Лекція 11. Перевірка правильності непараметричних гіпотез

Існують різні критерії, які не потребують значень параметрів генеральної сукупності. Вони називаються *непараметричними критеріями*.

Критерії узгодження, що найчастіше використовуються при перевірці гіпотез про вигляд розподілу:

- для перевірки гіпотези про нормальний розподіл генеральної сукупності:

– критерій згоди Романовського;

– критерій згоди Ястремського;

– критерій Шапіро-Уілка;

– критерій Пірсона $\left(\chi^2\right)$;

- для перевірки гіпотези про довільний розподіл генеральної сукупності:

– критерій серій;

– критерій Колмогорова;

– критерій Мізеса $\left(\omega^2\right)$;

- для перевірки гіпотези про однаковий розподіл двох генеральних сукупностей:

– критерій знаків;

– критерій Колмагорова-Смирнова;

– критерій інверсії (Вілкоксона).

11.1. Критерій згоди В. І. Романовського

За критерій (статистику) В. І. Романовський обрав величину:

$$Y_{Ром} = \frac{\chi^2 - k_{св}}{\sqrt{2k_{св}}}, \qquad (11.1)$$

де $k_{св}$ — число ступенів вільності.

Якщо $|Y_{Ром}| \le 3$, то несуттєвою є розбіжність між емпіричним і теоретичним розподілом і емпіричний розподіл можна вважати приблизно нормальним; якщо ж $|Y_{Ром}| > 3$, то нульова гіпотеза про близькість емпіричного і теоретичного розподілів відхиляється.

Приклад 11.1. Маємо згруповані дані про денний виторг у магазині електротоварів (тис. грн).

Сума продажу	Кількість одиниць продажу
190-200	10
200-210	26
210-220	56
220-230	64
230-240	30
240-250	14

Перевірити нульову гіпотезу H_0 про те, що сума продажу (X) є випадковою величиною, яка розподілена за нормальним законом. Рівень значущості α прийняти за 0,5.

Розв'язання:

$$\overline{x}_B = \frac{1}{n}\sum_{i=1}^{6} x_i^* \cdot n = \frac{1}{200} \cdot 44200 = 221.$$

Тоді $a = 221$, $D_B = \frac{1}{n}\sum_{i=1}^{6}\left(x_i^*\right)^2 \cdot n - \overline{x}_B{}^2 = \frac{30400}{200} = 152$, $\sigma_B = \sqrt{D_B} = \sqrt{152} \approx 12{,}33$.

Подальші обчислення для зручності занесемо у вигляді таблиці:

x_i^*	n_i	$x_i^* \cdot n_i$	$x_i^* - \overline{x}_B$	$\left(x_i^* - \overline{x}_B\right)^2 \cdot n_i$
195	10	1950	-26	6760
205	26	5330	-16	6656
215	56	12 040	-6	2016
225	64	14 404	16	1024
235	30	7050	14	5880
245	14	3430	24	8064
$\sum$	200	44 200	-	30 400

Тепер обчислимо теоретичні ймовірності p_i потрапляння випадкової величини $X \to N(221;152)$ у частинні інтервали $(x_i;\ \ x_{i+1})$ за формулою $p_i = \Phi(z_{i+1}) - \Phi(z_i)$, де $z_i = \dfrac{x_i - \overline{x}_B}{\sigma_B}$.

Зауважимо, що найменше значення $\dfrac{x_i - \overline{x}_B}{\sigma_B} = \dfrac{190-221}{12,33} = -2,514$ замінено на «-∞», а найбільше значення $\dfrac{x_i - \overline{x}_B}{\sigma_B} = \dfrac{250-221}{12,33} = -2,352$ замінено на «+∞».

Після цього обчислимо $\chi^2_{cn} = \sum_{i=1}^{6} \dfrac{(n_i - n'_i)^2}{n'_i}$

Розрахунки проводимо в таблиці:

Інтервали (S), $(x_i; x_{i+1})$	Частоти n	Нормовані інтервали $(z_i; z_{i+1})$, де $z_i = \dfrac{x_i - \overline{x}_B}{\sigma_B}$	$\Phi(z_i)$	$\Phi(z_{i+1})$	$p_i = \Phi(z_{i+1}) - \Phi(z_i)$
190-200	10	(-∞; -1,7)	-0,500	-0,4554	0,0446
200-210	26	(-1,7; -0,89)	-0,4554	-0,3133	0,1421
210-220	56	(-0,89; -0,08)	-0,3133	-0,0319	0,2814
220-230	64	(-0,08; 0,73)	-0,0319	0,2673	0,2992
230-240	30	(0,73; 1,54)	0,2673	0,4382	0,1709
240-250	14	(1,54; ∞)	0,4382	0,5	0,0618
$\sum$	$n = 200$	-	-	-	1

