DS-поток, 3 курс, осень 2024

**Порядок проведения зачета**

**Общее**

1. Доп. вопросы могут быть также по соответствующим темам курса Введение в анализ данных.
2. Вопросы могут быть и по другим предметам при наличии необходимости, если они тесно связаны с программой. *Например, если экзаменатор спросил что-то про состоятельность оценки, то и про сходимость по вероятности нужно уметь ответить.*
3. Зачет проходит очно в соответствии с графиком **(будет)**. Каждый приходит к своему времени и отвечает 25 минут.   
   **Аудитория: поточная Арктика**
4. Опоздание студента может стать причиной задержек экзаменатора по приему следующих студентов. **Поэтому в случае опоздания студента экзаменатор может закончить экзамен в момент окончания его ответа по расписанию, снизив оценку пропорционально времени задержки.**
5. Распределение по экзаменаторам определяется генерацией случайных чисел в момент захода группы студентов в аудиторию при наличии возможности.
6. В экзамен входит весь материал, который был на занятиях. В том числе доказательства различных утверждений и вывод формул.

**Порядок зачета**

1. В момент начала ответа группы студентов (по количеству свободных экзаменаторов) одним из экзаменаторов производится случайное распределение студентов по экзаменаторам с помощью генерации случайных чисел (при наличии возможности).
2. Все студенты из группы начинают отвечать сразу. Времени на подготовку к чему-либо после указанного в графике времени нет.
3. **Первый этап: быстрые вопросы**, на которые нужно давать ответ сразу без использования материалов. Вопроса два:
   * формулировка определения + пояснение
   * формулировка утверждения/теоремы + пояснение

Уточнения:

* + В качестве пояснения можно рассказать, например, чем полезно именно такое определение/утверждение, или почему оно интуитивно и согласуется с простой логикой.
  + 50% баллов баллов дает пояснение.
  + За пояснения нельзя получить баллы без верной формулировки.
  + Обдумывание более 30 секунд приравнивается к отсутствию ответа.
  + Общее время - 4 минуты.
  + *Если получено менее 30% баллов, зачет заканчивается 0 баллов.*

1. **Второй этап: обсуждение** одной темы из числа **жирных зеленых.** 
   * Конкретную тему экзаменатор определяет с помощью генерации случайных чисел (при наличии возможности).
   * Рассказывать нужно сразу, без подготовки и без использования материалов.
   * На листе бумаги необходимо записывать формулы, возникающие во время рассказа.
   * Общее время - 9 минут.
   * *Если получено менее 20% баллов, зачет заканчивается 0 баллов.*
2. **Третий этап: обсуждение** одной темы из числа простых зеленых.
   * Конкретные темы экзаменатор определяет с помощью генерации случайных чисел (при наличии возможности).
   * Общее время - 9 минут.
   * *Если получено менее 20% баллов, зачет заканчивается 0 баллов.*
   * *Можно пользоваться*
     + любыми бумажными материалами,
     + заранее скачанными электронными материалами,
     + страницей курса,
     + этим документом.
3. **На десерт:** вопрос на понимание.
   * Можно использовать материалы по списку выше.
   * Общее время - 3 минуты.
   * Обдумывание более 1 минуты приравнивается к отсутствию ответа.

На этапах 2 и 3 экзаменатор имеет право снизить баллы по своему усмотрению в случае если ответ студента слишком неспешный и за 9 минут рассказано не сильно больше половины вопроса. Волноваться не стоит, это скорее исключения, но они бывают.

**Замечания относительно устного ответа с использованием материалов**

* На прошлых зачетах и экзаменах некоторые студенты почти не пытались заранее разобраться в каком-либо материале, а пытались разобраться уже на ходу во время ответа. Принимающие это понимают. Смысл того, что мы разрешаем пользоваться материалами, не в том, чтобы вы на ходу пытались разобраться, а в том, чтобы вам не приходилось выучивать все формулы и прочие моменты. Но это значит, что по материалам вы должны сходу все пояснить.  
  Если экзаменатор поймет, что вы пытаетесь разобраться на ходу, он имеет право не засчитывать ответ.
* Ответ студента в стиле "ну тут понятно, тут все очевидно" может быть приравнен к отсутствию ответа.

**Программа**

**Принцип формирования программы**

1. Программа состоит из:
   1. Одно занятие со 2 курса
   2. Около половины лекций по статистике с 3 курса
   3. Немного из семинаров, что не попало в лекции.
2. Из таблицы тем 3 курса **к зачету относится только то**, что выделено з**е**л**е**н**ы**м**.**
   1. Те части программы, которые выделены **жирным зеленым**, необходимо уметь рассказывать без обращения к материалам.
   2. Остальные части программы, выделенные простым зеленым, необходимо понимать и уметь пояснять с использованием материалов.
3. **Программа примерная. Приведена только разметка, а не полный список того, что будет спрашиваться.** Полная программа - все, что было на лекциях по выделенным темам.
4. Вы слушали “единый” курс статистики, формально их два разных, по одному зачет, по другому - экзамен. Это сделано с точки зрения удобства изложения материала и его изучения вами. Однако, к курсу мат. статистики предъявляются некоторые требования, что сказывается на экзамене.
5. В ограничениях выше программа зачета построена таким образом, чтобы
   1. охватить все, что не войдет в экзамен;
   2. взять из программы экзамена то, что наиболее важно и логично дополняет программу зачета;
   3. некоторым образом “сгладить” программы зачета и экзамена, сделав более логичным их сочетание.
6. На будущее: тем самым программа экзамена - то, что не выделено зеленым, и наиболее важное из того, что выделено.

