Wydział Matematyki, Fizyki i Informatyki
Uniwersytet Marii Curie-Skłodowskiej w Lublinie

Luiza Pańczyk

Wybór optymalnych L–statystyk w estymacji kwantyli

Rozprawa doktorska napisana pod kierunkiem
dr. hab. Mariusza Bieńka, prof. UMCS

LUBLIN 2023

Składam serdeczne podziękowania Promotorowi
dr. hab. Mariuszowi Bieńkowi, prof. UMCS
za opiekę naukową, cenne wskazówki oraz pomoc
w trakcie przygotowania niniejszej pracy.

Spis treści

Wstęp

Jednym z najważniejszych problemów statystyki matematycznej i jej zastosowań jest wyznaczanie nieznanego rozkładu prawdopodobieństwa pewnej wielkości losowej lub jej charakterystyk liczbowych na podstawie obserwacji jej wartości danych w postaci próby losowej.

W rozprawie skupimy się na problemie estymacji kwantyli obserwowanej nieznanej zmiennej losowej. Kwantyle należą do najważniejszych charakterystyk rozkładów prawdopodobieństwa, i służą one między innymi do określenia różnych miar położenia i rozproszenia rozkładów (np. mediana, kwantyle, odstęp międzykwartylowy i inne). Bardzo dobre wprowadzenie do problemu estymacji kwantyli możemy znaleźć w artykułach przeglądowych [16], [36] oraz cytowanej tam literaturze. Teoretycznie wyznaczenie rozkładu F jest zawsze możliwe, dzięki twierdzeniu Glivienki-Cantelli'ego, które mówi, że dystrybuanty empiryczne zbiegają jednostajnie do F gdy rozmiar próby rośnie do nieskończoności. Jednakże praktyczne zastosowanie tego wyniku wymaga obserwacji próby bardzo dużego rozmiaru. Ponadto, nawet jeżeli funkcja F jest ciągła, to dystrybuanta empiryczna jest zawsze funkcją schodkową. Stąd twierdzenie to daje dość słabe przybliżenie w przypadku małych prób.

Jako estymatory kwantyli bardzo często używane są tak zwane L–statystyki, czyli kombinacje liniowe statystyk porządkowych próby. Rozważmy próbę losową $\mathbf{X} = (X_1, \ldots, X_n)$ rozmiaru n złożoną z niezależnych obserwacji zmiennej losowej X z nieznaną dystrybuantą F. Statystykami porządkowymi nazywamy wartości tej próby ustawione w porządku rosnącym czyli $X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$. Wtedy L–statystyką nazywamy każdą liniową kombinację statystyk porządkowych próby $\mathbf{X}$.

W literaturze poświęconej problemowi estymacji kwantyli albo brak prostych reguł jak wybierać optymalne estymatory w postaci L–statystyk, albo podawane reguły mają uzasadnienie jedynie intuicyjne. Najprostszym i najczęściej stosowanym wyborem estymatora kwantyla rzędu p jest odpowiednio dobrana pojedyncza statystyka porządkowa. Jest to tzw. kwantyl z próby $X_{\lfloor np \rfloor + 1:n}$, gdzie $\lfloor x \rfloor$ oznacza część całkowitą liczby rzeczywistej x. Wybór ten uzasadniony jest asymptotyczną mocną zgodnością kwantyla z próby, ale dla prób niewielkiego rozmiaru nie jest jasne dlaczego statystyka $X_{\lfloor np \rfloor + 1:n}$

jest lepszym estymatorem kwantyla niż na przykład $X_{\lfloor np \rfloor:n}$.

Innym często stosowanym wyborem są kombinacje liniowe dwóch sąsiednich statystyk porządkowych postaci $(1-g)X_{j:n} + gX_{j+1:n}$, gdzie parametry $g \in (0,1)$ oraz $j \in \{1,\ldots,n\}$ są odpowiednio dobrane w zależności od n oraz p. Inne przykłady wyboru g i j omówimy szczegółowo w podrozdziale 1.3.

Kolejnymi przykładami L–statystyk stosowanych do estymacji $x_p(F)$ są dwa estymatory zaproponowane przez Reissa [26]: *quasi–kwantyl z próby* (średnia arytmetyczna statystyk porządkowych o rzędach symetrycznych względem $\lfloor np \rfloor$) oraz tzw. *adaptacyjny quasi–kwantyl* (średnia ważona odpowiednio dobranych pięciu statystyk porządkowych). Bardziej skomplikowanymi przykładami L–statystyk są *estymator Harella–Davisa* [12], oraz *estymator Kaigha–Lachenbrucha* [15]. Zestawienie różnych L–statystyk używanych do estymacji kwantyli oraz porównanie ich obciążenia w estymacji kwantyli rozkładu normalnego można znaleźć m.in. w pracy Parisha [24].

W rozprawie wprowadzimy nowe kryterium optymalności, które pozwala wyznaczać względnie proste reguły wyboru jak najlepszych estymatorów oraz dowodzić analitycznie ich optymalności. Przy zadanym rzędzie kwantyla i rozmiarze próby będziemy wybierać L–statystyki (a dokładniej współczynniki kombinacji liniowych) tak, aby "jak najlepiej" estymować dany kwantyl. Mówiąc dokładniej, główne cele rozprawy są następujące:

1. wprowadzenie nowego kryterium optymalności L–statystyk jako estymatorów kwantyli, opartego na osiągalnych oszacowaniach obciążenia estymacji $x_p(F)$ przez ustaloną L–statystykę $L(\mathbf{c})$;

2. wyznaczenie jawnej postaci wyżej wspomnianych oszacowań;

3. wyznaczenie wektorów współczynników optymalnych L–statystyk w sposób jak najbardziej jawny;

4. w szczególności wybór najlepszej pojedynczej statystyki porządkowej oraz kombinacji liniowej dwóch sąsiednich statystyk porządkowych;

5. porównanie numeryczne zachowania wyznaczonych estymatorów kwantyli z wyżej wspomnianymi estymatorami znanymi z literatury.

Podkreślmy, że podejście do problemu zaprezentowane w rozprawie będzie w pełni nieparametryczne. Innymi słowy, zostaną wyznaczone estymatory, które optymalnie szacują kwantyl zadanego rzędu w sytuacji, gdy nie mamy żadnej uprzedniej wiedzy co do rozkładu badanej wielkości losowej. Ponadto interesuje nas jedynie podejście nieasymptotyczne, a więc rozważamy próby ustalonego rozmiaru.

W Rozdziale 1 przedstawiamy podstawowe pojęcia używane w rozprawie, takie jak kwantyl, funkcja kwantylowa oraz omawiamy przykładowe L–statystyki. W Rozdziale 2 wprowadzamy definicję nowego kryterium optymalności oraz własności oszacowań, które są niezbędne do wyznaczenia optymalnego estymatora. Następnie poruszamy problem optymalnego oszacowania obciążenia estymacji kwantyli, o którym jest mowa w Rozdziale 3. W Rozdziale 4 jest przedstawiony przypadek wyboru estymatora kwantyla w postaci odpowiednio dobranej statystyki porządkowej. Następnie w Rozdziale 5 przedstawiamy przypadek wyboru kombinacji liniowej dwóch statystyk porządkowych jako estymatora kwantyla. Rozdział 6 jest poświęcony analizie numerycznej błędu średnio–kwadratowego wyznaczonych estymatorów i porównaniu ze znanymi estymatorami. W Dodatku A są umieszczone dowody technicznych lematów, niezbędnych do udowodnienia głównych wyników rozprawy.

Wyniki rozdziałów 3 i 4 zostały opublikowane w pracy [4]. Wyniki rozdziałów 5 i 6 zostały zamieszczone w pracy [3].

Rozdział 1

Pojęcia wstępne

W pierwszym rozdziale omówimy szczegółowo najważniejsze pojęcia rozważane w niniejszej rozprawie. W podrozdziale 1.1 przedstawimy definicje i własności kwantyli i funkcji kwantylowych dowolnych rozkładów prawdopodobieństwa. Podrozdział 1.2 jest poświęcony statystykom porządkowym i ich wybranym własnościom. W podrozdziale 1.3 wprowadzamy pojęcie L–statystyk oraz oznaczenia z nimi związane. Prezentujemy również liczne przykłady L–statystyk stosowanych w literaturze jako estymatory kwantyli.

1.1 Kwantyle i funkcje kwantylowe

Niech F będzie dowolną dystrybuantą na prostej $\mathbb{R}$, odpowiadającą pewnej zmiennej losowej X. Przypomnijmy, że F jest funkcją niemalejącą, prawostronnie ciągłą oraz

$$\lim_{x\to-\infty} F(x) = 0, \qquad \lim_{x\to+\infty} F(x) = 1.$$

Ponadto, $F(x) = \mathbb{P}(X \leq x)$ dla $x \in \mathbb{R}$. *Kwantylem* rzędu $p \in (0,1)$ zmiennej losowej X (i rozkładu F) nazywamy każdą liczbę $x_p = x_p(F)$, dla której spełnione są nierówności

$$\mathbb{P}(X \leq x_p) \geq p, \quad \mathbb{P}(X \geq x_p) \geq 1 - p,$$

lub równoważnie $F(x_p^-) \leq p \leq F(x_p)$. Używamy tu standardowej notacji $f(a^-)$ oraz $f(a^+)$ na oznaczenie lewostronnej oraz prawostronnej granicy funkcji f w punkcie a. W szczególności, kwantyl rzędu $p = \frac{1}{2}$ nazywamy medianą.

Kwantyl x_p nie musi być wyznaczony jednoznacznie, dlatego definiujemy *górną* oraz *dolną funkcję kwantylową* rozkładu F odpowiednio wzorami

$$F^{\rightarrow}(p) = \sup\{x \in \mathbb{R} : F(x) \leq p\},$$

oraz

$$F^{\leftarrow}(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}. \tag{1.1}$$

Wtedy dla dowolnego rzędu kwantyla $p \in (0,1)$ wartości funkcji $F^{\rightarrow}(p)$ oraz $F^{\leftarrow}(p)$ są zdefiniowane jednoznacznie, i dla dowolnej wartości kwantyla $x_p(F)$ zachodzi nierówność $F^{\leftarrow}(p) \leq x_p(F) \leq F^{\rightarrow}(p)$.

Podamy teraz kilka ważnych własności funkcji kwantylowych, których dowody można znaleźć między innymi w podrozdziale 0.2 książki [27] lub 14.4 podręcznika [5]. Funkcje kwantylowe $F^{\rightarrow}$ oraz $F^{\leftarrow}$ są odpowiednio prawostronnie oraz lewostronnie ciągłe. Ponadto, jeśli X jest zmienną losową o rozkładzie F oraz U jest zmienną losową o rozdziale jednostajnym na przedziale $[0,1]$, to zmienne losowe $F^{\leftarrow}(U)$ oraz $F^{\rightarrow}(U)$ mają również taki sam rozkład F. Dlatego wartość oczekiwana i wariancja zmiennej X oraz dystrybuanty F mogą być przedstawione przy użyciu funkcji kwantylowej w postaci

$$\mu_F = \mathbb{E}_F X = \mathbb{E}_F F^{\leftarrow}(U) = \int_0^1 F^{\leftarrow}(u)du,$$

i podobnie

$$\sigma_F^2 = Var_F X = \int_0^1 \left(F^{\leftarrow}(u) - \mu_F\right)^2 du. \tag{1.2}$$

W powyższych wzorach funkcję $F^{\leftarrow}$ można zastąpić przez $F^{\rightarrow}$. Ta uwaga dotyczy również wzoru (1.6) poniżej.

Dodatkowo zauważmy, że skoro $F^{\leftarrow}$ jest funkcją niemalejącą lewostronnie ciągłą, to całka Riemanna-Stieltjesa $\int_0^1 F^{\leftarrow}(u)dG(u)$ ma sens jedynie dla funkcji G o wahaniu ograniczonym prawostronnie ciągłych (zob. np. [28], Zadanie 3, str. 138). W szczególności możemy napisać

$$F^{\leftarrow}(p) = \int_0^1 F^{\leftarrow}(u)d\mathbf{I}_{[p,1)}(u), \tag{1.3}$$

gdzie $\mathbf{I}_A$ oznacza indykator zbioru $A \subset \mathbb{R}$. Podobnie

$$F^{\rightarrow}(p) = \int_0^1 F^{\rightarrow}(u)d\mathbf{I}_{(p,1)}(u).$$

1.2 Statystyki porządkowe

Oznaczmy przez $(X_1, X_2, \ldots, X_n)$ próbę losową prostą złożoną z niezależnych zmiennych losowych o jednakowym rozkładzie z dystrybuantą F. Wtedy przez $X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$ oznaczamy statystyki porządkowe będące uszeregowaniem wartości próby w porządku od najmniejszej do największej. W dalszym ciągu przypomnimy znane własności rozkładów statystyk porządkowych (zob. np. [1], [8]). W tym miejscu podkreślmy, że w całej rozprawie interesuje nas podejście nieasymptotyczne, a więc rozmiar próby będzie ustalony. Czytelnika zainteresowanego podejściem asymptotycznym odsyłamy do obszernych monografii [26] oraz [31].

Niech $F_{X_{j:n}}$ oznacza dystrybuantę j–tej statystyki porządkowej $X_{j:n}$, a więc

$$F_{X_{j:n}}(x) = \mathbb{P}(X_{j:n} \leq x) = \sum_{i=j}^{n} \binom{n}{i} F^i(x)[1 - F(x)]^{n-i}. \tag{1.4}$$

Ponadto, jeśli F jest dystrybuantą absolutnie ciągłą o funkcji gęstości f, to funkcją gęstości zmiennej $X_{j:n}$ jest

$$f_{X_{j:n}}(x) = \frac{n!}{(j-1)!(n-j)!}(F(x))^{j-1}(1 - F(x))^{n-j} f(x). \tag{1.5}$$

W szczególności, jeśli $X = U$ ma rozkład jednostajny na przedziale $[0, 1]$ o funkcji gęstości $f(x) = \mathbf{1}_{[0,1]}(x)$ i dystrybuancie $F(x) = x$, $x \in [0, 1]$, to zamiast $X_{j:n}$ piszemy $U_{j:n}$. Na mocy (1.4) i (1.5) dystrybuantą $U_{j:n}$ jest

$$F_{j:n}(u) = \sum_{i=j}^{n} \binom{n}{i} u^i (1 - u)^{n-i}, \quad u \in [0, 1],$$

a funkcją gęstości $U_{j:n}$ jest

$$f_{j:n}(u) = \frac{n!}{(j-1)!(n-j)!} u^{j-1}(1 - u)^{n-j}, \quad u \in [0, 1].$$

Ponadto, zmienne losowe $X_{j:n}$ oraz $F^{\rightarrow}(U_{j:n})$ mają takie same rozkłady. Dlatego dla dowolnego rozkładu F wartość oczekiwana j-tej statystyki porządkowej $X_{j:n}$ wyraża się wzorem

$$\mathbb{E}_F X_{j:n} = \int_0^1 F^{\leftarrow}(u) f_{j:n}(u) du = \int_0^1 F^{\leftarrow}(u) dF_{j:n}(u). \tag{1.6}$$

W dalszym ciągu pracy używamy następujących oznaczeń. Dla $0 \leq j \leq n$ przez

$$B_{j,n}(p) = \binom{n}{j} p^j (1 - p)^{n-j}, \quad 0 \leq p \leq 1, \tag{1.7}$$

oznaczamy *wielomiany Bernsteina* rzędu n. Wielomiany te mają szereg interesujących własności, z których najważniejsze w tej pracy są własność VDP oraz nierówność Simmonsa. Zostaną one dokładniej przedstawione jako lematy A.1 oraz A.2 w Dodatku A.1.

