Статистический анализ данных / Python. Задание №5

Условия

- Дедлайн:
  - **Очное:** 20 декабря 23:59
  - **Дистанционное:** 24 декабря 23:59
- **ВНИМАНИЕ** Выполненную работу нужно отправить на **psad.homework@gmail.com**, указав тему письма "[MADE19] Фамилия Имя - Задание 5". Квадратные скобки обязательны
- Максимальный балл: 10 баллов. В случае, если в решении есть грубые ошибки, задание отправляется на доработку. За отправку на доработку начисляется штраф -2 балла ПЛЮС -1 балл за каждый день, пока задание находится на доработке.

Задание

Требуется подобрать и применить наилучший статистический метод, позволяющий ответить на вопрос прикладной задачи; обосновать выбор метода, его применимость и оптимальность. Не нужно использовать все возможные методы, нужно выбрать наиболее подходящий. Помимо выводов, касающихся математических особенностей решения, необходимо в терминах предметной области сформулировать выводы, которые могли бы быть понятны гипотетическому заказчику-нематематику.

Необходимо сдать: ipython notebook и сгенерированный по нему pdf-файл с подробным отчётом по проведённому исследованию, содержащий визуализацию исходных данных, описания и выводы каждого этапа анализа — используемые методы, обоснование их применимости, графики.

В данном задании воспользуйтесь **дисперсионным анализом**, либо известными вам **одномерными методами** статистического анализа. Не забывайте делать поправку на множественную проверку гипотез там, где это необходимо.

Данные для всех задач:
`https://www.dropbox.com/sh/zhieidbawsmpbmq/AAD8SzfpIBgI4ppDUf6JRK7pa?dl=0`

Варианты

Вариант задания определяется первой буквой вашей фамилии

Эхинацея и респираторные заболевания

Группа детей от 2 до 11 лет случайным образом была распределена по двум типам лечения ОРЗ — плацебо и экстракт эхинацеи. Доза эхинацеи подбиралась соответственно возрасту ребёнка по рекомендации производителя. В группе плацебо произошло 329 простудных заболеваний, в группе эхинацеи — 367. Для каждого заболевания известны тяжесть его протекания по мнению родителей и наличие сопутствующих симптомов (echinacea.xlsx).

А: эффективно ли лечение ОРЗ с помощью экстракта эхинацеи?

General Social Survey

General Social Survey — ежегодный социологический опрос нескольких тысяч граждан США; на сайте https://gssdataexplorer.norc.org доступны все данные с 1972 по 2014 год (GSS.xls).

Б: исследовать связь уровня счастья (General happiness) опрошенных 2014 года с демографическими признаками (пол, возраст, раса, семейное положение, количество детей, образование, сексуальная ориентация, занятость, доход).

В: исследовать связь веры в научность астрологии опрошенных 2014 года с признаками, описывающими отношение к религии (религиозные предпочтения, уверенность в существовании бога, религиозная и духовная самоидентификация, участие в религиозных активностях).

Г: для опрошенных 2014 года исследовать связь суммарного количества правильных ответов на 11 вопросов на знание базовых научных фактов с демографическими признаками (пол, возраст, раса, семейное положение, количество детей, образование, сексуальная ориентация, занятость, доход).

Качество воды в Миннесоте

Для 895 источников воды в Миннесоте известны водоносный горизонт, водоём, уровень и химические свойства воды (pH, щёлочность, содержание алюминия, мышьяка, хлора и свинца) (MnGroundwater.csv).

Д: сравнить свойства воды из разных водоёмов.

Задержка авиарейсов

Для 4029 рейсов, вылетавших из Нью-Йоркского аэропорта Ла-Гуардия, известны название авиакомпании-перевозчика, аэропорт назначения, диапазон планируемого времени вылета, день недели, месяц, продолжительность полёта и время задержки вылета (FlightDelays.csv).

Е: есть ли закономерности в задержках вылетов?

Биоразлагаемость молекул

1055 химических молекул описаны с помощью 41 признака (число атомов кислорода, нитратных групп, донорных связей с водородом, потенциал ионизации и т.д.); 355 из них биоразложимы (biodeg.xlsx).

