Звіт
до лабораторної роботи №6:
«Перевірка гіпотези про однорідність вибірок за допомогою статистики Петуніна»

студента 1-го курсу магістратури
факультету комп'ютерних наук та кібернетики
Кравця Олексія

Зміст

1 Теоретичні відомості

Статистика Петуніна (p-статистика) — міра близькості між вибірками, запропонована українським математиком Юрій Петуніним. Використовується для перевірки гіпотези про рівність функцій розподілу двох вибірок.

Розглянемо дві генеральні сукупності G, G' та відповідні функції розподілу F_G, $F_{G'}$.

Нехай задано дві вибірки $x = (x_1, x_2, \ldots, x_n) \in G$ та $x' = (x'_1, x'_2, \ldots, x'_m) \in G'$, а $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ та $x'_{(1)} \leq x'_{(2)} \leq \ldots \leq x'_{(m)}$ – відповідні порядкові статистики та необхідно визначити, чи вони належать однаковим розподілам. Припустимо, що $F_G(u) = F_{G'}(u)$, тоді

$$P(A_{ij}) = P\left(x'_k \in (x_{(i)}, x_{(j)})\right) = p_{ij} = \frac{j-i}{n+1}$$

Якщо маємо вибірку $x' \in (x'_{(1)}, x'_{(2)}, \ldots, x'_{(m)})$, можемо знайти частоту h_{ij} випадкової події A_{ij} та довірчі інтервали $(p_{ij}^{(1)}, p_{ij}^{(2)})$ для ймовірності p_{ij} при заданому рівні значущості β, тобто $B = \left\{p_{ij} \in (p_{ij}^{(1)}, p_{ij}^{(2)})\right\}$, $P(B) = 1 - \beta$. Тоді

$$p_{ij}^{(1)} = \frac{h_{ij}m + \frac{g^2}{2} - g\sqrt{h_{ij}(1-h_{ij})m + \frac{g^2}{4}}}{m + g^2}$$

$$p_{ij}^{(2)} = \frac{h_{ij}m + \frac{g^2}{2} + g\sqrt{h_{ij}(1-h_{ij})m + \frac{g^2}{4}}}{m + g^2}$$

де g задовольняє умову $\phi(g) = 1 - \frac{\beta}{2}$ ($\phi(g)$ – щільність нормального розподілу).Покладемо $g = 3$. Величина g визначає рівень значущості довірчого інтервалу $I_{ij}^{(n,m)} = (p_{ij}^{(1)}, p_{ij}^{(2)})$. В силу правила 3σ рівень значущості цього інтервалу не перевищує 0.05.

Позначимо через N кількість довірчих інтервалів $I_{i,j} = (p_{ij}^{(1)}, p_{ij}^{(2)})$, також $N = \frac{n(n-1)}{2}$. Позначимо L – кількість тих інтервалів I_{ij}, які містять ймовірність $p_{ij}^{(n)}$. Статистику $h_{ij} = \frac{L}{N}$ будемо називати p-статистикою і вона буде мірою близькості $\rho(x, x')$ між вибірками x, x'.

2 Практичні результати

Для тестування візьмемо вибірку з нормального розподілу $N(0, 1)$, розмір вибірки $m = 200$ елементів.

Перевіримо гіпотезу про рівність функцій розподілу для нормальних розподілів $N(\mu, 1)$, де μ змінюється від -2 до 2 з кроком 0.1. Розмір кожної вибірки $n = 100$. Отримаємо рисунок 1. Бачимо на осі абсцис значення μ на осі ординат значення p-статистики.