Przy użyciu wielomianów Bernsteina dla $1 \leq j \leq n$ oraz $p \in [0, 1]$ możemy zapisać funkcje $F_{j:n}$ i $f_{j:n}$ w postaci

$$F_{j:n}(p) = \sum_{i=j}^{n} B_{i,n}(p) \tag{1.8}$$

oraz $f_{j:n}(p) = nB_{j-1,n-1}(p)$. Z teoretycznego punktu widzenia ciekawe jest, że w dowodach używamy prostych zależności pomiędzy $F_{j:n}$ oraz rozkładem dwumianowym $\mathcal{B}(n, p)$ z parametrami $n \in \mathbb{N}$ oraz $p \in (0, 1)$. Ściślej rzecz ujmując, niech Y będzie zmienną o rozkładzie $\mathcal{B}(n, p)$, czyli

$$\mathbb{P}(Y = j) = \binom{n}{j} p^j (1 - p)^{n-j}, \quad 0 \leq j \leq n.$$

Wtedy

$$F_{j:n}(p) = \mathbb{P}(Y \geq j), \quad 1 \leq j \leq n,$$

oraz

$$1 - F_{j+1:n}(p) = \mathbb{P}(Y \leq j), \quad 0 \leq j \leq n-1.$$

W szczególności, mediana oraz moda dla rozkładu $\mathcal{B}\left(n, \frac{j}{n}\right)$ są równe j (zob. [14]). Dokładniej, w tym przypadku j jest silną medianą dla zmiennej losowej Y co oznacza, że $\mathbb{P}(Y > j) > \frac{1}{2}$ oraz $\mathbb{P}(Y < j) > \frac{1}{2}$, lub równoważnie

$$F_{j:n}\left(\frac{j}{n}\right) > \frac{1}{2} \quad \text{oraz} \quad F_{j+1:n}\left(\frac{j}{n}\right) < \frac{1}{2}. \tag{1.9}$$

1.3 L–statystyki

Głównym celem rozprawy jest wyznaczenie L–statystyk, które byłyby jak najlepszymi estymatorami kwantyla x_p zadanego rzędu $p \in (0,1)$. L–statystyką nazywamy kombinację liniową statystyk porządkowych $(X_{1:n}, X_{2:n}, \ldots, X_{n:n})$ postaci

$$L(\mathbf{c}) = \sum_{i=1}^{n} c_i X_{i:n},$$

gdzie $\mathbf{c} = (c_1, c_2, \ldots, c_n) \in \mathbb{R}^n$ oznacza wektor współczynników. Jako estymatory kwantyli będziemy w ogólności rozważać L–statystyki, dla których $\mathbf{c} \in \mathcal{C}_n$, gdzie

$$\mathcal{C}_n = \left\{ (c_1, c_2, \ldots, c_n) : \sum_{i=1}^{n} c_i = 1 \text{ oraz } c_i \geq 0 \text{ dla } 1 \leq i \leq n \right\}.$$

Określmy ponadto 3 szczególne podzbiory zbioru $\mathcal{C}_n$. Oznaczmy

$$\mathcal{C}_n^{(1)} = \left\{ \mathbf{c} \in \mathcal{C}_n : \bigvee_{1 \leq j \leq n} \quad c_j = 1 \right\}.$$

Zauważmy, że skoro dla $\mathbf{c} \in \mathcal{C}_n$ mamy $\sum_{j=1}^{n} c_j = 1$, to dla $\mathbf{c} \in \mathcal{C}_n^{(1)}$ dokładnie jeden ze współczynników jest równy 1, a wszystkie pozostałe wynoszą 0. Zatem odpowiadająca L–statystyka jest w istocie pojedynczą statystyką porządkową. Dalej niech

$$\mathcal{C}_n^{(2)} = \left\{ \mathbf{c} \in \mathcal{C}_n : \bigvee_{1 \leq j \leq n} \quad c_1 = \cdots = c_{j-1} = 0 \wedge c_{j+2} = \cdots = c_n = 0 \right\}.$$

Zauważmy, że wektory $\mathbf{c}$ należące do $\mathcal{C}_n^{(2)}$ mają co najwyżej dwie współrzędne niezerowe, a więc opisują L–statystyki postaci

$$(1-\alpha)X_{j:n} + \alpha X_{j+1:n},$$

gdzie $0 \leq \alpha < 1$ oraz $j \leq j < n$. Ostatnim z rozważanych podzbiorów zbioru $\mathcal{C}_n$ jest

$$\mathcal{C}_n^{(\wedge)} = \left\{ \mathbf{c} \in \mathcal{C}_n : \bigvee_{1 \leq j \leq n} \quad c_1 \leq \cdots \leq c_j \wedge c_{j+1} \geq \cdots \geq c_n \right\}.$$

Zauważmy, że opisuje on L–statystyki, w których współczynniki do pewnego miejsca rosną, a następnie maleją, a więc $c_1, \ldots, c_j$ rosną, a $c_{j+1}, \ldots, c_n$ maleją. Odpowiada to sytuacji, gdy statystykom porządkowym z indeksami bardziej oddalonymi od j przypisujemy coraz mniejsze wagi. Oczywiście $\mathcal{C}_n^{(1)} \subset \mathcal{C}_n^{(2)} \subset \mathcal{C}_n^{(\wedge)} \subset \mathcal{C}_n$.

W dalszym ciągu podamy przykłady L–statystyk stosowanych w estymacji kwantyli. Przykłady te zostały zaczerpnięte z artykułów [7], [12], [13], [15], [24], [26] oraz [33]. Będziemy przy tym używać tak zwanych funkcji podłogi $\lfloor \cdot \rfloor$ i sufitu $\lceil \cdot \rceil$ zdefiniowanych dla $x \in \mathbb{R}$ równościami

$$\lfloor x \rfloor = \max\{n \in \mathbb{Z} : n \leq x\},$$

oraz

$$\lceil x \rceil = \min\{n \in \mathbb{Z} : n \geq x\}.$$

Ponadto przez $\{x\}$ oznaczamy część ułamkową liczby $x \in \mathbb{R}$, czyli

$$\{x\} = x - \lfloor x \rfloor .$$

Funkcje te będą również pojawiać się często w dalszym ciągu rozprawy.

1.3.1 Proste estymatory

Najpierw przedstawimy L–statystyki oparte albo o pojedynczą albo o dwie kolejne statystyki porządkowe. Bardzo dobry przegląd takich estymatorów można znaleźć w artykule Hyndmana i Fana [13]. Jego autorzy porównują 9 estymatorów kwantyli, które są zaimplementowane w pakietach matematycznych i statystycznych między innymi programie Mathematica, w języku R oraz pakiecie NumPy języka Python (zob. [9], [21], [35]). Wszystkie te estymatory mają postać

$$Q_i(p) = (1 - \gamma) X_{j:n} + \gamma X_{j+1:n}, \tag{1.10}$$

gdzie $1 \leq i \leq 9$ oraz $\frac{j-m}{n} \leq p < \frac{j-m+1}{n}$ dla pewnych $m \in \mathbb{R}$, oraz $0 \leq \gamma \leq 1$. Wartość γ jest funkcją parametrów $k = \lfloor np + m \rfloor$ oraz $g = np + m - k$.

Trzy pierwsze estymatory Q_1, Q_2, Q_3 w istocie używają tylko jednej odpowiednio wybranej statystyki porządkowej. Estymator Q_1 odpowiada wyborowi $m = 0$ oraz $\gamma = 1$ jeżeli $g > 0$, lub $\gamma = 0$ jeżeli $g = 0$, a więc mamy $Q_1(p) = X_{\lceil np \rceil : n}$. Jest to najstarsza znana definicja estymatora kwantyla z próby, która sprowadza się do wyznaczenia dolnej funkcji kwantylowej dla dystrybuanty empirycznej. Dokładniej, jeżeli przez $\mathbf{I}_A$

oznaczymy indykator zdarzenia A, to dystrybuantę empiryczną próby $(X_1, \ldots, X_n)$ definiujemy jako

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}_{\{X_i \leq x\}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}_{\{X_{i:n} \leq x\}},$$

dla $x \in \mathbb{R}$. Wtedy na mocy wzoru (1.1) dostajemy $F^{\leftarrow}(p) = X_{\lceil np \rceil:n}$ dla każdego $p \in (0,1)$. Dalej, Q_2 otrzymujemy przez podstawienie $m = 0$ oraz $\gamma = \frac{1}{2}$ jeżeli $g = 0$ lub $\gamma = 1$ jeżeli $g > 0$. Stąd

$$Q_2(p) = \begin{cases} \frac{1}{2}(X_{j:n} + X_{j+1:n}), & \text{jeżeli } p = \frac{j}{n}, \\ X_{\lceil np \rceil:n}, & \text{jeżeli } p \neq \frac{j}{n}. \end{cases}$$

Ten estymator jest podobny do $Q_1(p)$, ale dla $p = \frac{j}{n}$ stosujemy średnią arytmetyczną $X_{j:n}$ i $X_{j+1:n}$. Estymator Q_3 otrzymujemy przez podstawienie $m = -\frac{1}{2}$ oraz $\gamma = 1$ jeżeli $g > 0$. Ponadto, przyjmujemy $\gamma = 0$ jeżeli $g = 0$ oraz j jest parzyste, a w przeciwnym przypadku $\gamma = 1$. Zatem

$$Q_3(p) = \begin{cases} X_{j:n}, & \text{jeżeli } p = \frac{j+0.5}{n} \text{ oraz } j \text{ jest liczbą parzystą}, \\ X_{j+1:n}, & \text{jeżeli } p = \frac{j+0.5}{n} \text{ oraz } j \text{ jest liczbą nieparzystą}, \\ X_{\lceil np-0.5 \rceil:n}, & \text{poza}. \end{cases}$$

Inaczej jest to statystyka porządkowa $X_{j:n}$, gdzie indeks j jest liczbą całkowitą najbliższą wartości np.

Pozostałe estymatory $Q_4, \ldots, Q_9$ są interpolacjami na przedziale $(0,1)$ pomiędzy punktami $(\pi_j, X_{j:n})$, $1 \leq j \leq n$, gdzie liczby $\pi_1, \ldots, \pi_n \in [0,1]$ są odpowiednio dobrane. W tym kontekście są one nazywane punktami wykresu kwantyli (ang. *plotting positions*). Zwykle używana jest interpolacja liniowa i wtedy problem wyznaczenia dobrego estymatora sprowadza się do wyboru właściwych punktów wykresu kwantyli. Intuicyjnym uzasadnieniem takiego wyboru najczęściej jest wspomniany już fakt, że dla dowolnej funkcji dystrybuanty F statystyka porządkowa $X_{j:n}$ ma taki sam rozkład jak $F^{\leftarrow}(U_{j:n})$. Co więcej, jeżeli dystrybuanta F jest ciągła, to $F(X_{j:n})$ ma taki sam rozkład jak $U_{j:n}$.

W naszym przypadku estymatory $Q_4, \ldots, Q_9$ są interpolacjami liniowymi pomiędzy punktami wykresu kwantyli zadanymi wzorem

$$\pi_j = \frac{j-a}{n-a-b-1}, \quad 1 \leq j \leq n,$$

gdzie wartości a oraz b są podane w Tabeli 1.1. Dokładna postać estymatorów $Q_i(p)$, $4 \leq i \leq 9$, jest dana wzorem

$$Q_i(p) = (1 - \{\ell\})X_{\lfloor \ell \rfloor:n} + \{\ell\}X_{\lfloor \ell \rfloor+1:n},$$

gdzie $\ell = \ell(i, n, p)$ jest odpowiednio wybrane. Zatem jest to postać (1.10) gdzie $j = \lfloor \ell \rfloor$ oraz $\gamma = \{\ell\}$. Wartości ℓ oraz p, dla których $Q_i(p)$ jest poprawnie określone są podane w Tabeli 1.1. Estymatory $Q_4, \ldots, Q_9$ możemy opisać jak następująco:

- estymator Q_4 jest klasyczną liniową interpolacją pomiędzy punktami $\left(\frac{j}{n}, X_{j:n}\right)$. W tym przypadku mamy $m = 0$, $\ell = np$ oraz $p \in \left[\frac{1}{n}, 1\right)$.

- estymator Q_5 jest zdefiniowany w oparciu o punkty $\pi_j = \frac{j-1/2}{n}$, a więc jest to przesunięta wersja estymatora Q_4. W tym przypadku mamy $m = \frac{1}{2}$, $\ell = np + 0.5$ oraz $p \in \left[\frac{1}{2n}, 1 - \frac{1}{2n}\right)$. Taka postać często jest stosowana w hydrologii.

- estymator Q_6 jest podejściem zaproponowanym przez Weibulla [34] i Gumbela [11] motywowanym faktem, iż w tym przypadku $\pi_j = \frac{j}{n+1}$, a więc $\pi_j = \mathbb{E}U_{j:n}$. Mamy wtedy $m = p$, $\ell = (n+1)p$ oraz $p \in \left[\frac{1}{n+1}, \frac{n}{n+1}\right)$.

- estymator Q_7 został zaproponowany przez Gumbela [11]. W tym przypadku $\pi_j = \frac{j-1}{n-1}$, co pokrywa się z modą $U_{j:n}$. W tym przypadku mamy $m = 1 - p$, $\ell = (n-1)p + 1$ oraz $p \in [0, 1)$.

- estymator Q_8 jest oparty na wyborze $\pi_j = \frac{j-1/3}{n+1/3}$, co jest dobrym przybliżeniem mediany $U_{j:n}$, która jest po prostu wartością liczby p_j (zob. np. [26]). W tym przypadku mamy $m = \frac{1}{3}(p+1)$, $\ell = (n+\frac{1}{3})p + \frac{1}{3}$ oraz $p \in \left[\frac{2}{3n+1}, \frac{n-1}{3n+1}\right)$. Zauważmy, że mediana statystyki porządkowej $U_{j:n}$ będzie odgrywać ważną rolę w dalszej części rozprawy (por. definicję liczby p_j w podrozdziale 4.2).

- estymator Q_9 jest oparty na wyborze $\pi_j = \frac{j-3/8}{n+1/4}$, co jest dobrym przybliżeniem wartości $F(\mathbb{E}X_{j:n})$ przy założeniu, że F jest dystrybuantą rozkładu normalnego. W tym przypadku mamy $m = \frac{1}{4}p + \frac{3}{8}$, $\ell = (n + \frac{1}{4})p + \frac{3}{8}$ oraz $p \in \left[\frac{5}{2(4n+1)}, \frac{8n-3}{2(4n+1)}\right)$ (zob. [6]).

1.3.2 Inne znane estymatory

Inne znane estymatory kwantyli w postaci L–statystyk używają zwykle więcej niż dwie statystyki porządkowe. Jednakże wspólną intuicją stojącą za ich definicjami jest to, aby największą wagę przypisać statystyce z indeksem bliskim np. Pozostałym statystykom przypisywane są wagi coraz mniejsze w zależności od oddalenia ich indeksu od np.

Estymator Kaigha-Lachenbrucha. Jest to estymator kwantyla rzędu p postaci

$$Q_{KL}^{(k)}(p) = \sum_{i=r}^{r+n-k} \binom{i-1}{r-1} \binom{n-i}{k-r} \binom{n}{k}^{-1} X_{i:n}, \tag{1.11}$$

i	a	b	m	ℓ	p
4	0	1	0	np	$\left[\frac{1}{n}, 1\right)$
5	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$np + 0.5$	$\left[\frac{1}{2n}, 1 - \frac{1}{2n}\right)$
6	0	0	p	$(n+1)p$	$\left[\frac{1}{n+1}, \frac{n}{n+1}\right)$
7	1	1	$1-p$	$(n-1)p+1$	$[0, 1)$
8	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}(p+1)$	$(n+\frac{1}{3})p + \frac{1}{3}$	$\left[\frac{2}{3n+1}, \frac{n-1}{3n+1}\right)$
9	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{4}p + \frac{3}{8}$	$(n+\frac{1}{4})p + \frac{3}{8}$	$\left[\frac{5}{2(4n+1)}, \frac{8n-3}{2(4n+1)}\right)$

Tabela 1.1: Wartości a oraz b oraz odpowiadające im wartości m oraz $\ell(i)$ dla estymatorów Q_i dla $4 \leq i \leq 9$.

gdzie $r = \lfloor (k+1)p \rfloor$ dla odpowiednio dobranego $k \leq n$ (zob. [15]).

