

**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:	Основы прикладной статистики
по направлению:	Информатика и вычислительная техника
профиль подготовки:	Системное программирование и прикладная математика Физтех-школа Прикладной Математики и Информатики кафедра дискретной математики
курс:	3
квалификация:	бакалавр

Семестр, формы промежуточной аттестации: 5 (осенний) - Дифференцированный зачет

Аудиторных часов: 60 всего, в том числе:

лекции: 30 час.

семинары: 30 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 75 час.

Всего часов: 135, всего зач. ед.: 3

Программу составил: И.В. Родионов, канд. физ.-мат. наук, доцент

Программа обсуждена на заседании кафедры дискретной математики 05.03.2020

Аннотация

Курс посвящен основным методам статистического анализа реальных данных, которые подкрепляются необходимой математической теорией, являясь дополнением до курса математической статистики. В курсе разбираются основа всех статистических инструментов, без которых невозможно корректно применять статистику на практике. Особое внимание уделяется особенностям работы с реальными данными, а так же их подготовке. В рамках курса предполагается практика на языке Python.

1. Цели и задачи

Цель дисциплины

изучение математи-ческих и теоретических основ современного статистического анализа, а также подготовка слушателей к дальнейшей самостоятельной работе в области анализа статистических задач прикладной математики, физики и экономики.

Задачи дисциплины

- ☐ изучение математических основ математической статистики;
- ☐ приобретение слушателями теоретических знаний в области современного статистического анализа.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-4 Способен осуществлять сбор и обработку научно-технической и (или) технологической информации для решения фундаментальных и прикладных задач	ОПК-4.1 Владеет методами научного поиска и интеллектуального анализа информации при решении задач профессиональной деятельности
	ОПК-4.2 Знает основные источники научно-технической и (или) технологической информации в области профессиональной деятельности
	ОПК-4.3 Умеет составлять аннотации, рефераты, библиографические перечни и обзоры информации в области своей профессиональной деятельности
	ОПК-4.4 Владеет навыками работы с компьютером и компьютерными сетями с целью получения, хранения и обработки научной (технической, технологической) информации
ОПК-5 Способен участвовать в проведении фундаментальных и прикладных исследований и разработок, самостоятельно осваивать новые теоретические, в том числе, математические методы исследований и работать на современной экспериментальной научно-исследовательской, измерительно-аналитической и технологической аппаратуре)	ОПК-5.1 Способен решать поставленные задачи в области теоретических и экспериментальных исследований и разработок
	ОПК-5.2 Обладает способностью к освоению новых знаний на основе изучения литературы, научных статей и других источников
	ОПК-5.3 Способен к профессиональной эксплуатации современной экспериментальной научно-исследовательской (измерительно-аналитической и технологической) аппаратуры
ПК-1 Способен ставить, формализовывать и решать задачи, в том числе разрабатывать и исследовать математические модели изучаемых явлений и процессов, системно анализировать научные проблемы, получать	ПК-1.2 Способен выдвигать гипотезы, строить математические модели для описания изучаемых явлений и процессов, оценивать качество разработанной модели
	ПК-1.1 Способен находить, анализировать и обобщать информацию об актуальных результатах исследований в рамках тематической области своей профессиональной деятельности

новые научные результаты	ПК-1.3 Способен применять теоретические и (или) экспериментальные методы исследований к конкретной научной задаче и интерпретировать полученные результаты
ПК-2 Способен самостоятельно или в качестве члена (руководителя) малого коллектива организовывать и проводить научные исследования и их апробацию	ПК-2.1 Знает принципы построения научной работы, методы сбора и анализа полученного материала, способы аргументации
	ПК-2.2 Способен планировать и проводить научные исследования самостоятельно или в качестве члена (руководителя) малого научного коллектива
	ПК-2.3 Способен проводить апробацию результатов научно-исследовательской работы посредством публикации научных статей и участия в конференциях

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- основные понятия математической статистики;
- основные подходы к сравнению оценок параметров неизвестного распределения;
- асимптотические и неасимптотические свойства оценок параметров неизвестного распределения;
- основные методы построения оценок с хорошими асимптотическими свойствами: метод моментов, метод максимального правдоподобия, метод выборочных квантилей;
- понятие эффективных оценок и неравенство информации Рао-Крамера;
- определение и главные свойства условного математического ожидания случайной величины относительно сигма-алгебры или другой случайной величины;
- определение общей линейной регрессионной модели и метод наименьших квадратов;
- многомерное нормальное распределение и его основные свойства;
- базовые понятия теории проверки статистических гипотез;
- лемму Неймана – Пирсона и теорему о монотонном отношении правдоподобия;
- критерий хи-квадрат Пирсона для проверки простых гипотез в схеме Бернулли.

