

Лекція 4 (продовження).

Перевірка статистичних гіпотез

4.3.3. Перевірка гіпотези про рівність середніх двох сукупностей нормально розподіленої генеральної сукупності при відомих дисперсіях

Нехай є дві незалежні генеральні сукупності X та Y , які мають нормальний розподіл. Потрібно перевірити гіпотезу $H_0 : M(X) = M(Y)$.

Розглянемо три випадки:

1) Обсяг вибірки великий ($n > 40$) і відомі значення D_x і D_y генеральних сукупностей.

З кожної генеральної сукупності роблять вибірку обсягами n_1 і n_2 відповідно. Статистичні розподіли:

x_i	x_1	x_2	...	x_k
n_{1i}	n_{11}	n_{12}	...	n_{1k}

y_j	y_1	y_2	...	y_m
n_{2j}	n_{21}	n_{22}	...	n_{2m}

$$\bar{x}_B = \frac{1}{n_1} \sum x_i \cdot n_{1i}; \quad \bar{y}_B = \frac{1}{n_2} \sum y_j \cdot n_{2j}.$$

За статистичний критерій візьмемо випадкову величину $Z = \frac{\bar{x}_B - \bar{y}_B}{\sqrt{\frac{D_x}{n_1} + \frac{D_y}{n_2}}}$,

яка має розподіл $Z \rightarrow N(0;1)$. Якщо дисперсії рівні $D_x = D_y = \sigma^2$, то

$$Z = \frac{\bar{x}_B - \bar{y}_B}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

В залежності від формулювання альтернативної гіпотези H_1 будується відповідно правостороння, лівостороння або двостороння критичні області. Спостережуване значення критерію відповідно обчислюється за формулами:

$$Z_{спост} = \frac{\bar{x}_B - \bar{y}_B}{\sqrt{\frac{D_x}{n_1} + \frac{D_y}{n_2}}}; \quad Z_{спост} = \frac{\bar{x}_B - \bar{y}_B}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

Приклад 4.4. Спостерігають за роботою двох автоматів, які мають виготовляти однакові втулки. Для першого автомата зробили вибірку $n_1 = 50$ втулок і отримали середній діаметр $\bar{x}_B = 20,1$ мм. Для другого автомата зробили таку саму вибірку $n_2 = 50$ і отримали середній діаметр $\bar{y}_B = 19,8$ мм. Генеральні дисперсії відомі: $D_x = 1,75$ мм² і $D_y = 1,375$ мм².

На рівні значущості $\alpha = 0,05$ перевірити нульову гіпотезу $H_0 : M(X) = M(Y)$ при альтернативній $H_1 : M(X) \neq M(Y)$.

Розв'язання. Маємо двосторонню критичну область. $z_{1кр} = -z_{2кр}$.

$\Phi(z_{2кр}) = \frac{1-\alpha}{2} = \frac{1-0,05}{2} = 0,475$. За таблицею значень функції

Лапласа $z_{2кр} = 1,96$, тоді $-z_{1кр} = -1,96$. Спостережене значення

критерію:
$$Z_{спост} = \frac{\bar{x}_B - \bar{y}_B}{\sqrt{\frac{D_x + D_y}{n}}} = \frac{20,1 - 19,8}{\sqrt{\frac{1,75 + 1,375}{50}}} = 1,2.$$

Оскільки $Z_{спост} \in [-1,96; 1,96]$, то нульова гіпотеза не відхиляється. Це означає, що середні вибіркowi відрізняються неістотно.

2) Якщо обсяг вибірки великий ($n_1 > 40$; $n_2 > 40$), але невідомі значення генеральних дисперсій, то застосовують їх точкові незміщені оцінки s_x^2 і s_y^2 . Припустимо, що генеральні дисперсії рівні. Тоді критерій перевірки нульової гіпотези можна записати так:

$$T = \frac{\bar{x}_B - \bar{y}_B}{\sqrt{\frac{(n_1 - 1)s_x^2 + (n_2 - 1)s_y^2}{n_1 + n_2 - 2}}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}; \quad T \rightarrow N(0; 1).$$

Критична область будується в залежності від виду альтернативної гіпотези.

