

Задача 1

Реализуйте критерий Джонкхиерра (Jonckheere). Если это возможно, то примените его в рамках своего исследования. Почитать о критерии - Кобзарь А.И. Прикладная математическая статистика, 2006, - 4.2.1.2.9, стр. 477.

Задача 2

Необходимо сдать: iрunb 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.

РАБОТА С РЕАЛЬНЫМИ ДАННЫМИ

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

Данные доступны по ссылке

<https://github.com/csmsupsad/psad2021/tree/main/%D0%94%D0%97/Task%203>

1. БАКЛАНОВА АНАСТАСИЯ ОЛЕГОВНА

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

Задача. Сравнить свойства воды из разных водоёмов.

Данные. MnGroundwater.csv.

2. БАЛАКОВА АННА СЕРГЕЕВНА

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

Задача. 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).

Данные. GSS.xls.

3. ВЕСЕЛОВ АРСЕНИЙ СЕРГЕЕВИЧ

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).

Задача Оценить влияние образования, пола и типа работы на годовой заработок.

Данные. workers.xls.

4. ДЕВБУНОВА ВИЛИАНА

Засеивание облаков и уровень осадков. Исследовалось воздействие засеивания облаков на обилие дождей. Измерения проводились в течение 108 периодов на пяти участках земли в Тасмании - участки обозначены в файле как западный, восточный, южный, северный и северо-восточный. В выборке содержатся данные об уровне осадков (в миллиметрах) на каждом из пяти участков, о времени года, к которому относится период, и о том, проводилось ли засеивание, проверить, как засеивание облаков повлияло на уровень осадков отдельно по каждому из пяти экспериментальных участков.

Задача. Проверить, одинаково ли проявляется эффект засеивания на каждом из них, или, возможно, он как-то зависит от исходного уровня осадков на участке?

Данные. cloudseeding.txt

5. ДЕЕВ АЛЕКСАНДР СЕРГЕЕВИЧ

Допустимость наказаний. Известно мнение двенадцати родителей о допустимости наказания их детей по результатам оценки в психогенном эксперименте; допустимость выражается в баллах. Чем ниже балл, тем менее допустимым участник исследования считает наказание. Имеются результаты о наказании самим родителем, бабушкой и учителем ребёнка.

Задача. Как зависит оценка допустимости наказания от наказывающего?

Данные. punishment.txt

6. КОРНИЛОВ МАКСИМ АЛЕКСЕЕВИЧ

Продажи платьев. Имеются данные по продажам 479 платьев на сайте alieexpress.com за полтора месяца осени 2013 года. Для каждого из платьев известны также стиль, ценовая категория, рейтинг, размер, сезон, ряд характеристик внешнего вида и индикатор участия в программе рекомендаций.

Задача. Исследовать, как каждый из признаков по отдельности влияет на уровень продаж.

Данные. alieexpress_dress_data.csv

7. КУЗЬМИН НИКИТА ВАЛЕРЬЕВИЧ

Пожертвования на благотворительность. Благотворительная организация разослала 4268 писем с предложением сделать пожертвование и получила отклик с пожертвованиями от 1707 адресатов. Для каждого адресата известны: индикатор ответа на предыдущее письмо, число недель, прошедших с момента предыдущего пожертвования, размеры текущего, предыдущего и среднего по всем предыдущим пожертвованиям в голландских гульденах, число писем, отправляемых адресату в год, доля писем, в ответ на которые приходят пожертвования.

Задача. Какие признаки отличают людей, совершающих пожертвования?

Данные. charity.xlsx

8. МАРКОВ ВЛАДИСЛАВ ОЛЕГОВИЧ

Дома престарелых Нью-Мексико. Для 52 лицензированных домов престарелых Нью-Мексико известны: число коек, суммарное годовое число дней в стационаре и койко-дней (в сотнях), суммарные годовые расходы на уход за пациентами, зарплату медсестёр и инфраструктуру (в сотнях долларов).