Estymator Harrella-Davisa. Jest to estymator postaci

$$Q_{HD}(p) = \sum_{i=1}^{n} W_{p,i,n} X_{i:n},$$

gdzie współczynniki $W_{p,i,n}$ są zadane wzorami

$$W_{p,i,n} = \frac{1}{B((n+1)p, (n+1)(1-p))} \int_{(i-1)/n}^{i/n} t^{(n+1)p-1}(1-t)^{(n+1)(1-p)-1} dt,$$

(zob. [12]) oraz $B(a,b)$ oznacza funkcję beta Eulera daną wzorem

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1} dt, \quad a > 0, b > 0.$$

Zauważmy, że współczynniki można zapisać przy pomocy niekompletnej funkcji beta

$$I_x(a,b) = \frac{1}{B(a,b)} \int_0^x t^{a-1}(1-t)^{b-1} dt, \quad x \in (0,1),$$

w postaci

$$W_{p,i,n} = I_{i/n}\Big((n+1)p, (n+1)(1-p)\Big) - I_{(i-1)/n}\Big((n+1)p, (n+1)(1-p)\Big).$$

Estymator Bernsteina. Jest to estymator postaci

$$Q_B(p) = \sum_{i=1}^{n} \left[\binom{n-1}{i-1} p^{i-1}(1-p)^{n-i} \right] X_{i:n}, \quad 0 < p < 1.$$

Zauważmy, że $Q_B(p)$ można zapisać w postaci

$$Q_B(p) = \sum_{i=1}^{n} B_{i-1,n-1}(p) X_{i:n}$$

gdzie $B_{i,n}$ oznaczają wielomiany Bernsteina zdefiniowane powyżej wzorem (1.7), co tłumaczy nazwę tego estymatora (zob. [7]).

Uwaga 1.1. Łatwo sprawdzić, że w powyższych trzech przypadkach wektory odpowiadających im współczynników $\mathbf{c}$ należą do zbioru $\mathcal{C}_n^{(\wedge)}$.

Quasikwantyle. Jest to klasa estymatorów wprowadzona przez Reissa [26], opartych o statystykę porządkową $X_{j:n}$, gdzie $j = \lfloor np \rfloor$ oraz o statystyki z indeksami $j \pm m$, $j \pm 2m$ itd. dla odpowiednio dobranego m. Najprostszym przykładem quasikwantyla jest

$$Q_{R1}^{(m)}(p) = \frac{1}{2}\left(X_{\lfloor np \rfloor - m:n} + X_{\lfloor np \rfloor + m:n}\right), \tag{1.12}$$

gdzie $p \in (0,1)$ oraz $1 \leq m \leq \min\{\lfloor np \rfloor - 1, n - \lfloor np \rfloor\}$. Zauważmy, że w tym przypadku wektor współczynników $\mathbf{c} \in \mathcal{C}_n$, ale $\mathbf{c} \notin \mathcal{C}_n^{(\wedge)}$. Innym przykładem jest

$$\begin{aligned} Q_{R2}^{(m)}(p) = &-\frac{2}{25}X_{\lfloor np \rfloor - 2m:n} + \frac{8}{25}X_{\lfloor np \rfloor - m:n} \\ &+ \frac{13}{25}X_{\lfloor np \rfloor:n} + \frac{8}{25}X_{\lfloor np \rfloor + m:n} - \frac{2}{25}X_{\lfloor np \rfloor + 2m:n}, \end{aligned} \tag{1.13}$$

gdzie $p \in (0,1)$ oraz $1 \leq m \leq \min\{(\lfloor np \rfloor - 1)/2, (n - \lfloor np \rfloor)/2\}$. W tym przypadku mimo, że suma współczynników wynosi 1, to $\mathbf{c} \notin \mathcal{C}_n$, gdyż niektóre współczynniki są ujemne.

Estymatory jądrowe. Załóżmy, że f jest funkcją gęstości symetryczną względem 0 oraz, że $h \to 0$ gdy $n \to \infty$. Niech $f_h(x) = h^{-1}f(x/h)$. Wtedy jedną z wersji estymatora jądrowego jest estymator $KQ(p)$ dany wzorem

$$KQ(p) = \sum_{i=1}^{n} \left[\int_{(i-1)/n}^{i/n} f_h(t-p)dt\right] X_{i:n},$$

(zob. np. [24]). W praktyce zamiast $KQ(p)$ często stosujemy jego przybliżenia, na przykład

$$KQ_1(p) = \frac{1}{n}\sum_{i=1}^{n} f_h\Big(\frac{i}{n} - p\Big)X_{i:n},$$

lub

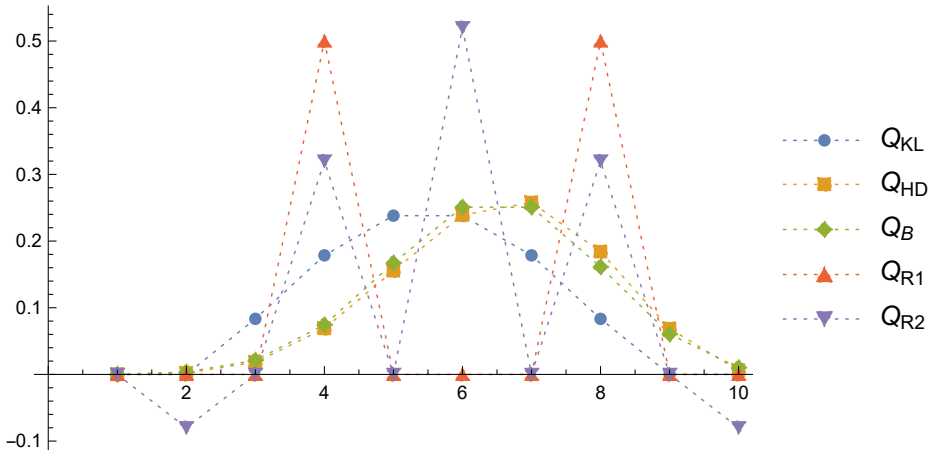
$$KQ_2(p) = \frac{\sum_{i=1}^{n} f_h\left(\frac{i-0.5}{n} - p\right) X_{i:n}}{\sum_{j=1}^{n} f_h\left(\frac{j-0.5}{n} - p\right)},$$

będące ustandaryzowaną wersją estymatora $KQ_1(p)$.

Przykład 1.1. Przykładowe wartości współczynników powyższych estymatorów dla $n = 10$ i $p = 0.6$ są podane w Tabeli 1.2 i zilustrowane na Rysunku 1.1. W przypadku estymatora $Q_{KL}^{(k)}$ przyjęto $k = 5$, co daje $r = 3$. W przypadku estymatorów $Q_{R1}^{(m)}$ i $Q_{R2}^{(m)}$ przyjęto $m = 2$.

j	1	2	3	4	5	6	7	8	9	10
$Q_{KL}^{(5)}$	0	0	0.0833	0.1786	0.2381	0.2381	0.1786	0.0833	0	0
Q_{HD}	0.00003	0.0020	0.0181	0.0689	0.1555	0.2389	0.2579	0.1841	0.0687	0.0059
Q_B	0.0003	0.0035	0.0212	0.0743	0.1672	0.2508	0.2508	0.1612	0.0605	0.0101
$Q_{R1}^{(2)}$	0	0	0	0.5	0	0	0	0.5	0	0
$Q_{R2}^{(2)}$	0	−0.08	0	0.32	0	0.52	0	0.32	0	−0.08

Tabela 1.2: Przykładowe wartości współczynników estymatorów $Q_{KL}^{(5)}$, Q_{HD}, Q_B, $Q_{R1}^{(2)}$, $Q_{R2}^{(2)}$ dla $n = 10$ i $p = 0.6$.



Rysunek 1.1: Wykres wartości współczynników estymatorów $Q_{KL}^{(5)}$, Q_{HD}, Q_B, $Q_{R1}^{(2)}$, $Q_{R2}^{(2)}$.

Rozdział 2

Nowe kryterium optymalności

W tym rozdziale wprowadzimy nowe kryterium optymalności L–statystyk jako estymatorów kwantyla zadanego rzędu $p \in (0,1)$ przy ustalonym rozmiarze próby n. Kryterium to oparte jest na optymalnych oszacowaniach obciążenia estymacji kwantyla rzędu p przez zadaną L–statystykę i ma jasne uzasadnienie intuicyjne. Pozwala ono wybrać optymalne estymatory w jednoznacznie określony sposób. W ten sposób uzyskujemy wyniki teoretyczne, które mają precyzyjne dowody analityczne, ale mogą być dość łatwo zastosowane w praktyce.

2.1 Definicja kryterium

W dalszym ciągu rozprawy rozważamy rozkłady prawdopodobieństwa nieznanych wielkości losowych opisane dystrybuantami F o skończonej wariancji danej wzorem (1.2). Klasę tych rozkładów oznaczamy przez $\mathcal{F}_2$. Podkreślmy, że w przeciwieństwie do większości autorów nie ograniczamy się do dystrybuant ciągłych i ściśle rosnących. W szczególności nie zakładamy jednoznaczności kwantyla $x_p(F)$.

Dla ustalonego wektora współczynników $\mathbf{c} \in \mathcal{C}_n$ oznaczmy oszacowania górne i dolne obciążenia L–statystyki $L(\mathbf{c})$, jako estymatora dolnej i górnej funkcji kwantylowej odpowiednio przez

$$\overline{B}_p(\mathbf{c}) = \sup_{F \in \mathcal{F}_2} \frac{\mathbb{E}_F L(\mathbf{c}) - F^{\leftarrow}(p)}{\sigma_F}, \tag{2.1}$$

oraz

$$\underline{B}_p(\mathbf{c}) = \inf_{F \in \mathcal{F}_2} \frac{\mathbb{E}_F L(\mathbf{c}) - F^{\rightarrow}(p)}{\sigma_F}. \tag{2.2}$$

Oszacowania te są optymalne w klasie $\mathcal{F}_2$ w tym sensie, że można wskazać rozkłady F, $G \in \mathcal{F}_2$, dla których zachodzą równości

$$\frac{\mathbb{E}_F L(\mathbf{c}) - F^{\leftarrow}(p)}{\sigma_F} = \overline{B}_p(\mathbf{c})$$

oraz

$$\frac{\mathbb{E}_G L(\mathbf{c}) - G^{\rightarrow}(p)}{\sigma_G} = \underline{B}_p(\mathbf{c}).$$

Zauważmy, że $F^{\leftarrow}(p) \leq F^{\rightarrow}(p)$, a więc oczywiście dla każdego $F \in \mathcal{F}_2$

$$\frac{\mathbb{E}_F L(\mathbf{c}) - F^{\rightarrow}(p)}{\sigma_F} \leq \frac{\mathbb{E}_F L(\mathbf{c}) - F^{\leftarrow}(p)}{\sigma_F}.$$

Zatem mamy $\underline{B}_p(\mathbf{c}) \leq \overline{B}_p(\mathbf{c})$. Niestety, jak zobaczymy w podrozdziale 3.3, wyrażenia dla $\overline{B}_p(\mathbf{c})$ i $\underline{B}_p(\mathbf{c})$ umiemy znaleźć tylko dla $\mathbf{c} \in \mathcal{C}_n^{(\wedge)}$. Nawet w tym przypadku uzyskane wyrażenia są tylko częściowo jawne, i dlatego nie są zbyt użyteczne przy dalszej analizie obciążenia estymacji kwantyli przez L–statystyki. Z tego względu będziemy poszukiwać przybliżeń dla $\overline{B}_p(\mathbf{c})$ i $\underline{B}_p(\mathbf{c})$, które będą ułatwiać dalsze rozważania. Okazuje się, że potrafimy wyznaczać jawne wzory na powyższe oszacowania dla $\mathbf{c} \in \mathcal{C}_n^{(1)}$, i wzory te są bardzo pomocne w dalszej analizie. Mianowicie dla $\mathbf{c} \in \mathcal{C}_n^{(1)}$, a więc $c_j = 1$ oraz $c_i = 0$ dla $i \neq j$, piszemy $\mathbf{c} = \boldsymbol{\delta}_j$ i oznaczamy przez

$$\overline{B}_{n,p}(j) = \sup_{F \in \mathcal{F}_2} \frac{\mathbb{E}_F X_{j:n} - F^{\leftarrow}(p)}{\sigma_F}, \tag{2.3}$$

oraz

$$\underline{B}_{n,p}(j) = \inf_{F \in \mathcal{F}_2} \frac{\mathbb{E}_F X_{j:n} - F^{\rightarrow}(p)}{\sigma_F}, \tag{2.4}$$

odpowiednio górne i dolne oszacowanie obciążenia pojedynczej statystyki porządkowej $X_{j:n}$ jako estymatora funkcji kwantylowych $F^{\leftarrow}(p)$ oraz $F^{\rightarrow}(p)$. Wtedy oczywiście

$$\overline{B}_{n,p}(j) = \overline{B}_p(\boldsymbol{\delta}_j) \quad \text{oraz} \quad \underline{B}_{n,p}(j) = \underline{B}_p(\boldsymbol{\delta}_j).$$

Pokażemy teraz, że pomiędzy zdefiniowanymi wyżej wielkościami zachodzą proste nierówności.

Lemat 2.1. *Dla $p \in (0,1)$ oraz $\mathbf{c} \in \mathcal{C}_n$ mamy*

$$\overline{B}_p(\mathbf{c}) \leq \sum_{j=1}^{n} c_j \overline{B}_{n,p}(j) \tag{2.5}$$

oraz

$$\underline{B}_p(\mathbf{c}) \geq \sum_{j=1}^{n} c_j \underline{B}_{n,p}(j). \tag{2.6}$$

Dowód. Dla dowolnego rozkładu $F \in \mathcal{F}_2$ oraz $\mathbf{c} \in \mathcal{C}_n$, a więc $\sum_{i=1}^{n} c_i = 1$, mamy równość

$$\frac{\mathbb{E}_F L(\mathbf{c}) - F^{\leftarrow}(p)}{\sigma_F} = \sum_{j=1}^{n} c_j \frac{\mathbb{E}_F X_{j:n} - F^{\leftarrow}(p)}{\sigma_F}.$$

Biorąc pod uwagę to, że $c_j \geq 0$ oraz stosując do niej własności kresów górnych otrzymujemy (2.5). Podobnie dowodzimy nierówności dla $\underline{B}_p(\mathbf{c})$. □

Oznaczmy prawe strony powyższych nierówności przez

$$\overline{b}_p(\mathbf{c}) = \sum_{j=1}^{n} c_j \overline{B}_{n,p}(j) \tag{2.7}$$

oraz

$$\underline{b}_p(\mathbf{c}) = \sum_{j=1}^{n} c_j \underline{B}_{n,p}(j). \tag{2.8}$$

Teraz określmy pomocniczą funkcję $r_p : \mathcal{C}_n \to \mathbb{R}$ wzorem

$$r_p(\mathbf{c}) = \overline{b}_p(\mathbf{c}) + \underline{b}_p(\mathbf{c}),$$

oraz funkcję $s_p : \mathcal{C}_n \to [0, \infty)$ wzorem

$$s_p(\mathbf{c}) = \sqrt{p(1-p)}\,|r_p(\mathbf{c})|. \tag{2.9}$$

Dodatkowo rozważamy takie wektory $\mathbf{c}$, które należą do pewnego domkniętego podzbioru $\mathcal{C} \subset \mathcal{C}_n$, który w tym kontekście będziemy nazywać klasą.

Definicja 2.1. Powiemy, że L–statystyka $L(\mathbf{c}_0)$ jest estymatorem optymalnym w klasie $\mathcal{C} \subset \mathcal{C}_n$ kwantyla rzędu p nieznanego rozkładu o skończonej wariancji, jeżeli

$$s_p(\mathbf{c}_0) = \min_{\mathbf{c} \in \mathcal{C}} s_p(\mathbf{c}).$$

W ogólności wektor $\mathbf{c}$ nie musi być wyznaczony jednoznacznie, ale oczywiście zależy od n oraz p. Zauważmy, że z Wniosku 3.3 poniżej wynika, że s_p jest funkcją ciągłą. Ponadto oczywiście $\mathcal{C}_n$ jest zbiorem zwartym w $\mathbb{R}^n$ jako zbiór domknięty i ograniczony. Dlatego funkcja s_p zawsze osiąga minimum na zbiorze $\mathcal{C}_n$. Zatem aby zapewnić istnienie $\mathbf{c}_0$ optymalnego w klasie $\mathcal{C}$ wystarczy założyć, że $\mathcal{C}$ jest domkniętym podzbiorem zbioru $\mathcal{C}_n$.