Схема перевірки нульової гіпотези:

1. Нульова гіпотеза $H_0 : M(X) = M(Y)$.

2. Альтернативна гіпотеза:

а) $H_1 : M(X) \neq M(Y)$;

б) $H_1 : M(X) > M(Y)$;

в) $H_1 : M(X) < M(Y)$.

3. При великих обсягах вибірок ($n_1 > 40$; $n_2 > 40$), $T \rightarrow N(0;1)$ і за тосовують таблицю для розподілу Стьюдента:

4. Критичні точки:

а) Двостороння критична область: критичні точки симетричні $t_{1кр} = -t_{2кр}$ і

знаходяться з формули: $\Phi(t_{2кр}) = \frac{1-\alpha}{2}$ за таблицею Лапласа.

б) Правостороння критична область: $\Phi(t_{кр}) = \frac{1-2\alpha}{2}$.

в) Лівостороння критична область: $\Phi(t_{кр}) = -\frac{1-2\alpha}{2}$.

5. Обчислюємо $T_{спост} = \frac{\bar{x}_B - \bar{y}_B}{\sqrt{\frac{(n_1-1)s_x^2 + (n_2-1)s_y^2}{n_1+n_2-2}}} \sqrt{\frac{n_1 n_2}{n_1+n_2}}$.

6. Правило прийняття рішень: гіпотеза $H_0 : M(X) = M(Y)$

відхиляється:

а) $|T_{спост}| \geq t_{кр}$;

б) $T_{спост} > t_{кр}$;

в) $T_{спост} < -t_{кр}$.

Приклад 4.5. Досліджувався прибуток програмістів на двох типових фірмах ІТ галузі. Одержали наступні статистичні розподіли:

x_i	140,8	160,8	180,8	200,8	220,8
n_{1i}	2	6	32	8	2

y_j	150,6	160,6	170,6	180,6	190,6
n_{2j}	12	28	40	18	2

Ознаки є незалежними і мають нормальний розподіл. Для рівня значущості $\alpha = 0,01$ перевірити правильність нульової гіпотези $H_0 : M(X) = M(Y)$ при альтернативній $H_1 : M(X) < M(Y)$.

Розв'язання. За даними обчислюємо вибіркові середні і вибіркові дисперсії: $\bar{x}_B = 181,6$; $\bar{y}_B = 167,6$; $D_x = 239,36$; $D_y = 93,0$. Оскільки дисперсії невідомі, знаходимо незміщені оцінки: $n_1 = 50$; $n_2 = 100$;

$$s_x^2 = \frac{n_1}{n_1 - 1} D_x = 244,24; \quad s_y^2 = \frac{n_2}{n_2 - 1} D_y = 93,94.$$

1. Нульова гіпотеза: $H_0 : M(X) = M(Y)$.
2. Альтернативна: $H_1 : M(X) < M(Y)$.
3. Критична область: лівостороння.
4. Критичні точки: $\Phi(t_{кр}) = -\frac{1-2\alpha}{2} = -0,49$; $t_{кр} = -2,33$.
5. Спостережене значення критерію:

$$T_{спост} = \frac{\bar{x}_B - \bar{y}_B}{\sqrt{\frac{(n_1 - 1)s_x^2 + (n_2 - 1)s_y^2}{n_1 + n_2 - 2}}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} = 6,74;$$

$$T_{спост} = 6,74 > t_{кр} = -2,33.$$

Висновок: нульова гіпотеза приймається.

Приклад 4.6. Проведено дві вибірки завантаженості серверів: вдень і вночі. В першому випадку при спостереженні 8 серверів вибіркова середня завантаженість склала 16,2 годин на добу, а середнє квадратичне відхилення — 3,2 год/доба; у другому випадку при спостереженні 9 серверів ті ж характеристики дорівнювали відповідно 13,9 год/доба і 2,1 год/доба. На рівні значущості $\alpha=0,05$ визначити вплив часу доби на завантаженість серверів.