Задача. Есть ли различия между сельскими и городскими домами престарелых? По каким признакам?

Данные. nursing_homes.txt

9. Мишустина МАРГАРИТА ВЛАДИМИРОВНА

Maryland's Pick-3 Lottery. Даны результаты розыгрыша лотереи Maryland's Pick-3 Lottery за 218 подряд идущих дней. Результатом является трёхзначное число.

Задача. Можно ли считать розыгрыш случайным?

Данные. lottery.txt.

10. Новосёлов ВАДИМ ВЛАДИМИРОВИЧ

Задержка авиарейсов. Для 4029 рейсов, вылетающих 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.

Задача. Есть ли закономерности в задержках вылетов?

Данные. FlightDelays.csv.

11. Ниничук МАРИНА МАТВЕЕВНА

Интеллект и размер головного мозга. Исследование проводилось среди студентов психологического факультета крупного университета. Все испытуемые должны были быть правшами, а также не иметь повреждений мозга, эпилепсии, алкоголизма и сердечных заболеваний. Участники предварительного этапа эксперимента прошли несколько IQ-тестов, после чего для дальнейшего участия было отобрано 20 мужчин и 20 женщин, имевших коэффициент интеллекта либо ниже 103, либо выше 130 баллов. Для каждого из отобранных при помощи магнитно-резонансной томографии были получены 18 снимков срезов головного мозга, и общее количество пикселей на всех 18 снимках было принято в качестве меры объёма мозга. Помимо этого, были собраны данные о росте и массе тела испытуемых.

Задача. Исследовать взаимосвязи между коэффициентами интеллекта и биологическими характеристиками испытуемых (пол, рост, вес, объём мозга).

Данные. brain.xlsx.

12. Осин ДМИТРИЙ ЭДУАРДОВИЧ

Обучение родителей воспитанию детей. 975 родителей участвовало в программе обучения воспитанию. Было проведено три опроса, в ходе которых родители отвечали на вопрос: «За последние несколько недель обращались ли дети к вам с проблемой или вопросом, который их беспокоил?» Первый опрос был проведён до начала обучения, второй - сразу после, и третий - по прошествии 6-8 недель после окончания обучения. Известен также уровень образования родителя.

Задача. Стали ли родители больше общаться с детьми в результате обучения? Проанализировать с учётом уровня образования родителей.

Данные. education.txt

13. Пойманов ДМИТРИЙ РОМАНОВИЧ

Продолжительность жизни больных онкологическими заболеваниями. Выборка состоит из 64 пациентов, у которых был диагностирован неизлечимый рак какого-либо органа. Всем им в качестве поддерживающей терапии был назначен приём витамина С (считалось, что он может способствовать выздоровлению раковых больных). Приведены данные об остаточной продолжительности жизни пациентов в днях.

Задача. Исследовать связь между остаточной продолжительностью жизни и типом рака.

Данные. cancer.txt

14. ПОНОМАРЁВ ВАЛЕРИЙ

Линька крабов. У 472 самок крабов *metacarcinus magister* измерена ширина панциря до и после линьки. Измерения были получены двумя способами: 1) 12000 крабов измеряли, помечали сигнальными маячками и выпускали обратно в естественную среду перед периодом линьки, затем часть крабов за вознаграждение возвращалась в лабораторию рыбаками, выловившими их с помощью стандартных ловушек; 2) крабов, выловленных на суше во время спаривания непосредственно перед линькой, приносили в лабораторию, измеряли, затем, через несколько дней после линьки, измеряли снова. Для второй категории известен год вылова.

Задача. Исследовать различия между изменениями размеров панциря особей, линька которых проходила в лабораторных условиях и в естественных. Для последних оценить влияние года вылова.

Данные. crabs.csv.

