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

```
df = pd.read_csv('/Volumes/MY_DRIVE/MA_CS/An_Dan/HW/HW3/StudentsPerformanc
df = df[df['gender'] == 'female']
df = df[df['parental level of education'] == "master's degree"]
df = df[df['test preparation course'] == 'none']


data = df[['math score', 'writing score']]
data.transpose().to_excel('students_performance.xlsx')
data.transpose()
```

	2	14	29	32	79	106	164	225	478	500	...	579	600	607	781	789	861	892	901
math score	90	50	62	56	62	87	81	45	55	74	...	53	54	40	77	52	81	50	73
writing score	93	58	75	65	68	100	91	54	70	82	...	68	63	54	84	61	87	73	74

2 rows × 22 columns

Решение

Решение: Пусть выборка $\{(X_1,Y_1),...,(X_n,Y_n)\}^T$, n = 22, порождена двумерным случайным вектором $W = (X ,Y)^T$, имеющим некоторое непрерывное распределение F_W (x, y). Проверим гипотезу H_0 о независимости случайных величин X и Y:

$$H_0: \tau_{XY} = 1- 2P\{(X_2 -X_1)(Y_2 -Y_1) < 0\} = 0$$

Против альтернативной гипотезы

$$H_A: \tau_{XY} = 1- 2P\{(X_2 -X_1)(Y_2 -Y_1) < 0\} \neq 0$$

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 r_i — реализация ранга элемента x_i в выборке $\{X_1,...,X_{22}\}^T$, а s_i — реализация ранга элемента y_i в выборке $\{Y_1,...,Y_{22}\}^T$.

Реализация статистики Спирмена

$$\hat{p}_s = 1 - \frac{6 * [(21 - 19)^2 + \cdots + (22 - 21{,}5)^2]}{\frac{1}{6} * (22^3 - 22)}$$

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

$$u_1 = {}^1\!/_{12} * (2^3 - 2) * 4 = 2 \qquad u_2 = {}^1\!/_{12} * ((2^3 - 2) * 3 \ + (3^3 - 3)) = 3$$

$$\hat{p}_s = 1 - \frac{6 * [(21 - 19)^2 + \cdots + (22 - 21{,}5)^2]}{\frac{1}{6} * (22^3 - 22) - 5} = 0.923$$

```
def spearman(data):
    ranges, bunches = count_range(data)
    total = len(data.values)
    sum_deltas = sum([(i - j)**2 for i, j in ranges])
    u1 = 1/12 * sum([i ** 3 - i for i in bunches[0]])
    u2 = 1/12 * sum([i ** 3 - i for i in bunches[1]])
    print('Т.к. в выборках есть связки, то: ', end = ' ')
    print(f'u1 = {u1}\tu2 = {u2}')
    return 1 - (sum_deltas / ((1 / 6) * (total ** 3 - total) - (u1 + u2)))



spearman(data)
```

```
Т.к. в выборках есть связки, то:  u1 = 2.0      u2 = 3.0
```

```
0.9227066817667045
```

Если в качестве альтернативной гипотезы выбрать H_A: $\tau \neq 0$, то критическая область будет иметь вид

$$[-1;\ z_{\frac{\alpha}{2},n}) \cup (z_{1-\frac{\alpha}{2},n};\ 1]$$

где $zz_{\frac{\alpha}{2},n}$ и $z_{1-\frac{\alpha}{2},n}$— квантили уровня α и 1−α распределения коэффициента ранговой корреляции Спирмена при справедливости гипотезы H_0 о независимости для выборки объема n. По таблицам находим $z_{0,975,\,22} = 0.428$, $z_{0,025,\,22} = -0.428$.

Таким образом, реализация статистики $\hat{p}_s = 0.923$ попадает в критическую область

$$[-1;\ -0.428) \cup (0.428;\ 1]$$

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

$$\widehat{\tau_{XY}} = \frac{\sum_{1\le i<j\le n} sign\{(X_i - X_j)(Y_i - Y_j)\}}{\sqrt{\frac{1}{2}n(n-1) - u_1}\sqrt{\frac{1}{2}n(n-1) - u_2}}$$

$$\sum_{i=1}^{21}\sum_{j=i+1}^{22} sign\{(X_i - X_j)(Y_i - Y_j)\} = 18 + 13 + 11 + 9 + 15 + 11 + 11 + 8 + 11 + 9 + 5 + 4 + 5 + 6 + 5 + 2 + 1 + 2 + 1 + 0 = 147$$

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

$$u_1 = {}^1\!/_2 * 2 * (2-1) * 4 = 4$$

$$u_2 = {}^1/_2 * \left(2 * (2-1) * 3 \ + 3 * (3-1)\right) = 5$$

Тогда реализация коэффициента согласованности

$$\widehat{\tau_{XY}} = \frac{147}{\sqrt{\frac{22*21}{2} - 4}\sqrt{\frac{22*21}{2} - 5}} = 0.649$$

```
def kendall(data):
    ranges, bunches = count_range(data)
    result = 0
    total = len(data.values)
    for idx in range(1, total - 1):
        tmp = 0
        for index in range(idx + 1, total - 1):
            xi, yi = data.values[idx]
            xj, yj = data.values[index]
            tmp += numpy.sign((xi - xj) * (yi - yj))
        result += tmp
    k = 0.5 * total * (total - 1)
    u1 = 1/2 * sum([i * (i - 1) for i in bunches[0]])
    u2 = 1/2 * sum([i * (i - 1) for i in bunches[1]])
    return result / (((k - u1) ** 0.5) * ((k - u2) ** 0.5))


kendall(data)
```

```
0.6490082038531315
```

Если в качестве альтернативной гипотезы выбрать H_A: $\tau \neq 0$, то критическая область будет иметь вид

$$[-1;\ z_{\frac{\alpha}{2},n}) \cup (z_{1-\frac{\alpha}{2},n};\ 1]$$

где $zz_{\frac{\alpha}{2},n}$ и $z_{1-\frac{\alpha}{2},n}$— квантили уровня α и 1−α распределения коэффициента ранговой корреляции Кендалла при справедливости гипотезы H_0 о независимости для выборки объема n. По таблицам находим $z_{0{,}975,\ 22} = 0.307$, $z_{0{,}025,\ 22} = -0.307$.

Полученное значение статистики $\widehat{\tau_{XY}} = \ 0.649$ входит в критическую область

$$[-1;\ -0.307) \cup (0.307;\ 1]$$

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

Таким образом, критерии Спирмена ($\hat{p}_s = 0.923$) и Кендалла ($\widehat{\tau_{XY}} = \ 0.649$) указывают на наличие зависимости случайных величин X (успеваемость по математике) и У (успеваемость по письму) у студентов.

- Вычисления: https://github.com/AnnaZhuravleva/AnDan_2021/blob/main/HW3/HW3.ipynb