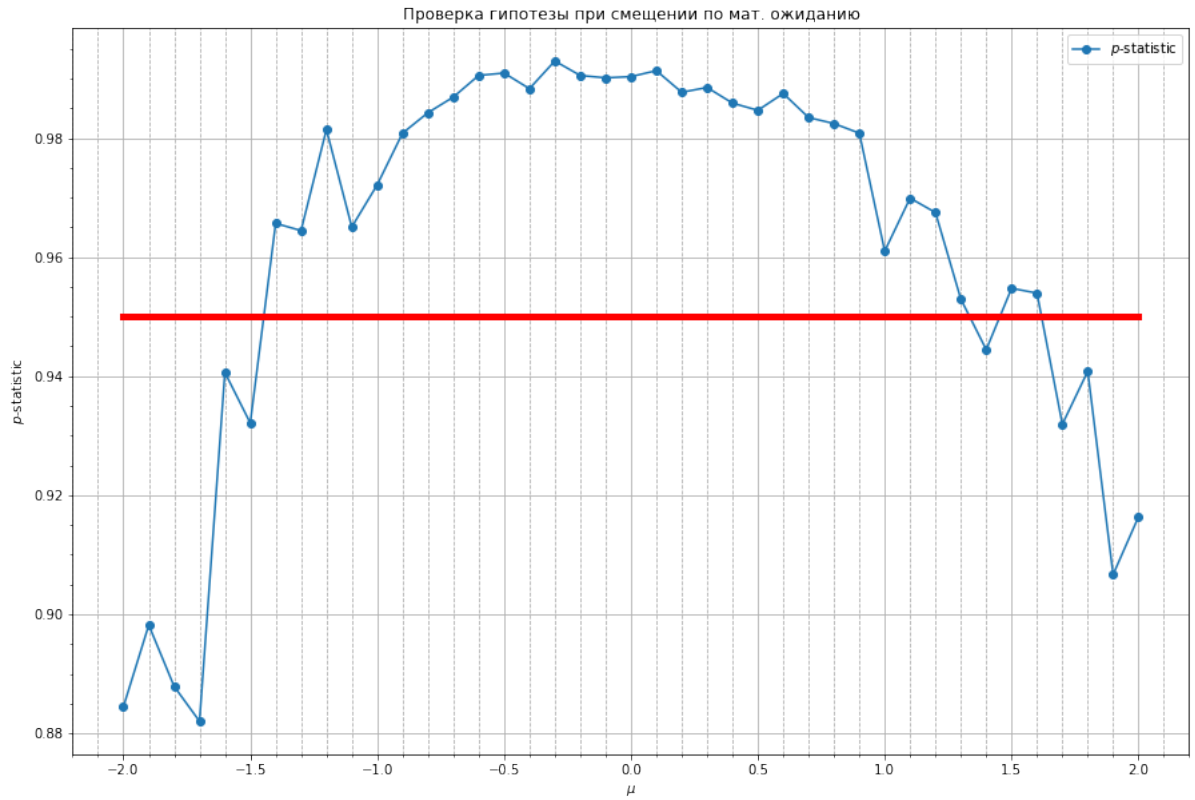

Рис. 1: Перевірка гіпотези при зміщенні по мат. сподіванню

Гіпотезу **не відхиляємо** при наступних значеннях μ:

```
array([-1.4, -1.3, -1.2, -1.1, -1. , -0.9, -0.8, -0.7, -0.6, -0.5, -0.4,
-0.3, -0.2, -0.1,  0. ,  0.1,  0.2,  0.3,  0.4,  0.5,  0.6,  0.7,
0.8,  0.9,  1. ,  1.1,  1.2,  1.3,  1.5,  1.6])
```

Перевіримо гіпотезу про рівність функцій розподілу для нормальних розподілів $N(0, \sigma)$, де σ змінюється від 0.1 до 3 з кроком 0.1. Розмір кожної вибірки $n = 100$. Отримаємо рисунок 2. Бачимо на осі абсцис значення μ на осі ординат значення p-статистики

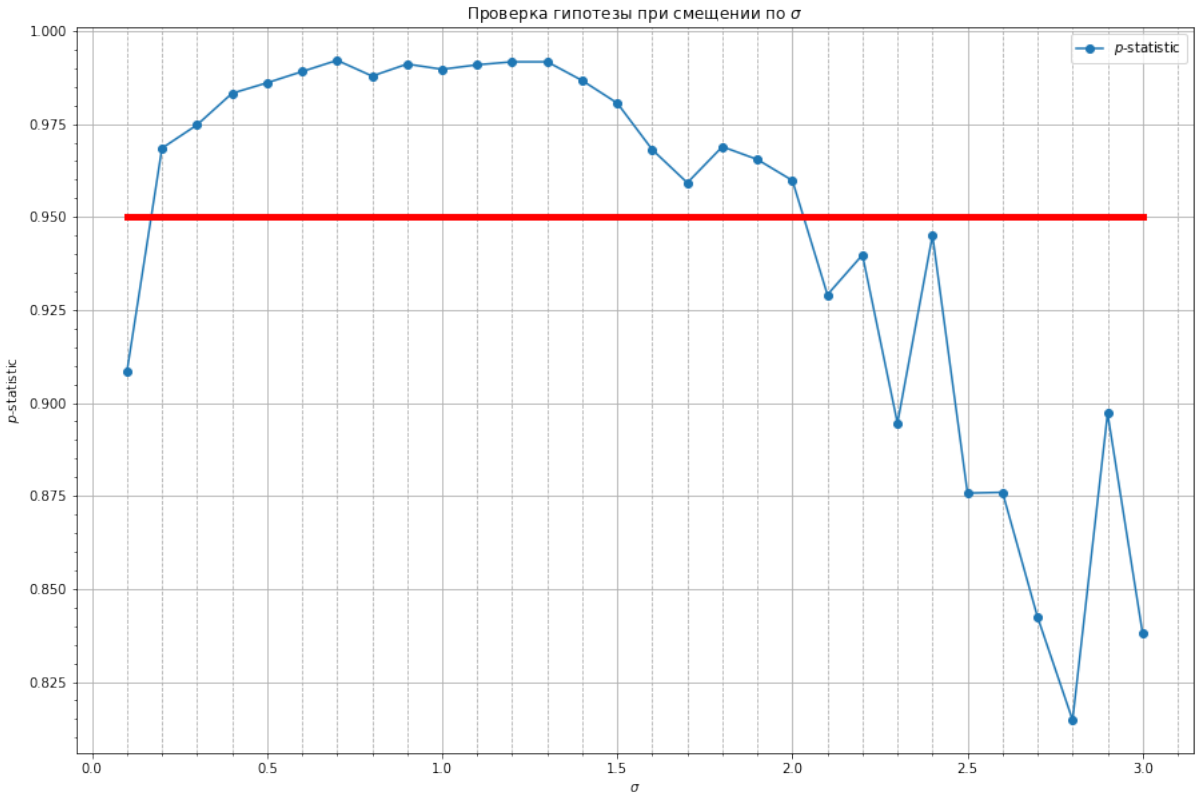

Рис. 2: Перевірка гіпотези при зміщенні σ

Гіпотезу **не відхиляємо** при наступних значеннях σ:

```
array([0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1. , 1.1, 1.2, 1.3, 1.4,
1.5, 1.6, 1.7, 1.8, 1.9, 2. ])
```

Література

[1] https://uk.wikipedia.org/wiki/Статистика_Петуніна

[2] Лекція 13. Непараметрична класифікація за допомогою p-статистики, Клюшин Дмитро Анатолійович