Приложение к рабочей программе дисциплины

МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ ПО ОРГАНИЗАЦИИ САМОСТОЯТЕЛЬНОЙ РАБОТЫ ОБУЧАЮЩИХСЯ

Наименование дисциплины	*Анализ данных и прикладная статистика*
Направление подготовки / Специальность	*38.03.01 Экономика*
Направленность (профиль) / Специализация	*Экономика и анализ данных* ОП *ВО*
Форма обучения	*очная*

Разработчик Гайдамак И.В., старший преподаватель кафедры фундаментальной математики и механики

1. Темы дисциплины для самостоятельного освоения обучающимися

Отсутствуют.

2. План самостоятельной работы:

№ п/п	Учебные встречи	Виды самостоятельной работы	Форма отчетности / контроля	Количество баллов	Рекомендуемый бюджет времени на выполнение (ак.ч.)
1.	Работа с табличными данными	1. Парсинг данных сайта	Представление кода в формате Jupyter Notebook	4	10
		2. Парсинг табличных данных	Представление кода в формате Jupyter Notebook	4	6
		3. Парсинг через API	Представление кода в формате Jupyter Notebook	4	8
2.	Генерация случайных величин	Изучение темы на онлайн-платформе ВШЭ	Тестирование на онлайн платформе	-	15
3.	Визуализация данных	Изучение темы на онлайн-платформе ВШЭ	Тестирование на онлайн платформе	-	15
4.	Точечные оценки, доверительные интервалы. Методы оценивания	Изучение темы на онлайн-платформе ВШЭ	Тестирование на онлайн платформе	-	20
5.	АБ-тесты. Проверка гипотез	Изучение темы на онлайн-платформе ВШЭ	Тестирование на онлайн платформе	-	20
6	Временные ряды	Изучение темы на онлайн-платформе ВШЭ	Тестирование на онлайн платформе	-	20
7	Введение в байесовскую статистику	Изучение темы на онлайн-платформе ВШЭ	Тестирование на онлайн платформе	-	15
8	Зачетная работа	Подготовка к зачетной работе	Контрольная работа	-	13
	Итого			12	142

3. Требования и рекомендации по подготовке отчетных документов по практике, критерии оценивания

Вид: парсинг данных сайта

Краткая характеристика: Требуется провести сбор данных с помощью языка python данных с сайта, содержащего статьи экономической направленности.

Рекомендации по выполнению:

Сайт подбирается студентом самостоятельно.

С помощью библиотеки BeautifulSoup осуществляется сбор данных по статьям: автор, название статьи, ссылка на статью, дата публикации и другая дополнительная информация.

Результаты сбора оформляются в виде таблицы типа DataFrame.

Результат работы представляется в виде файла с расширением ipynb. Название файла содержит фамилию студента и указание на вид работы. В файле код сопровождается комментариями.

Вид: парсинг табличных данных

Краткая характеристика: Требуется провести сбор данных с помощью языка python данных с сайта, содержащего в табличном виде информацию экономической направленности.

Рекомендации по выполнению:

Сайт подбирается студентом самостоятельно.

С помощью библиотеки BeautifulSoup осуществляется сбор табличных данных, результаты сбора оформляются в виде таблицы типа DataFrame.

Проводится базовый анализ полученных данных: рассчитываются числовые характеристики, данные визуализируются. На основе расчетов делаются выводы.

Название файла содержит фамилию студента и указание на вид работы. В файле код сопровождается комментариями.

Вид: парсинг через API

Краткая характеристика: Требуется провести сбор данных группы ВК, посвященной экономической тематике. Сбор данных осуществляется с помощью API (Application Programming Interface).

Рекомендации по выполнению:

Группа ВК подбирается студентом самостоятельно.

Студент регистрируется в ВК как разработчик, создает свое приложение, создает запрос на предоставление токена.

Студент определяет ID группы ВК, данные которой собирается парсить.

Проводится сбор данных: содержание постов, дата публикация, количество просмотров, количество «лайков». Результаты сбора оформляются в виде таблицы типа DataFrame.

Проводится базовый анализ полученных данных: строятся распределения, определяются числовые характеристики. В итоге формулируются выводы.

Название файла содержит фамилию студента и указание на вид работы. В файле код сопровождается комментариями.