Aby uzasadnić nasze kryterium zauważmy najpierw, że $\underline{b}_p(\mathbf{c}) < 0 < \overline{b}_p(\mathbf{c})$. Zatem wydaje się, że najbardziej rozsądnym kryterium jest minimalizacja różnicy $\overline{b}_p(\mathbf{c}) - \underline{b}_p(\mathbf{c})$ względem wektora $\mathbf{c}$. Jednakże taka różnica w większości przypadków wynosi $1/\sqrt{p(1-p)}$ lub więcej, a to wyrażenie nie zależy od wyboru wektora $\mathbf{c}$. Z drugiej strony, jeśli różnicę zastąpimy przez sumę (jak w definicji funkcji s_p), to mierzymy stopień symetrii oszacowań względem 0. Jeśli wartości s_p są bliskie 0, to wartości górnego i dolnego oszacowania są mniej więcej symetryczne względem 0. Innymi słowy, niezależnie od badanego rozkładu odpowiednio wybrany estymator średnio nie przeszacowuje ani nie niedoszacywuje kwantyla zadanego rzędu. Czynnik normalizujacy $1/\sqrt{p(1-p)}$ sprawia, że $\lim_{p\to 0^+} s_p(\mathbf{c}) = \lim_{p\to 1^-} s_p(\mathbf{c}) = 1$, a więc funkcja s_p jest ciągła i ograniczona na przedziale $[0, 1]$.

Oczywiście, na mocy definicji mamy dla $p \in (0,1)$, $\mathbf{c} \in \mathcal{C}_n$ oraz $F \in \mathcal{F}_2$

$$\underline{B}_p(\mathbf{c}) \leq \frac{\mathbb{E}_F L(\mathbf{c}) - F^{\rightarrow}(p)}{\sigma_F} \leq \frac{\mathbb{E}_F L(\mathbf{c}) - F^{\leftarrow}(p)}{\sigma_F} \leq \overline{B}_p(\mathbf{c})$$

i oszacowania te są osiągalne. Zatem stosując w definicji funkcji $r_p(\mathbf{c})$ sumę $\overline{B}_p(\mathbf{c}) + \underline{B}_p(\mathbf{c})$ zamiast $\overline{b}_p(\mathbf{c}) + \underline{b}_p(\mathbf{c})$ otrzymalibyśmy bardziej precyzyjne kryterium optymalności. Jednakże, z uwagi na trudności z wyznaczeniem jawnych wyrażeń na powyższe oszacowania, kryterium takie byłoby mało użyteczne. Z drugiej strony mamy na mocy Lematu 2.1

$$\underline{b}_p(\mathbf{c}) \leq \frac{\mathbb{E}_F L(\mathbf{c}) - F^{\rightarrow}(p)}{\sigma_F} \leq \frac{\mathbb{E}_F L(\mathbf{c}) - F^{\leftarrow}(p)}{\sigma_F} \leq \overline{b}_p(\mathbf{c}).$$

Mimo, że oszacowania te nie są osiągalne, to stosując je w definicjach $r_p(\mathbf{c})$ i $s_p(\mathbf{c})$ otrzymamy w wyniku spójną i dość rozwiniętą teorię, która może być łatwo zastosowana w praktyce. Ponadto, dla $\mathbf{c} \in \mathcal{C}_n^{(1)}$ mamy oczywiście $\overline{b}_p(\mathbf{c}) = \overline{B}_p(\mathbf{c})$ oraz $\underline{b}_p(\mathbf{c}) = \underline{B}_p(\mathbf{c})$ dla wszystkich $p \in (0,1)$. Dla $\mathbf{c} \in \mathcal{C}_n^{(2)}$ pokażemy, że nierówności te zachodzą dla większości wartości p (zob. równości (5.4) i (5.5)).

2.2 Własności oszacowań i pomocniczych funkcji

Aby efektywnie zastosować wprowadzone kryterium konieczna jest znajomość jawnej postaci wzorów na $\overline{b}_p(\mathbf{c})$ i $\underline{b}_p(\mathbf{c})$ oraz pewne własności pomocniczych funkcji $r_p(\mathbf{c})$ i $s_p(\mathbf{c})$. Jawne wzory będą konsekwencją wyników następnego rozdziału. W tym podrozdziale wprowadzimy wybrane własności tych funkcji, które można udowodnić bez znajomości jawnych wzorów.

Analogicznie do funkcji r_p i s_p określamy funkcję $r_{n,p} : \{1, \ldots, n\} \to \mathbb{R}$ wzorem

$$r_{n,p}(j) = \overline{B}_{n,p}(j) + \underline{B}_{n,p}(j), \tag{2.10}$$

oraz funkcję $s_{n,p} : \{1, \ldots, n\} \to [0, \infty)$ wzorem $s_{n,p}(j) = \sqrt{p(1-p)}|r_{n,p}(j)|$. Oczywiście $r_{n,p}(j) = r_p(\boldsymbol{\delta}_j)$ oraz $s_{n,p}(j) = s_p(\boldsymbol{\delta}_j)$. Zbadamy najpierw własności monotoniczności funkcji $r_{n,p}(j)$ ze względu na zmienne j oraz p.

Lemat 2.2. *Funkcja $r_{n,p}$ ma następujące własności:*

(a) przy ustalonym $p \in (0,1)$ mamy

$$r_{n,p}(j) \leq r_{n,p}(j+1), \quad 1 \leq j < n, \tag{2.11}$$

czyli $r_{n,p}$ jest funkcją niemalejącą względem j;

(b) przy ustalonym $j \in \{1, \ldots, n\}$ mamy

$$r_{n,p}(j) \geq r_{n,q}(j), \quad 0 < p < q < 1, \tag{2.12}$$

czyli $r_{n,p}(j)$ jest funkcją nierosnącą względem p.

Dowód. Z definicji statystyk porządkowych mamy $X_{j:n} \leq X_{j+1:n}$, a więc oczywiście

$$\frac{\mathbb{E}_F X_{j:n} - F^{\leftarrow}(p)}{\sigma_F} \leq \frac{\mathbb{E}_F X_{j+1:n} - F^{\leftarrow}(p)}{\sigma_F}.$$

Stąd otrzymujemy łatwo

$$\overline{B}_{n,p}(j) \leq \overline{B}_{n,p}(j+1) \quad \text{oraz} \quad \underline{B}_{n,p}(j) \leq \underline{B}_{n,p}(j+1). \tag{2.13}$$

Sumując stronami otrzymamy nierówność (2.11). Ponadto, funkcja $F^{\leftarrow}$ jest niemalejąca, a wiec dla $0 < p < q < 1$ mamy $F^{\leftarrow}(p) \leq F^{\leftarrow}(q)$. Zatem dla $1 \leq j \leq n$ oraz $F \in \mathcal{F}_2$ mamy oczywiście

$$\frac{\mathbb{E}_F X_{j:n} - F^{\leftarrow}(q)}{\sigma_F} \leq \frac{\mathbb{E}_F X_{j:n} - F^{\leftarrow}(p)}{\sigma_F}.$$

Stąd

$$\overline{B}_{n,p}(j) \geq \overline{B}_{n,q}(j). \tag{2.14}$$

Podobnie $F^{\rightarrow}$ jest niemalejąca i analogicznie dowodzimy nierówność $\underline{B}_{n,p}(j) \geq \underline{B}_{n,q}(j)$. Sumując te nierówności stronami otrzymamy wzór (2.12). □

Uwaga 2.1. Po wyznaczeniu jawnych wzorów na $\overline{B}_{n,p}(j)$ oraz $\underline{B}_{n,p}(j)$ pokażemy, że obydwie nierówności (2.11) i (2.12) są ostre, a więc $r_{n,p}$ jest funkcją ściśle monotoniczną względem j oraz p (zob. Wniosek 3.3).

W dalszym ciągu pokażemy, że pomiędzy górnymi i dolnymi oszacowaniami zachodzi własność symetrii

$$\underline{B}_{n,p}(j) = -\overline{B}_{n,1-p}(n-j+1),$$

gdzie $p \in (0,1)$, $1 \leq j \leq n$. W tym celu udowodnimy dwa pomocnicze lematy.

Lemat 2.3. *Dla dowolnej dystrybuanty F i dla dowolnego $p \in (0,1)$ zachodzi równość*

$$\inf\{x \in \mathbb{R} : F(x^-) \geq p\} = \inf\{x \in \mathbb{R} : F(x) \geq p\}.$$

Dowód. Funkcja F jest niemalejąca, a więc dla dowolnego $x \in \mathbb{R}$ mamy $F(x^-) \leq F(x)$. Zatem dla $p \in (0,1)$ oraz $x \in \mathbb{R}$

$$\{x : F(x^-) \geq p\} \subset \{x : F(x) \geq p\}.$$

Jeżeli zbiory te nie są równe, to istnieje $x_0 \in \mathbb{R}$ takie, że $F(x_0^-) < p \leq F(x_0)$. Skoro F jest niemalejąca, to taki punkt x_0 jest wyznaczony jednoznacznie. Zatem

$$\{x : F(x^-) \geq p\} = (x_0, \infty), \qquad \{x : F(x) \geq p\} = [x_0, \infty),$$

a więc zbiory te mają równe kresy dolne. □

Lemat 2.4. *Niech F będzie dowolną dystrybuantą oraz określmy*

$$G(x) = 1 - F((-x)^-), \quad x \in \mathbb{R}.$$

Wtedy dla $p \in (0,1)$

$$G^{\rightarrow}(p) = -F^{\leftarrow}(1-p).$$

Dowód. Z definicji funkcji kwantylowych otrzymujemy

$$\begin{aligned} G^{\rightarrow}(p) &= \sup\{x : G(x) \le p\} \\ &= \sup\{x : F((-x)^-) \ge 1-p\} \\ &= -\inf\{x : F(x^-) \ge 1-p\} \\ &= -\inf\{x : F(x) \ge 1-p\} = -F^{\leftarrow}(1-p), \end{aligned}$$

gdzie przedostatnia równość wynika z Lematu 2.3. □

Lemat 2.5. *Dla $n \ge 2$, $p \in (0,1)$ i $1 \le j \le n$ mamy*

$$\overline{B}_{n,1-p}(j) = -\underline{B}_{n,p}(n-j+1), \tag{2.15}$$

oraz

$$r_{n,1-p}(j) = -r_{n,p}(n-j+1). \tag{2.16}$$

Dowód. Rozważmy próbę $(X_1, \ldots, X_n)$ z rozkładu F i określmy $Y_i = -X_i$ dla $1 \le i \le n$. Wtedy $(Y_1, \ldots, Y_n)$ jest próbą z rozkładu $G(x) = 1 - F((-x)^-)$. Statystyki porządkowe tej próby to $Y_{j:n} = -X_{n-j+1:n}$. Oczywiście $\mathbb{E}_G Y_{j:n} = -\mathbb{E}_F X_{n-j+1:n}$ oraz $\sigma_G = \sigma_F$, a więc $G \in \mathcal{F}_2$ wtedy i tylko wtedy, gdy $F \in \mathcal{F}_2$. Zatem korzystając z Lematu 2.4, dostajemy

$$\frac{\mathbb{E}_F X_{j:n} - F^{\leftarrow}(1-p)}{\sigma_F} = -\frac{\mathbb{E}_G Y_{n-j+1:n} - G^{\rightarrow}(p)}{\sigma_G}.$$

Stąd otrzymujemy

$$\begin{aligned} \overline{B}_{n,1-p}(j) &= \inf_{F \in \mathcal{F}_2} \frac{\mathbb{E}_F X_{j:n} - F^{\leftarrow}(1-p)}{\sigma_F} = \\ &= -\sup_{G \in \mathcal{F}_2} \frac{\mathbb{E}_G Y_{n-j+1:n} - G^{\rightarrow}(p)}{\sigma_G} = -\underline{B}_{n,p}(n-j+1), \end{aligned}$$

co kończy dowód równości (2.15). Wzór (2.16) jest konsekwencją wzorów (2.10) i (2.15). □

Przejdźmy teraz do wyznaczenia własności funkcji $r_p(\mathbf{c})$. Oczywiście, z definicji $r_p(\mathbf{c})$ oraz $r_{n,p}(j)$ dostajemy łatwo równość

$$r_p(\mathbf{c}) = \sum_{j=1}^{n} c_j r_{n,p}(j). \tag{2.17}$$

Łącząc Lematy 2.2, 2.5 i wzór (2.17) wyprowadzamy własności funkcji r_p. Będziemy przy tym stosować następującą notację: dla $\mathbf{c} \in \mathcal{C}_n$ przez $\underline{\mathbf{c}}$ oznaczamy wektor $\underline{\mathbf{c}} = (\underline{c}_1, \ldots, \underline{c}_n)$, gdzie $\underline{c}_j = c_{n-j+1}$ dla $1 \leq j \leq n$.

Lemat 2.6. *Funkcja r_p ma następujące własności:*

(a) r_p jest malejąca względem zmiennej $p \in (0, 1)$;

(b) dla $\mathbf{c} \in \mathcal{C}_n$ mamy

$$r_{1-p}(\mathbf{c}) = -r_p(\underline{\mathbf{c}}); \tag{2.18}$$

(c) $r_{1/2}(\mathbf{c}) = -r_{1/2}(\underline{\mathbf{c}})$ oraz jeśli $\underline{\mathbf{c}} = \mathbf{c}$, to $r_{1/2}(\mathbf{c}) = 0$.

Dowód. Na mocy wzoru (2.14), jeżeli $\mathbf{c} \in \mathcal{C}_n$, to $r_p(\mathbf{c})$ jest wypukłą kombinacją liniową funkcji $r_{n,p}(j)$. Na mocy Lematu 2.2(b), funkcje te są malejące względem p, co dowodzi punktu (a). Własność (b) wynika ze wzoru (2.17) oraz symetrii (2.16). Własność (c) wynika łatwo z punktu (b) po podstawieniu $p = \frac{1}{2}$. □

Lemat 2.7. *Załóżmy, że $\mathcal{C} \subset \mathcal{C}_n$ jest zbiorem wektorów współczynników o własności*

$$\mathbf{c} \in \mathcal{C} \Rightarrow \underline{\mathbf{c}} \in \mathcal{C}. \tag{2.19}$$

Wtedy jeżeli $L(\mathbf{c})$ jest estymatorem optymalnym w klasie $\mathcal{C}$ kwantyla rzędu p, to estymatorem kwantyla rzędu $1 - p$ optymalnym w klasie $\mathcal{C}$ jest $L(\underline{\mathbf{c}})$.

Dowód. Z Lematu 2.6 oraz ze wzorów (2.9) oraz (2.18) otrzymujemy następującą symetrię $s_p(\mathbf{c}) = s_{1-p}(\underline{\mathbf{c}})$. Zatem z Definicji 2.1 otrzymujemy, że jeśli $\mathbf{c}$ jest optymalnym wektorem dla p, to $\underline{\mathbf{c}}$ jest optymalnym wektorem dla $1 - p$. □

Rozdział 3

Optymalne oszacowania obciążenia estymacji kwantyli

W tym rozdziale wyznaczone zostaną najpierw jawne wzory na górne i dolne oszacowania obciążenia estymacji funkcji kwantylowych przez pojedynczą statystykę porządkową. Dokładniej mówiąc, dla ustalonych $p \in (0,1)$, $n \geq 2$ oraz $1 \leq j \leq n$ wyznaczymy wartości $\overline{B}_{n,p}(j)$ i $\underline{B}_{n,p}(j)$ dane wzorami (2.3) oraz (2.4).

W szczególności, udowodnimy, że dla ustalonych wartości $2 \leq j \leq n-1$ istnieje przedział $(\theta_j(n), \xi_j(n))$ zawierający liczbę $\frac{j}{n}$ taki, że dla dowolnego $F \in \mathcal{F}_2$ oraz liczby $p \in (\theta_j(n), \xi_j(n))$ zachodzą nierówności

$$-\frac{F_{j:n}(p)}{\sqrt{p(1-p)}} \leq \frac{\mathbb{E}_F X_{j:n} - F^{\rightarrow}(p)}{\sigma_F} \leq \frac{\mathbb{E}_F X_{j:n} - F^{\leftarrow}(p)}{\sigma_F} \leq \frac{1-F_{j:n}(p)}{\sqrt{p(1-p)}}. \tag{3.1}$$

Ponadto, oszacowania te są osiągnięte dla odpowiednio dobranych rozkładów dwupunktowych. Dla p leżących poza tym przedziałem, jedno z oszacowań $\overline{B}_{n,p}(j)$ lub $\underline{B}_{n,p}(j)$ wyraża się bardziej skomplikowanym wzorem. W szczególności, jeśli przyjmiemy $p = \frac{j}{n}$, to otrzymamy poprawną wersję stwierdzenia z artykułu Okolewskiego i Rychlika [22], które mówi, że dla większości wartości j, n oraz $p = \frac{j}{n}$ nierówności

$$-\frac{F_{j:n}(p)}{\sqrt{p(1-p)}} \leq \frac{\mathbb{E}X_{j:n} - F^{\rightarrow}(p)}{\sigma_F} \leq \frac{1-F_{j:n}(p)}{\sqrt{p(1-p)}} \tag{3.2}$$

są optymalne. Nie jest to prawdą w przypadku górnej nierówności. Ponadto błędnym wnioskiem było to, że obydwie nierówności są osiągnięte dla tego samego rozkładu dwupunktowego. W Uwadze 3.1 pokażemy jak powinno wyglądać prawidłowe sformułowanie oraz pokazujemy dlaczego dowód nierówności (3.2) jest niepoprawny. Wyniki tego rozdziału zostały opublikowane w pracy [4].

