

Beata Zmyślona

BADANIE WPŁYWU ZABURZEŃ NA OBCIĄŻENIE ESTYMATORA PARAMETRU POŁOŻENIA ORAZ PARAMETRU SKALI W POPULACJI

1. Wstęp

W klasycznym podejściu do estymacji przyjmuje się, że obserwacje są realizacjami zmiennych losowych o pewnym rozkładzie prawdopodobieństwa indeksowanym parametrem θ , którego dystrybuantę oznacza się przez F_θ . Rodzinę rozkładów prawdopodobieństwa $\{F_\theta\}$ nazywa się modelem parametrycznym. Badając własności estymatora, rozpatruje się jego rozkład przy określonym modelu parametrycznym (por. np. [1; 4; 5]).

W teorii odpornych metod estymacji przyjmuje się, że model parametryczny jest jedynie pewną aproksymacją rzeczywistości. W procesie estymacji należy uwzględnić odstępstwa od tego modelu wynikające np. z istnienia niejednorodnych populacji, błędów powstałych na etapie pozyskiwania danych czy też z występowania obserwacji nietypowych. Idea odpornych metod estymacji polega na znajdowaniu takich estymatorów, które byłyby w bardzo małym stopniu podatne na wpływ nieznaczących odstępstw od założonego modelu. W artykule pokazano na przykładach, jaki jest wpływ odstępstw od modelu parametrycznego na obciążenie estymatorów oraz wspomniano o wpływie zaburzeń na ocenę wariancji estymatorów.

2. Rozkład zaburzony

W teorii odpornych metod estymacji uwzględnia się tzw. obserwacje zaburzające (nazywane w skrócie zaburzeniami) określane jako obserwacje pochodzące z innego rozkładu niż zasadnicza część obserwacji, czyli rdzeń. Istnienie w zbiorze danych obserwacji zaburzających wskazuje na niejednorodność badanego zjawiska

(niejednorodność populacji) lub występowanie pewnych błędów. Uwzględnianie obserwacji zaburzających wymaga wyspecyfikowania tzw. rozkładu zaburzonego. Rozkład zaburzony jest przedstawiony jako następująca mieszanka rozkładów (por. [1]):

$$H(x) = (1 - \lambda)F(x) + \lambda G(x). \quad (1)$$

Przez $\lambda \in (0, 1)$ określa się wagę związaną z zaburzeniem (lub inaczej udział obserwacji zaburzających). Jako $H(x)$, $F(x)$ oraz $G(x)$ oznacza się dystrybuantę odpowiednio rozkładu zaburzonego, rozkładu rdzenia oraz rozkładu zaburzającego rozkład rdzenia.

W poniższych przykładach za rozkłady zaburzające są przyjmowane albo rozkłady normalne, albo rozkłady skupione w jednym punkcie, inaczej rozkłady zdegenerowane, dla których $P(X = x) = 1$. Jeżeli za rozkład zaburzający przyjmuje się rozkład zdegenerowany, wówczas rozkład zaburzony może być przedstawiony w następujący sposób [1]:

$$H(x) = (1 - \lambda)F(x) + \lambda \Delta_x, \quad (2)$$

gdzie Δ_x jest miarą prawdopodobieństwa, która przypisuje masę 1 w punkcie x , natomiast λ jest udziałem zaburzenia punktowego. Przyjmuje się, że obserwacje z n -elementowej próby są realizacjami zmiennej losowej pochodzącej z rozkładu (1) lub jako (2), co można zapisać jako

$$X_j \sim H(x) \text{ dla } j = 1, 2, \dots, n.$$

Rozkład estymatora rozpatrywany jest w otoczeniu następującego modelu:

$$F = \{H(x) = (1 - \lambda)F(x) + \lambda G(x)\}.$$

Do badania odporności estymatorów na pewne odstępstwa od modelu parametrycznego wykorzystuje się rozmaite narzędzia, m.in. funkcję wpływu (*the influence function*) oraz funkcję zmiany wariancji (*the change-of-variance function*). Funkcja wpływu wykorzystywana jest do określania, jak zmieni się asymptotyczna wartość estymatora w przypadku wystąpienia nieznaczących odchyłeń od przyjętego modelu, natomiast funkcja zmiany wariancji wskazuje, jak zmieni się asymptotyczna ocena wariancji estymatora przy pewnych zaburzeniach. W artykule ograniczono się do prezentacji funkcji wpływu, która dla estymatorów odpornych charakteryzuje się pewnymi własnościami (szerzej zob. w [2; 3; 4]).

3. Wpływ zaburzeń na obciążenie estymatora

Poniżej zostały przedstawione przykłady wpływu zaburzeń na obciążenie estymatorów oraz na ocenę wariancji estymatora.

Przykład 1

Przyjmuje się, że obserwacje z rdzenia pochodzą z rozkładu normalnego parametrami μ oraz σ . Rozkład rdzenia jest zaburzony rozkładem normalnym parametrami $\mu + a$ (dla $a \neq 0$) oraz σ . Przyjmuje się, że wartość parametru σ jest znana. Zatem rozkład zaburzony jest mieszkanką dwóch rozkładów normalnych. Funkcja gęstości tego rozkładu wyraża się następującym wzorem:

$$h(x) = (1 - \lambda) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} + \lambda \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu-a)^2}{2\sigma^2}} \quad (3)$$

Nieobciążonym estymatorem parametru położenia μ jest średnia arytmetyczna

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

Spełniona jest następująca równość $E(\bar{X}) = \mu$.