Вид: изучение темы на онлайн-платформе ВШЭ

Краткая характеристика: Студенты проходят три курса на онлайн платформе: Сбор и анализ данных в Python, Статистические методы анализа данных, Математическая статистика и A/B тестирование.

Рекомендации по выполнению:

В конце каждого раздела студенты проходят неоцениваемое тестирование на общее понимание темы, а также тестирование, которые проверяет навыки работы с данными на языке python.

Вид: подготовка к зачетной работе

Краткая характеристика: задания направлены на проверку знаний, умений и навыков анализа данных и проверки гипотез.

Рекомендации по выполнению:

Для подготовки к зачетной работе рекомендуется:

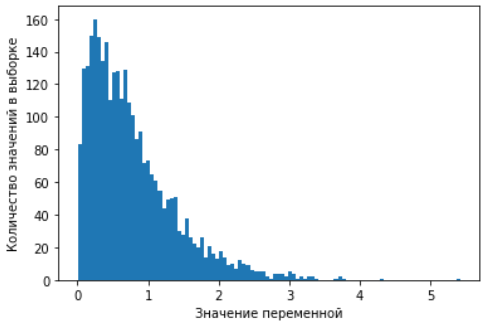
- повторить лекционный материал по пройденным темам;
- решать задачи по пройденным темам;
- провести самопроверку, решив демонстрационный вариант.

Демонстрационный вариант зачетной работы

Часть А.

1. Исследователь Анатолий хочет построить 95%-ый доверительный интервал для среднего значения некоторой переменной и размышляет, может ли он воспользоваться нормальным приближением. Предположим, что Анатолий работает с выборкой малого размера. Какой тип графика ему стоит использовать, чтобы наиболее просто оценить корректность применения нормального приближения?
    a. Корреляционная карта переменной с самой собой.
    b. Диаграмма рассеяния.
    c. Гистограмма.
    d. Одномерный линейный график.

2. На рисунке ниже изображена гистограмма некоторой переменной. Выберите неверное утверждение.



    a. Ряд значений переменной содержит не менее одной моды.
    b. Мода и среднее значение переменной совпадают.

c. Медиана переменной меньше её среднего значения.
d. Ряд значений переменной содержит только неотрицательные числа.

3. Аналитик Арсений специализируется в применениях методов машинного обучения и статистики для областей сельского хозяйства. Он решает задачу классификации типа почвы и хочет построить модель для разделения чернозёма и прочих типов. Для этого он использует данные по содержанию гуминовых кислот в кДа (acidG) и фульвокислот в кДа (acidF) в образцах почвы. Он оценил логистическую регрессию на обучающей выборке. Оказалось, что уравнение модели задаётся следующим образом:

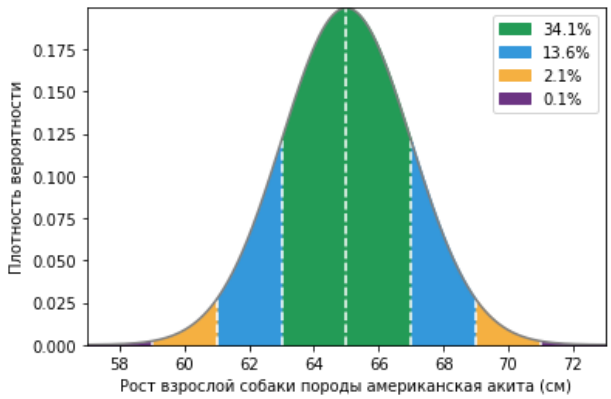
$$\hat{P}(\mathrm{y_i}=\mathrm{Chernozem}) = \sigma(0.00 - 0.01 \times \mathrm{acidG_i} + 0.2 \times \mathrm{acidF_i}),$$

где $\mathrm{y_i}$ – значение целевой переменной для наблюдения i, $\mathrm{acidG_i}$ и $\mathrm{acidF_i}$ – значения признаков для наблюдения i, $\sigma(\cdot)$ – логистическая функция.

Пусть имеется образец почвы с содержанием гуминовых кислот 15 кДа и фульвокислот 4 кДа. Найдите оценку вероятности принадлежности этого образца к чернозёму, которую выдаст модель Арсения. Ответ округлите до сотых.

a. 0.71
b. 0.54
c. 0.66
d. 0.52

4. Предположим, что рост взрослой собаки породы американская акита является нормальной случайной величиной со средним 65 см и стандартным отклонением 2 см. Плотность распределения этой случайной величины изображена на рисунке ниже. Выберите все верные утверждения.



