Testy statystyczne służące do wykrywania wartości odstających

Wartość odstająca x_o (ang. *outlier*) – obserwacja odległa od pozostałych elementów próby i posiadająca nietypową (zbyt małą lub zbyt dużą) wartość liczbową. Obserwacje odstające mogą być rezultatem przypadku, świadczyć o błędnym pomiarze i/lub pomyłkach przy wprowadzaniu informacji do bazy danych. Zazwyczaj $x_o = x_{min}$ i/lub $x_o = x_{max}$.

Test Q Dixon'a – opiera się na definicji rozstępu: $R = x_{max} - x_{min}$. Przed wykonaniem testu dane należy uporządkować niemalejąco: $x_1, x_2, x_3, \dots, x_{N-1}, x_N$. Test ten polega na obliczeniu różnicy między **wartością wątpliwą** x_w, a najbliższym wynikiem w serii pomiarowej i podzieleniu tej różnicy przez rozstęp.

a) $x_w = x_{min}$

$$Q = \frac{x_2 - x_{min}}{R} = \frac{x_2 - x_1}{R}$$

b) $x_w = x_{max}$

$$Q = \frac{x_{max} - x_{N-1}}{R} = \frac{x_N - x_{N-1}}{R}$$

Następnie, wyznaczony parametr Q jest porównywany z wartością krytyczną Q_K testu Dixon'a, którą należy odczytać z tabeli. Uwaga! Wartość krytyczna Q_K zależy zarówno od liczebności serii pomiarów jak i założonego poziomu ufności. Jeżeli $Q \geq Q_K$ to wynik wątpliwy należy uznać za wynik odstający ($x_w \rightarrow x_o$) i odrzucić. Gdy $Q < Q_K$ to nie ma podstaw do odrzucenia wyniku wątpliwego.

Test Grubbs'a – przed wykonaniem testu dla serii pomiarowej należy wyznaczyć średnią $\langle x \rangle$ oraz odchylenie standardowe Δs (metodą $N-1$) a dane uporządkować niemalejąco: $x_1, x_2, x_3, \dots, x_{N-1}, x_N$.

a) $x_w = x_{min}$

$$G = \frac{\langle x \rangle - x_{min}}{\Delta s}$$

b) $x_w = x_{max}$

$$G = \frac{x_{max} - \langle x \rangle}{\Delta s}$$

Następnie, wyznaczony parametr G jest porównywany z wartością krytyczną G_K testu Grubbs'a, którą należy odczytać z tabeli. Uwaga! Wartość krytyczna G_K zależy zarówno od liczebności serii pomiarów jak i założonego poziomu ufności. Jeżeli $G \geq G_K$ to wynik wątpliwy x_w należy uznać za wynik odstający ($x_w \rightarrow x_o$) i odrzucić. Gdy $G < G_K$ to nie ma podstaw do odrzucenia wyniku wątpliwego.

Metoda odchylenia standardowego – przed wykonaniem testu dla serii pomiarowej należy wyznaczyć średnią $\langle x \rangle$, średnią $\langle x \rangle_w$ bez uwzględnienia wyniku wątpliwego oraz odchylenie standardowe Δs_w (metodą $N-1$) bez uwzględnienia wyniku wątpliwego. Jeżeli zachodzi nierówność: $|\langle x \rangle - \langle x \rangle_w| \geq \Delta s_w$ to wynik wątpliwy należy odrzucić, tzn. $x_w \rightarrow x_o$. Uwaga! Metoda odchylenia standardowego jest wiarygodna gdy liczebność serii pomiarowej spełnia nierówność: $N \geq 30$.

Liczba pomiarów	Wartość krytyczna Q_k		Wartość krytyczna G_k		Liczba pomiarów	Wartość krytyczna Q_k		Wartość krytyczna G_k	
(n)	95 %	99 %	95 %	99 %	(n)	95 %	99 %	95 %	99 %
3	0,970	0,994	1,153	1,155	22	0,468	0,544	0,438	2,939
4	0,829	0,926	1,463	1,492	23	0,459	0,535	2,624	2,963
5	0,710	0,821	1,672	1,749	24	0,451	0,626	2,644	2,987
6	0,628	0,740	1,822	1,944	25	0,443	0,517	2,663	3,009
7	0,569	0,680	1,938	2,097	26	0,436	0,510	2,681	3,029
8	0,608	0,717	2,032	2,221	27	0,429	0,502	2,698	3,049
9	0,564	0,672	2,110	2,323	28	0,423	0,495	2,714	3,068
10	0,530	0,635	2,176	2,410	29	0,417	0,489	2,730	3,085
11	0,502	0,605	2,234	2,485	30	0,412	0,483	2,745	3,103
12	0,479	0,579	2,285	2,550	31	0,407	0,477	2,759	3,11 9
13	0,611	0,697	2,331	2,607	32	0,402	0,472	2,773	3,135
14	0,586	0,670	2,371	2,659	33	0,397	0,467	2,786	3,150
15	0,565	0,647	2,409	2,705	34	0,393	0,462	2,799	3,164
16	0,546	0,627	2,443	2,747	35	0,388	0,458	2,811	3,178
17	0,529	0,610	2,475	2,785	36	0,384	0,454	2,823	3,191
18	0,514	0,594	2,504	2,821	37	0,381	0,450	2,835	3,204
19	0,501	0,580	2,532	2,854	38	0,377	0,446	2,846	3,216
20	0,489	0,567	2,557	2,884	39	0,374	0,442	2,857	3,228
21	0,478	0,555	0,442	2,912	40	0,371	0,438	2,866	3,240

