*Анна Журавлева*

**ДЗ№3 по курсу «Современные методы анализа данных»**

**Источник данных** - <https://www.kaggle.com/spscientist/students-performance-in-exams>.

**Описание данных**

Имеются данные об успеваемости студентов по различным дисциплинам. База данных содержит оценки студентов, полученных на экзамене по математике, чтению и письму.

Для анализа в рамках данной задачи были отобраны студенты женского пола, чьи родители имели уровень образования "master's degree", а параметр 'test preparation course' (проходил ли курс подготовки перед тестом) соответствует значению ‘none’ (не проходил). В результате были получены данные по 22 студентам. Проверим, можно ли считать, что успеваемость студентов по математике и письму связаны.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** | **19** | **20** | **21** | **22** |
| **math score** | 90 | 50 | 62 | 56 | 62 | 87 | 81 | 45 | 55 | 74 | 40 | 64 | 53 | 54 | 40 | 77 | 52 | 81 | 50 | 73 | 78 | 92 |
| **writing score** | 93 | 58 | 75 | 65 | 68 | 100 | 91 | 54 | 70 | 82 | 54 | 66 | 68 | 63 | 54 | 84 | 61 | 87 | 73 | 74 | 96 | 100 |

Изображение выглядит как текст

Автоматически созданное описание

**Решение**

Решение: Пусть выборка {(X1,Y1),...,(Xn,Yn)}T, n = 22, порождена двумерным случайным вектором W = (X ,Y)T , имеющим некоторое непрерывное распределение FW (x, y). Проверим гипотезу H0 о независимости случайных величин X и Y:

H0: τXY = 1− 2P{(X2 −X1)(Y2 −Y1) < 0} = 0

Против альтернативной гипотезы

HA: τXY = 1− 2P{(X2 −X1)(Y2 −Y1) < 0} ≠ 0

Критерий Спирмена

Применим критерий Спирмена. Для вычисления статистики критерия составим таблицу рангов

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** | **19** | **20** | **21** | **22** |
| math score | 21 | 4,5 | 11,5 | 10 | 11,5 | 20 | 18,5 | 3 | 9 | 15 | 1,5 | 13 | 7 | 8 | 1,5 | 16 | 6 | 18,5 | 4,5 | 14 | 17 | 22 |
| writing score | 19 | 4 | 14 | 7 | 9,5 | 21,5 | 18 | 2 | 11 | 15 | 2 | 8 | 9,5 | 6 | 2 | 16 | 5 | 17 | 12 | 13 | 20 | 21,5 |

в которой ri — реализация ранга элемента xi в выборке {X1,...,X22}T, а si — реализация ранга элемента yi в выборке {Y1,...,Y22}T .

Реализация статистики Спирмена

Так как в выборках имеются связки, то при вычислении коэффициента следует внести поправку. В реализации выборки соответствующей случайной величине X есть четыре связка размера 2, а в выборке, соответствующей случайной величине Y — две связки размера 2 и одна связка размера 3. Тогда

3

Изображение выглядит как текст

Автоматически созданное описание

Если в качестве альтернативной гипотезы выбрать HA: τ ≠ 0, то критическая область будет иметь вид

где z и — квантили уровня α и 1−α распределения коэффициента ранговой корреляции Спирмена при справедливости гипотезы H0 о независимости для выборки объема n. По таблицам находим z0,975, 22 = 0.428, z0,025, 22 = -0.428.

Таким образом, реализация статистики 0.923 попадает в критическую область

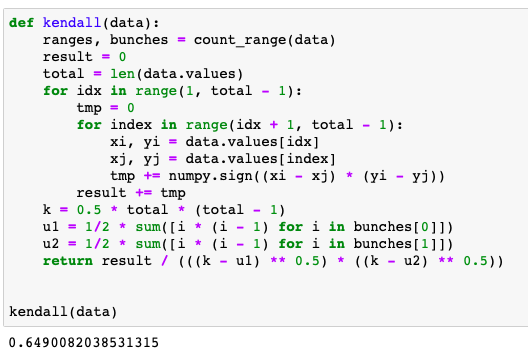
Таким образом, гипотеза о независимости случайных величин X и Y отвергается на уровне значимости 0,05 в пользу альтернативы.

Критерий Кендалла

Применим теперь критерий Кендалла. Поскольку в наблюдениях имеются связки, то для вычисления коэффициента τ надо воспользоваться формулой

В выборке иксов есть четыре связка размера 2, а в выборке, соответствующей случайной величине Y — две связки размера 2 и одна связка размера 3. Тогда

Тогда реализация коэффициента согласованности

****

Если в качестве альтернативной гипотезы выбрать HA: τ ≠ 0, то критическая область будет иметь вид

где z и — квантили уровня α и 1−α распределения коэффициента ранговой корреляции Кендалла при справедливости гипотезы H0 о независимости для выборки объема n. По таблицам находим z0,975, 22 = 0.307, z0,025, 22 = -0.307.

Полученное значение статистики входит в критическую область

Таким образом, гипотеза о независимости случайных величин X и Y отвергается на уровне значимости α = 0,05 в пользу альтернативы.

Таким образом, критерии Спирмена (0.923) и Кендалла ( указывают на наличие зависимости случайных величин Х (успеваемость по математике) и У (успеваемость по письму) у студентов.

* Вычисления: <https://github.com/AnnaZhuravleva/AnDan_2021/blob/main/HW3/HW3.ipynb>