уметь:

- обосновывать асимптотические свойства оценок с помощью применения предельных теорем теории вероятностей;
- строить оценки с хорошими асимптотическими свойствами для параметров неизвестного распределения по заданной выборке из него;
- находить байесовские оценки по заданному априорному распределению;
- вычислять условные математические ожидания с помощью условных распределений;
- находить оптимальные оценки с помощью полных достаточных статистик;
- строить точные и асимптотические доверительные интервалы, и области для параметров неизвестного распределения;
- находить оптимальные оценки и доверительные области в гауссовской линейной модели;
- строить равномерно наиболее мощные критерии в случае параметрического семейства с монотонным отношением правдоподобия;
- строить F-критерий для проверки линейных гипотез в линейной гауссовской модели.

владеть:

- основными методами математической статистики построения точечных и доверительных оценок: методом моментов, выборочных квантилей, максимального правдоподобия, методом наименьших квадратов, методом центральной статистики.
- навыками асимптотического анализа статистических критериев;
- навыками применения теорем математической статистики в прикладных задачах физики и экономики.

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	Статистики и оценки. Метод максимального правдоподобия. Методы построения доверительных интервалов.	4	4		12
2	Методы множественной проверки гипотез: Бонферрони, Холма, Шидака, Шидака-Холма, Бенджамини-Хохберга, Бенджамини-Иекутиели.	4	4		12
3	Коэффициенты корреляции Пирсона, Стьюдента, Кендалла и их свойства. Частная и множественная корреляция.	4	4		12
4	Двухфакторная модель. Взаимодействие факторов, его интерпретация. Двухфакторный нормальный анализ.	6	6		12
5	Регуляризация в линейной регрессии, свойства решений. Связь с байесовскими оценками.	6	6		12
6	Дополнительные библиотеки анализа данных на Python: pandas, seaborn, ipywidgets, qgrid, plotly.	6	6		15
Итого часов		30	30		75
Подготовка к экзамену		0 час.			
Общая трудоёмкость		135 час., 3 зач.ед.			

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

Семестр: 5 (Осенний)

1. Статистики и оценки. Метод максимального правдоподобия. Методы построения доверительных интервалов.

Байесовские оценки, полный байесовский вывод. Сопряженные распределения.

Проверка статистических гипотез. Лемма Неймана-Пирсона и критерий монотонного отношения правдоподобия. Достигаемый уровень значимости. Статистическая и практическая значимость. Кривые мощности.

2. Методы множественной проверки гипотез: Бонферрони, Холма, Шидака, Шидака-Холма, Бенджамини-Хохберга, Бенджамини-Иекутиели.

Критерии согласия: Колмогорова (и критерии на его основе), хи-квадрат, Шапиро-Уилка. Квантиль-квантиль график.

Перестановочные критерии для среднего. Метод бутстрепа. Метод множественной проверки гипотез на основе метода бутстрепа. Метод случайного леса.

3. Коэффициенты корреляции Пирсона, Стьюдента, Кендалла и их свойства. Частная и множественная корреляция.

Таблицы сопряженности. Проверка гипотезы независимости с помощью критерия хи-квадрат и критерия Фишера.

Однофакторная модель дисперсионного анализа. Независимые выборки: критерии Фишера, Краскела-Уоллиса, Джонкхиера. Связанные выборки: критерии Фишера, Фридмана и Пейджа. Модель с фиксированным эффектом, уточнение различий: методы LSD и HSD, критерии Неменьи и Даннета. Проверка гипотезы о равенстве дисперсий: критерии Бартлета и Флайнера-Киллиана.

4. Двухфакторная модель. Взаимодействие факторов, его интерпретация. Двухфакторный нормальный анализ.

Линейная регрессия. Остаточная сумма квадратов, коэффициент детерминации. Мультиколлинеарность. Доверительные интервалы для дисперсии шума, коэффициентов регрессии, прогнозируемого значения отклика.

Анализ регрессионных остатков: визуальный анализ, проверка гипотез несмещённости, гомоскедастичности, нормальности. Обработка выбросов, расстояние Кука. Метод Бокса-Кокса для преобразования отклика. Устойчивая оценка дисперсии Уайта, её модификации.

5. Регуляризация в линейной регрессии, свойства решений. Связь с байесовскими оценками.

Методы снижения размерности: PCA, t-SNE.

Последовательный анализ в задачах проверки гипотез о значениях параметра.

6. Дополнительные библиотеки анализа данных на Python: pandas, seaborn, ipywidgets, qgrid, plotly.

Язык R и его библиотеки.

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

Стандартная учебная аудитория.