Ж: какие признаки отличают биоразложимые молекулы от устойчивых?

Годовой заработок

Опрос US Bureau of Labor Statistics 2002 года содержит данные о годовом заработке 55729 участников; известны также их пол (1 = male, 2 = female), возраст, уровень образования (1 = no high school, 2 = some high school, 3 = high school diploma, 4 = some college, 5 = bachelor's degree, 6 = postgraduate degree) и тип работы (5 = private sector, 6 = government, 7 = self-employed) (workers.xls).

З: оценить влияние образования и пола на годовой заработок.

И: оценить влияние пола и типа работы на годовой заработок.

Риск сердечно-сосудистых заболеваний

В рамках исследования риска сердечно-сосудистых заболеваний изучалось влияние регулярных занятий спортом на частоту сердечных сокращений при выполнении шестиминутного упражнения на беговой дорожке. Выборка состоит из 800 испытуемых двух групп. Половина из них пробегала не меньше 15 миль в неделю, половина вела малоподвижный образ жизни. Известен пол испытуемых и их пульс по окончании упражнения (heartrate.txt).

К: оценить влияние регулярных занятий спортом на частоту сердечных сокращений после нагрузки с учётом пола.

Содержание азота в кормовых растениях

Четыре вида растений (Leucaena leucocephala, Acacia saligna, Prosopis juliflora, Eucalyptus citriodora), которые могут расти в долине реки Иордан, где климат достаточно засушливый, выращивались в лаборатории при разных условиях доступа к воде. Количество воды, получаемой растением в сутки, менялось в разных группах от 50 до 650 мм с шагом 100 мм. Исследователей интересовала возможность их использования для кормления сельскохозяйственных животных, для чего необходимо высокое содержание азота. Для 9 растений в каждой группе известно содержание азота (plants.xlsx).

Л: как содержание азота меняется для разных видов растений при разных условиях выращивания? Какие растения лучше всего подходят для сельскохозяйственного использования?

Продолжительность жизни и активность размножения самцов дрозофилы

Для изучения влияния активности размножения самцов дрозофилы на продолжительность их жизни был организован следующий эксперимент. По 25 самцов в пяти группах содержались в одинаковых условиях, за исключением одного отличия: в первой группе к каждому самцу ежедневно подсаживалась готовая к размножению самка, во второй – восемь готовых к размножению самок, в третьей и четвёртой - соответственно, одна и восемь беременных самок, не готовых к размножению, наконец, к самцам пятой группы не подсаживали никого. Для каждого самца измерена продолжительность жизни, длина грудной клетки и доля времени, проводимого во сне (fly.txt).

М: исследовать связь между продолжительностью жизни самцов дрозофилы и наличием самок разного типа и количества.

Курение и болезнь Альцгеймера

Ретроспективное исследование влияния курения на болезнь Альцгеймера включает пациентов с болезнью Альцгеймера, другими формами деменции и другими диагнозами; известны статус курения и пол (alzheimer.txt).

Н: как курение и пол связаны с различными формами снижения умственной деятельности?

Открытие депозита

Имеются результаты обзвона 4119 клиентов португальского банка, которым предлагалось завести депозит. Известны социально-демографические характеристики клиентов, история предыдущих коммуникаций, социально-экономические показатели на момент совершения звонка (deposit.xlsx).

О: какие значения характеристик клиента повышают вероятность открытия им депозита?

Биомаркеры рака груди

В эксперименте принимали участие 24 человек, у которых не было рака груди (normal), 25 человек, у которых это заболевание было диагностировано на ранней стадии (early neoplasia), и 23 человека с сильно выраженными симптомами (cancer). Секвенирование — это определение степени активности генов в анализируемом образце с помощью подсчёта количества соответствующей каждому гену РНК; именно эта количественная мера активности каждого из 15748 генов для каждого из 72 человек записана в данных (gene_high_throughput_sequencing.csv).