Розв’язання. Гіпотеза, яка перевіряється, $H_0: \bar{x}_0 = \bar{y}_0$, тобто середні значення завантаженості сервера вдень і вночі рівні. В якості конкуруючої гіпотези беремо гіпотезу $H_1: \bar{x}_0 > \bar{y}_0$, прийняття якої означає значний вплив часу доби на завантаженість сервера. Значення статистики критерію, який фактично спостерігається

$$T = \frac{16,2 - 13,9}{\sqrt{\frac{8 \cdot 3,2^2 + 9 \cdot 2,1^2}{8 + 9 - 2} \left(\frac{1}{8} + \frac{1}{9} \right)}} = 1,62$$

Критичне значення статистики для односторонньої області визначається при числі ступенів свободи $k = n_1 + n_2 - 2 = 9 + 8 - 2 = 15$ із умови $\theta(t, k) = 1 - 2 \cdot 0,05 = 0,9$, звідки за таблицею II додатків підручника [1] $t_{0,9;15} = 1,75$. Оскільки $t = 1,62 < t_{0,9;15} = 1,75$, то гіпотеза H_0 приймається. Це означає, що вибірккові дані на 5%-вому рівні значущості не дозволяють вважати, що час доби суттєво впливає на завантаженість серверів. ►

4.4. Перевірка гіпотез про числові значення параметрів

Гіпотези про числові значення зустрічаються в області ІТ технологій найчастіше. В загальному випадку гіпотези такого типу мають вигляд $H_0: \theta = \Delta_0$, де θ — деякий параметр розподілу, який досліджується, а Δ_0 — область його конкретних значень, яка в частинному випадку складається із одного значення. Під час перевірки гіпотези вказаного типу можна використовувати той самий підхід, що і раніше. Відповідні критерії перевірки гіпотез про числові значення параметрів нормального закону надані в табл. 4.2.

Таблиця 4.2

Гіпотези про числові значення параметрів у відповідності до критеріїв перевірки

Нульова гіпотеза	Припущення	Статистика критерію	Альтернативна гіпотеза	Критерій відхилення гіпотези
$a = a_0$	σ^2 відома	$t = \frac{\bar{x} - a_0}{\sigma/\sqrt{n}}$	$a = a_1 > a_0$ $a = a_1 < a_0$ $a = a_1 \neq a_0$	$ t > t_{1-2\alpha}$ $ t > t_{1-\alpha}$
	σ^2 невідома	$t = \frac{\bar{x} - a_0}{s/\sqrt{n-1}}$	$a = a_1 > a_0$ $a = a_1 < a_0$ $a = a_1 \neq a_0$	$ t > t_{1-2\alpha; n-1}$ $ t > t_{1-\alpha; n-1}$
$\sigma_0^2 = \sigma^2$	a невідомо	$\chi^2 = \frac{ns^2}{\sigma_0^2}$	$\sigma^2 = \sigma_1^2 > \sigma_0^2$ $\sigma^2 = \sigma_1^2 < \sigma_0^2$ $\sigma^2 = \sigma_1^2 \neq \sigma_0^2$	$\chi^2 > \chi_{\alpha; n-1}^2$ $\chi^2 < \chi_{1-\alpha; n-1}^2$ або $\begin{cases} \chi^2 > \chi_{\alpha/2; n-1}^2 \\ \chi^2 < \chi_{1-\alpha/2; n-1}^2 \end{cases}$

$p=p_0$	n достатньо велике	$t = \frac{w - p_0}{\sqrt{p_0 q_0 / n}}$	$\left. \begin{array}{l} p = p_1 > p_0 \\ p = p_1 < p_0 \\ p = p_1 \neq p_0 \end{array} \right\}$	$ t > t_{1-2\alpha}$ $ t > t_{1-\alpha}$
---------	----------------------	--	---	---

Приклад 4.7. В ідеалі середній час відгуку сервера даного виробника складає $a_0 = 120$ мілісекунд. Вибіркова перевірка 10 серверів дала середнє значення відгуку $\bar{x} = 135$ мілісек., а середнє квадратичне відхилення відгуку $s = 20$ мілісек. На рівні значущості 0,05: а) визначити, чи можна прийняти за істину ідеальний відгук; б) знайти потужність критерію, використаного в п.а); в) знайти мінімальне число серверів, яке варто перевірити, щоб забезпечити потужність критерію 0,975.

Розв’язання. а) Гіпотеза, яка перевіряється, $H_0: \bar{x}_0 = a_0 = 120$.