6. Перечень рекомендуемой литературы

Основная литература

1. Математическая статистика [Текст] : [учебник для вузов] / А. А. Боровков .— [3-е изд., испр.] .— М. : Физматлит, 2007 .— 704 с.
2. Введение в математическую статистику [Текст] : [учебник для вузов] / Г. И. Ивченко, Ю. И. Медведев .— М. : ЛКИ, 2010, 2014, 2015 .— 600 с.
3. Наглядная математическая статистика [Текст] : учеб. пособие для вузов / М. Б. Лагутин .— 2-е изд., испр. — М. : Бинوم. Лаб. знаний, 2009 .— 472 с.

Дополнительная литература

1. Курс теории вероятностей и математической статистики [Текст] : [учеб. пособие для вузов] / Б. А. Севастьянов .— М. ; Ижевск : Ин-т компьютерных исследований, 2004 .— 272 с.

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.

В процессе самостоятельной работы обучающихся возможно использование таких программных средств, как Mathcad, MATLAB, Maple и др.

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

1. Рекомендуется успешно сдавать контрольные работы, так как это упрощает итоговую аттестацию по предмету.
2. Для подготовки к итоговой аттестации по предмету лучше всего пользоваться материалами лекций.

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению:	Информатика и вычислительная техника
профиль подготовки:	Системное программирование и прикладная математика Физтех-школа Прикладной Математики и Информатики кафедра дискретной математики
курс:	3
квалификация:	бакалавр

Семестр, формы промежуточной аттестации: 5 (осенний) - Дифференцированный зачет

Разработчик: И.В. Родионов, канд. физ.-мат. наук, доцент

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-4 Способен осуществлять сбор и обработку научно-технической и (или) технологической информации для решения фундаментальных и прикладных задач	ОПК-4.1 Владеет методами научного поиска и интеллектуального анализа информации при решении задач профессиональной деятельности
	ОПК-4.2 Знает основные источники научно-технической и (или) технологической информации в области профессиональной деятельности
	ОПК-4.3 Умеет составлять аннотации, рефераты, библиографические перечни и обзоры информации в области своей профессиональной деятельности
	ОПК-4.4 Владеет навыками работы с компьютером и компьютерными сетями с целью получения, хранения и обработки научной (технической, технологической) информации
ОПК-5 Способен участвовать в проведении фундаментальных и прикладных исследований и разработок, самостоятельно осваивать новые теоретические, в том числе, математические методы исследований и работать на современной экспериментальной научно-исследовательской, измерительно-аналитической и технологической аппаратуре)	ОПК-5.1 Способен решать поставленные задачи в области теоретических и экспериментальных исследований и разработок
	ОПК-5.2 Обладает способностью к освоению новых знаний на основе изучения литературы, научных статей и других источников
	ОПК-5.3 Способен к профессиональной эксплуатации современной экспериментальной научно-исследовательской (измерительно-аналитической и технологической) аппаратуры
ПК-1 Способен ставить, формализовывать и решать задачи, в том числе разрабатывать и исследовать математические модели изучаемых явлений и процессов, системно анализировать научные проблемы, получать новые научные результаты	ПК-1.2 Способен выдвигать гипотезы, строить математические модели для описания изучаемых явлений и процессов, оценивать качество разработанной модели
	ПК-1.1 Способен находить, анализировать и обобщать информацию об актуальных результатах исследований в рамках тематической области своей профессиональной деятельности
	ПК-1.3 Способен применять теоретические и (или) экспериментальные методы исследований к конкретной научной задаче и интерпретировать полученные результаты
ПК-2 Способен самостоятельно или в качестве члена (руководителя) малого коллектива организовывать и проводить научные исследования и их апробацию	ПК-2.1 Знает принципы построения научной работы, методы сбора и анализа полученного материала, способы аргументации
	ПК-2.2 Способен планировать и проводить научные исследования самостоятельно или в качестве члена (руководителя) малого научного коллектива
	ПК-2.3 Способен проводить апробацию результатов научно-исследовательской работы посредством публикации научных статей и участия в конференциях

2. Показатели оценивания компетенций

В результате изучения дисциплины «Основы прикладной статистики» обучающийся должен:

знать:

- основные понятия математической статистики;
- основные подходы к сравнению оценок параметров неизвестного распределения;
- асимптотические и неасимптотические свойства оценок параметров неизвестного распределения;
- основные методы построения оценок с хорошими асимптотическими свойствами: метод моментов, метод максимального правдоподобия, метод выборочных квантилей;
- понятие эффективных оценок и неравенство информации Рао-Крамера;
- определение и главные свойства условного математического ожидания случайной величины относительно сигма-алгебры или другой случайной величины;
- определение общей линейной регрессионной модели и метод наименьших квадратов;
- многомерное нормальное распределение и его основные свойства;
- базовые понятия теории проверки статистических гипотез;
- лемму Неймана – Пирсона и теорему о монотонном отношении правдоподобия;
- критерий хи-квадрат Пирсона для проверки простых гипотез в схеме Бернулли.