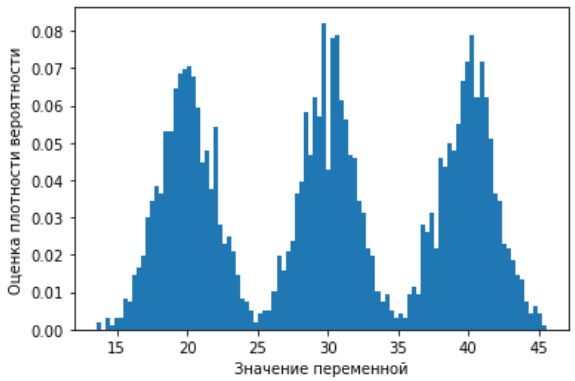
a. Обязательно найдётся взрослая собака породы американская акита с ростом 65 см.
b. Мода распределения роста взрослой собаки породы американская акита равна 65 см.

c. Медиана распределения роста взрослой собаки породы американская акита равна 65 см.
d. 0.1% всех взрослых собак породы американская акита имеет рост не менее 71 см.

5. Отважная исследовательница Анастасия занимается исследованием гигантских кальмаров (giant squids). Какая из собранных ей выборок будет более репрезентативной, чем другие?
   a. Данные о кальмарах, обитающих на дне Марианской впадины.
   b. Данные о кальмарах, обитающих вблизи подводных вулканов.
   c. Данные о кальмарах, обитающих в Тихом океане.
   d. Данные о кальмарах, обитающих в наиболее экологически чистых регионах.

6. Экономист Степан хочет понять, как годовой валовый внутренний продукт (ВВП) некоторой страны зависит от среднего годового уровня безработицы (УБ) этой страны (обе величины являются непрерывными). Оказалось, что выборочный коэффициент корреляции между ВВП и УБ равен -0.12. Выберите все верные утверждения.
   a. Между переменными ВВП и УБ существует отрицательная линейная взаимосвязь.
   b. Между переменными ВВП и УБ не существует положительной нелинейной связи.
   c. ВВП и УБ являются независимыми величинами.
   d. Рассматривая только коэффициент корреляции, нельзя однозначно определить, существует ли причинно-следственная связь между ВВП и УБ.

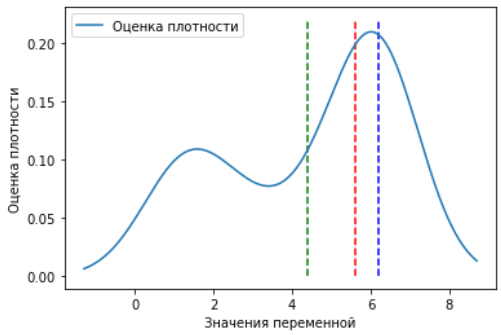
7. Какой переменной может соответствовать гистограмма, приведённая на рисунке ниже?



   a. Количество рабочих часов в неделю работающих взрослых людей в возрасте от 20 до 40 лет.

b. Количество рабочих часов в неделю взрослых людей в возрасте от 18 до 40 лет.
c. Количество видимых невооружённым глазом созвездий в разное время суток.
d. Среднее количество детей в российских семьях в 2008 году.

8. На рисунке изображена оценка функции плотности некоторой переменной. Цветными линиями обозначены выборочные медиана, среднее и мода распределения этой переменной. Соотнесите линию и меру центральной тенденции, которую она показывает.



Варианты: мода, медиана, среднее. Красная – …, синяя – …, зелёная – …

Ответ: Красная – медиана, синяя – мода, зелёная – среднее

9. Специалист по машинному обучению Светлана тестирует гипотезу о том, что MSE, полученная после оценки некоторой модели на тренировочной выборке, равна 542. Для этого она использует t-тест. Оказалось, что p-value равно 0.241. Выберите все верные утверждения.
    a. Основная гипотеза не отвергается на уровне значимости 1%.
    b. Основная гипотеза не отвергается на уровне значимости 5%.
    c. Основная гипотеза не отвергается на уровне значимости 10%.
    d. Основная гипотеза отвергается на любом разумном уровне значимости.
10. Соотнесите ситуацию и тип ошибки (I или II рода).