Разница в уровнях экспрессии гена между группами считается практически значимой, если средние уровни в группах отличаются более, чем в полтора раза; таким образом, необходимо посчитать величину fold change:

$$F_c(C,T) = \begin{cases} \frac{T}{C}, & T > C, \\ -\frac{C}{T}, & T < C, \end{cases}$$

где C, T — средние значения экспрессии гена в сравниваемых группах; практически значимыми считаются те отличия, для которых $|F_c(C,T)| > 1.5$.

П: по каким генам имеются статистически и практически значимые отличия в уровнях экспрессии между здоровыми испытуемыми и испытуемыми с ранней стадией рака? Между здоровыми и испытуемыми с сильно выраженными симптомами?

Заживление ран

На 26 пациентах было испытано экспериментальное лекарство, способствующее заживлению ран; для сравнения ещё к 26 пациентам применялась стандартная терапия. Измерялась площадь раны до начала терапии, после курса лечения и на заключительном визите через длительное время после завершения лечения. Кроме того, приведена субъективная оценка изменения состояния раны пациентом и врачом (wounds.csv).

Р: отличается ли эффективность экспериментального лекарства от эффективности стандартного?

Прочность промышленных вентиляторов

Измерен разрушающий крутящий момент 64 промышленных вентиляторов; для каждого известны тип отверстия, форма барабана и метод соединения (fans.txt).

С связан ли разрушающий крутящий момент с характеристиками вентилятора?

Вакцина против папилломавируса

Собраны данные по 1413 пациенткам клиник при университете Джона Хопкинса, проходившим с 2006 по 2008 вакцинацию против папилломавируса человека препаратом Гардасил. Рекомендуемый курс — три укола в течение года — был пройдён только 469 пациентками. Производитель препарата исследует, в каких демографических группах и каком способе получения вакцины проведение полного курса наиболее вероятно (gardasil.xls).

Т: что отличает пациентов, прошедших полный курс вакцинации в течение года, от тех, кто его не прошёл?

Нарушения ПДД

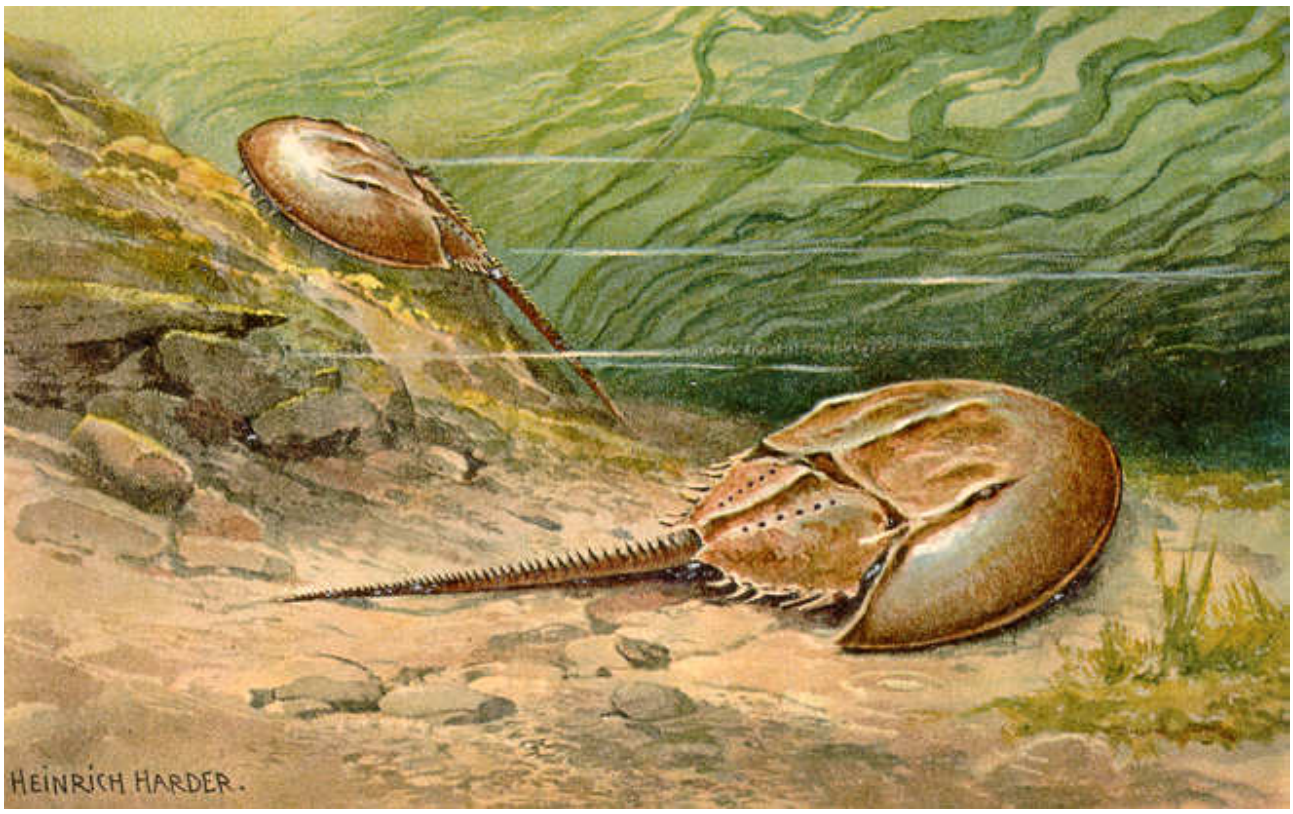
В исследовании влияния обучения подростков вождению на число инцидентов с нарушениями ПДД контрольная группа состоит из 2409 человек. По каждому из них данные собираются на протяжении четырёх лет (traffic_violation.txt).

