Лекція 14. Перевірка правильності непараметричних гіпотез про належність двох вибірок до однієї генеральної сукупності

14.1. Критерій серій

Серією називається частинна послідовність елементів одного виду, що входить в упорядковану послідовність елементів двох видів.

Нехай є $n_1 + n_2$ елементів, серед яких n_1 елементів виду «*а*» та n_2 елементів виду «*в*».

Порівнюючи кожне значення з вибірковою медіаною, можемо розбити всі значення на два типи (їх позначимо «*а*» та «*в*»), використовуючи те, що є значення більші або менші від медіани.

Перевіримо гіпотезу H_0: досліджувана вибірка є випадковою.

Спочатку спостережувані значення вибірки розташуємо в зростаючому порядку:

$$x_1 \le x_2 \le x_3 \le \ldots \le x_{n-1} \le x_n .$$

Потім визначимо медіану Me і кожному елементу $x_i \left(i = \overline{1,n}\right)$ припишемо символ «*а*», якщо $x_i < Me$, і символ «*в*», якщо $x_i > Me$.

Одержимо послідовність із літер «*а*» та «*в*». Позначимо через випадкову величину k — кількість серій отриманої послідовності. Доведено, що якщо нульова гіпотеза справедлива, то для досить великих $n\left(n > 10\right)$ розподіл кількості серій k є нормальним, $N\left(M\left(k\right);\sigma^2\left(k\right)\right)$ де:

$$M\left(k\right) = \frac{2n_1n_2}{n_1 + n_2};\ \sigma\left(k\right) = \frac{2n_1n_2}{\left(n_1 + n_2\right)^{3/2}}.$$

Тоді нормована випадкова величина:

$$Z = \frac{k - \dfrac{2n_1n_2}{n_1 + n_2}}{\dfrac{2n_1n_2}{\left(n_1 + n_2\right)^{3/2}}} \to N\left(0;1^2\right)$$

Якщо ж із двох генеральних сукупностей з довільними функціями розподілу $F_1(x)$ і $F_2(x)$ витягнуті вибірки обсягами n_1 і n_2, то за допомогою критерію серій можна перевірити гіпотезу:

$$H_0 : F_1(x) = F_2(x),$$

тобто дві генеральні сукупності мають ту саму функцію розподілу.

Отже, маємо таку схему застосування критерію серій:

• формулюється нуль—гіпотеза; а) H_0: вибірка випадкова; б) H_0: дві вибірки вибрані з однієї й тієї самої генеральної сукупності $\left(F_1(x) = F_2(x)\right)$.

• Альтернативна гіпотеза H_1, як правило, і у випадку а) і у випадку б) не формулюється при застосуванні критерію серій.

• Рівень значущості α для критерію, як правило, вибирають $\alpha = 0,05$, або $\alpha = 0,01$.

• Критерій (критеріальна статистика):

$$Z_{стат} = \frac{k - \frac{2n_1 n_2}{n_1 + n_2}}{\frac{2n_1 n_2}{(n_1 + n_2)^{3/2}}},$$

де k - кількість серій.

• Критичні точки залежать від α. Оскільки альтернативні гіпотези не формулюються, то застосовується z-критерій з двобічною критичною областю. Маємо межі $\pm z_{\frac{\alpha}{2}}$, які відокремлюють критичні області від області прийняття гіпотези H_0. Якщо $\alpha = 0,05$, то критичні точки дорівнюють ± 1,96; коли $\alpha = 0,01$, то — ± 2,575. Для інших значень α критичні точки беруть за додатком 2.

Гіпотеза H_0 приймається, якщо:

а) $|z_{сп}| < z_{\frac{\alpha}{2}}$;

б) $|z_{сп}| < z_{\frac{\alpha}{2}}$.