Ситуация 1: Студент получил оценку «неудовлетворительно» за взятый им курс. Предположим, что он формулировал нулевую гипотезу так: «Это достаточно простой курс, за который легко получить отличную оценку».

Ситуация 2: Элизабет вышла на прогулку без зонта, однако вскоре пошёл сильный дождь. Предположим, что она формулировала нулевую гипотезу так: «Сегодня будет дождь».

Ответ: Ситуация 1 – II род, Ситуация 2 – I род.

11. Исследователь Михаил занимается поиском взаимосвязей в документах, закодированных векторами длины 10^8. Михаил хочет визуализировать полученные векторы на двумерной плоскости, чтобы наглядно увидеть, не образуют ли документы тематические облака. Какую задачу машинного обучения решает Михаил?
    a. Регрессия.
    b. Классификация.
    c. Ранжирование.
    d. Снижение размерности.

12. Метеоролог Степан хочет оценить зависимость количества осадков (в мм) perc от температуры воздуха (в градусах Цельсия) temp и скорости ветра (в м/с) speed при помощи линейной модели. Жизненный опыт подсказывает Степану, что в модель стоит включить не абсолютные значения переменных, а их натуральный логарифмы, и именно в такой спецификации он и проводит оценку модели. Оценённое уравнение регрессии выглядит следующим образом:

$$\widehat{\ln \text{perc}}_i = 1 - 5 \times \ln(\text{temp}_i) + 17.2 \times \ln(\text{speed}_i),$$

    где perc_i – значение целевой переменной для наблюдения i, temp_i и speed_i – значения признаков для переменной i. Выберите верное утверждение об интерпретации оценённой модели.
    a. При увеличении температуры воздуха на 1% количество осадков уменьшается на 5%.
    b. При увеличении температуры воздуха на 1 градус Цельсия количество осадков уменьшается на 5%.
    c. При увеличении скорости ветра на 1 м/с температура воздуха уменьшается на 17.2%.
    d. При уменьшении температуры воздуха на 1% количество осадков увеличивается на 5 м/с.

13. Выберите верное утверждение о метриках качества, применяемых к модели линейной регрессии без константного признака, обучаемой при помощи минимизации MSE (среднеквадратичной ошибки).

    Пояснение: MAE – средняя абсолютная ошибка.

a. При добавлении нового признака в такую модель MSE на тестовой выборке обязательно уменьшится.
b. MSE и MAE на тестовой выборке для этой модели совпадают.
c. Для оценки качества этой модели на тестовой выборке можно использовать как MSE, так и MAE.
d. Оценки коэффициентов в такой модели невозможно корректно интерпретировать.

Часть B.

14. Винни-Пух в течение 150 дней фиксировал изменения количества пчёл в улье. Он уверен, что полученные наблюдения являются выборкой независимых одинаково распределённых нормальных случайных величин. Оказалось, что среднее количество пчёл равно 25000, а выборочная дисперсия равна 1300. Постройте 95%-ый доверительный интервал для математического ожидания количества пчёл в улье и выпишите в ответ его нижнюю границу, округлённую до сотых.

    Пример ответа: 1500.00

15. Ниже приведены данные об уровнях осадков в двух различных регионах России, измеренные за одинаковые промежутки времени. Предполагая, что все необходимые предпосылки выполнены, дисперсии генеральных совокупностей равны, а выборки независимы, проверьте гипотезу о равенстве средних уровней осадков ($H_0: \mu_1 = \mu_2$) при помощи t-теста на уровне значимости 5%. Выберите верное утверждение.

    Регион 1: [103.01, 101.99, 105.21, 106.80, 112.70, 106.13, 110.48, 109.26, 100.44, 100.28].

    Регион 2: [107.38, 106.31, 106.00, 105.27, 105.27, 104.66, 103.70, 105.07, 105.12, 104.74].

    a. Основная гипотеза не отвергается.
    b. Основная гипотеза отвергается.

16. Выберите две переменные, между которыми возможно рассчитать интерпретируемый выборочный коэффициент корреляции Пирсона, и вычислите этот коэффициент по приведённым данным. Ответ округлите до сотых.