| **Дата** | **Лекции** | **Семинары** |
| --- | --- | --- |
| 01.09 | 1\.1 Основная задача математической статистики. 1.2 Вероятностно-статистическая модель. 1.3 Виды подходов к статистике. Примеры.  Численное интегрирование: метод прямоугольников, метод Монте-Карло. | Повторение теории вероятностей. |
| 08.09 | 2\.1 Статистики и оценки, определения, примеры. 2.2 Несмещенность, примеры. Несмещенные оценки в линейной модели регрессии. 2.3 Состоятельность и сильная состоятельность. 2.4 Доверительные интервалы, точные д.и., асимптотические д.и. Особенности реализации дов. интервала. Асимптотическая нормальность, пример. Асимптотические интервалы Вальда, примеры. Дельта-метод, пример. | Оценки и их свойства. Несмещенность, состоятельность, ассимптотическая нормальность. Наследование свойств, дельта-метод. Метод моментов.  Визуализация состоятельности и ассимптотической нормальности |
| 15\.09 | 2\.5 Наследование свойств. Теорема о наследовании асимптотической нормальности, пример, наследование (сильной) состоятельности. Лемма Слуцкого, теорема о производной (док-во для d=1), пример. Дельта-метод (док-во), пример. Док-во корректности интервалов Вальда.  2\.6 Точные доверительные интервалы и доверительные области. Центральный интервал (метод центральной функции), пример. Точные доверительные интервалы в нормальной модели для среднего и дисперсии при известном и неизвестном другом параметре. | Интервальное оценивание. Методы построения доверительных интервалов. Метод центральной функции. Доверительные интервалы в распределении Бернулли  Сравнение доверительных интервалов Вальда и Вилсона |
| 22\.09 | 3\.1 Метод максимального правдоподобия, его смысл, свойства. Задачи для разных типов распределений. 3.2 Задача оценки сдвига в модели распределения Коши (гамма-котики), поведение выборочного среднего (док-во). Метод выборочных квантилей, асимптотическая нормальность выборочной квантили и выборочной медианы.  4\.1 Достаточные статистики, их смысл и применение для последовательно поступающих данных. Пример для бернуллиевского распределения. Критерий факторизации Неймана-Фишера, пример. | Метод максимального правдоподобия, оценки ОМП.  Оценки параметров сдвига и масштаба в распределении Коши |
| 29\.09 | 4\.2 Экспоненциальный класс распределений, естественная параметризация. Пример для нормального распределения. Достаточная статистика для семейства из эксп. класса, существование достаточных статистик фиксированного размера для произвольных семейств распределений. Матем. ожидание и дисперсия достаточных статистик эксп. класса (док-во), свойства ОМП для него (док-во). 4.3 Сравнение оценок. Функция потерь и функция риска, примеры. Подходы к сравнению оценок: равномерный, байесовский, минимаксный, асимптотический. Пример, в котором нет наилучшей оценки в равномерном подходе. Bias-variance разложение (док-во). | Сравнение оценок. Равномерный, Байесовский, мин-максный, ассимптотический подходы.  Оценки ОМП и Ходжеса-Лемана для выборки из распределения Бернулли |
| 6\.10 | 4\.4 Метод Ньютона и поиск ОМП с его помощью. Одношаговые оценки, идея док-ва d=1. Одношаговая оценка в модели распределения Коши.  4\.5 Робастность, асимптотическая толерантность как мера робастности, примеры. Усеченное среднее и медиана средних Уолша их асимптотическая толерантность и асимптотическая эффективность.  5\.1 **Эмпирическое распределение, его свойства. Эмпирическая функция распределения**, теоремы Гливенко-Кантелли, Вапника-Червоненкиса, Колмогорова-Смирнова. 5.2 Метод подстановки, примеры. | Достаточные статистики  Робастные оценки  Приближенный поиск ОМП |
| 13\.10 | Пример про коэффициенты асимметрии и эксцесса. 5.3 Сложность вычисления оценки дисперсии оценки методом подстановки. **Схема бутстрепа, способ генерации бутстрепных выборок, примеры. Особенности бутстрепа.** Бутстрепные интервалы: нормальный, центральный, квантильный, примеры. 5.4 **Ядерная оценка плотности** и ее свойства. Связь с ЭФР. Теорема об асимптотической нормальности. Теорема об оптимальной ширине ядра для MSE. Оптимальное ядро по MSE. Подбор ширины ядра по выборке.  5\.5 Методы классификации и регрессии на основе ближайших соседей. **kNN и взвешенный kNN**. Приближенный поиск ближайших соседей с помощью LSH. Ядерная регрессия, теорема о сходимости, подбор ширины ядра. Локальная линейная регрессионная модель. | Повторение эмпирического распределения и метода подстановки. Бутстреп: алгоритм, размеры и количества бутстрепных выборок, плюсы и минусы бутстрепа. |