Приклад 14.1. Лікар рекомендував своїм пацієнтам, які мають зайву вагу, ліки «*а*» та «*в*». При цьому щоразу фіксував початкову вагу пацієнта в кг. У результаті він дістав таблицю:

а	66,5	83,0	67,8	75,6	81,6	98,0	57,6	100,7	59,7	73,3	100,3	92,1
в	81,4	73,1	71,0	70,1	66,3	59,4	73,8	72,2	73,5	102,1	71,8	-

Перевірити за допомогою критерію серій для рівнів значущості $\alpha = 0,05$ та $\alpha = 0,1$ гіпотезу про випадковість у призначенні ліків, тобто про вплив ліків «а» та «в» на зміну ваги пацієнтів.

Розв'язання:

За заданою вибіркою маємо, що обсяг вибірки за критерієм «*а*» дорівнює 12 $(n_1 = 12)$, а за критерієм «*в*» - 11 $(n_2 = 11)$.

Крок 1. Сформулюємо гіпотезу H_0: досліджувана вибірка випадкова.

Крок 2. Утворимо спадну послідовність, у якій символом «*а*» будемо позначати значення з першої вибірки, символом «*в*» — значення із другої вибірки:

№ п/п	1	2	3	4	5	6	7	8	9	10	11	12
x_i	57,6	59,4	59,7	66,3	66,5	67,8	70,1	71,0	71,8	72,2	73,1	73,5
а, в	*а*	*в*	*а*	*в*	*а*	*а*	*в*	*в*	*в*	*в*	*в*	*в*
	1	2	3	4	5		6					
№ п/п	13	14	15	16	17	18	19	20	21	22	23	
x_i	73,7	73,8	75,6	81,4	81,6	83,0	92,1	98,0	100,3	100,7	102,1	
а, в	*а*	*в*	*а*	*в*	*а*	*а*	*а*	*а*	*а*	*а*	*в*	
	7	8	9	10	11						**12**	

Загальна кількість серій у вибірці $k = 12$.

Крок 3. За формулою критерію обчислимо спостережуване його значення:

$$Z_{сп} = \frac{k - \frac{2n_1n_2}{n_1 + n_2}}{\frac{2n_1n_2}{(n_1 + n_2)^{3/2}}} = \frac{12 - \frac{2 \cdot 12 \cdot 11}{12 + 11}}{\frac{2 \cdot 12 \cdot 11}{(12 + 11)^{\frac{3}{2}}}} \approx 0,22.$$

Крок 4. Для $\alpha = 0,05$: згідно додатку 2:

$\Phi(x) = 0,5 \cdot \gamma = 0,5 \cdot (1-\alpha) = 0,5 \cdot (1-0,05) = 0,5 \cdot 0,95 = 0,475 \quad \Rightarrow \quad x = 1,96,$

отже $Z_{kp}(0,05) = 1,96$.

Для $\alpha = 0,1$: згідно додатку 2:

$\Phi(x) = 0,5 \cdot \gamma = 0,5 \cdot (1-\alpha) = 0,5 \cdot (1-0,1) = 0,5 \cdot 0,9 = 0,45 \quad \Rightarrow \quad x = 1,645$, отже

$Z_{kp}(0,1) = 1,645.$

Оскільки в обох випадках $|Z_{cn}| < Z_{kp}$, тому немає підстав відхиляти гіпотезу H_0 про випадковість вибірки, отже, вважаємо, що лікар призначав ліки «*а*» і «*в*» випадковим чином.

14.2. Критерій знаків

Передбачається, що випадкові величини X і Y, значення яких спостерігаються в i-тому випробуванні, незалежні одна від одної, а послідовні n спостережень незалежні між собою. Вибірки ранжирувані, тобто $x_1 < x_2 < \ldots < x_n$ і $y_1 < y_2 < \ldots < y_n$. Позначимо $x_i - y_i = r_i$. Чи можна вважати розбіжність між x_i і y_i значущою, істотною? Різниці $r_i = 0$ виключаємо.

Досліджуємо знаки різниць r_i і знайдемо кількість тих знаків, яких менше. Нехай їх виявилося r.