Liczba zaburzeń w n -elementowej próbie jest traktowana jako realizacja zmiennej losowej R o rozkładzie dwumianowym z parametrem λ równym udziałowi obserwacji zaburzających, czyli

$$R \sim B(n, \lambda). \quad (4)$$

Jeżeli w próbie występują pewne zaburzenia, wówczas średnia arytmetyczna wyraża się w następujący sposób

$$\bar{X} = \frac{1}{n} \left(\sum_{i=1}^R X_i + \sum_{i=R+1}^n X_i \right). \quad (5)$$

W celu ustalenia wpływu zaburzeń na obciążenie estymatora parametru położenia rozkładu normalnego należy wyznaczyć wartość oczekiwaną estymatora (5), która jest obliczana ze względu na zmienną R (czyli liczbę zaburzeń w próbie) w następujący sposób [1]:

$$\begin{aligned} E(\bar{X}) &= E_R \left(\frac{1}{n} \left(\sum_{i=1}^R X_i + \sum_{i=R+1}^n X_i \right) \right) = E_R \left(\frac{1}{n} (R(\mu + a) + (n - R)\mu) \right) = \\ &= E_R \left(\mu + \frac{R}{n} a \right) = \mu + \frac{a}{n} E_R(R). \end{aligned}$$

Pierwszy moment zmiennej losowej R wynosi $E(R) = n\lambda$, zatem wartość oczekiwana średniej arytmetycznej dla rozkładu (3) wynosi

$$E(\bar{X}) = \mu + \lambda a. \quad (6)$$

Obciążenie estymatora $b(\mu) = \lambda a$ zależy od udziału obserwacji zaburzających λ oraz od wartości stałej a . Na przykład, jeżeli rozkład zaburzony jest mieszanką dwóch rozkładów normalnych, tzn. rozkładu rdzenia $F: N(2, 1)$ oraz rozkładu zaburzającego $G: N(12, 1)$, to dla $\lambda = 0,01$ wartość oczekiwana średniej arytmetycznej wynosi 2,1, natomiast dla $\lambda = 0,1$ przyjmuje wartość 3.

Przykład 2

Drugi przykład pokazuje wpływ zaburzeń na obciążenie estymatora parametru σ^2 . Przyjmuje się, że rozkładem rdzenia jest rozkład normalny z parametrami μ oraz σ , który jest zaburzony rozkładem normalnym z parametrami μ oraz $\sqrt{b}\sigma$, gdzie $b > 1$. Funkcja gęstości rozkładu zaburzonego wyraża się następującym wzorem:

$$h(x) = (1 - \lambda) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} + \lambda \frac{1}{\sqrt{2\pi}\sigma\sqrt{b}} e^{-\frac{(x-\mu)^2}{2b\sigma^2}} \quad (7)$$

Nieobciążonym estymatorem parametru σ^2 jest wariancja z próby: $S^2 = \frac{1}{n-1} \left(\sum_{i=1}^n (X_i - \bar{X})^2 \right)$. Wiadomo, że $E(S^2) = \sigma^2$. Wariancja z próby w przypadku wystąpienia w próbce zaburzeń wyraża się następującym wzorem:

$$S^2 = \frac{1}{n-1} \left(\sum_{i=1}^R X_i^2 + \sum_{i=R+1}^n X_i^2 - nE(\bar{X}^2) \right).$$

Wartość oczekiwana estymatora wyznaczana ze względu na liczbę zaburzeń w próbce w przypadku rozkładu zaburzonego (7) wyraża się w następujący sposób [1]:

$$\begin{aligned} E_R(S^2) &= E_R \left(\frac{1}{n} \left((n-R)\sigma^2 + Rb\sigma^2 \right) \right) = \frac{\sigma^2}{n} E_R((n-R) + Rb) = \\ &= \sigma^2 + \frac{\sigma^2}{n} E_R(R(b-1)) = \sigma^2 + \sigma^2 \lambda(b-1). \end{aligned}$$

Obciążenie estymatora $b(\sigma^2) = \sigma^2 \lambda(b-1)$ zależy od udziału obserwacji zaburzających oraz od wartości stałej b . Jeżeli rozkład zaburzony jest mieszanką dwóch rozkładów normalnych, tzn. rozkładu rdzenia $F: N(2, 1)$ oraz rozkładu zaburzającego $G: N(2, \sqrt{10})$, to dla $\lambda = 0,01$ wartość oczekiwana wariancji z próby wynosi 1,09, natomiast dla $\lambda = 0,1$ przyjmuje wartość 1,9.