15. СЕНИН АЛЕКСАНДР НИКОЛАЕВИЧ

Оптимальные условия размножения штаммов золотистого стафилококка. При подозрении на инфекционное заболевание для правильной постановки диагноза часто бывает важно из взятых у пациентов образцов вырастить как можно более многочисленную колонию бактерий, чтобы её было удобнее исследовать. Считается, что оптимальные параметры для размножения штаммов стафилококка в лабораторных условиях следующие: температура 35 градусов, концентрация триптона в питательном растворе 1.0%, время выдержки 24 часа. Для проверки оптимальности этих условий было проведено 30 экспериментов над пятью различными штаммами стафилококка. Для каждого из экспериментов известны время выдержки, температура, концентрация триптона, а также измеренное по окончании выдержки число колониеобразующих единиц (КОЕ) бактерий каждого штамма.

Задача. Оценить зависимость итогового числа КОЕ каждого штаммов стафилококка от внешних условий; одинакова ли эта зависимость?

Данные. Staphylococcus aureus.txt.

16. СУХОДОЕВА АННА ЕВГЕНЬЕВНА

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.

Задача. Отличается ли экспрессия белков у здоровых и трисомных мышей в каких-нибудь из экспериментальных подгрупп?

Данные. memantine.xls.

17. ТРАВНИКОВА АРИНА СЕРГЕЕВНА

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.

Задача. Отличается ли эффективность экспериментального лекарства от эффективности стандартного?

Данные. wounds.csv.

18. ТРОШКОВ АРТЁМ ИГОРЕВИЧ

Массовая доля жира в организме. Массовая доля жира, важная характеристика здоровья, рассчитывается через плотность тела, измеряемую при помощи взвешивания в воде. Для 252 мужчин проведены такие расчёты.

Задача. Имеются также данные антропометрии (возраст, рост, вес, обхват грудной клетки и т.д.) как связаны простые антропометрические показатели (возраст, рост, вес, ИМТ) с обхватами?

Данные. fat.xls.

19. УСКОВА ЕКАТЕРИНА ДМИТРИЕВНА

Одеяла с электрообогревом. Одеяла с электрообогревом применяются в хирургии для восстановления температуры тела пациента после операции. Имеются четыре вида одеяла: стандартный, B0, и три экспериментальных — B1, B2, B3. Для 41 пациента известно время, за которое нормальная температура тела восстанавливается при использовании одеяла одного из видов.

Задача. Отличаются ли экспериментальные одеяла от стандартного?

Данные. blanket.txt

20. УШАКОВ ДАНИИЛ ЕВГЕНЬЕВИЧ

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.

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 control и treatment группах соответственно, и считать практически значимыми те отличия, для которых $|F_c(C, T)| > 1.5$.

Задача. 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?

Данные. gene_high_throughput_sequencing.csv

21. ФЁДОРОВ ИЛЬЯ СЕРГЕЕВИЧ

Содержание азота в кормовых растениях. Четыре вида растений (Leucaena leucosephala, 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.

Задача. Как содержание азота меняется для разных видов растений при разных условиях выращивания? Какие растения лучше всего подходят для сельскохозяйственного использования?

Данные. plants.xlsx

22. ШАРЫПОВ РУСЛАН РИНАТОВИЧ

Рейтинг шоколадок. Для каждой шоколадки известно: % какао-бобов, страна-производитель, страна-источник какао и оценка от 20 до 100.

Задача. Влияют ли эти признаки на рейтинг шоколада?

Данные. chocolate_ratings.csv.

23. ШВЕЦ ПАВЕЛ ИГОРЕВИЧ

Рак лёгких в Китае. Для участников исследования, проживающих в одном из восьми городов Китая, известно, курят ли они и больны ли раком лёгких.

Задача. Как связаны риск заболевания раком лёгких, курение и город проживания участников исследования?

Данные. china_smoking.xls.

24. ШИБАНОВ АЛЕКСЕЙ ЕВГЕНЬЕВИЧ

Информация о датасетах с kaggle. Для датасетов известно: их размер, лицензия распространения, формат данных и количество загрузок.

Задача. От чего зависит число загрузок?

Данные. kaggle_datasets.csv.