W podrozdziale 3.3 pokażemy jak wyznaczać wyrażenia na oszacowania $\overline{B}_p(\mathbf{c})$ i $\underline{B}_p(\mathbf{c})$ określone wzorami (2.1) i (2.2). Następnie, uzyskamy ich częściowo jawną postać w szczególnym przypadku, gdy $\mathbf{c} \in \mathcal{C}_n^{(\wedge)}$.

3.1 Pomocnicze liczby θ_j i ξ_j

Zdefiniujmy najpierw pomocnicze liczby θ_j oraz ξ_j dla $1 \leq j \leq n$. Liczby te są niezbędne do podania wzorów na oszacowania (2.3) oraz (2.4) oraz na funkcje $r_{n,p}$ i $s_{n,p}$. Na przykład, jeżeli $p \in (\theta_j, \xi_j)$, to funkcja $r_{n,p}(j)$ wyraża się prostym wzorem w przeciwieństwie do wartości p leżących poza tym przedziałem (zob. Wniosek 3.1).

Dla $2 \leq j \leq n-1$ zdefiniujmy liczbę $\theta_j = \theta_j(n)$ jako jedyne rozwiązanie równania

$$1 - F_{j:n}(\theta_j) = (1-\theta_j) f_{j:n}(\theta_j) \tag{3.3}$$

na przedziale $\left(0, \frac{j-1}{n-1}\right)$. Podobnie, zdefiniujmy liczbę $\xi_j = \xi_j(n)$ jako jedyne rozwiązanie równania

$$F_{j:n}(\xi_j) = \xi_j f_{j:n}(\xi_j)$$

na przedziale $\left(\frac{j-1}{n-1}, 1\right)$. Jednoznaczność rozwiązań powyższych równań jest dobrze znanym faktem, ale dla kompletności rozważań zostanie ona wykazana w Dodatku A.1. Oczywiście $\theta_j < \xi_j$ dla $2 \leq j \leq n-1$. Ponadto, przyjmujemy, że $\theta_1 = \xi_1 = 0$ oraz $\theta_n = \xi_n = 1$. Używając prostych zależności

$$1 - F_{j:n}(u) = F_{n-j+1:n}(1-u) \tag{3.4}$$

oraz $f_{j:n}(u) = f_{n-j+1:n}(1-u)$ otrzymujemy, że

$$\xi_j = 1 - \theta_{n-j+1}, \quad 1 \leq j \leq n. \tag{3.5}$$

3.2 Wartości oszacowań i ich własności

W tym podrozdziale wyznaczymy jawne wzory na $\overline{B}_{n,p}(j)$ oraz $\underline{B}_{n,p}(j)$. Są one bardzo ważne w dalszych rozważaniach, gdyż występują w definicjach funkcji $r_p(\mathbf{c})$ i $s_p(\mathbf{c})$. Dodatkowo, jest to bardzo dobry przypadek modelowy dla ogólnego przypadku $\mathbf{c} \in \mathcal{C}_n$ rozważanego w podrozdziale 3.3.

Idea dowodu głównego twierdzenia pochodzi z artykułu [22], w którym została użyta tak zwana nierówność Moriguti'ego (zob. [20]). Jest to szczególny przypadek tzw. *metody projekcji* opisanej szczegółowo w monografii [29].

Lemat 3.1 (Nierówność Moriguti'ego)**.** *Niech $\Phi : [0,1] \to \mathbb{R}$ będzie funkcją o wahaniu ograniczonym ciągłą na obu końcach przedziału. Wtedy nierówność*

$$\int_0^1 x(t)\, d\Phi(t) \leq \int_0^1 x(t)\overline{\varphi}(t)\, dt$$

zachodzi dla dowolnej niemalejącej funkcji $x : [0,1] \to \mathbb{R}$, dla której całki są skończone, gdzie $\overline{\varphi}$ jest prawostronną pochodną największej wypukłej minoranty $\overline{\Phi}$ funkcji Φ.

Równość zachodzi wtedy i tylko wtedy, gdy funkcja x jest stała na każdym z przedziałów gdzie zachodzi nierówność $\overline{\Phi}(t) < \min\{\Phi(t^-), \Phi(t^+)\}$, oraz jeżeli Φ nie jest ciągła w pewnym punkcie t_0, to

(a) jeżeli $\Phi(t_0^-) < \Phi(t_0^+)$, to x jest prawostronnie ciągła;

(b) jeżeli $\Phi(t_0^-) > \Phi(t_0^+)$, to x jest lewostronnie ciągła.

Teraz przedstawimy główne twierdzenie tego rozdziału (zob. [4]).

Twierdzenie 3.1. *Niech $n \geq 2$ oraz $p \in (0,1)$.*

(a) Załóżmy, że $2 \leq j \leq n-1$. Jeżeli $p \geq \theta_j$, to

$$\overline{B}_{n,p}(j) = \frac{1 - F_{j:n}(p)}{\sqrt{p(1-p)}}, \tag{3.6}$$

oraz równość zachodzi dla rozkładu dwupunktowego postaci

$$\mathbb{P}\left(X = \mu - \sigma\sqrt{\frac{1-p}{p}}\right) = p = 1 - \mathbb{P}\left(X = \mu + \sigma\sqrt{\frac{p}{1-p}}\right). \tag{3.7}$$

W przeciwnym wypadku, jeżeli $p < \theta_j$, to

$$\overline{B}_{n,p}(j) = \overline{B}_j = \left(\frac{(1 - F_{j:n}(p))^2}{p} + \int_p^{\theta_j} f_{j:n}^2(x)dx + \frac{(1 - F_{j:n}(\theta_j))^2}{1-\theta_j}\right)^{1/2}. \tag{3.8}$$

Oszacowanie (3.8) *jest osiągnięte dla dystrybuanty F postaci*

$$F(x) = \begin{cases} 0, & \text{jeśli } \frac{x-\mu}{\sigma} < -\frac{1-F_{j:n}(p)}{p\overline{B}_j}, \\ p, & \text{jeśli } -\frac{1-F_{j:n}(p)}{p\overline{B}_j} \leq \frac{x-\mu}{\sigma} < \frac{f_{j:n}(p)}{\overline{B}_j}, \\ f_{j:n}^{-1}\left(\overline{B}_j \frac{x-\mu}{\sigma}\right), & \text{jeśli } \frac{f_{j:n}(p)}{\overline{B}_j} \leq \frac{x-\mu}{\sigma} < \frac{1-F_{j:n}(\theta_j)}{(1-\theta_j)\overline{B}_j}, \\ 1, & \text{jeśli } \frac{x-\mu}{\sigma} \geq \frac{1-F_{j:n}(\theta_j)}{(1-\theta_j)\overline{B}_j}. \end{cases} \tag{3.9}$$

(b) Dla $j = 1$ oraz $p \in (0,1)$ mamy

$$\overline{B}_{n,p}(1) = \frac{1 - F_{1:n}(p)}{\sqrt{p(1-p)}} = (1-p)^{n-1}\sqrt{\frac{1-p}{p}},$$

a równość jest osiągnięta dla rozkładu dwupunktowego danego wzorem (3.7).

(c) Dla $j = n$ oraz $p \in (0,1)$ mamy

$$\overline{B}_{n,p}(n) = \overline{B}_n = \left(\frac{(1-p^n)^2}{p} + \frac{n^2}{2n-1}(1 - p^{2n-1})\right)^{1/2}. \tag{3.10}$$

Oszacowanie (3.10) *jest osiągnięte dla następującej dystrybuanty*

$$F(x) = \begin{cases} 0, & \text{jeśli } \frac{x-\mu}{\sigma} < -\frac{1-p^n}{p\overline{B}_n}, \\ p, & \text{jeśli } -\frac{1-p^n}{p\overline{B}_n} \le \frac{x-\mu}{\sigma} < \frac{np^{n-1}}{\overline{B}_n}, \\ \left(\overline{B}_n \frac{x-\mu}{\sigma n}\right)^{\frac{1}{n-1}}, & \text{jeśli } \frac{np^{n-1}}{\overline{B}_n} \le \frac{x-\mu}{\sigma} < \frac{n}{\overline{B}_n}, \\ 1, & \text{jeśli } \frac{x-\mu}{\sigma} \ge \frac{n}{\overline{B}_n}. \end{cases} \tag{3.11}$$

Dowód. Korzystając ze wzorów (1.6) oraz (1.3) na $\mathbb{E}_F X_{j:n}$ oraz $F^{\leftarrow}(p)$ otrzymujemy

$$\mathbb{E}_F X_{j:n} - F^{\leftarrow}(p) = \int_0^1 [F^{\leftarrow}(u) - \mu_F]\, dH_{j:n}(u), \tag{3.12}$$

gdzie

$$H_{j:n}(u) = \begin{cases} F_{j:n}(u), & \text{dla } u < p, \\ F_{j:n}(u) - 1, & \text{dla } u \ge p. \end{cases}$$

Oczywiście wartości funkcji $H_{j:n}$ zależą również od p, ale dla uproszczenia oznaczeń pomijamy tą zmienną w notacji.

(a) Załóżmy, że $2 \le j \le n-1$. Wtedy funkcja $F_{j:n}$ jest ściśle rosnąca na przedziale $[0,1]$, wypukła na przedziale $\left(0, \frac{j-1}{n-1}\right)$ oraz wklęsła na przedziale $\left(\frac{j-1}{n-1}, 1\right]$. Ponieważ mamy $\theta_j < \frac{j-1}{n-1}$, to $F_{j:n}$ jest wypukła na przedziale $(0, \theta_j)$, a na przedziale $(\theta_j, 1)$ jej wykres leży w całości powyżej prostej przechodzącej przez punkty $(\theta_j, F_{j:n}(\theta_j))$ oraz $(1,1)$. Zatem funkcja $H_{j:n}$ jest rosnąca od wartości 0 w punkcie 0 do jej lewostronnej granicy $H_{j:n}(p^-) = F_{j:n}(p) > 0$, a następnie $H_{j:n}(p) = F_{j:n}(p) - 1 < 0$ i $H_{j:n}$ rośnie na przedziale $[p,1]$ do 0. Ponadto, ponieważ $p \ge \theta_j$, $H_{j:n}$ jest albo wklęsła albo wypukła-wklęsła na przedziale $(p,1]$, ale jej wykres w całości leży nad prostą przechodzącą przez punkty $(p, F_{j:n}(p)-1)$ oraz $(1,0)$. Dlatego, jej największa wypukła minoranta jest krzywą łamaną składającą się z dwóch odcinków o końcach w punktach $(0,0)$, $(p, F_{j:n}(p)-1)$ oraz $(1,0)$. Zatem dana jest ona wzorem

$$\overline{H}_{j:n}(u) = \begin{cases} \frac{F_{j:n}(p)-1}{p} u, & \text{dla } u < p, \\ \frac{1-F_{j:n}(p)}{1-p}(u-1), & \text{dla } u \ge p. \end{cases} \tag{3.13}$$

Jej prawostronna pochodna jest postaci

$$\overline{H}'_{j:n}(u) = \begin{cases} \frac{F_{j:n}(p)-1}{p}, & \text{dla } u < p, \\ \frac{1-F_{j:n}(p)}{1-p}, & \text{dla } u \ge p, \end{cases}$$

a więc jest prawostronnie ciągła. Oczywiście

$$\|\overline{H}'_{j:n}\| = \frac{1 - F_{j:n}(p)}{\sqrt{p(1-p)}},$$

gdzie $\|\cdot\|$ oznacza tradycyjną normę w przestrzeni $\mathcal{L}^2$ funkcji całkowalnych z kwadratem względem miary Lebesgue'a na przedziale $[0,1]$. Z Lematu 3.1 i nierówności Schwarz'a dla każdej dystrybuanty $F \in \mathcal{F}_2$ mamy

$$\mathbb{E}_F X_{j:n} - F^{\leftarrow}(p) \leq \int_0^1 [F^{\leftarrow}(u) - \mu_F]\overline{H}'_{j:n}(u)du \leq \sigma_F \|\overline{H}'_{j:n}\|, \tag{3.14}$$

co dowodzi, że $\overline{B}_{n,p}(j) \leq \frac{1-F_{j:n}(p)}{\sqrt{p(1-p)}}$. Równość w drugiej nierówności we wzorze (3.14) zachodzi wtedy i tylko wtedy, gdy

$$F^{\leftarrow}(u) - \mu = \sigma \frac{\overline{H}'_{j:n}(u)}{\|\overline{H}'_{j:n}\|}, \quad u \in (0,1), \quad u \neq p.$$

Korzystając z powyższego oraz z Lematu 3.1 łatwo zauważyć, że równość w obu nierównościach w (3.14) jest osiągnięta wtedy i tylko wtedy, gdy

$$\frac{F^{\leftarrow}(u) - \mu}{\sigma} = \begin{cases} -\sqrt{\frac{1-p}{p}}, & \text{dla } u \leq p, \\ \sqrt{\frac{p}{1-p}}, & \text{dla } u > p. \end{cases}$$

To daje nam równości (3.6) oraz (3.7).

Jeżeli $p < \theta_j$, to funkcja $H_{j:n}$ jest wypukła na przedziale (p, θ_j), a na przedziale $(\theta_j, 1)$ jej wykres leży w całości powyżej prostej łączącej punkty $(\theta_j, F_{j:n}(\theta_j) - 1)$ oraz $(1,0)$. W tym przypadku największa wypukła minoranta dana jest wzorem

$$\overline{H}_{j:n}(u) = \begin{cases} \dfrac{F_{j:n}(p) - 1}{p}\, u, & \text{dla } 0 \leq u \leq p, \\ H_{j:n}(u), & \text{dla } p < u \leq \theta_j, \\ \dfrac{1 - F_{j:n}(\theta_j)}{1 - \theta_j}\,(u-1), & \text{dla } \theta_j < u \leq 1. \end{cases} \tag{3.15}$$

Teraz stosując analogiczne rozumowanie jak w przypadku $p \geq \theta_j$ dowodzimy wzorów (3.8) oraz (3.9).

(b) Dla $j = 1$ funkcja $F_{1:n}(u) = 1 - (1-u)^n$ jest wklęsła na przedziale $[0,1]$. Dlatego największa wypukła minoranta $\overline{H}_{1:n}$ jest dana wzorem (3.13) dla $j = 1$. Pozostała część dowodu jest analogiczna jak w punkcie (a) dla przypadku $p \geq \theta_j$.

(c) Dla $j = n$ mamy $\theta_n = 1$ i największa wypukła minoranta $\overline{H}_{n:n}$ wyraża się wzorem (3.15) z pominięciem trzeciego przypadku. Zatem dowód punktu (c) jest analogiczny jak wzorów (3.8) oraz (3.9). □

Uwaga 3.1. Używając porównania wartości liczb θ_{j+1}, ξ_j oraz $\frac{j}{n}$ (zob. Wniosek 4.1) dla $p = \frac{j}{n}$ otrzymamy poprawną wersję wyników artykułu [22] (zob. również Twierdzenie 16 w monografii [29]). Błąd w tej pracy spowodowany jest usterką w dowodzie

Twierdzenia 1. Dokładniej, zamiast równości (1.3) autorzy użyli równości

$$F^{\rightarrow}(p) = \int_0^1 F^{\rightarrow}(u) d\mathbf{I}_{[p,1)}(u),$$

która nie może być prawdziwa. Problem polega na tym, że taka całka Riemanna-Stieltjesa nie istnieje, ponieważ obie funkcje $F^{\rightarrow}$ oraz $\mathbf{I}_{[p,1)}$ są prawostronnie ciągłe w punkcie p. Natomiast całka występująca we wzorze (1.3) istnieje, ponieważ funkcja $F^{\leftarrow}$ jest lewostronnie ciągła, a funkcja $\mathbf{I}_{[p,1)}$ jest prawostronnie ciągła. Ponadto, autorzy artykułu [22] stwierdzili, że obydwa oszacowania we wzorze (3.2) są osiągnięte dla tego samego rozkładu danego wzorem (3.7), co jest oczywiście niemożliwe.