$n'_i = n \cdot p_i$	$(n_i - n'_i)$	$\dfrac{(n_i - n'_i)^2}{n'_i}$
8,92	1,08	0,1308
28,42	-2,42	0,206
56,28	-0,28	0,0014
59,84	4,16	0,2892
34,18	-4,18	0,5112
12,36	1,64	0,2176
$\sum$ 200		$\chi^2_{cn} = 1,3562$

Оскільки, $\chi^2_{сп} = 1{,}3562$, а число ступенів вільності $k_{св} = \nu = S - r - 1 = 6 - 2 - 1 = 3$, то за критерієм Романовського:

$$|Y_{Ром}| = \left|\frac{\chi^2 - k_{св}}{\sqrt{2k_{св}}}\right| = \left|\frac{1{,}3562 - 3}{\sqrt{6}}\right| = \left|\frac{-1{,}6438}{2{,}4495}\right| \approx 0{,}67 < 3,$$

тому немає підстав відхилити нульову гіпотезу. Отже, математичною моделлю заданого вибіркового розподілу можна вважати нормальний закон розподілу.

Зауваження. Відношення Романовського має підґрунтям те, що $M\left(\chi^2\right) = k_{св}$, а $D\left(\chi^2\right) = \sigma^2 = 2 \cdot k_{св}$. Тому ймовірність відхилення χ^2 на $\sqrt{2 \cdot k_{св}}$ близька до 1.

11.2. Критерій згоди Б. С. Ястремського

Як і критерій Романовського, критерій Ястремського:

$$Y_{ястр} = \frac{|C - k|}{\sqrt{2k + 4 \cdot \Theta}} \qquad (11.2)$$

застосовується без звернення до таблиць розподілу χ^2. У формулі (7.2) k — кількість груп, Θ — величина, яка залежить від k і C. Якщо $k < 20$ і $C = \sum_{i=1}^{k} \frac{(n_i - n_i')}{n_i' \cdot (1 - p_i)}$, тоді Θ=0,6.

Якщо $|Y_{ястр}| > 3$, то гіпотеза H_0 відхиляється; якщо ж $|Y_{ястр}| \le 3$, то H_0 приймається.

Приклад 11.2. Маємо згруповані дані про денний виторг у магазині електротоварів (тис. грн).

Сума продажу	Кількість одиниць продажу
190-200	10
200-210	26
210-220	56
220-230	64

230-240	30
240-250	14

Перевірити нульову гіпотезу H_0 про те, що сума продажу (X) є випадковою величиною, яка розподілена за нормальним законом. Рівень значущості α прийняти за 0,5.

Розв'язання: Для цього розрахунки виконуємо в наступній таблиці, де p_i — знайдені під час розв'язання прикладу 7.3.

p_i	$1-p_i$	$n_i' = n \cdot p_i$	$\frac{(n_i - n_i')^2}{n_i'}$	$\frac{(n_i - n_i')^2}{n_i' \cdot (1-p_i)}$
0,0446	0,9554	8,92	0,1308	0,1369
0,1421	0,8579	28,42	0,206	0,2401
0,2814	0,7186	56,28	0,0014	0,0019
0,2992	0,7008	59,84	0,2892	0,4127
0,1709	0,8291	34,18	0,5112	0,6166
0,0618	0,9382	12,36	0,2176	0,2319
$\sum$		-	-	$C = 1,64$

Оскільки кількість груп $k = 6$, та $k = 6 < 20$, то Θ=0,6.

Тоді $Y_{ястр} = \frac{|C-k|}{\sqrt{2k+4\Theta}} \Rightarrow |Y_{ястр}| = \left|\frac{1,64-6}{\sqrt{12+4\cdot 0,6}}\right| = \left|\frac{-4,36}{3,7947}\right| = 1,143 < 3.$

Отже, немає підстав відхилити нульову гіпотезу про нормальний закон розподілу.

11.3. Критерій згоди Шапіро-Уілка (статистичний аналіз даних вимірювань)

Критерій узгодження Шапіро-Уілка використовують для перевірки гіпотези про нормальний розподіл генеральної сукупності. Він базується на відношенні оптимальної лінійної незсунутої оцінки дисперсії до її звичайної оцінки методом найбільшої правдоподібності.

Статистика критерію має вигляд

$$W = \frac{1}{S^2}\left(\sum_{i=1}^{n} a_{n-i+1} \cdot (x_{n-i+1} - x_i)\right)^2$$

Коефіцієнти a та критичні значення $W_{кр}$ беруться з таблиць. Критерій надійний при $8 \le n \le 50$ (існує модифікований критерій Шапіро-Франчича, який можна застосовувати при n до 2000). Критерій є найбільш ефективним, оскільки він має найбільшу потужність порівняно з іншими критеріями перевірки на нормальність.