У випадку, коли нульова гіпотеза справедлива, різниці $x_i - y_i = r_i$ будуть симетрично розподілені відносно нуля, тобто знаки «+» та «–» рівноймовірні:

$$P("+") = P("-") = \frac{1}{2}, \text{ або } P(x - y > 0) = P(x - y < 0) = \frac{1}{2}.$$

Зрозуміло, що число r - дискретна випадкова величина, розподілена за біноміальним законом з параметрами n і $p = \frac{1}{2}$, тобто:

$$P_n(r) = C_n^r \left(\frac{1}{2}\right)^r \cdot \left(\frac{1}{2}\right)^{n-r} = C_n^r \left(\frac{1}{2}\right)^n = C_n^r \cdot 2^{-n}. \qquad (14.1)$$

Критична область будується правобічна, якщо $H_1 : \bar{X} > \bar{Y}$; лівобічна, якщо $H_1 : \bar{X} < \bar{Y}$; двобічна, якщо $H_1 : \bar{X} \neq \bar{Y}$.

Позначимо через r_α − найменше значення r, для якого $P_n(r) \le \alpha$. Тоді гіпотеза H_0 відхиляється, коли $r_{cn} \le r_\alpha$, а для $r_{cn} > r_\alpha$ немає підстав відхилити гіпотезу H_0.

Для перевірки гіпотез застосовується спеціальна таблиця критичних значень кількості знаків r, що відповідають заданому рівню α і обсягу вибірки n (додаток 10).

Величина r — розподілена за біноміальним законом. Виходить, що $M(r) = np = \frac{n}{2}$ і $D(r) = npq = \frac{n}{4}$.

Оскільки для $n \to \infty$ біноміальний розподіл згідно з теоремою Муавра—Лапласа наближається до нормального розподілу, то $r \to N\left(\frac{n}{2}; \frac{n}{4}\right)$.

Для невеликих значень n і r імовірність (14.1) легко обчислити безпосередньо за формулою, але для більших n $(n > 30)$ і r можна використовувати нормальний розподіл.

Приклад 14.2. У таблиці наведено результати обстеження 20 родин, що мають однаковий дохід. З'ясувалося, яку частину доходу кожна родина витрачає на транспорт і культурні потреби.

Необхідно для рівня значущості 0,05 перевірити гіпотезу про те, що в генеральній сукупності середні частки витрат на транспорт і культурні потреби однакові. Завдання виконати за допомогою критерію знаків.

№ з/п	Частка витрат на транспорт, %	Частка витрат на культурні потреби, %	№ з/п	Частка витрат на транспорт, %	Частка витрат на культурні потреби, %
1	5,3	6,8	11	4,9	8,3
2	5,1	6,7	12	4,7	10,5
3	10,9	4,1	13	5,6	8,1

4	4,7	8,3	14	8,3	2,1
5	4,3	5,9	15	6,6	7,2
6	5,7	6,1	16	7,3	12,1
7	5,4	12,0	17	4,2	6,1
8	12,6	10,9	18	5,0	7,8
9	6,2	13,1	19	4,9	10,5
10	3,1	6,8	20	8,3	5,9

Розв'язання:

Сформулюємо нульову гіпотезу H_0: $F(x) = F(y)$, тобто частки витрат кожної родини на транспорт і культурні потреби однакові.

Позначимо знаком «+» більшу частку витрат на культурні потреби, а знаком «–» більшу частку витрат на транспорт:

№ з/п	Частка витрат на транспорт, %	Частка витрат на культурні потреби, %		№ з/п	Частка витрат на транспорт, %	Частка витрат на культурні потреби, %	
1	5,3	6,8	+	11	4,9	8,3	+
2	5,1	6,7	+	12	4,7	10,5	+
3	10,9	4,1	-	13	5,6	8,1	+
4	4,7	8,3	+	14	8,3	2,1	-
5	4,3	5,9	+	15	6,6	7,2	+
6	5,7	6,1	+	16	7,3	12,1	+
7	5,4	12,0	+	17	4,2	6,1	+
8	12,6	10,9	-	18	5,0	7,8	+
9	6,2	13,1	+	19	4,9	10,5	+
10	3,1	6,8	+	20	8,3	5,9	-

Кількість знаків «–» дорівнює 4, тобто $r = 4$.