Uwaga 3.2. Załóżmy, że funkcja F jest ciągłą i ściśle rosnącą dystrybuantą na pewnym przedziale $(a, b) \subset \mathbb{R}$ taką, że $F(a) = 0$ oraz $F(b) = 1$. Wtedy obydwie funkcje kwantylowe $F^{\leftarrow}$ i $F^{\rightarrow}$ są równe i pokrywają się z F^{-1}, czyli zwyczajną funkcją odwrotną do funkcji F. Wtedy oczywiste jest, że środkowa nierówność we wzorze (3.1) stanie się równością. Zatem powstaje pytanie czy górne i dolne oszacowania we wzorze (3.1) mogą być poprawione w klasie rozkładów ciągłych i ściśle rosnących ze skończoną wariancją. Jednakże analizując dokładnie dowód Twierdzenia 3.1 możemy łatwo skonstruować ciąg ciągłych i ściśle monotonicznych funkcji kwantylowych, dla których oszacowania we wzorze (3.1) są osiągnięte w granicy.

Z Twierdzenia 3.1 używając symetrii oszacowań opisanej tożsamością (2.15) oraz własności (3.5) otrzymujemy wartości dolnych oszacowań $\underline{B}_{n,p}(j)$ oraz wzór na funkcję $r_{n,p}$ zdefiniowaną wzorem (2.10).

Wniosek 3.1. *Dla* $2 \le j \le n-1$ *oraz* $p \le \xi_j$ *mamy*

$$\underline{B}_{n,p}(j) = -\frac{F_{j:n}(p)}{\sqrt{p(1-p)}},$$

a dla $p > \xi_j$ *mamy*

$$\underline{B}_{n,p}(j) = -\left(\frac{(F_{j:n}(\xi_j))^2}{1-\xi_j} + \int_{\xi_j}^{p} f_{j:n}^2(x)dx + \frac{(F_{j:n}(p))^2}{1-p} \right)^{1/2}.$$

Ponadto,

$$\underline{B}_{n,p}(n) = \frac{(F_{n:n}(p))^2}{\sqrt{p(1-p)}} = -p^{n-1}\sqrt{\frac{p}{1-p}},$$

oraz

$$\underline{B}_{n,p}(1) = \underline{B}_{n,p}(1) = -\left(\int_0^p f_{1:n}^2(x)dx + \frac{(F_{1:n}(p))^2}{1-p} \right)^{1/2} =$$
$$= -\left(\frac{[1-(1-p)^n]^2}{1-p} + \frac{n^2}{2n-1}[1-(1-p)^{2n-1}] \right)^{1/2}.$$

Dowód. Na mocy (2.15) mamy $\underline{B}_{n,p}(j) = -\overline{B}_{n,1-p}(n-j+1)$. Ponadto na mocy równości (3.5) mamy $\xi_j = 1-\theta_{n-j+1}$, a więc $1-p \geq \theta_{n-j+1}$ wtedy i tylko wtedy, gdy $p \leq \xi_j$. $\square$

Wniosek 3.2. *Dla $j = 1$ oraz $p \in (0,1)$ mamy*

$$r_{n,p}(1) = \frac{1 - F_{1:n}(p)}{\sqrt{p(1-p)}} - \left(\frac{(F_{1:n}(p))^2}{1-p} + \int_0^p f_{1:n}^2(x)dx \right)^{1/2},$$

oraz analogicznie $r_{n,p}(n) = -r_{n,1-p}(1)$. Dla $2 \leq j \leq n-1$ mamy

(a) jeżeli $p < \theta_j$, to

$$r_{n,p}(j) = \overline{B}_j - \frac{F_{j:n}(p)}{\sqrt{p(1-p)}},$$

gdzie $\overline{B}_j$ jest zdefiniowane wzorem (3.8);

(b) jeżeli $\theta_j \leq p \leq \xi_j$, to

$$r_{n,p}(j) = \frac{1 - 2F_{j:n}(p)}{\sqrt{p(1-p)}};$$

(c) jeżeli $p > \xi_j$, to

$$r_{n,p}(j) = \frac{1 - F_{j:n}(p)}{\sqrt{p(1-p)}} - \left(\frac{(F_{j:n}(\xi_j))^2}{\xi_j} + \int_{\xi_j}^p f_{j:n}^2(x)dx + \frac{(F_{j:n}(p))^2}{1-p} \right)^{1/2}.$$

Teraz rozważmy własności oszacowań, z których wynikają dalsze własności funkcji $r_{n,p}$. Pierwszy wniosek mówi, że własności monotoniczności $r_{n,p}$ opisane w Lemacie 2.2 są ścisłe.

Wniosek 3.3. *Funkcja $r_{n,p}$ jest ściśle rosnącą funkcją zmiennej j, czyli*

$$r_{n,p}(j) < r_{n,p}(j+1), \quad 1 \leq j \leq n-1,$$

oraz ciągłą i ściśle malejącą funkcją zmiennej $p \in (0,1)$, czyli

$$r_{n,q}(j) < r_{n,p}(j), \quad 0 < p < q < 1.$$

Dowód. Z uwagi na definicję $r_{n,p}$ i symetrię (2.15) wystarczy udowodnić własności opisane w Lemacie jedynie dla funkcji $\overline{B}_{n,p}(j)$. Dla $p \geq \theta_{j+1}$ nierówność

$$\overline{B}_{n,p}(j) < \overline{B}_{n,p}(j+1), \quad 1 \leq j \leq n-1, \tag{3.16}$$

jest oczywista, gdyż $F_{j:n} > F_{j+1:n}$ na przedziale $(0,1)$. Dla $p \in (0, \theta_{j+1})$ używamy pierwszej nierówności we wzorze (2.13) oraz faktu, że oszacowania $\overline{B}_{n,p}(j)$ oraz $\overline{B}_{n,p}(j+1)$ nie mogą być równe, ponieważ są osiągnięte dla różnych rozkładów. To dowodzi nierówności (3.16) dla wszystkich $p \in (0,1)$.

Dalej, dla $j = 1$ lub $j = n$ ciągłość $\overline{B}_{n,p}(j)$ jest oczywista na mocy części (b) i (c) Twierdzenia 3.1. Dla ustalonego $j \in \{2, \ldots, n-1\}$ ciągłość $\overline{B}_{n,p}(j)$ ze względu na p na przedziałach $(0, \theta_j)$ oraz $(\theta_j, 1)$ wynika wprost ze wzorów odpowiednio (3.6) i (3.8). Również na mocy tych wzorów łatwo zauważyć, że granice lewostronne i prawostronne funkcji $\overline{B}_{n,p}(j)$ gdy $p \to \theta_j^-$ oraz $p \to \theta_j^+$ są równe $\frac{1-F_{j:n}(\theta_j)}{\sqrt{\theta_j(1-\theta_j)}}$.

Ścisła monotoniczność $\overline{B}_{n,p}(j)$ ze względu na p wynika z faktu, że $\overline{B}_{n,p}(j) \geq \overline{B}_{n,q}(j)$ dla $0 < p < q < 1$ (zob. (2.14)) oraz oszacowania te nie mogą być równe, gdyż są osiągnięte dla różnych rozkładów. □

Ostatnia własność oszacowań stosowana w dalszym ciągu pracy jest podana w następującym lemacie.

Lemat 3.2. *Jeżeli $p \in (\xi_j, 1)$ dla $1 \leq j \leq n-1$, to*

$$\underline{B}_{n,p}(j) < -\frac{F_{j:n}(p)}{\sqrt{p(1-p)}},$$

lub równoważnie, jeżeli $p \in (0, \theta_j)$ dla $2 \leq j \leq n$, to

$$\overline{B}_{n,p}(j) > \frac{1 - F_{j:n}(p)}{\sqrt{p(1-p)}}.$$

W konsekwencji, dla wszystkich $1 \leq j \leq n$ oraz $p \in (0,1)$ mamy

$$\underline{B}_{n,p}(j) \leq -\frac{F_{j:n}(p)}{\sqrt{p(1-p)}} \quad \textit{oraz} \quad \overline{B}_{n,p}(j) \geq \frac{1 - F_{j:n}(p)}{\sqrt{p(1-p)}}. \tag{3.17}$$

Dowód. Przypomnijmy, że $\xi_1 = 0$. Rozważmy funkcję

$$k_j(t) = \int_{\xi_j}^{t} f_{j:n}^2(x)dx + \xi_j f_{j:n}^2(\xi_j) - \frac{F_{j:n}^2(t)}{t}, \quad 0 \leq t \leq 1,$$

Ponieważ $F_{j:n}(\xi_j) = \xi_j f_{j:n}(\xi_j)$, to mamy $k_j(\xi_j) = 0$. Ponadto

$$k_j'(t) = f_{j:n}^2(t) - \frac{2t f_{j:n}(t) F_{j:n}(t) - F_{j:n}^2(t)}{t^2},$$

a więc po elementarnych obliczeniach otrzymujemy

$$k_j'(t) = \frac{1}{t^2}(t f_{j:n}(t) - F_{j:n}(t))^2 \geq 0$$

dla wszystkich $0 \leq t \leq 1$, a równość zachodzi dla $t = \xi_j$. Zatem k_j jest ściśle rosnąca na przedziale $[0,1]$. W szczególności dla $p > \xi_j$ mamy $k_j(p) > k_j(\xi_j) = 0$ lub

$$\int_{\xi_j}^{p} f_{j:n}^2(x)dx + \xi_j f_{j:n}^2(\xi_j) > \frac{(F_{j:n}(p))^2}{p}.$$

Łącząc ten wzór ze wzorem (3.8) otrzymujemy

$$(\underline{B}_{n,p}(j))^2 > \left(\frac{F_{j:n}(p)}{\sqrt{p(1-p)}}\right)^2.$$

Ponieważ $\underline{B}_{n,p}(j) < 0$, to kończy dowód. □

Dla ustalonych $n \geq 3$ oraz $p \in (0,1)$ definiujemy funkcję $t_{n,p}$ następująco

$$t_{n,p}(j) = \frac{1 - 2F_{j:n}(p)}{\sqrt{p(1-p)}}, \quad 1 \leq j \leq n. \tag{3.18}$$

Oczywiście, $t_{n,p}$ jest funkcją ciągłą względem zmiennej $p \in (0,1)$ oraz $t_{n,p}$ jest ściśle rosnącą funkcją zmiennej j, na mocy nierówności $F_{j:n}(p) > F_{j+1:n}(p)$. Prostą konsekwencją Wniosku 3.2 oraz Lematu 3.2 jest następujące porównanie wartości funkcji $r_{n,p}$ oraz $t_{n,p}$.

Wniosek 3.4.

(a) Dla wszystkich $p \in (0,1)$ mamy $r_{n,p}(1) < t_{n,p}(1)$ oraz $r_{n,p}(n) > t_{n,p}(n)$.

(b) Dla $2 \leq j \leq n-1$ mamy $r_{n,p}(j) \leq t_{n,p}(j)$ dla $p \in (\theta_j, 1)$ oraz $r_{n,p}(j) \geq t_{n,p}(j)$ dla $p \in (0, \xi_j)$.

3.3 Oszacowania obciążenia ogólnych L–statystyk

Dla ustalonego $\mathbf{c} \in \mathcal{C}_n$ określamy pomocniczą funkcję

$$F_{\mathbf{c}}(u) = \sum_{j=1}^{n} c_j F_{j:n}(u), \quad 0 \leq u \leq 1.$$

Oczywiście $F_{\mathbf{c}}$ jest dystrybuantą taką, że $F_{\mathbf{c}}(0) = 0$, $F_{\mathbf{c}}(1) = 1$, a jej gęstością jest

$$f_{\mathbf{c}}(u) = \sum_{j=1}^{n} c_j f_{j:n}(u),$$

dla $u \in [0,1]$ oraz 0 poza. Wtedy na mocy wzoru (1.6) mamy

$$\mathbb{E}_F L(\mathbf{c}) = \int_0^1 F^{\leftarrow}(u) dF_{\mathbf{c}}(u).$$

Zatem podobnie jak wzór (3.12) otrzymujemy wzór

$$\mathbb{E}_F L(\mathbf{c}) - F^{\leftarrow}(p) = \int_0^1 \left[F^{\leftarrow}(u) - \mu_F\right] dH_{\mathbf{c}}(u),$$

gdzie $H_{\mathbf{c}}(u) = F_{\mathbf{c}}(u) - \mathbf{I}_{[p,1)}(u)$. Podobnie jak (3.14) z nierówności Moriguti'ego i Schwarza otrzymujemy

$$\mathbb{E}_F L(\mathbf{c}) - F^{\leftarrow}(p) \leq \sigma_F \|\overline{H}'_{\mathbf{c}}\|,$$

a więc $\overline{B}_p(\mathbf{c}) = \|\overline{H}'_{\mathbf{c}}\|$, gdzie $\overline{H}_{\mathbf{c}}$ oznacza największą wypukłą minorantę funkcji $H_{\mathbf{c}}$, a $\overline{H}'_{\mathbf{c}}$ oznacza jej pochodną.

Aby wyznaczyć te funkcje potrzebna jest znajomość przedziałów wypukłości i wklęsłości funkcji $F_{\mathbf{c}}$, a więc zmian znaku pochodnej $f'_{\mathbf{c}}$. Niestety da się je wyznaczyć efektywnie jedynie dla $\mathbf{c} \in \mathcal{C}_n^{(\wedge)}$. Istotnie, mamy $f_{j:n}(u) = nB_{j-1,n-1}(u)$, a więc

$$f_{\mathbf{c}}(u) = n\sum_{j=1}^{n} c_j B_{j-1,n-1}(u) = n\sum_{j=0}^{n-1} c_{j+1} B_{j,n-1}(u).$$

Stosując podany w Dodatku wzór (A.7) na pochodną wielomianu Bernsteina otrzymujemy po prostych przekształceniach

$$f'_{\mathbf{c}}(u) = n(n-1)\sum_{j=0}^{n-2}(c_{j+2} - c_{j+1}) B_{j,n-2}(u).$$

W ogólnym przypadku, gdy $\mathbf{c} \in \mathcal{C}_n$, wyznaczenie zmian znaku tej funkcji wydaje się zadaniem wręcz niewykonalnym. Jednakże, dzięki szczególnej własności wielomianów Bernsteina, jest to możliwe dla $\mathbf{c} \in \mathcal{C}_n^{(\wedge)}$. Zauważmy, że jeśli $\mathbf{c} \in \mathcal{C}_n^{(\wedge)}$, gdzie $n \geq 2$ to zachodzi jeden z przypadków:

(A) $n \geq 4$ oraz istnieje $k \in \{2, \ldots, n-2\}$ takie, że $c_1 \leq \cdots \leq c_k$ oraz $c_{k+1} \geq \cdots \geq c_n$ przy czym w każdej nierówności zachodzi przynajmniej jedna nierówność ostra, lub $n = 3$ oraz $c_1 < c_2$ i $c_2 > c_3$;

(B) $n \geq 2$ oraz $c_1 \geq \cdots \geq c_n$;

(C) $n \geq 2$ oraz $c_1 \leq \cdots \leq c_n$ i przynajmniej jedna nierówność jest ostra.