4. Funkcja wpływu

Narzędziem umożliwiającym ocenę wpływu zaburzeń na obciążenie estymatora jest funkcja wpływu. Ze względu na to, że rozkład estymatora rozpatrywany jest w otoczeniu pewnej dystrybuanty F , wygodnie jest przyjąć, że rolę estymatorów pełnią funkcjonały oznaczane przez $T(F)$. Funkcjonały określane są także mianem miar opisowych. Funkcjonały umożliwiają wyznaczenie estymatora bez konieczności specyfikacji rozkładu F . Przykładem funkcjonału jest wartość oczekiwana $E(F) = \int x dF(x)$, wariancja $V(F) = \int (x - \mu)^2 dF(x)$ oraz mediana $Me(F) = F^{-1}(0,5) \int x dF(x)$. Jeśli za dystrybuantę F zostanie podstawiona dystrybuanta empiryczna F_n , wtedy otrzymuje się średnią arytmetyczną $E(F_n) = \int x dF_n(x) = \sum_{i=1}^n \frac{X_i}{n} = \bar{X}$, wariancję z próby $V(F_n) = \int (x - \mu)^2 dF_n(x) = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n} = S^2$ oraz medianę z próby $Me(F_n) = F_n^{-1}(0,5)$. Funkcjonały indukują naturalny, nieparametryczny estymator $T(F_n)$ (por. [4; 5]).

Miarę opisową określa się jako funkcję rzeczywistą określoną na pewnej rodzinie dystrybuant. Do dziedziny miar opisowych należą wszystkie dystrybuanty F , dla których funkcjonal T jest zdefiniowany, oraz dystrybuanty empiryczne F_n .

Przyjmuje się, że dla $n \rightarrow \infty$ ciąg estymatorów jest zbieżny z miarą opisową, co można zapisać w następujący sposób [4]:

$$T(F) = \lim_{n \rightarrow \infty} T(F_n).$$

Ponadto zakłada się, że rozkład różnicy $T(F_n) - T(F)$ dla $n \rightarrow \infty$ jest zbieżny z rozkładem normalnym, co symbolicznie zapisuje się jako [4]

$$\sqrt{n}(T(F_n) - T(F)) \xrightarrow{n \rightarrow \infty} N\left(0, \sqrt{V_{T(F)}}\right). \quad (8)$$

Wariancja $V_{T(F)}$ jest asymptotyczną wariancją estymatora.

Pewne funkcjonały spełniają dodatkowo warunek zgodności Fishera, czyli

$$T(F_\theta) = \theta.$$

W teorii dotyczącej funkcji wpływu przyjmuje się, że rozkład zaburzający jest rozkładem zdegenerowanym, a zatem rozkład zaburzony jest postaci (2). Funkcja wpływu jest pochodną kierunkową funkcjonału T w punkcie F w kierunku Δ_x (warto przypomnieć, że Δ_x jest miarą prawdopodobieństwa, która przypisuje masę 1 w punkcie x). Funkcję wpływu wyraża się następującym wzorem [4]:

$$IF_{T(F)}(x) = \lim_{\lambda \rightarrow 0} \frac{T((1-\lambda)F + \lambda\Delta_x) - T(F)}{\lambda}. \quad (9)$$

Funkcja (9) mierzy efekt nieznaczących zaburzeń w punkcie x na wartość estymatora, czyli mierzy asymptotyczne obciążenie. Jeśli dla danego funkcjonału spełnione jest założenie (8) dotyczące asymptotycznej normalności, wtedy zachodzi następująca zależność [4]

$$\int IF_{T(F)} dF(x) = 0,$$

natomiast asymptotyczną wariancję estymatora można wyrazić za pomocą następującej całki [4]:

$$V_{T(F)} = \int IF_{T(F)}(x)^2 dF(x).$$

Dla funkcji wpływu wyznaczone są następujące charakterystyki: miara wrażliwości estymatora na tzw. duże błędy (*the gross-error sensitivity*), punkt odrzucenia (*the rejection point*) oraz miara wrażliwości na lokalne przesunięcia (*the local-shift sensitivity*).

Definicja 1. Miarą wrażliwości estymatora na duże błędy nazywamy supremum funkcji wpływu $IF_{T(F)}$. Miara wrażliwości wyraża się następującym wzorem [4]:

$$\gamma_{T(F)}^* = \sup_{x \in \mathcal{X}} |IF_{T(F)}(x)|, \quad (10)$$

gdzie $\mathcal{X}$ jest przestrzenią prób. Charakterystyka $\gamma_{T(F)}^*$ mierzy najgorszy wpływ na wartość estymatora, który może być spowodowany najmniejszym udziałem zaburzeń. Pożądaną własnością z punktu widzenia odporności jest przyjmowanie przez miarę $\gamma_{T(F)}^*$ skończonych wartości. W przypadku gdy $\gamma_{T(F)}^* < \infty$, estymator nazywany jest B -odpornym (*B-robust*), co oznacza, że niewielki udział zaburzeń w punkcie x ma ograniczony wpływ na wartość estymatora. Jeśli $\gamma_{T(F)}^* = \infty$, wtedy nawet niewielki udział zaburzeń ma nieograniczony wpływ na obciążenie estymatora i taki estymator nie jest estymatorem odpornym.

