МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение
высшего образования
«Дальневосточный федеральный университет»
(ДВФУ)

ИНСТИТУТ МАТЕМАТИКИ И КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ

Лабораторная работа №8
Тестирование гипотезы о значимости коэффициента корреляции

Дисциплина «Теория вероятностей и математическая статистика»

Студент группы Б9123-01.03.02ии

Моттуева Уруйдана Михайловна

г. Владивосток

2025

1. Напишите функцию, вычисляющую коэффициент корреляции Пирсона и определяющую его значимость. На вход функции подаются 2 выборки. Возвращает функция p-значение и значение коэффициента.

Даны две выборки $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$, представляющие пары значений случайных величин X и Y.

Гипотезы:
$H_0: \rho_{X,Y} = 0$,
$H_1: \rho_{X,Y} \neq 0$.

1. Коэффициент корреляции Пирсона

Выборочный коэффициент корреляции Пирсона $\hat{\rho}_{X,Y}$ вычисляется по формуле:

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

где:
- X_i, Y_i — значения выборок,
- $\bar{X}, \bar{Y}$ — средние значения выборок.

Замечания:

- Показывает линейную зависимость между X и Y;
- Стоит использовать в случае нормального распределения выборок;
- $\left|\hat{\rho}_{X,Y}\right| \leq 1$.

Для проверки значимости коэффициента используется t-статистика:

$$T_{n-2} = \hat{\rho}_{X,Y}\sqrt{\frac{n-2}{1-\hat{\rho}_{X,Y}{}^2}} \sim T(n-2),$$

где n — размер выборки. Затем рассчитывается p-значение для проверки гипотезы $H_0: \hat{\rho}_{X,Y} = 0$ (отсутствие корреляции).

p-значение:

$$p = 2F_{T_{n-2}}(-|t|)$$

Моя реализация

```
def pearson_corr(x, y):
    n = len(x)
    mean_x = np.mean(x)
    mean_y = np.mean(y)

    cov = np.sum((x - mean_x) * (y - mean_y))
    std_x = np.sqrt(np.sum((x - mean_x)**2))
    std_y = np.sqrt(np.sum((y - mean_y)**2))

    r = cov / (std_x * std_y)

    # Расчет p-value
    if n > 2:
        t = r * np.sqrt((n - 2) / (1 - r**2))
        p_value = 2 * (1 - stats.t.cdf(abs(t), df=n-2))
    else:
        p_value = np.nan

    return r, p_value
```

Код с использованием scipy

```
pearsonr(X_table, Y_table)
```

2. Коэффициент корреляции Спирмена

2.1. Ранги

Составим вариационный ряд $X_{(1)} \le X_{(2)} \le \dots \le X_{(n)}$ для выборки $X_1, X_2, \dots, X_n$.

Если $X_{(i-1)} < X_{(i)} = \dots = X_{(i+k-1)} < X_{(i+k)}$, то ранг:

$$R(X_{(i)}) = \frac{1}{k}\sum_{j=0}^{k-1}(i+j),$$

где:

- k — количество одинаковых значений,

- i — начальная позиция первого из этих значений в отсортированном массиве.

2.2. Коэффициент корреляции

Пусть имеется выборка пар $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ из (X, Y). Вычислим ранги в выборке из X: $R(X_1), R(X_2), \dots, R(X_n)$ и в выборке из Y: $R(Y_1), R(Y_2), \dots, R(Y_n)$.

Тогда коэффициент корреляции Спирмена равен $\hat{\rho}_{R(X),R(Y)}$.
$R(X)$ и $R(Y)$ — ранги выборок X и Y.

Замечания:

- Показывает монотонную зависимость между переменными
- Стоит использовать в случае не нормального распределения выборок

Он вычисляется как коэффициент Пирсона для рангов данных.

- $R(X_i), R(Y_i)$— ранги элементов выборок.

Значимость проверяется аналогично коэффициенту Пирсона.