Альтернативна - гіпотеза $H_1: \bar{x}_0 = a_1 = 135$. Оскільки генеральна дисперсія σ^2 невідома, то використаємо t -критерій Стьюдента. Статистика критерію у відповідності до табл. 6.2 дорівнює

$$t = \frac{\bar{x} - a_0}{s/\sqrt{n}} = \frac{135 - 120}{20/\sqrt{10-1}} = 2,25. \quad \text{Критичне значення статистики}$$

$t_{1-2\cdot 0,05; 10-1} = t_{0,9,9} = 1,83$. Оскільки $|t| > t_{0,9,9}$ ($2,25 > 1,83$), то гіпотеза H_0 відкидається, тобто на 5%-вому рівні значущості зроблений прогноз повинен бути відкинтий.

б) Оскільки $a_1 = 135 > a_0 = 120$, то критична область правостороння і критичне значення вибіркового середнього

$$\bar{x}_{\text{кр.}} = \bar{x}_0 + t_{1-2\alpha; n-1} \frac{s}{\sqrt{n-1}} = a + t_{0,9,9} \frac{s}{\sqrt{n-1}} = 120 + 1,83 \frac{20}{\sqrt{10-1}} = 132,2$$

(мілісекунд). Тобто критична область значень для x — інтервал $(132,2; +\infty)$.

Потужність критерію дорівнює

$$P = P(132,2 < \bar{x} < +\infty) = \frac{1}{2} - \frac{1}{2} \theta(t, n-1),$$

де $\theta(t, n-1)$ — функція, яка виражає ймовірність попадання випадкової величини, яка має t -розподіл Стюдента, на відрізок $(-t, t)$

$$t = \frac{\bar{x} - a_1}{s/\sqrt{n-1}} = t = \frac{132,2 - 135}{20/\sqrt{10-1}} = -0,42.$$

За таблицею III додатків підручника [1]:

$$\theta(-0,42; 9) = -\theta(0,42; 9) \approx -0,31.$$

Отже,
$$P = \frac{1}{2} - \frac{1}{2} \theta(-0,42; 9) = \frac{1}{2} (1 + 0,31) \approx 0,66.$$

в) Skorистаємось розв'язком прикладу 6.1 б). Оскільки у нас σ^2 невідома, а відома лише її вибіркова оцінка s^2 , то статистика критерію $t = \frac{\bar{x} - a_0}{s/\sqrt{n-1}}$

має t -розподіл Стюдента (див. табл. 4.2), і відповідна скорегована формула для n прийме вигляд:

$$n = \frac{(t_{1-2\alpha; n-1} + t_{1-2\beta; n-1})^2 s^2}{(a_1 - a_0)^2}.$$

При $\alpha = 0,05$, $\beta = 0,025$ (тому що за умовою потужність критерію $1 - \beta = 0,975$), $a_0 = 120$, $a_1 = 135$, $s = 20$ отримаємо:

$$n = \frac{16}{9} (t_{0,9; n-1} + t_{0,95; n-1})^2.$$

Оскільки права частина рівності сама залежить від невідомого значення n , то n знаходиться наближено підбором. Так, при $n = 20$ і при $n = 30$, рівність не виконується (наприклад, при $n=20$

$$20 \neq \frac{16}{9} (t_{0,9;19} + t_{0,95;19})^2 = \frac{16}{9} (1,73 + 2,09)^2 = 24,7), \quad \text{а при } n=25$$

$$25 \approx \frac{16}{9} (t_{0,9;24} + t_{0,95;24})^2 = \frac{16}{9} (1,71 + 2,06)^2 = 25,3.$$

Таким чином, необхідно перевірити 25 серверів. ►

Аналогічно перевіряються і інші гіпотези про числові значення параметрів у відповідності до критеріїв перевірки, наведених у табл. 6.2.

4.5. Побудова теоретичного закону розподілу за дослідними даними.

Перевірка гіпотез про закон розподілу

Однією з найважливіших задач математичної статистики є **встановлення теоретичного закону розподілу випадкової величини**, яка характеризує ознаку, яка вивчається, за дослідним (емпіричним) розподілом, представленим варіаційним рядом. Для розв'язання цієї задачі необхідно визначити вид та параметри закону розподілу. Припущення про **вид закону розподілу** може бути висунуте виходячи з теоретичних міркувань.