Definicja 2. Punkt odrzucenia ρ^* jest definiowany jako następujące infimum [4]:

$$\rho^* = \inf\{r > 0; IF_{T(F)} = 0, \text{ gdy } |x| > r\}. \quad (11)$$

Jeśli infimum (11) nie istnieje, przyjmuje się, że $\rho^* = \infty$. Punkt odrzucenia wyznacza taki obszar wartości x , dla których funkcja wpływu przyjmuje wartości zero ($IF_{T(F)}(x) = 0$), co oznacza, że obserwacje przyjmujące wartości większe co do

wartości bezwzględnej niż punkt odrzucenia nie mają wpływu na estymator. Wszystkie estymatory, dla których $\rho^* < \infty$, są B -odporne. Zatem z punktu widzenia odporności pożądaną cechą jest, aby punkt odrzucenia przyjmował skończone wartości.

Definicja 3. Miara wrażliwości na lokalne przesunięcia, określająca jaki jest wpływ na wartość estymatora nieznacznego przesunięcia obserwacji z punktu x_1 do sąsiedniego punktu x_2 , jest definiowana jako [4]:

$$\beta^* = \frac{\sup_{x_1 \neq x_2} |IF_{T(F)}(x_2) - IF_{T(F)}(x_1)|}{|x_2 - x_1|}. \quad (12)$$

Z trzech opisanych powyżej charakterystyk funkcji wpływu najistotniejsza jest miara wrażliwości estymatora na tzw. duże błędy.

5. Wyznaczanie funkcji wpływu dla wartości oczekiwanej

W przypadku estymacji parametru położenia jako funkcjonal przyjmuje się wartość oczekiwaną ($T(F) = E(F)$). Funkcję wpływu wyznacza się zgodnie ze wzorem (9) w następujący sposób (por. np. [1; 4]):

$$\begin{aligned} IF_{T(F)}(x) &= \lim_{n \rightarrow \infty} \frac{T((1-\lambda)F + \lambda\Delta_x) - T(F)}{\lambda} = \lim_{n \rightarrow \infty} \frac{(1-\lambda)E(F) + \lambda x - E(F)}{\lambda} = \\ &= \lim_{n \rightarrow \infty} \frac{E(F) + \lambda E(F) + \lambda x - E(F)}{\lambda} = x - E(F). \end{aligned} \quad (13)$$

Funkcja (13) jest liniową funkcją zaburzenia w punkcie x i ma następujące charakterystyki: miara wrażliwości estymatora na tzw. duże błędy oraz punkt odrzucenia przyjmują nieskończoną wartość ($\gamma_{E(F)}^* = \sup_{x \in \mathcal{X}} |IF_{E(F)}(x)| = \infty$; $\rho^* = \infty$), natomiast

miara wrażliwości na lokalne przesunięcia wynosi 1 ($\beta^* = 1$).

Poniżej zostaną przedstawione dwa przykłady, w których wyznaczona jest funkcja wpływu dla estymatorów parametru położenia.

Przykład 3

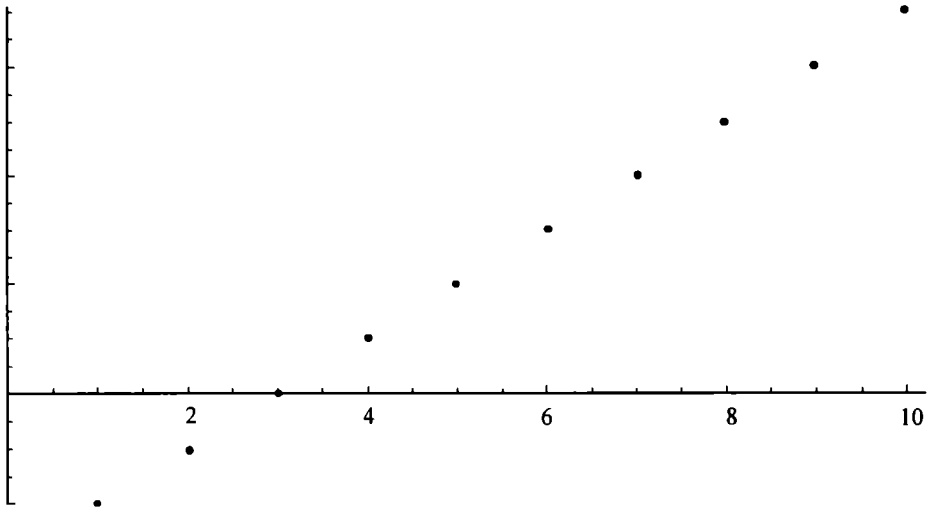
Przykład ten dotyczy wyznaczania funkcji wpływu dla estymatora parametru rozkładu Poissona. Przyjmuje się, że obserwacje w próbie są realizacjami zmiennej losowej o rozkładzie Poissona z parametrem θ , (czyli $X_1, X_2, \dots, X_n \sim P(\theta)$, przestrzeń parametrów $\Theta = (0, \infty)$, przestrzeń prób $X = \{0, 1, 2, \dots\}$). Funkcję

gęstości względem miary liczącej można zapisać jako $f_{\theta}(k) = \frac{\theta^k e^{-\theta}}{k!}$. Za estymator parametru θ przyjmuje się średnią arytmetyczną $\bar{X}$, która jako funkcjonal jest wartością oczekiwaną rozkładu Poissona, czyli $E(F_{\theta}) = \sum_{k=0}^{\infty} k f_{\theta}(k)$. Funkcjonał ten spełnia warunek zgodności Fishera: $E(F_{\theta}) = \sum_{k=0}^{\infty} k \frac{\theta^k e^{-\theta}}{k!} = \theta \sum_{k=0}^{\infty} \frac{\theta^{k-1} e^{-\theta}}{(k-1)!} = \theta$.