Знайдемо із дод. 10 критичних значень кількості знаків за заданим рівнем значущості $\alpha = 0,05$ та обсягу вибірки $n = 20$ критичне значення:

$r_{\alpha;n} = r_{0,05;\,20} = 5$.

Оскільки $r_{cn} = 4 < 5$, то нульова гіпотеза відхиляється на користь альтернативної, тобто це означає, що частка витрат кожної родини на культурні потреби перевищує частку витрат на транспортні послуги.

14.3. Критерій Колмогорова-Смирнова

Даний критерій застосовують для статистичної перевірки гіпотези про належність двох вибірок одній генеральній сукупності. Для цього застосовують формулу:

$$\lambda'_{спост} = \max\left|F_1^*(x) - F_2^*(x)\right|,$$

де $F_1^*(x)$ та $F_2^*(x)$ - емпіричні функції розподілу, побудовані за двома вибірками об'ємом n_1 та n_2.

Критичне значення обчислюється згідно формули:

$$\lambda'_{кр} = \lambda_{кр} \cdot \sqrt{\frac{n_1 + n_2}{n_1 \cdot n_2}},$$

де $\lambda_{кр}$ критична точка, що визначається відповідно до заданого рівня значущості (α) згідно дод. 9 або за формулою $\lambda_{кр} = \sqrt{-\ln\left(\frac{\alpha}{2}\right) \cdot \frac{1}{2}}$.

Гіпотеза H_0 відхиляється, якщо фактично спостережене значення статистики $\lambda'_{спост}$ більше від критичного $\lambda'_{кр}$, тобто $\lambda'_{спост} > \lambda'_{кр}$, інакше – приймається.

Приклад 14.3. Протягом місяця вибірково перевіряли овочі торгові точки міста. Результати двох перевірок з недоважування покупцям по одному виду овочів представлені у таблиці:

Номер інтервалу, i	Інтервали недоважування, $x_i - x_{i+1}$	Частоти	
		Для вибірки 1, n_1	Для вибірки 2, n_2
1	0-10	3	5
2	10-20	10	12
3	20-30	15	8
4	30-40	20	25
5	40-50	12	10

6	50-60	5	8
7	60-70	25	20
8	70-80	15	7
9	80-90	5	5
Об’єм вибірки		$n_1 = 110$	$n_2 = 100$

Розв’язання:

Підрахуємо накопичені частоти обох вибірок та значення їх емпіричних функцій розподілу. Результати занесемо до розрахункової таблиці:

x_i^*	$n_1^{нак}$	$n_2^{нак}$	$F_1^*(x_i)$	$F_2^*(x_i)$	$\left\|F_1^*(x_i) - F_2^*(x_i)\right\|$
5	3	5	0,027	0,050	0,023
15	13	17	0,118	0,170	0,052
25	28	25	0,254	0,250	0,004
35	48	50	0,436	0,500	0,064
45	60	60	0,545	0,600	0,055
55	65	68	0,591	0,680	**0,089**
65	90	88	0,818	0,880	0,072
75	105	95	0,955	0,950	0,005
85	110	100	1,000	1,000	0,000

З останнього стовпчика таблиці знаходимо:

$$\lambda'_{спост} = \max\left|F_1^*(x) - F_2^*(x)\right| = 0,089,$$

де $F_1^*(x)$ та $F_2^*(x)$ - емпіричні функції розподілу, побудовані за двома вибірками об’ємом $n_1 = 110$ та $n_2 = 100$.

