Лабораторная работа 8
Тестирование гипотезы о значимости коэффициента корреляции

Гузовская Александра Чеславовна
Б9123-01.03.02сп

26 мая 2025 г.

Матчасть

Коэффициент корреляции Пирсона

Для выборки пар $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ коэффициент корреляции Пирсона вычисляется по формуле:

$$\hat{\rho}_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \cdot \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

где:

- $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ – выборочные средние
- n – объем выборки

Ранжирование данных

Для коэффициента Спирмена требуется вычисление рангов. Для выборки $X_1, X_2, \ldots, X_n$:

$$R(X_{(i)}) = \begin{cases} i, & \text{если } X_{(i-1)} < X_{(i)} < X_{(i+1)} \\ \frac{1}{k}\sum_{j=0}^{k-1}(i+j), & \text{если } X_{(i)} = \cdots = X_{(i+k-1)} \end{cases}$$

Коэффициент корреляции Спирмена

После вычисления рангов $R(X_i)$ и $R(Y_i)$:

$$\hat{\rho}_S = \frac{\sum_{i=1}^{n}(R(X_i) - \overline{R_X})(R(Y_i) - \overline{R_Y})}{\sqrt{\sum_{i=1}^{n}(R(X_i) - \overline{R_X})^2 \cdot \sum_{i=1}^{n}(R(Y_i) - \overline{R_Y})^2}}$$

или

$$\hat{\rho}_S = \hat{\rho}_{R(X_i),R(Y_i)}$$

Проверка на нормальность

Используем критерий Шапиро-Уилка (выборки меньше 30 элементов) В коде используем встроенную в scipy.stats функцию shapiro()
Если p-value больше уровня значимости α (стандартно $\alpha = 0.05$), нормальность не отвергается. Чем ближе значение статистики к 1, тем ближе предполагаемое распределение для данной выборки к нормальному

Статистические гипотезы для корреляции

Проверяется гипотеза H_0 против одной из возможных H_1:

$$H_0 : \rho = 0 \quad \text{(корреляция отсутствует)}$$

$$H_1 : \rho \neq 0 \quad \text{(корреляция существует, двусторонняя проверка)}$$

$$H_1 : \rho > 0 \quad \text{(проверка на положительную корреляцию)}$$

$$H_1 : \rho < 0 \quad \text{(проверка на отрицательную корреляцию)}$$

Проверка значимости

Для коэффициента Пирсона статистика критерия:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

где:

- r - выборочный коэффициент корреляции
- n - объём выборки

- t_{n-2} - t-распределение с $n-2$ степенями свободы

Для коэффициента Спирмена используется аналогичный подход после ранжирования данных.

NB:

- Для нормально распределённых данных предпочтительнее коэффициент Пирсона
- При отклонении от нормальности следует использовать коэффициент Спирмена
- Всегда проверять статистическую значимость полученных коэффициентов
- При проверке гипотез:
  - p-value $< 0.05 \rightarrow$ отвергаем H_0 (корреляция значима)
  - p-value $\geq 0.05 \rightarrow$ нет оснований отвергать H_0

Код программы

```
import numpy as np
import scipy.stats as sps

def pearson_(x, y):
    """
    Расчёт коэффициента корреляции Пирсона
    """
    n = len(x)
    if n != len(y):
        raise ValueError("Выборки должны быть одинаковой длины")

    mean_x = sum(x) / n
    mean_y = sum(y) / n

    numerator = sum((xi - mean_x) * (yi - mean_y) for xi, yi in zip(x, y))
    denominator_x = sum((xi - mean_x)**2 for xi in x)
    denominator_y = sum((yi - mean_y)**2 for yi in y)

    if denominator_x == 0 or denominator_y == 0:
```

```
        r = 0.0
    else:
        r = numerator / np.sqrt(denominator_x * denominator_y)

    if r == 1 or r == -1:
        p_value = 0.0
    else:
        t_value = r * np.sqrt(n - 2) / np.sqrt(1 - r**2)
        p_value = 2 * (1 - sps.t.cdf(abs(t_value), df=n-2))

    return r, p_value

def spearman_(x, y):
    """
    Расчёт коэффициента корреляции Спирмена
    """
    n = len(x)
    if n != len(y):
        raise ValueError("Выборки должны быть одинаковой длины")

    def compute_ranks(data):
        ranked = {}
        sorted_data = sorted((val, i) for i, val in enumerate(data))
        i = 0
        while i < n:
            j = i
            while j < n and sorted_data[j][0] == sorted_data[i][0]:
                j += 1

            rank = (i + 1 + j) / 2.0

            for k in range(i, j):
                ranked[sorted_data[k][1]] = rank
            i = j
        return [ranked[i] for i in range(n)]

    rank_x = compute_ranks(x)
    rank_y = compute_ranks(y)

    return pearson_(rank_x, rank_y)
```

Задачи

задача 1

По выборке объема $n = 100$, извлеченной из двумерной нормальной генеральной совокупности (X, Y), получена корреляционная табл. 16.