W przypadku (A), jeśli dla $0 \leq j \leq n-2$ określimy

$$d_j = c_{j+2} - c_{j+1},$$

to

$$d_0, \ldots, d_{k-2} \geq 0, \quad d_k, \ldots, d_{n-2} \leq 0.$$

Ponadto przynajmniej jedna z liczb $d_0, \ldots, d_{k-2}$ jest dodatnia, i co najmniej jedna z liczb $d_k, \ldots, d_{n-2}$ jest ujemna. Na mocy własności VDP wielomianów Bernsteina, sformułowanej w Dodatku A jako Lemat A.1, funkcja $f'_{\mathbf{c}}$ zmienia znak tylko raz, z dodatniego na ujemny. Zatem gęstość $f_{\mathbf{c}}$ najpierw rośnie, a później maleje, a dystrybuanta $F_{\mathbf{c}}$ jest ściśle rosnąca, najpierw wypukła, a potem wklęsła. Dla takich funkcji istnieje dokładnie jedno rozwiązanie $\theta_{\mathbf{c}}$ równania

$$1 - F_{\mathbf{c}}(\theta) = (1-\theta) f_{\mathbf{c}}(\theta)$$

oraz dokładnie jedno rozwiązanie $\xi_{\mathbf{c}}$ równania

$$F_{\mathbf{c}}(\xi) = \xi f_{\mathbf{c}}(\xi).$$

Zauważmy, że funkcja $F_{\mathbf{c}}$ ma własności wypukłości analogiczne jak funkcja $F_{j:n}$, $2 \leq j \leq n-1$. Zatem aby wyznaczyć wzór na $\overline{B}_p(\mathbf{c}) = \|\overline{H}'_{\mathbf{c}}\|$ możemy stosować argumenty z dowodu Twierdzenia 3.1(a). Otrzymamy wtedy, że jeśli $p \geq \theta_{\mathbf{c}}$, to

$$\overline{B}_p(\mathbf{c}) = \frac{1 - F_{\mathbf{c}}(p)}{\sqrt{p(1-p)}}, \tag{3.19}$$

a oszacowanie jest osiągalne dla rozkładu dwupunktowego (3.7). W przeciwnym razie, gdy $p < \theta_{\mathbf{c}}$, to otrzymamy

$$\overline{B}_p(\mathbf{c}) = \left(\frac{(1 - F_{\mathbf{c}}(p))^2}{p} + \int_p^{\theta_{\mathbf{c}}} (f_{\mathbf{c}}(u))^2 du + \frac{(1 - F_{\mathbf{c}}(\theta_{\mathbf{c}}))^2}{1 - \theta_{\mathbf{c}}} \right)^{1/2}.$$

Oszacowanie to jest osiągnięte dla rozkładu typu (3.9) z odpowiednimi zmianami.

W przypadku (B) ciąg $\{c_j\}$ jest nierosnący, a więc wszystkie liczby $d_0, \ldots, d_{n-2}$ są niedodatnie. Zatem, na mocy własności VDP pochodna $f'_{\mathbf{c}}$ jest niedodatnia, a więc $f_{\mathbf{c}}$ jest nierosnąca. Stąd $F_{\mathbf{c}}$ jest wklęsła i możemy stosować rozumowanie z dowodu Twierdzenia 3.1(b). Zauważmy, że jeśli $c_1 = \cdots = c_n = \frac{1}{n}$, to $F_{\mathbf{c}}(u) = u$ jest funkcją liniową. W każdym przypadku otrzymujemy oszacowanie

$$\overline{B}_p(\mathbf{c}) = \frac{1 - F_{\mathbf{c}}(p)}{\sqrt{p(1-p)}},$$

które jest osiągnięte dla rozkładu (3.7).

Analogicznie, w przypadku (C) ciąg $\{c_j\}$ jest niemalejący, a więc w konsekwencji $F_{\mathbf{c}}$ jest funkcją wypukłą. Stosując rozumowanie z dowodu Twierdzenia 3.1(c) otrzymamy oszacowanie

$$\overline{B}_p(\mathbf{c}) = \left(\frac{(1 - F_{\mathbf{c}}(p))^2}{p} + \int_p^1 (f_{\mathbf{c}}(u))^2 du \right)^{1/2},$$

które jest osiągnięte dla rozkładu typu (3.11).

Aby wyznaczyć wartości dolnych oszacowań $\underline{B}_p(\mathbf{c})$ wystarczy skorzystać z symetrii $\underline{B}_p(\mathbf{c}) = -\overline{B}_{1-p}(\underline{\mathbf{c}})$, którą łatwo udowodnić postępując podobnie jak w dowodzie Lematu 2.5. Ponadto, należy skorzystać z równości $\xi_{\mathbf{c}} = 1 - \theta_{\underline{\mathbf{c}}}$ (por. (3.5)). Na przykład w przypadku (A) dla $p \leq \xi_{\mathbf{c}}$ mamy

$$\underline{B}_p(\mathbf{c}) = -\frac{F_{\mathbf{c}}(p)}{\sqrt{p(1-p)}}. \tag{3.20}$$

Uwaga 3.3. Dla wektorów $\mathbf{c}$, które nie należą do klasy $\mathcal{C}_n^{(\wedge)}$, nie umiemy precyzyjnie ustalić wypukłości funkcji $F_{\mathbf{c}}$, a więc nie umiemy podać jawnych wzorów na $\overline{B}_p(\mathbf{c})$ i $\underline{B}_p(\mathbf{c})$.

Uwaga 3.4. Nawet jeśli $\mathbf{c} \in \mathcal{C}_n^{(\wedge)}$, to dla ustalonego p nie umiemy ustalić, która z nierówności $p \geq \theta_{\mathbf{c}}$ lub $p < \theta_{\mathbf{c}}$ zachodzi. Jednakże jest to możliwe w większości przypadków jeśli $\mathbf{c} = (0, \ldots, 0, 1-\alpha, \alpha, 0, \ldots, 0)$.

Podsumowując, widzimy, że stosowanie w dalszych rozważaniach oszacowań $\overline{B}_p(\mathbf{c})$ i $\underline{B}_p(\mathbf{c})$ znacznie by je utrudniło. Dlatego w dalszym ciągu zamiast nich stosujemy $\overline{b}_p(\mathbf{c})$ i $\underline{b}_p(\mathbf{c})$ określone wzorami (2.7) i (2.8).

Rozdział 4

Wybór pojedynczej statystyki porządkowej

Stosując kryterium wprowadzone w Rozdziale 2, w tym rozdziale wyznaczymy optymalny estymator kwantyla zadanego rzędu $p \in (0,1)$ w postaci pojedynczej statystyki porządkowej. Przy ustalonym rozmiarze próby n pokażemy, która ze statystyk $X_{\lfloor np \rfloor:n}$, $X_{\lceil np \rceil:n}$ lub $X_{\lceil np \rceil+1:n}$ jest lepsza. Dokładnie mówiąc, udowodnimy, że istnieją jednoznacznie określone liczby $a_{j,n}$, $1 \leq j < n$, takie, że jeżeli $p \in (a_{j-1,n}, a_{j,n})$, to optymalnym wyborem jest $X_{j:n}$, a jeżeli $p = a_{j,n}$, to obydwie statystyki $X_{j:n}$ oraz $X_{j+1:n}$ są równie dobre (zob. Twierdzenie 4.4). Ponadto liczby te spełniają warunki $a_{j,n} < \frac{j}{n} < a_{j+1,n}$ dla $1 \leq j < \frac{n}{2}$ oraz $a_{n-j,n} = 1 - a_{j,n}$. W szczególności, w odróżnieniu od klasycznego wyboru pokażemy, że najlepszym estymatorem kwantyla $x_{k/n}$ jest $X_{k:n}$ jeżeli $1 \leq k < \frac{n}{2}$ lub $X_{k+1:n}$ jeżeli $\frac{n}{2} < k \leq n$. W przypadku gdy $k = \frac{n}{2}$, obie statystyki $X_{n/2:n}$ oraz $X_{n/2+1:n}$ są równie dobrymi estymatorami.

W podrozdziale 4.4 rozważymy również inne kryterium optymalności pojedynczej statystyki porządkowej oparte o minimalizację maksymalnego obciążenia. Pokażemy, że dla większości wartości parametru p obydwa rozważane kryteria są równoważne. W ostatnim podrozdziale przedyskutujemy otrzymane wyniki i podamy przykłady numeryczne.

Wyniki tego rozdziału zostały zamieszczone w pracy [4].

4.1 Kryterium optymalności

Tak jak w podrozdziale 2.1 rozważmy klasę rozkładów $\mathcal{F}_2$ dystrybuant F z wartością oczekiwaną μ_F oraz skończoną wariancją σ_F^2. Ponadto przypomnijmy, że

$$\mathcal{C}_n^{(1)} = \{\mathbf{c} \in \mathcal{C}_n : \exists j \quad c_j = 1\},$$

oraz $\boldsymbol{\delta}_j$ oznacza wektor $\mathbf{c} \in \mathcal{C}_n$ taki, że $c_j = 1$ oraz $c_i = 0$ dla $i \neq j$. Oczywiście wtedy $\mathcal{C}_n^{(1)} = \{\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_n\}$. Rozważamy pojedyncze statystyki porządkowe, a więc L–statystyki $L(\mathbf{c})$, gdzie $\mathbf{c} \in \mathcal{C}_n^{(1)}$. W tym przypadku oczywiście $\overline{b}_p(\mathbf{c}) = \overline{B}_p(\mathbf{c})$ oraz $\underline{b}_p(\mathbf{c}) = \underline{B}_p(\mathbf{c})$, a więc funkcja $s_p(\mathbf{c})$ przyjmuje wtedy postać

$$s_{n,p}(j) = \sqrt{p(1-p)}\left|\overline{B}_{n,p}(j) + \underline{B}_{n,p}(j)\right|.$$

Zatem Definicja 2.1 przyjmuje następującą postać.

Definicja 4.1. Statystyka porządkowa $X_{j_0:n}$ jest optymalnym estymatorem kwantyla rzędu p rozkładu $F \in \mathcal{F}_2$ jeżeli

$$s_{n,p}(j_0) = \min_{1<j<n} s_{n,p}(j).$$

Zatem statystyka porządkowa $X_{k:n}$ jest optymalnym estymatorem kwantyla x_p jeżeli k minimalizuje funkcję $s_{n,p}$ względem zmiennej $j \in \{1, \ldots, n\}$.

Twierdzenie 4.1. *Optymalnym wyborem liczby k dla której funkcja $s_{n,p}$ przyjmuje wartość najmniejszą jest:*

(a) jeżeli $|r_{n,p}(1)| < r_{n,p}(2)$, to $k = 1$, oraz jeżeli $r_{n,p}(n-1) > |r_{n,p}(n)|$, to $k = n$;

(b) w przeciwnym wypadku $k = j$ lub $k = j+1$, gdzie $j \in \{2, \ldots, n-2\}$ jest jedynym rozwiązaniem równania

$$r_{n,p}(j) \leq 0 < r_{n,p}(j+1); \tag{4.1}$$

(c) w szczególności, jeżeli

$$r_{n,p}(j) = -\,r_{n,p}(j+1), \tag{4.2}$$

to $s_{n,p}(j) = s_{n,p}(j+1)$ oraz obie wartości j oraz $j+1$ są równie dobre.

Dowód. Jeżeli $r_{n,p}(1) \geq 0$ to z faktu, że funkcja $r_{n,p}$ jest rosnąca i monotoniczna, optymalnym wyborem jest indeks $k = 1$. Jeżeli $r_{n,p}(1) < 0 < r_{n,p}(2)$ ale $-r_{n,p}(1) < r_{n,p}(2)$ to znów optymalnym wyborem indeksu jest $k = 1$. To dowodzi punktu (a). Jeżeli j jest określone równaniem (4.1), to $s_{n,p}$ jest funkcją nieujemną, która najpierw ściśle maleje, a potem ściśle rośnie, co dowodzi punktów (b) i (c). □

Zdefiniujmy funkcję $k(n,p)$, której wartościami są podzbiory zbioru $\{1, 2, \ldots, n\}$ zawierające wszystkie wartości indeksów minimalizujących funkcję $s_{n,p}$. Zatem

$$k(n,p) = \left\{1 \leq j_0 \leq n : s_{n,p}(j_0) = \min_{1\leq j\leq n} s_{n,p}(j)\right\}.$$

Powyższe twierdzenie pokazuje, że wartościami funkcji $k(n,p)$ są zbiory jednoelementowe, o ile nie jest spełniony warunek (4.2). W takim przypadku mamy $k(n,p) = \{j, j+1\}$. Dla uproszczenia notacji będziemy stosować zapis $k(n,p) = j$ zamiast $k(n,p) = \{j\}$.

Będziemy porównywać wartości indeksów k dla ustalonych wartości n oraz różnych wartości p, i w szczególności piszemy $k(n,p) \leq k(n,q)$, jeśli nierówność ta jest spełniona w zwykłym sensie dla wszystkich wartości funkcji $k(n,p)$ oraz $k(n,q)$. Piszemy również $k(n,p) \leq j$ jeśli każdy element zbioru $k(n,p)$ jest nie większy niż j. Następny lemat nieformalnie mówi, że $k(n,p)$ jest niemalejącą funkcją zmiennej $p \in (0,1)$.

Lemat 4.1. *Dla $0 < p < q < 1$ mamy nierówność $k(n,p) \leq k(n,q)$.*

Dowód. Załóżmy, że $0 < p < q < 1$. Na mocy Wniosku 3.3 mamy $r_{n,q}(j) < r_{n,p}(j)$ dla $1 \leq j \leq n$. Załóżmy, że dla pewnych liczb k oraz ℓ zachodzą nierówności

$$r_{n,p}(k) \leq 0 < r_{n,p}(k+1)$$

oraz $r_{n,q}(\ell) \leq 0 < r_{n,q}(\ell+1)$. Gdyby zachodziła nierówność $\ell < k$, to $\ell + 1 \leq k$, a więc otrzymalibyśmy

$$r_{n,q}(\ell+1) \leq r_{n,q}(k) < r_{n,p}(k) \leq 0,$$

co daje sprzeczność. Zatem pokazaliśmy, że jeśli k jest minimalizuje funkcję $s_{n,p}$ oraz ℓ jest wartością, która minimalizuje funkcję $s_{n,q}$, to zachodzi nierówność $k \leq \ell$. Ponadto, jeżeli $k = 1$ jest optymalnym wyborem dla rzędu kwantyla q, to $\ell = 1$ jest optymalnym wyborem dla rzędu kwantyla p. □

Z powyższego lematu dla dowolnie ustalonego $n \geq 2$ otrzymujemy istnienie zbioru $n-1$ liczb $a_{1,n}, \ldots, a_{n-1,n}$ o następujących własnościach:

(a) $0 = a_{0,n} < a_{1,n} < \ldots < a_{n-1,n} < a_{n,n} = 1$;

(b) $k(n, a_{j,n}) = \{j, j+1\}$ dla $1 \leq j \leq n-1$;

(c) $k(n,p)$ ma stałą wartość $j+1$ dla $p \in (a_{j,n}, a_{j+1,n})$ oraz $0 \leq j \leq n-1$.

Zatem problem znalezienia najbardziej optymalnego estymatora w postaci pojedynczej statystyki porządkowej sprowadza się do (a) wyznaczenia liczb $a_{j,n}$ oraz (b) określenia ich położenia względem liczb $\frac{j}{n}$, gdzie $1 \leq j \leq n$.

Na początku pokażemy, że $a_{j,n} = 1 - a_{n-j,n}$ dla $0 \leq j \leq n$. Jest to prosty wniosek z następującego lematu.

Lemat 4.2. *Dla $p \in (0,1)$ mamy:*

(a) $k(n,p) = j$ wtedy i tylko wtedy, gdy $k(n, 1-p) = n - j + 1$;

(b) $k(n,p) = \{j, j+1\}$ wtedy i tylko wtedy, gdy $k(n, 1-p) = \{n-j, n-j+1\}$.

Dowód. Tezy lematu wynikają łatwo z Lematu 2.7, gdyż zbiór $\mathcal{C}_n^{(1)}$ oczywiście spełnia warunek (2.19). □

Teraz pokażemy, że jeżeli rozmiar próby n jest liczbą parzystą, to $a_{n/2,n} = \frac{1}{2}$, zaś jeżeli n jest liczbą nieparzystą to $a_{(n-1)/2,n} < \frac{1}{2} < a_{(n+1)/2,n}$. Jest to prosty wniosek wynikający z wartości $k(n, 1/2)$, które są podane w lemacie poniżej.

Lemat 4.3. *Dla $n \geq 2$ mamy*

$$k\left(n, \frac{1}{2}\right) = \begin{cases} \dfrac{n+1}{2}, & \text{jeżeli } n \text{ jest liczbą nieparzystą,} \\ \left\{\dfrac{n}{2}, \dfrac{n}{2}+1\right\}, & \text{jeżeli } n \text{ jest liczbą parzystą.} \end{cases}$$

Dowód. Załóżmy najpierw, że n jest nieparzyste i rozważmy wektor $\mathbf{c} = \boldsymbol{\delta}_{\frac{n+1}{2}}$. Wtedy $\mathbf{c}$ jest jedynym wektorem w zbiorze $\mathcal{C}_n^{(1)}$, dla którego $\underline{\mathbf{c}} = \mathbf{c}$. Z Lematu 2.6(c) mamy $r_{1/2}\left(\boldsymbol{\delta}_{\frac{n+1}{2}}\right) = 0$ i jest to jedyny wektor o tej własności.

