографического поиска, опыт работы с научными источниками

2. Показатели оценивания компетенций

В результате изучения дисциплины «Причинно-следственные выводы в статистике» обучающийся должен:

знать:

- основы базовой статистики и регрессионного анализа,

уметь:

- понять поставленную задачу;
- оценивать корректность постановок задач;
- строго доказывать или опровергать утверждение;
- самостоятельно находить алгоритмы решения, в том числе и нестандартных, и проводить их анализ;
- самостоятельно видеть следствия полученных результатов;
- давать экспертную оценку финальным результатам решения.

владеть:

- навыками освоения большого объема информации и решения задач;
- навыками самостоятельной работы и освоения новых дисциплин;
- культурой постановки, анализа и решения математических и прикладных задач, требующих для своего решения использования математических подходов;
- предметным языком прикладной статистики и навыками грамотного описания решения задач и представления полученных результатов.
- навыками компьютерной обработки информации.

3. Перечень типовых (примерных) вопросов, заданий, тем для подготовки к текущему контролю

1. Асимптотика big-p. «Проклятие размерности». Свойства пространств высокой размерности.
2. Выбор модели, регуляризация.
3. Неравномерная сходимость распределения постселекционных оценок. Пример: слабые инструменты.
4. Двойной выбор в структурной ЧЛМ.
5. Ортогональность по Нейману и двойное машинное обучение
6. (опционально) Гетерогенные эффекты воздействия

4. Перечень типовых (примерных) вопросов и тем для проведения промежуточной аттестации обучающихся

1. Стандартные модели причинно-следственного анализа
2. Контрфактуальность. Причинно-следственный анализ и предсказание как задачи анализа данных.
3. Причинно-следственная идентификация в статистике: модель потенциальных исходов Рубина и предположение о стабильности эффекта влияния (SUTVA). Структурные частично-линейные модели (ЧЛМ).
4. Модели на направленных ациклических графах (DAG). Критерий обходного пути. Причинно-следственный поиск.
5. Инструментальные переменные, разрывный регрессионный дизайн и синтетические контрольные переменные в моделях DAG.
6. Асимптотика big-n. Итеративная оценка ЧЛМ. Ядерная плотность, разложение в ряды.
7. Бутстрап
8. Асимптотика big-p. «Проклятие размерности». Свойства пространств высокой размерности.
9. Выбор модели, регуляризация.
10. Неравномерная сходимость распределения постселекционных оценок. Пример: слабые инструменты.
11. Двойной выбор в структурной ЧЛМ.
12. Ортогональность по Нейману и двойное машинное обучение
13. (опционально) Гетерогенные эффекты воздействия

Примеры экзаменационных билетов:

Билет № 1

1. Модели на направленных ациклических графах (DAG). Критерий обходного пути. Причинно-следственный поиск.
2. Выбор модели, регуляризация.

Билет № 2

1. Асимптотика big-n. Итеративная оценка ЧЛМ. Ядерная плотность, разложение в ряды.
2. Двойной выбор в структурной ЧЛМ.

Критерии оценивания

Оценка «отлично (10)» выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений;

Оценка «отлично (9)» выставляется студенту, показавшему систематизированные, глубокие знания учебной программы дисциплины и умение применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений;

Оценка «отлично (8)» выставляется студенту, показавшему систематизированные, знания учебной программы дисциплины и умение применять их на практике при решении конкретных задач, правильное обоснование принятых решений;

Оценка «хорошо (7)» выставляется студенту, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности;

Оценка «хорошо (6)» выставляется студенту, если он твердо знает материал, грамотно излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности;

Оценка «хорошо (5)» выставляется студенту, если он знает материал, грамотно излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности;

Оценка «удовлетворительно (4)» выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он владеет основными разделами учебной программы, необходимыми для дальнейшего обучения и может применять полученные знания по образцу в стандартной ситуации;

Оценка «удовлетворительно (3)» выставляется студенту, показавшему фрагментарный характер знаний, недостаточно правильные формулировки базовых понятий, но при этом он владеет основными разделами учебной программы, необходимыми для дальнейшего обучения и может применять полученные знания по образцу в стандартной ситуации;

Оценка «неудовлетворительно (2)» выставляется студенту, который не знает большей части основного содержания учебной программы дисциплины, допускает грубые ошибки в формулировках основных понятий дисциплины и не умеет использовать полученные знания при решении типовых практических задач.

Оценка «неудовлетворительно (1)» выставляется студенту, который не знает основного содержания учебной программы дисциплины, допускает грубые ошибки в формулировках основных понятий дисциплины и не умеет использовать полученные знания при решении типовых практических задач.

5. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности

Во время проведения экзамена обучающиеся могут пользоваться программой дисциплины.