    Переменные:
    - Цвет автомобиля, закодированный числами (1 – синий, 2 – красный, 3 – зелёный): [1, 1, 2, 3, 2]
    - Истинный объём бака автомобиля (л): [50.1, 53.2, 55.0, 55.0, 50.2]
    - Пройденный километраж (тыс. км): [15.2, 4.75, 1.2, 1.9, 9.3]

    a. 0.21
    b. -0.59
    c. -0.41
    d. -0.94

17. Бабушка Афросинья хочет проверить, существует ли зависимость между сортом баклажана и типом почвы, в которых сорт был высажен. Для этого она собрала данные по 150 сортам, высаженных в чернозём, и 150 сортам, высаженных в каштановую почву. Данные представлены в таблице ниже.

	Алмаз	Матросик	Снежный
Чернозёмная почва	75	20	55
Каштановая почва	20	120	10

    Рассчитайте статистику χ^2 критерия согласия Пирсона и выпишите её в ответ, округлив до сотых.

    Пример ответа: 101.02

Часть С.

Файл «c.csv» содержит данные о рейтингах видео-игр по данным Metacritic (источник}). Набор данных содержит следующие переменные:

- Name – название игры.
- Platform – платформа для запуска игры.
- Year_of_Release – год запуска игры.
- Genre – жанр игры.
- Publisher – компания, выпустившая игру.
- NA_Sales – продажи в Северной Америке (миллионы копий).

- EU_Sales – продажи в Европейском Союзе (миллионы копий).
- JP_Sales – продажи в Японии (миллионы копий).
- Other_Sales – продажи в прочих странах (миллионы копий).
- Global_Sales – общие продажи по миру (миллионы копий).
- Critic_Score – агрегированный рейтинг команды Metacritic.
- Critic_Count – количество экспертов, участвовавших в расчёте Critic_Score.
- User_Score – агрегированный рейтинг пользователей Metacritic.
- User_Count – количество пользователей, участвовавших в расчёте User_Score.
- Developer – разработчик игры.
- Rating – рейтинг ESRB (Everyone, Teen, Adults Only, ...).

Во всех заданиях этой части вам предстоит работать с этим набором данных.

При записи ответов дроби следует округлять до сотых и отделять дробную часть точкой, целые числа следует выписывать с указанием .00 в дробной части (смотрите примеры ответов).

18. Вычислите среднее по продажам в Северной Америке (в миллионах копий) – переменная NA_Sales.

    Пример ответа: 13.20

19. Вычислите минимальное значение по продажам в Северной Америке (в миллионах копий) – переменная NA_Sales.

    Пример ответа: 12.00

20. Вычислите стандартное отклонение по продажам в Северной Америке (в миллионах копий) – переменная NA_Sales.

    Пример ответа: 20.04

21. Определите наиболее часто встречающий жанр игр.

    Пример ответа: Simulation

22. Добавьте в таблицу новый признак Platform_Coded, который будет представлять собой закодированное название платформы для запуска игры. Кодировку поведите

следующим образом: если платформа – это PS3, то код равен 1, а если любая другая, то код равен 0. Выведите среднее по признаку Platform_Coded, округлённое до сотых.

Пример ответа: 12.14

23. Вычислите средний агрегированный рейтинг команды Metacritic для игр, выпущенных (published) компаниями Tecmo Koei или Wanadoo. Выпишите полученное значение, округлённое до сотых.

    Пример ответа: 17.20

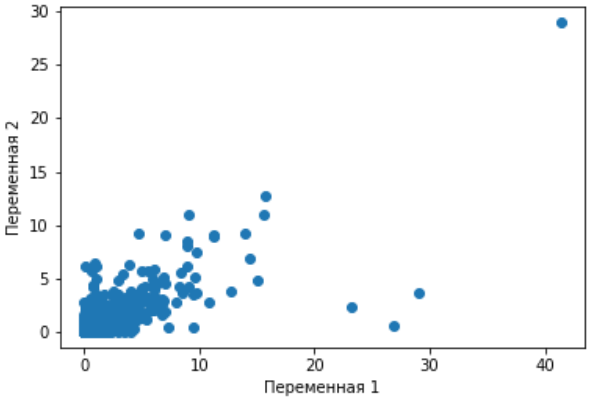
24. Определите количество пропущенных значений признака, содержащего информацию о названиях разработчиков игр (переменная Developer). Выпишите найденное количество пропущенных значений.

    Пример ответа: 100.00

25. Определите количество пропущенных значений признака, содержащего информацию о названиях разработчиков игр (переменная Developer). Замените пропущенные значения этого признака на моду по этому признаку и сохраните полученный признак без пропущенных значений в отдельную переменную, не включаемую в исходную таблицу. Выпишите количество значений этой переменной, равных её моде.