Funkcję wpływu, korzystając z warunku zgodności Fishera, wyznacza się w następujący sposób [4]:

$$\begin{aligned} IF_{E(F_{\theta})} &= \lim_{n \rightarrow \infty} \frac{\sum_{k=0}^{\infty} k \left[(1-\lambda) f_{\theta}(k) + \lambda 1_{\{x\}}(k) \right] - \sum_{k=0}^{\infty} k f_{\theta}(k)}{\lambda} = \\ &= \lim_{n \rightarrow \infty} \frac{\lambda \sum_{k=0}^{\infty} \lambda 1_{\{x\}}(k) - \lambda \sum_{k=0}^{\infty} k f_{\theta}(k)}{\lambda} = x - \theta. \end{aligned} \quad (14)$$

Funkcja wpływu (14) jest szczególnym przypadkiem funkcji (13). Zatem wszystkie charakterystyki funkcji (14) są takie same jak charakterystyki funkcji (13). Na rysunku 1 przedstawiono wykres funkcji wpływu przy założeniu, że parametr θ jest równy 3.



Rys. 1. Funkcja wpływu dla rozkładu Poissona z parametrem $\theta = 3$

W tym przypadku funkcja wpływu jest zbiorem punktów.

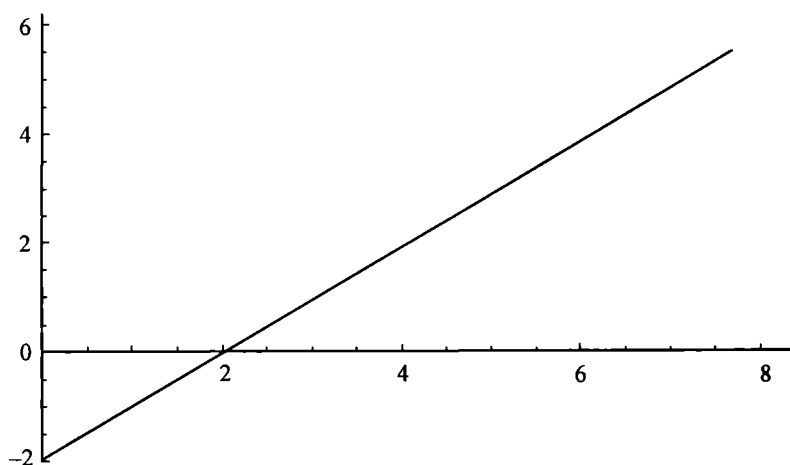
Przykład 2

W drugim przykładzie wyznaczono funkcję wpływu dla estymatora parametru położenia rozkładu normalnego. Przyjmuje się, że obserwacje w próbie są realizacjami zmiennej losowej o rozkładzie normalnym z parametrami μ oraz znaną wartością parametru σ^2 , (czyli $X_1, X_2, \dots, X_n \sim N(\mu, \sigma)$, przestrzeń parametrów $\Theta = (-\infty, \infty)$, przestrzeń prób $X = (-\infty, \infty)$). Funkcja gęstości względem miary

Lebesgue'a wyraża się następującym wzorem $f_\mu(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. Za estymator parametru μ przyjmuje się średnią arytmetyczną $\bar{X}$, która jako funkcjonal jest wartością oczekiwaną rozkładu normalnego, czyli $E(F_\mu) = \int x f_\mu(x) dx$. Dla rozkładu normalnego wartość oczekiwana spełnia warunek zgodności Fishera: $E(F_\mu) = \mu$.

Funkcję wpływu, korzystając z warunku zgodności Fishera, wyznacza się w analogiczny sposób:

$$\begin{aligned} IF_{E(F_\mu)}(x) &= \lim_{n \rightarrow \infty} \frac{(1-\lambda)E(F_\mu) + \lambda x - E(F_\mu)}{\lambda} = \\ &= \lim_{n \rightarrow \infty} \frac{\mu + \lambda\mu + \lambda x - \mu}{\lambda} = x - \mu. \end{aligned} \quad (15)$$



Rys. 2. Funkcja wpływu dla wartości oczekiwanej dla rozkładu normalnego z parametrami $\mu = 2$; $\sigma^2 = 1$

Źródło: opracowanie własne.

Funkcja (15) jest szczególnym przypadkiem funkcji (13) i ma takie same charakterystyki jak ta funkcja. Na rysunku 2 przedstawiono wykres funkcji wpływu dla rozkładu normalnego z parametrami $\mu = 2$ oraz $\sigma^2 = 1$.