Tabela. Wartości krytyczne dla testu Q Dixon'a (Q_K) i testu Grubbs'a (G_K) przy zadanym poziomie ufności.

Zadanie 1. Korzystając z testu Q Dixon'a proszę sprawdzić czy w następującej serii pomiarowej (7, 4, 5, 4, 4, 5, 5, 2) znajduje się wartość odstająca.

Tworzymy szereg statystyczny uporządkowany niemalejąco: (2, 4, 4, 4, 5, 5, 5, 7). Zakładamy, że $x_w = x_{max} = 7$. Następnie wyznaczamy rozstęp R i parametr Q Dixon'a

$$Q = \frac{x_{max} - x_{N-1}}{R} = \frac{7-5}{5} = \frac{2}{5} = 0{,}4$$

Wartość krytyczna Q_K na poziomie ufności 99% dla $N = 8$ wynosi $Q_K = 0{,}717$. Zachodzi nierówność: $Q < Q_K$ – nie ma więc podstaw do odrzucenia wyniku wątpliwego ($x_w \neq x_o$).

Zadanie 2. Korzystając z testu Grubbsa proszę sprawdzić czy w serii pomiarowej (3, 2, 8, 3) znajduje się wartość odstająca.

Tworzymy szereg statystyczny uporządkowany niemalejąco oraz zakładamy, że $x_w = 8$. Następnie wyznaczamy $\langle x \rangle$, odchylenie standardowe Δs metodą $N - 1$ oraz parametr Grubbs'a

$$G = \frac{x_{max} - \langle x \rangle}{\Delta s} = \frac{8 - 4}{\sqrt{\frac{22}{3}}} = 1{,}477$$

Wartość krytyczna G_K na poziomie ufności 95% dla $N = 4$ wynosi $G_K = 1{,}463$. Zachodzi nierówność: $Q > Q_K$ – wynik wątpliwy należy odrzucić na poziomie ufności 95% ($x_w \rightarrow x_o$). Wartość krytyczna G_K na poziomie ufności 99% dla $N = 4$ wynosi $G_K = 1{,}492$. Zachodzi nierówność: $Q < Q_K$ – nie ma więc podstaw do odrzucenia wyniku wątpliwego na poziomie ufności 99% ($x_w \neq x_o$).

Kowariancja

W celu uproszczenia obliczeń rachunkowych przyjmijmy, że i-ta wartość (jednowymiarowej) zmiennej losowej X występuje w analizowanej zbiorowości statystycznej tylko raz, tzn. $p_i = N^{-1}$, gdzie N to liczebność zbiorowości statystycznej. Zgodnie z definicją wariancji:

$$(\Delta x)^2 = \frac{1}{N}\sum_{i=1}^{N}[x_i - \langle x\rangle]^2 = \frac{1}{N}\sum_{i=1}^{N}[x_i - \langle x\rangle][x_i - \langle x\rangle]$$

Dla zmiennej losowej dwuwymiarowej wprowadźmy definicję **kowariancji** C_{xy}

$$C_{xy} = \frac{1}{N}\sum_{i=1}^{N}[x_i - \langle x\rangle][y_i - \langle y\rangle]$$

Uwaga! Wariancja stanowi szczególny przypadek kowariancji gdy $y_i = x_i$.

$$C_{xy} = \frac{1}{N}\sum_{i=1}^{N}[x_i - \langle x\rangle][y_i - \langle y\rangle] = \frac{1}{N}\sum_{i=1}^{N}[x_i y_i - \langle y\rangle x_i - \langle x\rangle y_i + \langle x\rangle\langle y\rangle]$$

$$\downarrow$$

$$C_{xy} = \frac{1}{N}\sum_{i=1}^{N} x_i\, y_i - \frac{1}{N}\sum_{i=1}^{N}\langle y\rangle x_i - \frac{1}{N}\sum_{i=1}^{N}\langle x\rangle y_i + \frac{1}{N}\sum_{i=1}^{N}\langle x\rangle\langle y\rangle \cdot 1$$

$$\downarrow$$

$$C_{xy} = \frac{1}{N}\sum_{i=1}^{N} x_i\, y_i - \frac{\langle y\rangle}{N}\sum_{i=1}^{N} x_i - \frac{\langle x\rangle}{N}\sum_{i=1}^{N} y_i + \frac{\langle x\rangle\langle y\rangle}{N}\sum_{i=1}^{N} 1$$

$$\downarrow$$

$$C_{xy} = \frac{1}{N}\sum_{i=1}^{N} x_i\, y_i - \langle y\rangle\langle x\rangle - \langle x\rangle\langle y\rangle + \frac{\langle x\rangle\langle y\rangle N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i\, y_i - \langle y\rangle\langle x\rangle - \langle x\rangle\langle y\rangle + \langle x\rangle\langle y\rangle$$

$$\downarrow$$

$$C_{xy} = \frac{1}{N}\sum_{i=1}^{N} x_i\, y_i - \langle y\rangle\langle x\rangle$$