Критичне значення обчислюється згідно формули:

$$\lambda'_{кр} = \lambda_{кр} \cdot \sqrt{\frac{n_1 + n_2}{n_1 \cdot n_2}} = 1,358 \cdot \sqrt{\frac{110 + 100}{110 \cdot 100}} = 0,6223,$$

де $\lambda_{кр}=1{,}358$ критична точка, що визначається відповідно до заданого рівня значущості $(\alpha=0{,}05)$ згідно дод. 9 або за формулою $\lambda_{кр}=\sqrt{-\ln\left(\frac{\alpha}{2}\right)\cdot\frac{1}{2}}=\sqrt{-\ln\left(\frac{0{,}05}{2}\right)\cdot\frac{1}{2}}=1{,}3581$.

Оскільки $0{,}089=\lambda'_{спост}<\lambda'_{кр}=0{,}6223$, то нульова гіпотеза, про те що дві випадкові вибірки відносяться до одного розподілу – приймається.

14.4. Критерій Вілкоксона

Цей критерій служить для перевірки, чи відносяться дві вибірки до однієї й тієї самої генеральної сукупності, іншими словами, нульова гіпотеза H_0 запевнює, що $F_X(x)=F_Y(y)$.

Значення елементів обох вибірок $\{x\}_{n_1}$, $\{y\}_{n_2}$ розподіляють разом в одну загальну вибірку у порядку їх зростання значень $\{x_1,y_1,y_2,y_3,x_2,y_4,\ldots\}$. Пара значень $(x_i;y_j)$ утворюють *інверсію*, якщо $y_j<x_i$.

Як критерій використовується величина $U=U_{спост}$ - повне число інверсій.

Якщо гіпотеза вірна, значення $U_{спост}$ не повинне дуже відхилятися від свого математичного сподівання $M(U)=\frac{n_1+n_2}{2}$. Якщо величина розподілена за законом Вілкоксона, то гіпотезу відхиляємо у тому випадку, коли $U_{спост}>U_{кр}$. $U_{кр}$ можна взяти із таблиці 15 для заданого рівня значущості. Для більших об'ємів вибірки $n_1>25$ та $n_2>25$ критичне значення знаходимо згідно формули:

$$U_{кр}=z_\alpha\cdot\sqrt{\frac{n_1\cdot n_2\cdot(n_1+n_2+1)}{12}},$$

де $z_\alpha=\arg\Phi\left(\frac{1-\alpha}{2}\right)$ - значення аргумента функції Лапласа, $\Phi(z_\alpha)=\frac{1-\alpha}{2}$.

Наприклад 14.4. Планується провести педагогічний експеримент в групі з 7 студентів і порівняти його результат з показниками контрольної групи, яка складається з 6 студентів. На початку експерименту оцінки x_i студентів контрольної групи і оцінки y_j студентів експериментальної групи є наступними:

x_i	76	81	81	82	83	86	
y_j	74	75	76	78	81	82	85

Для об'єктивності дослідження на початку експерименту групи мають бути однорідними. Відтак, за рівнем значущості $\alpha = 0,05$ перевіримо гіпотезу про належність вибірок до однієї генеральної сукупності.

Розв'язання: Перетворимо задану таблиці до зручного вигляду для встановлення кількості інверсій:

Значення елементів обох вибірок $\{x\}_{n_1}$, $\{y\}_{n_2}$ розподіляють разом в одну загальну вибірку у порядку їх зростання значень:

74	75	76	76	78	81	81
y_1	y_2	x_1	y_3	y_4	x_2	x_3
81	82	82	83	85	86	
y_5	x_4	y_6	x_5	y_7	x_6	

Кількість інверсій – 7, отже $U_{спост} = 7.$ За визначенням, кількість інверсій не повинно дуже відхилятися від свого математичного сподівання $M(U) = \frac{n_1 + n_2}{2}$, тобто $M(U) = \frac{6+7}{2} = 6,5$. Згідно табл. 15 знайдемо критичне значення при $n_1 = 6$ та $n_2 = 7$ і заданого рівня значущості: $U_{кр} = 30$. Оскільки $U_{спост} < U_{кр}$, то нульову гіпотезу приймаємо.