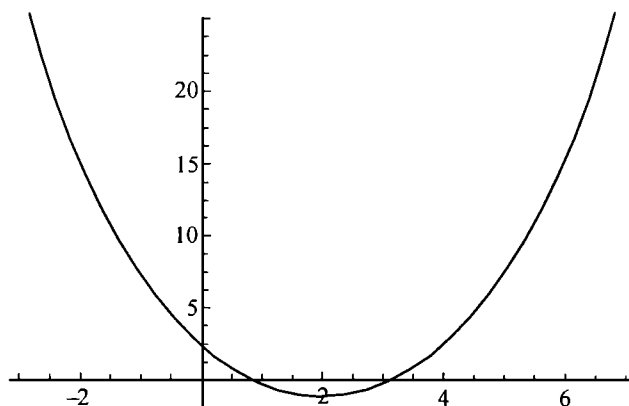
Z powyższych przykładów wynika, że średnia arytmetyczna nie jest odpornym estymatorem parametru położenia bez względu na przyjęty rozkład rdzenia F . Nieskończona wartość punktu odrzucenia wskazuje, że wszystkie obserwacje wykorzystywane są do wyznaczania wartości średniej arytmetycznej.

6. Wyznaczanie funkcji wpływu dla wariancji

Za estymator parametru skali przyjmuje się wariancję, która w postaci funkcjonalu wyraża się jako $T(F) = V(F)$. Funkcja wpływu dla wariancji wyraża się następującym wzorem:

$$\begin{aligned} IF_{V(F)} &= \lim_{\lambda \rightarrow 0} \frac{\left((1-\lambda)(E(F)^2 + V(F)) + \lambda x^2 \right) - ((1-\lambda)E(F) + \lambda x)^2 - V(F)}{\lambda} = \\ &= \lim_{\lambda \rightarrow 0} \frac{E(F)^2 - \lambda E(F)^2 - \lambda V(F) - (1-\lambda)^2 E(F)^2 - 2(1-\lambda)E(F)x - \lambda^2 x^2}{\lambda} = (16) \\ &= E(F)^2 - V(F) + x^2 - 2E(F)x = (x - E(F))^2 - V(F). \end{aligned}$$

Funkcja (16) jest kwadratową funkcją zaburzenia w punkcie x . Miara wrażliwości estymatora na duże błędy jest nieskończona ($\gamma_{V(F)}^* = \sup_{x \in X} |IF_{V(F)}| = \infty$), czyli wariancja nie jest odpornym estymatorem parametru skali. Punkt odrzucenia



Rys. 3. Funkcja wpływu dla wariancji dla rozkładu normalnego z parametrami $\mu = 2$; $\sigma^2 = 1$

Źródło: opracowanie własne.

przyjmuje również nieskończoną wartość ($\rho^* = \infty$), co jest równoznaczne z tym, że wszystkie obserwacje z próby są wykorzystywane do wyznaczania estymatora parametru skali.

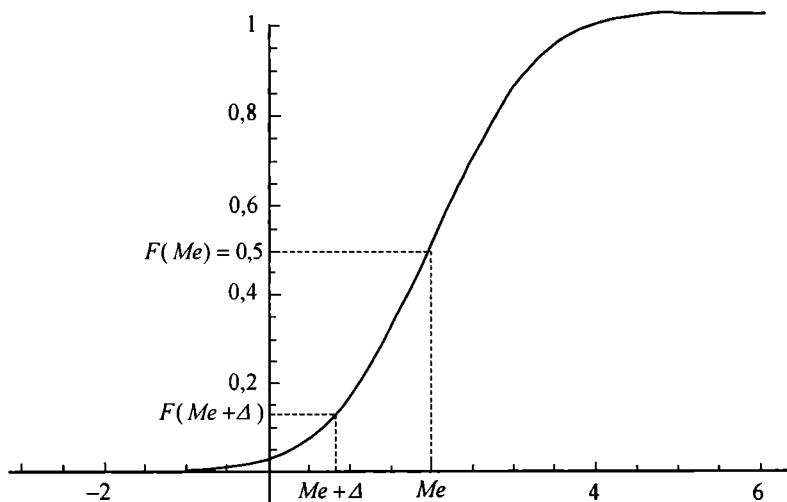
Na rysunku 3 przedstawiono wykres funkcji wpływu dla estymatora parametru skali dla rozkładu normalnego z parametrami $\mu = 2$ oraz $\sigma^2 = 1$.

7. Wyznaczanie funkcji wpływu dla mediany

Za estymator parametru położenia można przyjąć medianę, którą jako funkcjonal wyraża się w następujący sposób: $Me(F) = F^{-1}(0,5)$. Mediana dla rozkładu zaburzonego równa jest medianie dla rozkładu rdzenia powiększonej o pewną wartość Δ , czyli $Me(H) = Me((1 - \lambda)F + \lambda G) = Me + \Delta$. Wyznaczenie funkcji wpływu dla mediany sprowadza się do policzenia następującej granicy (por. np. [4]):

$$IF_{Me(F)} = \lim_{\lambda \rightarrow 0} \frac{Me + \Delta - Me}{\lambda} = \lim_{\lambda \rightarrow 0} \frac{\Delta}{\lambda}. \quad (17)$$

Wartość dystrybuanty rozkładu zaburzonego H w punkcie $Me + \Delta$, w przypadku gdy $\Delta < 0$, znajduje się pomiędzy wartością 1 oraz wartością dystrybuanty



Rys. 4. Dystrybuenta przykładowego rozkładu rdzenia dla $\Delta < 0$

Źródło: opracowanie własne.