Параметри розподілу, як правило, невідомі, тому їх заміняють найкращими оцінками за вибіркою.

Критерії узгодження: нехай необхідно перевірити нульову гіпотезу H_0 про те, що досліджувана випадкова величина X підпорядковується певному закону розподілу. Для перевірки гіпотези H_0 обирають деяку випадкову величину U , яка характеризує ступінь розходження теоретичного та емпіричного розподілів. Закон розподілу U при достатньо великих n відомий та практично не залежить від закону розподілу випадкової величини X .

Якщо відомий закон розподілу U , то можна знайти ймовірність того, що U прийняла значення не менше, ніж u , яке фактично спостерігається у досліді, тобто $U \geq u$. Якщо $P(U \geq u) = \alpha$ мала, то це означає (у відповідності з принципом практичної впевненості), що такі, як в досліді, та більші відхилення практично неможливі. В цьому випадку гіпотезу H_0 відкидають. Якщо ж ймовірність $P(U \geq u) = \alpha$ не мала, розходження між емпіричним та теоретичним розподілами не істотне та гіпотезу H_0 можна вважати правдоподібною чи такою, що не суперечить дослідним даним.

χ^2 -критерій Пірсона

В критерії χ^2 -Пірсона, який найчастіше використовується на практиці, в якості міри розходження U береться величина χ^2 , яка дорівнює сумі квадратів відхилень частостей (статистичних ймовірностей) w_i від гіпотетичних p_i , розрахованих за передбачуваним розподілом, взятих з деякими вагами c_i :

$$U = \chi^2 = \sum_{i=1}^m c_i (w_i - p_i)^2.$$

Ваги c_i вводяться таким чином, щоб при одних і тих самих відхиленнях $(w_i - p_i)^2$ більшу вагу мали відхилення, при яких p_i мала, та меншу вагу – при яких p_i велика. Цього вдається досягнути, якщо взяти c_i обернено пропорційним ймовірностям p_i . Якщо взяти в якості ваг $c_i = \frac{n}{p_i}$, можна

довести, що при $n \rightarrow \infty$ статистика $U = \chi^2 = \sum_{i=1}^m \frac{n}{p_i} (w_i - p_i)^2$, або

$$U = \chi^2 = \sum_{i=1}^m \frac{(n_i - np_i)^2}{np_i} \text{ має } \chi^2\text{-розподіл з } k = m - r - 1 \text{ ступенями}$$

свободи, де m – число інтервалів емпіричного розподілу (варіаційного ряду); r – число параметрів теоретичного розподілу, обчислених за експериментальними даними.

Числа $n_i = nw_i$ та np_i називають відповідно **емпіричними** та **теоретичними частотами**.

Схема застосування критерію χ^2 для перевірки гіпотези H_0 :

1. Визначається міра розходження емпіричних та теоретичних частот χ^2

за формулою $U = \chi^2 = \sum_{i=1}^m \frac{(n_i - np_i)^2}{np_i}$.

2. Для обраного рівня значущості α за таблицею χ^2 -розподілу знаходять критичне значення $\chi_{\alpha;k}^2$ при числі ступенів свободи $k = m - r - 1$.
3. Якщо значення χ^2 , яке фактично спостерігається, більше критичного, $\chi^2 > \chi_{\alpha;k}^2$, то гіпотеза H_0 відкидається; якщо $\chi^2 \leq \chi_{\alpha;k}^2$, то гіпотеза H_0 не суперечить експериментальним даним.

Зауваження. Статистика $\chi^2 = \sum_{i=1}^m \frac{(n_i - np_i)^2}{np_i}$ має χ^2 -розподіл лише

при $n \rightarrow \infty$, тому необхідно, щоб в кожному інтервалі була достатня кількість спостережень, якнайменше 5 спостережень. Якщо в будь-якому інтервалі кількість спостережень менше за 5, доцільно об'єднати сусідні інтервали, щоб в об'єднаних інтервалах n_i було не менше 5.