| 20\.10 | 6\.1. **Гипотеза, различные случаи. Критерии для проверки гипотез, частый случай - правосторонний критерий, для него: статистика критерия, критическое значение. Результаты тестирования гипотез, стат. значимость, связь с презумпцией невиновности. Ошибки 1 и 2 рода, уровень значимости. Альтернативные гипотезы, мощность. Пример. 6.2. Критерий Вальда, его мощность, замечания: виды альтернатив, сложная H\_0, дов. интервалы.** | Критерии: определения, критерий Вальда, критерии отношения правдоподобия (материал на доске) |
| 3\.11 | 6\.3 **Пример: критерий для AB-тестирования для бернуллиевских выборок. Величина p-value, основное правило, пример. Утверждение о распределении p-value. Общий случай p-value. Примеры того, что не есть p-value. 6.4 Поведение критического значения при росте размера выборки, выводы с точки зрения практики. Размер эффекта, практическая значимость. Соотношение практ. и стат. значимостей. Подбор необходимого размера выборки до эксперимента.** 6.5 Критерии согласия. Критерий Колмогорова, его свойства, другие подобные критерии. QQ plot. Критерии проверки нормальности - Жарка-Бера, Шапиро-Уилка, особенности проверки нормальности. Критерий хи-квадрат. Обобщенный критерий хи-квадрат, пример. | Оценка вероятности ошибки первого рода. p-value. Критерии хи-квадрат и обобщенный хи-квадрат (материал на доске) |
| 10\.11 | 6\.6. Пример про экстрасенсов. **FWER, методы Бонферрони (док-во) и Холма.** FDR, методы Бенджамини-Хохберга, Бенджамини-Иекутиели. Корректировка p-value. Примеры. Рекомендации по использованию. | Множественная проверка гипотез: понятия FWER и FDR, их соотношение, методы контролирующие FWER: Бонферрони, Холма, Шидака и Шидака-Холма |
| 17\.11 | 6\.7 Анализ зависимостей. **Коэффициенты корреляции Пирсона, Спирмена, Кендалла,** их свойства. Точный тест Фишера для таблиц 2x2. Критерий хи-квадрат для таблиц сопряженности общего вида. Коэффициенты взаимосвязи для категориальных признаков. Важность признаков у моделей на основе решающих деревьев. | Корреляции, важности признаков |
| 24\.11 | 7\.1 **Гауссовская линейная модель.** Доверительный интервал для дисперсии шума. Доверительный интервал для коэффициента и гипотеза о незначимости признака. Доверительная область для вектора коэффициентов. Во всех случаях - доказательства.  Критерий о незначимости группы признаков. | Гауссовская линейная модель: таблица в statsmodels, случай линейных гипотез |
| 01\.12 | 7\.2 Анализ остатков. **Неоднородность дисперсии при гомоскедастичности. Визуальный анализ.** Критерии проверки на гомоскедастичность. Работа с гетероскедастичностью, **устойчивые оценки дисперсии и их свойства**. Вывод критерия для проверки гипотезы о незначимости коэффициента при гетероскедастичности. 7.3 Линеаризация в пуассоновской и логистической регрессиях, обоснование. Обобщенная линейная модель, ее свойства, доверительные интервалы. Проверка линейности логита в логистической регрессии.  8\.1 Информация Фишера. Два вида определения и их пояснение, свойства (док-во). 8.2 Энтропия, кросс-энтропия, дивергенция Кульбака-Лейблера. Пояснение на примере теории кодирования. Свойства энтропии и дивергенции, связь с ОМП (док-во). | Информация Фишера, многомерный случай. Подсчет информации Фишера в случае логистической и пуассоновской регрессии. Калибровка |
| 08\.12 | 8\.3 Экстремальное свойство правдоподобия (док-во). Состоятельность ОМП (док-во). Асимптотическая нормальность (док-во) и асимптотическая эффективность ОМП. 8.4 Мотивация введения натурального градиента на примере градиентного спуска для поиска ОМП параметров нормального распределения и соображений близости распределений. Использование КЛ-дивергенции. Определение натурального градиента. Градиентный спуск для параметров нормального распределения через натуральный градиент, пример с логистической регрессией, сравнение с IRLS.  Доказательства теорем. Теорема о наследовании сходимостей. Критерий факторизации Неймана-Фишера. Лемма Неймана-Пирсона и несмещенность критерия. Торема Гливенко-Кантелли. Теорема Пирсона для критерия хи-квадрат.  Оптимальные оценки, связь с эффективными оценками, напоминание достаточных статистик. Теорема Колмогорова-Блекуэлла-Рао, следствия. И т.д. |  |

**Материалы со 2 курса, которые тоже будут на зачете**

Постановка задачи классификации, байесовские классификаторы. Наивный Байес.