Моя реализация

```
def spearman_corr(x, y):
    n = len(x)

    rank_x = stats.rankdata(x)
    rank_y = stats.rankdata(y)
    d = rank_x - rank_y

    rho, p_value = pearson_corr(rank_x, rank_y)

    return rho, p_value
```

Код с использованием scipy

```
spearmanr(X_table, Y_table)
```

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 create_samples_from_table():
    x_values = [100, 105, 110, 115, 120, 125]
    y_values = [35, 45, 55, 65, 75]
    frequencies = [
        [4, 0, 6, 7, 8, 3],   # Y=35
        [5, 5, 2, 10, 0, 0],  # Y=45
        [6, 7, 0, 0, 2, 3],    # Y=55
        [0, 6, 5, 4, 0, 2],    # Y=65
        [5, 1, 2, 4, 3, 0]     # Y=75
    ]

    X, Y = [], []
    for y_idx, y in enumerate(y_values):
        for x_idx, freq in enumerate(frequencies[y_idx]):
            if freq > 0:
                X.extend([x_values[x_idx]] * freq)
                Y.extend([y] * freq)
    return np.array(X), np.array(Y)

X_table, Y_table = create_samples_from_table()
```

```
[ ]  shapiro_x = shapiro(X_table)
     shapiro_y = shapiro(Y_table)
     print(f"Таблица 16: Shapiro-Wilk p-value (X): {shapiro_x.pvalue:.3f}")
     print(f"Таблица 16: Shapiro-Wilk p-value (Y): {shapiro_y.pvalue:.3f}")
```

```
Таблица 16: Shapiro-Wilk p-value (X): 0.000
Таблица 16: Shapiro-Wilk p-value (Y): 0.000
```

распределение в первой задаче не нормальное

Используем коэффициент Спирмена.

```python
# задача 1
r_pearson, p_pearson = pearson_corr(X_table, Y_table)
r_spearman, p_spearman = spearman_corr(X_table, Y_table)
r_pearson_lib, p_pearson_lib = pearsonr(X_table, Y_table)
r_spearman_lib, p_spearman_lib = spearmanr(X_table, Y_table)

print(f"мой Пирсон: r={r_pearson:.3f}, p={p_pearson:.3f}")
print(f"scipy Пирсон: r={r_pearson_lib:.3f}, p={p_pearson_lib:.3f}")
print(f"\nмой Спирмен: r={r_spearman:.3f}, p={p_spearman:.3f}")
print(f"scipy Спирмен: r={r_spearman_lib:.3f}, p={p_spearman_lib:.3f}")
```

```
мой Пирсон: r=-0.162, p=0.107
scipy Пирсон: r=-0.162, p=0.107

мой Спирмен: r=-0.184, p=0.067
scipy Спирмен: r=-0.184, p=0.067
```

Полученные p-значения коэффициента Спирмена (> 0.05) больше уровня значимости $\alpha = 0.05$. Гипотеза H_0 не отвергается — корреляция не значима.

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

```python
[85] # Задача 2
X_scores = np.array([98, 94, 88, 80, 76, 70, 63, 61, 60, 58, 56, 51])
Y_scores = np.array([99, 91, 93, 74, 78, 65, 64, 66, 52, 53, 48, 62])
```

```python
[86] shapiro_x_scores = shapiro(X_scores)
shapiro_y_scores = shapiro(Y_scores)
print(f"Оценки: Shapiro-Wilk p-value (X): {shapiro_x_scores.pvalue:.3f}")
print(f"Оценки: Shapiro-Wilk p-value (Y): {shapiro_y_scores.pvalue:.3f}")
```

```
Оценки: Shapiro-Wilk p-value (X): 0.299
Оценки: Shapiro-Wilk p-value (Y): 0.394
```

распределение в задаче 2 нормальное

Используем коэффициент Пирсона.

```python
# задача 2
r_pearson_scores, p_pearson_scores = pearson_corr(X_scores, Y_scores)
r_spearman_scores, p_spearman_scores = spearman_corr(X_scores, Y_scores)
r_pearson_lib_scores, p_pearson_lib_scores = pearsonr(X_scores, Y_scores)
r_spearman_lib_scores, p_spearman_lib_scores = spearmanr(X_scores, Y_scores)

print(f"мой Пирсон: r={r_pearson_scores:.3f}, p={p_pearson_scores:.3f}")
print(f"scipy Пирсон: r={r_pearson_lib_scores:.3f}, p={p_pearson_lib_scores:.3f}")
print(f"\nмой Спирмен: r={r_spearman_scores:.3f}, p={p_spearman_scores:.3f}")
print(f"scipy Спирмен: r={r_spearman_lib_scores:.3f}, p={p_spearman_lib_scores:.3f}")
```

```
мой Пирсон: r=0.935, p=0.000
scipy Пирсон: r=0.935, p=0.000

мой Спирмен: r=0.916, p=0.000
scipy Спирмен: r=0.916, p=0.000
```

Полученные p-значения корреляции Пирсона (< 0.001) меньше уровня значимости $\alpha = 0.05$. Гипотеза H_0 отвергается — корреляция значима.