Домашнее задание: Проверка статистических гипотез в R

Цель: Изучить различные методы проверки статистических гипотез с использованием среды разработки R. Вы научитесь выбирать подходящий тест в зависимости от типа данных и формулировать гипотезы.

Задача: Найти подходящую базу данных (из открытых источников или своей профессиональной области) и выполнить анализ по следующим этапам.

Описание задания

1. Вам необходимо выбрать один из предложенных наборов данных и сформулировать гипотезы для анализа. Гипотеза должна быть **глобальной,** далее разделяться на субгипотезы. Вы проводите полноценное исследование, поэтому важно избегать «островков» анализа. ***Работа должна быть последовательна, идти к определенному выводу.***
2. Используя разные типы данных (количественные и категориальные), проведите тесты, соответствующие вашим гипотезам.
3. Отчет должен содержать:
    - Постановку задачи.
    - Формулировку нулевой (H_0) и альтернативной (H_a) гипотез.
    - Проведение анализа в R с выводом необходимых статистик.
    - Интерпретацию результатов: примите или отвергните H_0.

Этапы выполнения задания

1. Описание выборки и предварительный анализ данных

Шаг 1: Выбор набора данных

Вам предлагаются следующие наборы данных:

- **`iris`** (из R, о морфологии растений).
- **`mtcars`** (данные об автомобилях).
- **`airquality`** (качество воздуха в Нью-Йорке).
- **`Titanic`** (в библиотеке `titanic`, выживаемость пассажиров).
- Собственный набор данных.

Шаг 2: Формулировка гипотез

Для каждой гипотезы подберите тест, опираясь на тип данных и цель исследования.
Привожу примеры тестов – все более ваш ответственный выбор[1]!

1. **Сравнение средних значений (t-тест или ANOVA):**

[1] Для каждого теста вы должны знать нулевую гипотезу. Как говорится «пока не доказано – не волнует, что сказано». Придерживаемся этого принципа.

- *Пример*: «Есть ли различия в длине чашелистиков (`Sepal.Length`) между видами ириса (`Species`) в данных `iris`?»
- Тест: t-тест или однофакторный ANOVA.

2. **Сравнение дисперсий (тест Левена или Барта):**
   - Пример: «Равны ли дисперсии мощности (`hp`) для автомобилей с разным числом цилиндров (`cyl`) в данных `mtcars`?»
   - Тест: тест Левена или тест Барта.
3. **Сравнение медиан (тест Манна-Уитни или Краскела-Уоллиса):**
   - Пример: «Отличаются ли медианы температуры (`Temp`) для месяцев в данных `airquality`?»
   - Тест: тест Краскела-Уоллиса.
4. **Многомерный анализ дисперсий (MANOVA):**
   - Пример: «Зависят ли одновременно длина и ширина лепестков (`Petal.Length`, `Petal.Width`) от вида ириса в данных `iris`?»
   - Тест: MANOVA.
5. **Анализ категориальных данных (хи-квадрат, точный тест Фишера):**
   - Пример: «Существует ли связь между классом пассажира (`Pclass`) и выживаемостью (`Survived`) в данных `Titanic`?»
   - Тест: хи-квадрат или точный тест Фишера.
6. **Корреляционный анализ (Пирсон, Спирмен, Кендалл):**
   - Пример: «Есть ли линейная связь между расходом топлива (`mpg`) и массой автомобиля (`wt`) в данных `mtcars`?»
   - Тест: корреляция Пирсона или Спирмена.
7. **Тест на нормальность распределения (тест Шапиро-Уилка или Колмогорова-Смирнова):**
   - Пример: «Следует ли распределение длины лепестков (`Petal.Length`) в данных `iris` нормальному закону?»
   - Тест: Шапиро-Уилк.

3. Требования к отчету

1. Оформите гипотезы для каждой задачи: нулевая и альтернативная.
2. Используйте подходящие тесты и обоснуйте их выбор.
3. Приведите код выполнения анализа в R.
4. Отразите выводы по каждому тесту: принимайте или отвергайте H_0 с указанием уровня значимости ($\alpha=0.05$). При желании можете варьировать уровень значимости в вашу пользу.
5. Для каждого теста необходимо посчитать мощность критерия, и естественно знать, что это такое.

Дополнительные задания (для углубленного анализа, а также для доп баллов)

- Проведите факторный анализ на наборе данных, интерпретируя выделенные факторы. Про факторный анализ необходимо знать все главное (методы отбора в факторы – метод главных компонент и т.д). Также нужно знать про методы вращения: варимакс, квартимакс, прямой облимин и многое другое.
- Выполните кластеризацию в данных, сравнив группы по средним значениями.

Также нужно будет знать разницу между кластерным и факторным анализами.

Советы по выбору данных

Подходящие темы для исходных данных:

- **Социологические опросы** – топ. Можно разобрать в качестве анализа успеваемость студентов в ВУЗах, посмотреть разницу между теми, кто срывается в СДВГ[2] и проанализировать средний балл. Таже проанализировать у каких студентов (отличников или не очень) более высокий балл по депрессии. Знайте – числовые тесты для этого не работают, это полностью категориальные тесты.
- **Продажи на маркетплейсах**: анализ продаж продуктов, зависимости выручки от скидок, сезона или рейтинга товаров. Внедрение ИИ в выбор тех или иных товаров.
- **Спорт**: какие сеты выбирать, чтобы накачаться. Анализ среднего обхвата бицепса. Что влияет на снижение веса.
- **Щепетильные темы** – алкоголь, анализ лайков в инстаграме, тиндере и все на ваш вкус.
- **Все то, что не РЖД**. Ну, без комментариев…

Критерии оценки работы

- **Постановка задачи и формулировка гипотез**
  - Четкость и логичность формулировки нулевой (H_0) и альтернативной (Ha) гипотез для всех задач.
  - Обоснование выбора каждого статистического теста в соответствии с типом данных и поставленной гипотезой.

- **Выбор и использование статистических тестов**
  - Корректность выбора тестов для анализа.
  - Проверка необходимых предположений тестов (нормальность, равенство дисперсий, независимость и т.д.).
  - Соответствие выбранных тестов поставленным задачам.

- **Интерпретация результатов**
  - Четкое описание выводов по каждому тесту с обоснованием принятия или отклонения нулевой гипотезы.
  - Указание уровня значимости (α) и интерпретация статистик (p-значения, F-статистика, хи-квадрат и т.д.).
  - Связь результатов анализа с поставленной задачей.

- **Оформление отчета**
  - Логичность структуры отчета и последовательность изложения.
  - Применение графиков и таблиц для наглядности результатов.

---

[2] Надеюсь, из-за моей дисциплины вы ее не словите. Просто хочу сделать вас чуть более умными.

Обращение к студентам

Дорогие студенты!

Вот и подошел к концу наш курс. Я искренне надеюсь, что он был для вас полезным, интересным и вдохновляющим. Для меня было важно не только передать вам знания, но и показать, как они могут быть применимы в реальных задачах. По крайней мере если ваша жизнь будет зависеть от покера, и вы попадете в плен мафиози, то скажете, что ваше 2+3 это фулл на его столе.

Спасибо вам за вашу вовлеченность, старание и интерес. Я был рад быть частью вашего профессионального пути.

С уважением,

Баташов Р.А.