rozkładu rdzenia w punkcie $Me + \Delta$, co zostało zobrazowane na rys. 4. Zatem wartość $H(Me + \Delta)$ można przedstawić jako następującą kombinację liniową:

$$H(Me + \Delta) = (1 - \lambda)F(Me + \Delta) + \lambda, \quad (18)$$

a zatem spełniona jest następująca równość:

$$0,5 = (1 - \lambda)(0,5 + \Delta f(Me)) + \lambda, \quad (19)$$

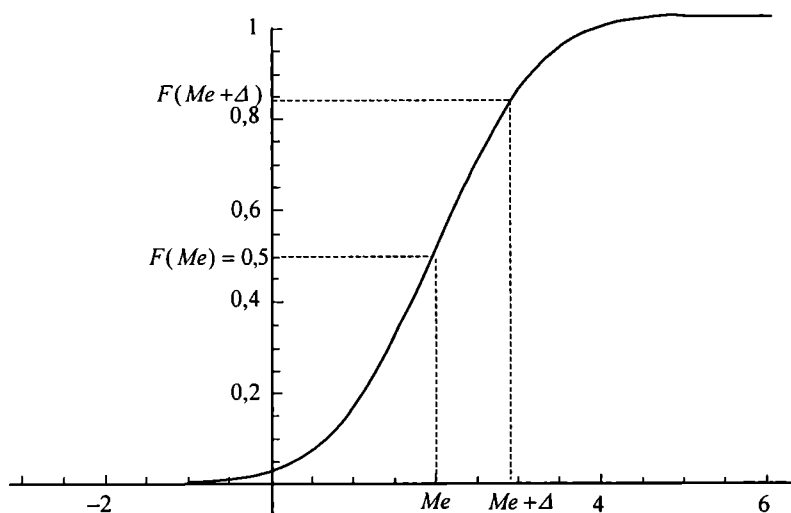
gdzie $f(Me)$ jest wartością funkcji gęstości rozkładu rdzenia w punkcie $x = Me$.

Z równości (19) można wyznaczyć wartość przyrostu $\Delta = -\frac{\lambda}{2f(Me)}$. Lewostronna granica funkcji (17) wynosi:

$$IF_{Me(F)} = \lim_{\lambda \rightarrow 0^-} \frac{\Delta}{\lambda} = \lim_{\lambda \rightarrow 0^-} \frac{-\lambda}{2f(Me)\lambda} = -\frac{1}{2f(Me)}. \quad (20)$$

W przypadku przyrostu dodatniego $\Delta > 0$ wartość dystrybucyjnego rozkładu zaburzonego H w punkcie $Me + \Delta$ znajduje się pomiędzy wartością 0 oraz wartością dystrybucyjnego rozkładu rdzenia w punkcie $Me + \Delta$, co zostało pokazane na rys. 5. Wartość mediany dla rozkładu zaburzonego $H(Me + \Delta)$ można przedstawić jako następującą kombinację liniową:

$$H(Me + \Delta) = (1 - \lambda)F(Me + \Delta) + 0. \quad (21)$$



Rys. 5. Dystrybucyjny przykładowy rozkład rdzenia dla $\Delta > 0$

Źródło: opracowanie własne.

W tym przypadku spełniona jest następująca równość:

$$0,5 = (1 - \lambda)(0,5 + \Delta f(Me)), \quad (22)$$

z której można wyznaczyć wartość przyrostu

$$\Delta = -\frac{\lambda}{(1-\lambda)2f(Me)}. \quad (23)$$

Prawostronna granica funkcji (17) wynosi:

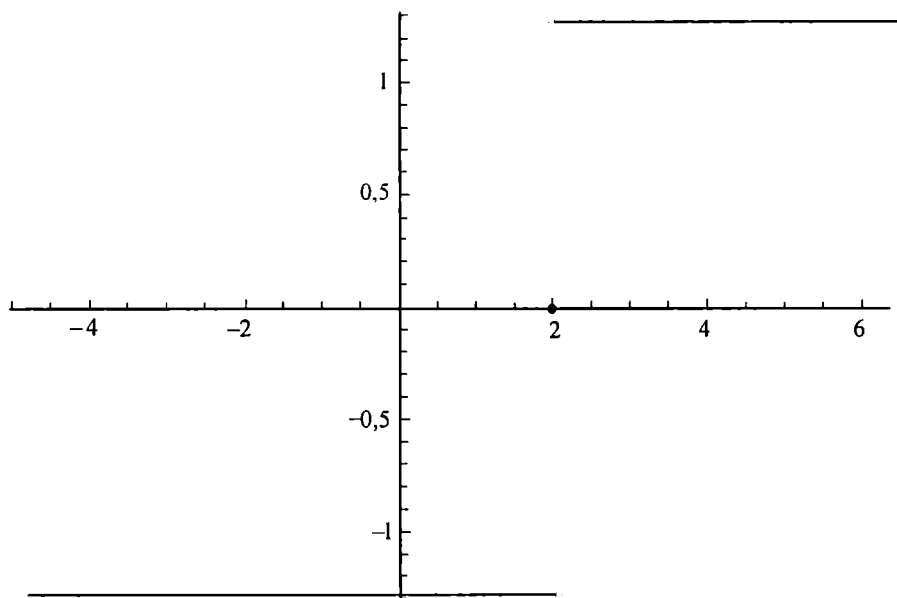
$$IF_{Me(F)} = \lim_{\lambda \rightarrow 0^+} \frac{\Delta}{\lambda} = \lim_{\lambda \rightarrow 0^+} \frac{\lambda}{(1-\lambda)2f(Me)\lambda} = \frac{1}{2f(Me)}. \quad (24)$$