    Пример ответа: 100.00

26. Постройте диаграмму рассеяния для переменных Critic_Score (на горизонтальной оси) и EU_Sales (на вертикальной оси). Выберите все верные утверждения.
    a. Между переменными Critic_Score и EU_Sales, скорее всего, существует сильная положительная линейная взаимосвязь.
    b. Построенная диаграмма показывает, что в выборке, возможно, есть выбросы.
    c. Между переменными Critic_Score и EU_Sales, скорее всего, существует сильная отрицательная линейная взаимосвязь.
    d. Изменение переменной Critic_Score совершенно не влияет на изменение переменной EU_Sales.
    e.
27. Определите, какой график изображён на рисунке ниже.



a. Диаграмма рассеяния переменных NA_Sales и EU_Sales.
b. Диаграмма рассеяния переменных Critic_Score и Critic_Count.
c. Диаграмма рассеяния переменных JP_Sales и User_Count.
d. Гистограмма переменной User_Count.

28. Выведите корреляционную матрицу для числовых переменных. Найдите переменную, которая имеет наибольшую по модулю корреляцию с переменной EU_Sales (не включая саму EU_Sales). В ответ выпишите модуль найденного значения корреляции.

    Пример ответа: 0.99

29. Оцените линейную регрессию, уравнение которой имеет следующий вид:
    $Global\_Sales_i = \widehat{w_0} + w_1 \ln(1 + \widehat{\text{NA\_Sales}_i})$,

    где нижние индексы обозначают значения соответствующих переменных для наблюдения i. В ответ выпишите оценку коэффициента при свободном члене.

    Пример ответа: -2.40

30. Оцените линейную регрессию, уравнение которой имеет следующий вид:
    $Global\_Sales_i = \widehat{w_0} + w_1 \ln(1 + \widehat{\text{NA\_Sales}_i})$,

    где нижние индексы обозначают значения соответствующих переменных для наблюдения i. В ответ выпишите значение среднеквадратичной ошибки (MSE) на обучающей выборке.

    Пример ответа: 2.40

4. Рекомендации по самоподготовке к промежуточной аттестации по дисциплине

Вопросы для самопроверки к дифференцированному зачету

1 Работа с таблицами в pandas.

2 Конструирование сложных признаков: groupby, join.

3 Случайные величины. Их распределения и характеристики.

4 ЦПТ и ЗБЧ. Генерация случайных величин в python. Метод Монте-Карло.

5 Решение различных практических задач с помощью генераций. От случайных величин к описательной статистике. Её применение для анализа признаков и поиска аномалий.

6 Визуализация данных: matplotlib, seaborn, wordcloud.

7 Хорошие и плохие графики, как рассказать историю с помощью визуализации.

8 Репрезентативность. Точечные оценки. Свойства оценок: несмещённость, состоятельность и эффективность.

9 Кейсы про не очень удачные выборки и исследования.

10. Методы построения точечных оценок: метод моментов, метод максимального правдоподобия.

11 От доверительных интервалов к проверке гипотез. Ошибки 1 и 2 рода. P-значение. Параметрические критерии.

12 Строительство асимптотических критериев на основе ЦПТ. Гипотезы о долях, среднем и разбросе. Тест отношения правдоподобий. Непараметрические критерии. Критерии согласия. Бустрап. АБ-тесты.

13 Планирование эксперимента, обсчёт результатов. Бизнес-метрики и анализ влияния изменений на продукт. Корреляции: Пирсона, ранговая, Мэтьюса. Проверка гипотез про корреляции. Эндогенность

14. Какими бывают данные: кросс-секция, временные ряды, панельные. Временные ряды как тип данных. Стационарность.

15 Гипотеза единичного корня и её проверка. ARIMA-модели. Кросс-валидация на временных рядах. Сглаживание, фильтры.

16. Разложение временных рядов на тренд, сезонную, циклическую и нерегулярную компоненты. Что такое коинтеграция, VAR и VECM.

17. Разница между Байесовской и частотной статистикой. Байесовский вывод, Байесовские точечные оценки и доверительные интервалы.

18. Про сопряженные распределения и MCMC. Язык описания вероятностных моделей STAN, его использование с python.