Sumę iloczynów $x_i y_i$, która pojawia się w definicji kowariancji nazwijmy **średnią mieszaną**

$$\langle xy\rangle = \frac{1}{N}\sum_{i=1}^{N} x_i\, y_i$$

Ostatecznie: $C_{xy} = \langle xy\rangle - \langle x\rangle\langle y\rangle$. Widzimy, że kowariancja C_{xy} stanowi różnicę między średnią mieszaną a iloczynem średnich arytmetycznych zmiennych X i Y. Kowariancja może być dodatnia, ujemna lub równa zero. Uwaga! Kowariancja C_{xy} jest **podwójnie mianowana**, tzn. jest wyrażona zarówno w jednostkach X jak i Y. Aby pozbyć podwójnego miana możemy podzielić C_{xy} przez iloczyn odchyleń standardowych $\Delta x \Delta y$ zmiennych X i Y. Uwaga! Podobny efekt daje dzielenie C_{xy} przez iloczyn średnich $\langle x\rangle\langle y\rangle$ zmiennych X i Y - w tym przypadku ograniczamy się jednak do zbiorowości statystycznych, dla których: $\langle x\rangle \neq 0$ i $\langle y\rangle \neq 0$. Iloraz kowariancji i iloczynu $\Delta x \Delta y$ nazywamy **współczynnikiem korelacji liniowej Pearsona** r, tzn.

$$r = \frac{C_{xy}}{\Delta x \Delta y} = \frac{\langle xy\rangle - \langle x\rangle\langle y\rangle}{\Delta x \Delta y}$$

Uwaga! Współczynnik korelacji liniowej Pearsona r jest znormalizowaną kowariancją, a zatem wielkością niemianowaną. Gdy $\Delta x \neq 0$ i $\Delta y \neq 0$ iloczyn $\Delta x \Delta y > 0$, a więc znak współczynnika korelacji liniowej Pearsona r $(-1 \leq r \leq +1)$ jest taki sam jak znak kowariancji C_{xy}. Dla $r = 0$ między zmiennymi X i Y nie występuje zależność liniowa, dla $r = \pm 1$ zależność między zmiennymi X i Y jest zupełna, tzn. $Y = f(X)$. Współczynnik korelacji liniowej Pearsona r jest parametrem, który mierzy jedynie liniową zależność między zmiennymi X i Y.

W związku z tym, może się zdarzyć że $r = 0$, a między zmiennymi występuje silna zależność nieliniowa.

Zadanie 3. Dla zmiennych losowych $X = \{2, 4, 6, 8\}$ oraz $Y = \{1, 3, 5, 7\}$ proszę wyznaczyć kowariancję C_{xy} oraz współczynnik korelacji liniowej Pearsona r.

Liczba par zmiennych: $N = 4$.

$$\langle x \rangle = 5$$

$$\langle y \rangle = 4$$

$$\langle xy \rangle = \frac{1}{N}\sum_{i=1}^{N} x_i\, y_i = \frac{(2 \cdot 1) + (4 \cdot 3) + (6 \cdot 5) + (8 \cdot 7)}{4} = 25$$

$$C_{xy} = \langle xy \rangle - \langle x \rangle \langle y \rangle = 25 - 20 = 5$$

$$\langle x^2 \rangle = 30$$

$$\langle y^2 \rangle = 21$$

$$(\Delta x)^2 = \langle x^2 \rangle - \langle x \rangle^2 = 30 - 25 = 5 \;\rightarrow\; \Delta x = \sqrt{5}$$

$$(\Delta y)^2 = \langle y^2 \rangle - \langle y \rangle^2 = 21 - 16 = 5 \;\rightarrow\; \Delta y = \sqrt{5}$$

$$r = \frac{C_{xy}}{\Delta x \Delta y} = \frac{5}{\sqrt{5} \cdot \sqrt{5}} = 1$$

Między zmiennymi X oraz Y występuje liniowa zależność funkcyjna, ponieważ $r = 1$.

Zadanie 4. Dla zmiennych losowych $X = \{2, 4, 6, 8\}$ oraz $Y = \{7, 5, 3, 1\}$ proszę wyznaczyć kowariancję C_{xy} oraz współczynnik korelacji liniowej Pearsona r.