Ф: меняется ли в контрольной группе число инцидентов с годами? Если да, то как?

Эффективность тромболитической терапии

Собраны данные по 206 пациентам второго кардиологического отделения московской городской клинической больницы №25. Имеются результаты 14 анализов, а также 8 дополнительных признаков, описывающих пациента (пол, возраст, курение, наличие диабета и т.д.) (cardio.xls).

Х: оценить влияние курения на вероятности выздоровления и возникновения осложнений, а также на результаты 14 анализов.



Внешний вид и привлекательность самок мечехвостов

Изучалось влияние внешних характеристик самок морских ракообразных ракообразных мечехвостов на их привлекательность для самцов. Выборка состоит из данных о наблюдениях над 173 особями и содержит закодированные данные о размере самок, их весе, цвете, состоянии панциря, а также о количестве спутников (horseshoe crab.txt).

У: сравнить по всем имеющимся признакам самок, имеющих хотя бы одного спутника, с самками, не имеющими ни одного.

Комментарии в блогах

Для 60021 постов в блогах, опубликованных не более, чем за 72 часа до базового времени, собрана информация о количестве комментариев, времени публикации, длине и количестве каждого из 200 часто встречающихся слов (blog_feedback.xlsx).

Ц: чем отличаются сообщения, не получившие ни одного комментария, от тех, которые хотя бы кто-нибудь прокомментировал?

Белки в коре мозга мышей

В 1080 образцах коры мозга мышей измерен уровень экспрессии 77 белков. Часть образцов взята от трисомных мышей (лабораторная модель синдрома Дауна), часть — от здоровых; в эксперименте перед получением образцов некоторые мыши получали стимул к обучению, а некоторые — нет; наконец, части мышей вводился Мемантин, а части — физраствор. Цель эксперимента — проверить, восстанавливает ли Мемантин способность к обучению у трисомных мышей (memantine.xls).

Ш: влияет ли Мемантин на экспрессию белков здоровых мышей?

Щ: отличается ли экспрессия белков у здоровых и трисомных мышей в каких-нибудь из экспериментальных подгрупп?

Размеры черепа древних египтян

Было измерено 150 черепов, найденных при раскопках в Египте. Находки относятся к пяти различным временным периодам. Для каждого черепа известны: максимальная ширина, базибрегматическая высота, базиальвеолярная длина, высота носа, примерная дата формирования. Была выдвинута гипотеза о том, что изменение этих параметров со временем может свидетельствовать о скрещивании египтян с другими популяциями. (skulls.txt)

Э-Ю-Я: проверить, есть ли различия между размерами черепов различных временных периодов, если есть, то какие периоды отличаются друг от друга.