Приклад 4.8. Тестується нова програма для розпізнавання прихованих зображень. Було протестовано 100 картинок із прихованим зображенням. За даними перших двох граф табл. 1.1 в прикладі 1.1 підібрати відповідний теоретичний розподіл та на рівні значущості $\alpha=0,05$ перевірити гіпотезу про узгодженість двох розподілів за допомогою критерію χ^2 .

Розв’язання. За видом гістограми розподілу часу роботи програми (рис. 1.2) можна передбачити нормальний закон розподілу ознаки. Параметри нормального закону μ та σ^2 , які є відповідно математичним сподіванням та дисперсією випадкової величини X , невідомі, тому заміняємо їх «найкращими» оцінками за вибіркою — незміщеними та ґрунтовними оцінками відповідно вибіркового середнім \bar{x} та «виправленою» вибірковою дисперсією \hat{s}^2 . Оскільки число спостережень $n=100$ достатньо велике, то замість «виправленої» \hat{s}^2 можна взяти «звичайну» вибірку дисперсію s^2 . У прикладі 1.1 обчислені $\bar{x}=119,2(xв), s^2=87,48, s=9,35(xв)$.

Таким чином, висунута гіпотеза H_0 : випадкова величина X — час дешифрування зображення — розподілена нормально з параметрами $\mu=119,2; \sigma^2=87,48$, тобто $X \rightarrow N(119,2; 87,48)$.

Для обчислення ймовірностей p_i попадання випадкової величини X в інтервал $[x_i, x_{i+1}]$ використовуємо функцію Лапласа у відповідності з властивістю нормального розподілу:

$$\begin{aligned}
p_i(x_i \leq X \leq x_{i+1}) &= \left[\Phi\left(\frac{x_{i+1} - a}{\sigma}\right) - \Phi\left(\frac{x_i - a}{\sigma}\right) \right] \approx \\
&\approx \left[\Phi\left(\frac{x_{i+1} - 119,2}{9,35}\right) - \Phi\left(\frac{94 - 119,2}{9,35}\right) \right] = \\
&= [\Phi(-2,05) - \Phi(-2,69)] = (-0,4798 + 0,4964) = 0,0166
\end{aligned}$$

Для визначення статистики χ^2 зручно скласти таблицю 6.3.

Враховуючи, що в емпіричному розподілі частоти першого та останнього інтервалів ($n_1 = 3, n_8 = 2$) менші за 5, при використанні критерію χ^2 -Пірсона у відповідності із зауваженням, доцільно об'єднати вказані інтервали із сусідніми (див. табл. 4.3). Таким чином, значення статистики, яке фактично спостерігається, $\chi^2 = 2,27$. Оскільки нова кількість інтервалів (враховуючи об'єднання крайніх) $m = 6$, а нормальний закон розподілу визначається $r = 2$ параметрами, то кількість ступенів свободи $k = m - r - 1 = 3$. Відповідне критичне значення статистики χ^2 за таблицею $\chi^2_{0,05;3} = 7,82$. Оскільки $\chi^2 < \chi^2_{0,05;3}$, то гіпотеза про обраний теоретично нормальний закон $N(119,2; 87,48)$ узгоджується з експериментальними даними.

Таблиця 4.3.

i	Інтервал $[x_i, x_{i+1}]$	Емпіричні частоти n_i	Ймовірності p_i	Теоретичні частоти np_i	$(n_i - np_i)^2$	$\frac{(n_i - np_i)^2}{np_i}$
1	94–100	3 } 10 7 }	0,017	1,7 } 7,6 5,9 }	5,76	0,758
2	100–106		0,059			
3	106–112	11	0,141	14,1	9,61	0,682
4	112–118	20	0,228	22,8	7,84	0,344

5	118–124	28	0,247	24,7	10,89	0,441
6	124–130	19	0,182	18,2	0,64	0,035
7	130–136	$\left. \begin{matrix} 10 \\ 2 \end{matrix} \right\} 12$	0,087	$\left. \begin{matrix} 8,7 \\ 2,9 \end{matrix} \right\} 11,6$	0,16	0,014
8	136–142		0,029			
	Σ	100	0,990	99,0	–	$\chi^2 = 2,27$

