ROCZNIKI POLSKIEGO TOWARZYSTWA MATEMATYCZNEGO
SERIA III: MATEMATYKA STOSOWANA III (1974)

R. BARTOSZYŃSKI (Warszawa)

J. Neyman, E. Scott: O odrzucaniu elementów odstających*

W praktyce statystycznej stosunkowo często spotykamy się z sytuacją, gdy jeden z wyników jest na tyle większy (mniejszy) od pozostałych, że zachodzi podejrzenie, iż jest on rezultatem omyłki (przeoczenia eksperymentatora, przesunięcia przecinka przy zestawianiu danych, itp.). W omawianej pracy Neyman i Scott proponują następującą procedurę rozstrzygania w jakich przypadkach podejrzenia takie można uznać za uzasadnione.

Niech $y_1, \ldots, y_n$ $(n \geqslant 3)$ będzie próbą prostą z populacji o rozkładzie F, i niech x_k oznacza k-ty co do wielkości spośród $y_1, \ldots, y_n$. Mamy zatem $x_1 \leqslant x_2 \leqslant \ldots \leqslant x_n$.

Dla oceny w jakim stopniu x_n jest elementem „odstającym" od pozostałych, rozważmy iloraz[1]

$$w = \frac{x_n - x_{n-1}}{x_{n-1} - x_1}$$

($w = \infty$ jeżeli $x_{n-1} = x_n$ i $x_n - x_{n-1} > 0$).

Intuicyjnie, im większe w, tym bardziej wydaje się to świadczyć, że element maksymalny x_n jest „obcy" w próbce.

Powiemy, że próbka zawiera element (k, n)-*odstający* (ang. (k, n)-outlier), jeżeli $w \geqslant k$, czyli

$$x_n - x_{n-1} \geqslant k(x_{n-1} - x_1) \tag{1}$$

(ściślej biorąc, należałoby tu mówić o „elementach (k, n)-odstających z prawej strony". Ponieważ określenie elementów (k, n)-odstających ze strony przeciwnej jest identyczne i wszystkie rozważania przenoszą się bez zmian, w dalszym ciągu będzie mowa jedynie o „odstawaniu" elementów największych).

Niech $p(k, n, F)$ oznacza prawdopodobieństwo, że n-elementowa próba prosta z rozkładu F zawiera element (k, n)-odstający. Dla wyznaczenia $p(k, n, F)$ zauważmy, że (1) jest równoważne nierówności

$$x_{n-1} \leqslant \frac{x_n + kx_1}{k+1}.$$

---

*J. Neyman and E. Scott, *Outlier proneness of phenomena and of related distributions;* w książce *Optimizing Methods in Statistics*, New York 1971.

[1] Wszystkie znane metody opierają się na porównaniu odległości x_n od zbioru $x_1, \ldots, x_{n-1}$ (określonej na przykład jako $x_n - x_{n-1}$, $x_n - \bar{x}$, itp.) z jakąś oceną odchylenia standardowego (np. s, $x_n - x_1$, itp.). Por. np. R. Zieliński, *Tablice statystyczne*, Warszawa 1972, str. 61–64.

Tak więc, pod warunkiem $x_1 = x$, $x_n = y$ $(x \leqslant y)$, próba będzie zawierać element (k, n)-odstający, jeżeli $x_2, x_3, \ldots, x_{n-1}$ znajdą się w przedziale między x i $(y + kx)/(k + 1)$. Zakładając dla uproszczenia, że F jest rozkładem typu ciągłego, prawdopodobieństwo ostatniego zdarzenia wynosi

$$\left[F\left(\frac{y + kx}{k + 1}\right) - F(x)\right]^{n-2},$$

skąd całkując otrzymujemy

$$p(k, n, F) = \int_{-\infty}^{+\infty} \int_{x}^{+\infty} \left[F\left(\frac{y + kx}{k + 1}\right) - F(x)\right]^{n-2} dF(y)\, dF(x).$$

W praktyce rozkład F jest zwykle nieznany; wiadomo natomiast na ogół, że jest on jednym z rozkładów pewnej rodziny $\mathcal{F}$ rozkładów prawdopodobieństwa.

Oznaczmy

$$\pi(k, n, \mathcal{F}) = \sup_{F \in \mathcal{F}} p(k, n, F)$$

i wprowadźmy następujące definicje:

Rodzina $\mathcal{F}$ jest *odporna* lub *nieodporna* na (k, n)-odstawanie (ang. outlier resistant oraz outlier prone), w zależności od tego, czy $\pi(k, n, \mathcal{F}) < 1$ czy $\pi(k, n, \mathcal{F}) = 1$. Jeżeli $\pi(k, n, \mathcal{F}) = 1$ dla wszystkich $k > 0$ oraz $n \geqslant 3$, to rodzinę $\mathcal{F}$ nazwiemy *całkowicie nieodporną na odstawanie* (completely outlier prone).

Neyman i Scott dowodzą następujących twierdzeń:

Niech F będzie dowolnym rozkładem ciągłym, i niech

$$\mathcal{F}_1 = \{F_m : F_m(x) \equiv F(x - m)\},$$

$$\mathcal{F}_2 = \{F_\sigma : F_\sigma(x) \equiv F(x/\sigma)\}.$$

Wówczas rodziny $\mathcal{F}_1$ i $\mathcal{F}_2$ są odporne na (k, n)-odstawanie przy dowolnych $k > 0$ i $n \geqslant 3$. Wynika stąd w szczególności, że rodzina rozkładów normalnych jest odporna na (k, n)-odstawanie przy każdym $k > 0$ i $n \geqslant 3$.

Najbardziej zaskakujące są jednak twierdzenia orzekające, że rodzina wszystkich rozkładów gamma, oraz rodzina wszystkich rozkładów logarytmo-normalnych są całkowicie nieodporne na odstawanie.

Wniosek praktyczny z dwóch ostatnich twierdzeń jest taki, że jeżeli o badanym zjawisku wiemy, że rządzone jest przez jakiś rozkład gamma (lub log-normalny), to nawet dla najbardziej „dziwnie" wyglądających wyników, z $x_n - x_{n-1}$ dowolnie wiele razy przekraczającym $x_{n-1} - x_1$, nie mamy podstaw do odrzucenia elementu x_n jako obarczonego błędem (ponieważ istnieje zawsze rozkład gamma (log-normalny), przy którym takie, lub jeszcze bardziej „dziwne" konfiguracje mają prawdopodobieństwo pojawienia się dowolnie bliskie 1).