Po uwzględnieniu lewo- i prawostronnej granicy (wyrażonej wzorami (20) oraz (24)) funkcja wpływu dla mediany przyjmuje następującą postać:

$$IF_{Me(F)} = \frac{\text{sgn}(x - Me)}{2f(Me)}. \quad (25)$$

Miara wrażliwości na duże błędy dla funkcji (25) przyjmuje skończoną wartość ($\gamma_{Me(F)}^* < \infty$), a zatem mediana jest B -odpornym estymatorem parametru położenia. Punkt odrzucenia jest nieskończony $\rho^* = \infty$, co wskazuje na to, że do wyznaczania estymatora wykorzystuje się wszystkie obserwacje z próby.

Na rysunku 6 przedstawiono wykres funkcji wpływu dla rozkładu normalnego z parametrami $\mu = 2$ oraz $\sigma^2 = 1$.



Rys. 6. Funkcja wpływu dla mediany dla rozkładu normalnego z parametrami $\mu = 2$; $\sigma^2 = 1$

Źródło: opracowanie własne.

Miara wrażliwości estymatora na duże błędy wynosi w tym przypadku $\gamma_{Me(F)}^* = \frac{\sqrt{2\pi}}{2} \approx 1,25$. Można udowodnić, że jest to najmniejsza z możliwych wartości miary wrażliwości (por. np. [4]).

Jak zostało już wspomniane na wstępie, zaburzenia mają nie tylko wpływ na obciążenie estymatora, ale także na ocenę jego wariancji. Pokazane to zostanie na przykładzie. Przyjmuje się, że funkcja gęstości rozkładu zaburzonego wyrażona wzorem (7) jest mieszanką dwóch rozkładów normalnych. Rozkładem rdzenia jest rozkład normalny z parametrami μ oraz σ^2 , a rozkładem zaburzającym jest rozkład normalny z parametrami μ oraz $b\sigma^2$, gdzie $b > 1$. Za estymator parametru μ przyjmuje się średnią arytmetyczną. Wariancja średniej arytmetycznej jest równa $\frac{\sigma^2}{n}$. W celu określenia wpływu zaburzeń na wariancję estymatora wyznacza się wariancję średniej w następujący sposób [1]:

$$\begin{aligned} V(\bar{X}) &= \frac{1}{n} E_R \left(\frac{1}{n} (n-R)\sigma^2 + Rb\sigma^2 \right) = \frac{\sigma^2}{n^2} (n - n\lambda + bn\lambda) = \frac{\sigma^2}{n} (1 - \lambda + b\lambda) = \\ &= \frac{\sigma^2}{n} (1 + (b-1)\lambda). \end{aligned}$$

Wariancja estymatora zwiększa się o czynnik $(1 + (b-1)\lambda)$, który jest zależny od udziału zaburzeń oraz wartości stałej b . Do określania stopnia wpływu zaburzeń na ocenę wariancji estymatora wykorzystywana jest funkcja zmiany wariancji (por. [2; 3; 4]).

8. Podsumowanie

Celem odpornych metod estymacji jest znajdowanie takich estymatorów, które będą mało wrażliwe na zmiany wartości estymatora oraz zmiany oceny jego wariancji w przypadku wystąpienia nieznacznych zaburzeń od przyjętego modelu. Poszukiwane są estymatory o korzystnych własnościach funkcji wpływu oraz funkcji wariancji. Dodatkową zaletą estymatora z punktu widzenia odporności jest przyjmowanie przez punkt odrzucenia skończonych wartości. Tę zaletę mają estymatory typu *redescending* (por. [3; 5]).

Literatura

- [1] Barnett V., Lewis T., *Outliers in Statistical Data*, J. Wiley & Sons, 1995.
- [2] Genton M., Rousseeuw P., *The Change-of-Variance Function of M-estimators of Scale under General Contamination*, „Journal of Computational and Applied Mathematics” 1995 vol. 64, s. 69-80.

- [3] Hampel F., Rousseeuw P., Ronchetti E., *The Change of Variance Curve and Optimal Redescending M-estimators*, JASA, vol. 76, 1981, s. 643-648.
- [4] Hampel F., Ronchetti E., Rousseeuw P., Stahel W., *Robust Statistics. The Approach Based on Influence Function*, J. Wiley & Sons, 1986.
- [5] *Statystyczne metody analizy danych*, red. W. Ostasiewicz, AE, Wrocław 1998.

THE OUTLIERS INFLUENCE ON ESTIMATE BIAS OF LOCATION AND SCALE PARAMETERS IN POPULATION RESEARCH

Summary

Outliers have an influence on the estimator bias and estimator variance. We have two tools to study this influence, namely the influence function and the change-of-variance function. The influence function has three characteristics: the gross-error sensitivity, the rejection point and the local-shift sensitivity.

In this paper we present some examples of influence function calculation and its characteristics for some estimates: the arithmetic mean, the sample variance and the median.