уметь:

- обосновывать асимптотические свойства оценок с помощью применения предельных теорем теории вероятностей;
- строить оценки с хорошими асимптотическими свойствами для параметров неизвестного распределения по заданной выборке из него;
- находить байесовские оценки по заданному априорному распределению;
- вычислять условные математические ожидания с помощью условных распределений;
- находить оптимальные оценки с помощью полных достаточных статистик;
- строить точные и асимптотические доверительные интервалы, и области для параметров неизвестного распределения;
- находить оптимальные оценки и доверительные области в гауссовской линейной модели;
- строить равномерно наиболее мощные критерии в случае параметрического семейства с монотонным отношением правдоподобия;
- строить F-критерий для проверки линейных гипотез в линейной гауссовской модели.

владеть:

- основными методами математической статистики построения точечных и доверительных оценок: методом моментов, выборочных квантилей, максимального правдоподобия, методом наименьших квадратов, методом центральной статистики.
- навыками асимптотического анализа статистических критериев;
- навыками применения теорем математической статистики в прикладных задачах физики и экономики.

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

Примеры задач:

Пусть X_1, \dots, X_n --- выборка из распределения Лапласа со сдвигом θ , то есть плотность имеет вид $p_{\theta}(x) = \frac{1}{2} e^{-|x-\theta|}$. Сравните в асимптотическом подходе выборочное среднее \overline{X} , усеченное среднее \overline{X}_{α} , выборочную медиану $\widehat{\mu}$, медиану средних Уолша W .

Найдите асимптотическую толерантность медианы средних Уолша W .

Пусть $X = (X_1, \dots, X_n)$ --- выборка из неизвестного распределения \mathcal{P} с плотностью $p(x)$ и $\widetilde{p}_n(x)$ --- построенная по ней ядерная оценка плотности. Пусть также ξ --- случайная величина из распределения \mathcal{P} , независимая с выборкой X . Рассмотрим ожидаемую среднеквадратичную ошибку $E \left(\widetilde{p}_n(\xi) - p(\xi) \right)^2$. Какова асимптотика оптимальной ширины ядра, минимизирующей эту ошибку?

Пусть $X_j = (X_{j1}, \dots, X_{jn})$, $j \in \{1, 2\}$ --- две выборки. Для каждой выборки проверяются гипотезы H_j vs. H'_j с помощью критерия S_j уровня значимости α . Предположим, гипотезы H_j верны. Какие значения может принимать FWER? Чему соответствуют пограничные значения, а также случай независимости выборок?

Пусть X_1, \dots, X_n --- выборка из распределения Бернулли с параметром θ . Вычислите $MSE_{\widehat{\theta}}(\theta)$, где $\widetilde{\theta} = \overline{X} + \frac{1}{1+\sqrt{n}}$ $\left(\frac{1}{2} - \overline{X} \right)$ --- оценка Ходжеса-Лемана.

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

Перечень вопросов:

1. Сгенерировать выборку из какого-то распределения, добавить к ней шум. Построить ящики с усами. Посчитать усеченное среднее и медиану средних Уолша.
2. Исследовать поведение усеченного среднего или медианы средних Уолша, если количество выбросов в выборке больше, чем асимптотическая толерантность. Распределения нормальное, Стьюдента, Коши, Лапласа.
3. Исследовать поведение выборочной медианы, если распределение симметрично с носителем $(-2, -1) \cup (1, 2)$.
4. Построить среднеквадратичный риск от тета для выборочного среднего, выборочной медианы, усеченного среднего, медианы средних Уолша. Распределения нормальное, Стьюдента, Коши, Лапласа. Что можно сказать в разных подходах к сравнению оценок?
5. Построить РНМК для нормальных распределений для $H_0: \theta=0$ vs $H_1: \theta>0$. С какой вероятностью H_0 будет отвергнута, если она верна, но в данных присутствует шум из Коши? Ответ дать в зависимости от доли шума.
6. Построить критерий Вальда для пуассоновских распределений для $H_0: \theta=\theta_0$ vs $H_1: \theta>\theta_0$. Посчитать реальный уровень значимости, а также мощность критерия при разных альтернативах.
7. Построить критерий проверки нормальности (с произвольными параметрами) на основе критерия хи-квадрат. Аналогично для других разбиений. Сравнить его мощность с другими критериями.

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 всесторонние, 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 всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение применять их на практике при решении конкретных задач, и 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 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 (5)» выставляется студенту, если он знает материал, и по существу 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, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он владеет фрагментарно 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 формулировок основных понятий дисциплины.

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

Во время проведения дифференцированного зачета обучающиеся могут пользоваться программой дисциплины.