Таблица 16

Y	X						
	100	105	110	115	120	125	n_y
35	4	—	6	7	8	3	28
45	5	5	2	10	—	—	22
55	6	7	—	—	2	3	18
65	—	6	5	4	—	2	17
75	5	1	2	4	3	—	15
n_x	20	19	15	25	13	8	$n = 100$

```
def task1():
    x_values = []
    y_values = []

    x_bins = [100, 105, 110, 115, 120, 125]
    y_bins = [35, 45, 55, 65, 75]
    frequencies = [
        [4, 0, 6, 7, 8, 3],
        [5, 5, 2, 10, 0, 0],
        [6, 7, 0, 0, 2, 3],
        [0, 6, 5, 4, 0, 2],
        [5, 1, 2, 4, 3, 0]
    ]

    for i in range(len(y_bins)):
        for j in range(len(x_bins)):
            count = frequencies[i][j]
            if count > 0:
```

```
            x_values.extend([x_bins[j]] * count)
            y_values.extend([y_bins[i]] * count)

r_pearson, p_pearson = pearson_(x_values, y_values)
rho_spearman, p_spearman = spearman_(x_values, y_values)

print("Задача 1:")
print(f"Коэффициент корреляции Пирсона:
        r = {r_pearson}, p = {p_pearson}")
print(f"Коэффициент корреляции Спирмена:
        rho = {rho_spearman}, p = {p_spearman}")
r_lib, p_lib = sps.pearsonr(x_values, y_values)
rho_lib, p_rho_lib = sps.spearmanr(x_values, y_values)
print("Scipy (Пирсон):", r_lib, p_lib)
print("Scipy (Спирмен):", rho_lib, p_rho_lib)
shapiro_test_x = sps.shapiro(x_values)
shapiro_test_y = sps.shapiro(y_values)
print(f"Проверка на нормальность:
        (для x) W = {shapiro_test_x.statistic} и
                p = {shapiro_test_x.pvalue:.10f},
        (для y) W = {shapiro_test_y.statistic} и
                p = {shapiro_test_y.pvalue:.10f}")
```

- p-value Пирсона $= 0.107 > 0.05 \rightarrow$ не отвергаем H_0
- p-value Спирмена $= 0.067 > 0.05 \rightarrow$ не отвергаем H_0
- Вывод: статистически значимая корреляция не обнаружена

задача 2

Два преподавателя оценили знания 12 учащихся по стобалльной системе и выставили им следующие оценки (в первой строке указано количество баллов, выставленных первым преподавателем, а во второй — вторым):

98	94	88	80	76	70	63	61	60	58	56	51
99	91	93	74	78	65	64	66	52	53	48	62

```
def task2():
    teacher1 = [98, 94, 88, 80, 76, 70, 63, 61, 60, 58, 56, 51]
    teacher2 = [99, 91, 93, 74, 78, 65, 64, 66, 52, 53, 48, 62]

    r_pearson, p_pearson = pearson_(teacher1, teacher2)
    rho_spearman, p_spearman = spearman_(teacher1, teacher2)

    print("\nЗадача 2:")
    print(f"Коэффициент корреляции Пирсона: r = {r_pearson},
            p = {p_pearson:.10f}")
    print(f"Коэффициент корреляции Спирмена: rho = {rho_spearman},
            p = {p_spearman:.10f}")
    r_lib, p_lib = sps.pearsonr(teacher1, teacher2)
    rho_lib, p_rho_lib = sps.spearmanr(teacher1, teacher2)
    print(f"Scipy (Пирсон): r = {r_lib}, p = {p_lib:.10f}")
    print(f"Scipy (Спирмен): rho = {rho_lib}, p = {p_rho_lib:.10f}")
    shapiro_test_x = sps.shapiro(teacher1)
    shapiro_test_y = sps.shapiro(teacher2)
    print(f"Проверка на нормальность:
            (для учителя 1) W = {shapiro_test_x.statistic} и
                    p = {shapiro_test_x.pvalue},
            (для учителя 2) W = {shapiro_test_y.statistic} и
                    p ={shapiro_test_y.pvalue}")
```

- p-value Пирсона $< 0.05 \rightarrow$ отвергаем H_0
- p-value Спирмена $< 0.05 \rightarrow$ отвергаем H_0
- Вывод: существует статистически значимая корреляция

Вывод программы

```
Задача 1:
Коэффициент корреляции Пирсона: r = -0.16210378679454263, p = 0.10710692680952527
Коэффициент корреляции Спирмена: rho = -0.18392692537117117, p = 0.06698033678777149
Scipy (Пирсон): -0.16210378679454265 0.10710692680952469
Scipy (Спирмен): -0.18392692537117117 0.0669803367877715
Проверка на нормальность:
(для x) W = 0.9134126535978283 и p = 0.0000064005,
(для y) W = 0.8731933560935314 и p = 0.0000000945


Задача 2:
Коэффициент корреляции Пирсона: r = 0.9352773674245594, p = 0.0000080186
Коэффициент корреляции Спирмена: rho = 0.916083916083916, p = 0.0000284280
Scipy (Пирсон): r = 0.9352773674245595, p = 0.0000080186
Scipy (Спирмен): rho = 0.9160839160839163, p = 0.0000284280
Проверка на нормальность:
(для учителя 1) W = 0.9215796832254396 и p = 0.2992682032342635,
(для учителя 2) W = 0.9312935536852316 и p =0.39402483634132013
```