Jeżeli natomiast n jest parzyste, to $\underline{\boldsymbol{\delta}}_{\frac{n}{2}} = \boldsymbol{\delta}_{\frac{n}{2}+1}$, a więc z Lematu 2.6(c) dostajemy $r_{1/2}\left(\boldsymbol{\delta}_{\frac{n}{2}+1}\right) = -r_{1/2}\left(\boldsymbol{\delta}_{\frac{n}{2}}\right)$. Zatem

$$r_{n,1/2}\left(\frac{n}{2}\right) < 0 < r_{n,1/2}\left(\frac{n}{2}+1\right),$$

i na mocy Twierdzenia 4.1(c) otrzymujemy tezę Lematu. □

Uwaga 4.1. Jeżeli n jest liczbą nieparzystą, to nasze kryterium optymalności daje klasyczny estymator mediany $X_{\frac{n+1}{2}:n}$. Jeżeli n jest liczbą parzystą, to obydwa wybory $X_{\frac{n}{2}:n}$ oraz $X_{\frac{n}{2}+1:n}$ jako estymatora mediany są równie dobre. W tym przypadku używając estymatorów w postaci jednej statystyki porządkowej nie otrzymamy klasycznego estymatora dla mediany, który jest średnią arytmetyczną dwóch środkowych statystyk porządkowych. Dlatego właśnie w następnym rozdziale zajmujemy się wyznaczeniem estymatorów w postaci kombinacji liniowej dwóch sąsiednich statystyk porządkowych.

Gdy wiemy już jaki jest najlepszy estymator kwantyla $x_{1/2}$, to możemy założyć, że $p \neq \frac{1}{2}$. Z Lematu 4.2 widzimy, że jest wystarczy rozważać $k(n, p)$ tylko dla $p \in \left(0, \frac{1}{2}\right)$.

Lemat 4.4. *Dla $p \in \left(0, \frac{1}{2}\right)$ zachodzi nierówność $k(n, p) \leq \left\lfloor \frac{n+1}{2} \right\rfloor$.*

Dowód. Z Lematu 4.3 otrzymujemy $k(n, 1/2) \subset \left\{\left\lfloor \frac{n+1}{2} \right\rfloor, \left\lfloor \frac{n+1}{2} \right\rfloor + 1\right\}$. Korzystając z Lematu 4.1 dostajemy tezę lematu. □

Na mocy powyższych lematów widzimy, że rozwiązanie problemu optymalnego wyboru pojedynczej statystyki porządkowej jako estymatora kwantyla sprowadza się do wyznaczenia liczb $a_{k,n}$ dla $1 \leq k \leq \frac{n}{2}$.

Zauważmy, że w trywialnym przypadku gdy $n = 2$ mamy $a_{1,2} = \frac{1}{2}$, więc $k(n, p) = 1$ dla $p \in (0, \frac{1}{2})$ oraz $k(n, p) = 2$ dla $p \in (\frac{1}{2}, 1)$. Innymi słowy, w tym przypadku optymalnym estymatorem kwantyla x_p rzędu $p \neq \frac{1}{2}$ jest $X_{\lceil np \rceil:n}$. Jest to $X_{1:2}$ dla $p \in (0, \frac{1}{2})$ oraz $X_{2:2}$ dla $p \in (\frac{1}{2}, 1)$. Zatem w dalszym ciągu możemy założyć, że $n \geq 3$.

4.2 Pomocnicze liczby p_j i q_j

Zdefiniujemy teraz pomocnicze funkcje oraz liczby, które będą używane w rozwiązaniu głównego problemu tego rozdziału. Dla ustalonych wartości $n \geq 2$ oraz $1 \leq j \leq n$ zdefiniujmy $p_j = p_j(n)$ jako jedyne rozwiązanie równania

$$F_{j:n}(p) = \frac{1}{2}.$$

Zauważmy, że p_j jest oczywiście medianą j–tej statystyki porządkowej z rozkładu jednostajnego $U_{j:n}$. Jednoznaczność wyboru p_j wynika łatwo z faktu, że $F_{j:n}$ jest funkcją ciągłą i ściśle rosnącą i $F_{j:n}(0) = 0$ oraz $F_{j:n}(1) = 1$. Ponieważ $F_{j:n} > F_{j+1:n}$ na przedziale $(0, 1)$, to jest oczywiste, że $p_1 < \ldots < p_n$. Ze wzoru (3.4) wynika łatwo równość

$$p_{n-j+1} = 1 - p_j. \tag{4.3}$$

Te liczby są ważne w dalszych rozważaniach, gdyż na mocy (3.18) dla ustalonego $1 \leq j \leq n$ mamy $t_{n,p_j}(j) = 0$. Ponadto $t_{n,p}(j) > 0$ dla $p \in (0, p_j)$ oraz $t_{n,p}(j) < 0$ dla $p \in (p_j, 1)$.

Naszym głównym problemem jest znalezienie liczb $a_{j,n}$, $1 \leq j \leq \left\lfloor \frac{n-1}{2} \right\rfloor$, zdefiniowanych po Lemacie 4.1. Pokażemy, że dla $3 \leq j \leq n-2$ wartości liczb $a_{j,n}$ są równe wartościom liczb $q_j(n)$, które są zdefiniowane następująco. Dla $1 \leq j \leq n-1$ niech $q_j = q_j(n)$ będzie jedynym rozwiązaniem równania

$$|1 - 2F_{j:n}(q)| = |1 - 2F_{j+1:n}(q)| \tag{4.4}$$

w przedziale (p_j, p_{j+1}). Aby pokazać jednoznaczność liczb q_j zauważmy, że jeżeli

$$Q_{j,n}(p) = |1 - 2F_{j:n}(p)| - |1 - 2F_{j+1:n}(p)| \tag{4.5}$$

to funkcja $Q_{j,n}$ jest równa

$$Q_{j,n}(p) = \begin{cases} -2B_{j+1,n}(p), & \text{jeśli } 0 \leq p < p_j, \\ B_{j+1,n}(p), & \text{jeśli } p_{j+1} < p \leq 1, \\ 2(F_{j:n}(p) + F_{j+1:n}(p) - 1), & \text{jeśli } p_j \leq p \leq p_{j+1}. \end{cases} \tag{4.6}$$

Z definicji liczb p_j otrzymujemy $Q_{j,n}(p_j) < 0 < Q_{j,n}(p_{j+1})$. Ponieważ $Q_{j,n}$ jest ciągła i ściśle rosnąca na przedziale $[p_j, p_{j+1}]$ to implikuje, że q_j jest zdefiniowane jednoznacznie jako rozwiązanie równania $F_{j:n}(q_j) + F_{j+1:n}(q_j) = 1$. Ponadto ze wzoru (3.4) otrzymujemy równość

$$q_{n-j} = 1 - q_j. \tag{4.7}$$

Teraz zbadamy kilka własności zdefiniowanych powyżej liczb oraz ich wzajemne relacje z liczbami θ_j i ξ_j zdefiniowanymi w podrozdziale 3.1. Naszym celem jest pokazanie, że dla większości przypadków mamy nierówność $\theta_{j+1} < q_j < \xi_j$. Powodem,

dla którego chcemy otrzymać taką nierówność jest fakt, że wynika z niej, że wartości funkcji $s_{n,q_j}(j)$ oraz $s_{n,q_j}(j+1)$ mają prostą analityczną postać. Dzięki temu możemy łatwo je porównywać. Dowody kolejnych lematów są dość techniczne zatem zostały zamieszczone w Dodatku A.1.

Na początek zbadamy zależności pomiędzy liczbami p_j, q_j oraz $\frac{j}{n}$. Zatem lemat ten będzie pomocny do ustalenia położenia punktów $a_{j,n}$ względem liczb $\frac{j}{n}$.

Lemat 4.5. *Dla $1 \leq j \leq n-1$ mamy*

(a) jeżeli $1 \leq j < \frac{n}{2}$, to

$$p_j < \frac{j}{n} < q_j < p_{j+1}; \tag{4.8}$$

(b) jeżeli $\frac{n}{2} < j \leq n-1$, to

$$p_j < q_j < \frac{j}{n} < p_{j+1}; \tag{4.9}$$

(c) jeżeli $j = \frac{n}{2}$, to

$$p_{n/2} < q_{n/2} = \frac{1}{2} < p_{n/2+1}.$$

Następnym krokiem jest zbadanie nierówności pomiędzy liczbami θ_j, ξ_j oraz p_j i q_j. Na początku zauważmy, że ze wzorów (3.4) oraz (3.5) dla $2 \leq j \leq n-1$, łatwo otrzymujemy równoważność

$$\theta_j < p_j \Leftrightarrow p_{n-j+1} < \xi_{n-j+1}. \tag{4.10}$$

Lemat 4.6.

(a) Dla ustalonego $n \geq 3$ ciągi liczb $\theta_j(n)$ oraz $\xi_j(n)$ są ściśle rosnące względem zmiennej $j = 1, \ldots, n$.

(b) Dla $j = 3$ oraz $n \geq 7$ lub $j = 2$ oraz $n = 5$ mamy $q_j(n) < \xi_j(n)$, ale dla $j = 2$ oraz $n = 6$ mamy $\xi_2(6) < q_2(6)$.

(c) Dla $j = 2$ oraz $n \geq 7$ mamy $\xi_2(n) < \frac{2}{n}$, ale dla $n = 6$ mamy $\frac{2}{6} < \xi_2(6)$.

(d) Dla $n \geq 3$ mamy $p_2(n) < \xi_2(n)$ oraz dla $n \geq 5$ mamy $p_3(n) < \xi_3(n)$.

(e) Dla $n \geq 3$ mamy $\xi_2(n) < p_3(n)$ oraz dla $n \geq 10$ mamy $\xi_3(n) < p_4(n)$, ale dla $n \leq 9$ mamy $\xi_3(n) > p_4(n)$.

(f) Dla $4 \leq j \leq n-1$ mamy $p_{j+1}(n) < \xi_j(n)$.

(g) Dla $1 \leq j \leq n-4$ mamy $\theta_{j+1}(n) < p_j(n)$.

Aby ułatwić korzystanie z Lematu 4.6, sformułujemy następujący wniosek.

Wniosek 4.1.

(a) $\left[\frac{1}{n}, q_1\right] \subset (p_1, p_2) \subset (\theta_2, \xi_2)$ *dla* $n \geq 3$*;*

(b) $\frac{2}{6} \in (\theta_3, \xi_2)$ *oraz* $q_2 \in (\xi_2, \xi_3) \subset (\theta_3, \xi_3)$ *dla* $n = 6$*;*

(c) $\left[\frac{2}{n}, q_2\right] \subset (\xi_2, p_3) \subset (p_2, p_3) \subset (\theta_3, \xi_3)$ *dla* $n \geq 7$*;*

(d) $q_j \in (\theta_{j+1}, \xi_j) \cap (p_j, p_{j+1})$ *dla* $n = 5$ *oraz* $j = 2$ *lub* $n \geq 7$ *oraz* $3 \leq j \leq \left\lfloor \frac{n-1}{2} \right\rfloor$.

Dowód. (a) Kładąc $j = 1$ w Lemacie 4.5(a) otrzymujemy $p_1 < \frac{1}{n} < q_1 < p_2$. Ponadto z części (g) Lematu 4.6 mamy $\theta_2 < p_1$, a z części (d) mamy $\xi_2 > p_2$. Łącząc te nierówności otrzymujemy

$$\theta_2 < p_1 < \frac{1}{n} < q_1 < p_2 < \xi_2.$$

(b) Z własności liczb q_j mamy $q_2 < p_3$. Z części (d) Lematu 4.6 dostajemy $p_3 < \xi_3$, a z części (g) $\theta_3 < p_2$. Ponadto $p_2 < \frac{2}{6}$, a z części (b) i (c) Lematu 4.6 mamy odpowiednio $\xi_2 < q_2$ oraz $\frac{2}{6} < \xi_2$. Te nierówności dają

$$\theta_3 < \frac{2}{6} < \xi_2 \quad \text{oraz} \quad \theta_3 < \xi_2 < q_2 < \xi_3.$$

(c) Z Lematu 4.6(c) oraz (d) mamy $p_2 < \xi_2 < \frac{2}{n}$ oraz $p_3 < \xi_3$. Stosując ponownie Lemat 4.6(g) dostajemy $\theta_3 < p_2$. Z Lematu 4.5(a) mamy znowu $\frac{2}{n} < q_2 < p_3$. Łącząc te nierówności dostajemy

$$\theta_3 < p_2 < \xi_2 < \frac{2}{n} < q_2 < p_3 < \xi_3.$$

(d) Z definicji liczb q_j mamy $p_j < q_j < p_{j+1}$, czyli $q_j \in (p_j, p_{j+1})$. Z Lematu 4.6(b) mamy $q_2(5) < \xi_2(5)$ oraz $q_j(n) < \xi_j(n)$ dla $j = 3$ i $n \geq 7$. Ponadto dla $j \geq 4$ z Lematu 4.6(f) oraz (g) mamy $\theta_{j+1} < p_j < p_{j+1} < \xi_j$. Zatem w każdym przypadku $\theta_{j+1} < q_j < \xi_j$, czyli $q_j \in (\theta_{j+1}, \xi_j)$. □

4.3 Rozwiązanie problemu optymalnego wyboru

Teraz jesteśmy gotowi, aby zająć się problemem wyznaczenia liczb $a_{j,n}$. Główna idea rozwiązania naszego problemu jest następująca. Z Wniosku 3.4 mamy

$$s_{n,p}(j) \geq \sqrt{p(1-p)}\,|t_{n,p}(j)|$$

dla wszystkich $p \in (0, 1)$ oraz równość zachodzi wtedy i tylko wtedy, gdy $p \in (\theta_j, \xi_j)$. Zatem problem minimalizacji funkcji $s_{n,p}$ sprowadza się do analizy funkcji $t_{n,p}$. Na początek rozważmy przypadek $j \geq 3$, który okazuje się najprostszy.

Twierdzenie 4.2. *Dla $n \geq 7$ oraz $3 \leq j \leq \left\lfloor \frac{n-1}{2} \right\rfloor$ mamy $k(n, q_j) = \{j, j+1\}$, a więc $a_{j,n} = q_j(n)$. Ponadto, $k(5, q_2) = \{2, 3\}$, a więc $a_{2,5} = q_2(5)$.*

Dowód. Ponieważ $t_{n,p}$ jest ściśle rosnącą funkcją zmiennej j, zatem wystarczy udowodnić, że $s_{n,q_j}(j) = s_{n,q_j}(j+1)$ biorąc pod uwagę tezę Wniosku 4.1(d). Ponieważ $q_j \in (p_j, p_{j+1})$, to najpierw wnioskujemy, że $t_{n,q_j}(j) < t_{n,p_j}(j) = 0$ oraz $t_{n,q_j}(j+1) > t_{n,p_{j+1}}(j+1) = 0$. Następnie, ponieważ $q_j \in (\theta_j, \xi_j) \cap (\theta_{j+1}, \xi_{j+1})$, to otrzymujemy

$$s_{n,q_j}(j) = -\sqrt{p(1-p)}t_{n,q_j}(j) = 2F_{k:n}(q_j) - 1,$$

oraz

$$s_{n,q_j}(j+1) = \sqrt{p(1-p)}t_{n,q_j}(j+1) = 1 - 2F_{j+1:n}(q_j).$$

Ale z definicji liczby q_j (zob. (4.4)) prawe strony obu równości są równe. □

Teraz rozważmy bardziej skomplikowane przypadki $j = 1$ oraz $j = 2$. Dowód następnego twierdzenia jest podobny, ale znacznie bardziej techniczny. Dlatego został przeniesiony do Dodatku A.2.

Twierdzenie 4.3.

(a) Dla $n \geq 3$ mamy $k\left(n, \frac{1}{n}\right) = 1$ oraz $k(n, q_1) = 2$, a więc $a_{1,n} \in \left(\frac{1}{n}, q_1\right)$ jest jedynym rozwiązaniem równania $s_{n,p}(1) = s_{n,p}(2)$ względem zmiennej p.

(b) Dla $n \geq 6$ mamy $k\left(n, \frac{2}{n}\right) = 2$ oraz $k(n, q_2) = 3$, a więc $a_{2,n} \in \left(\frac{2}{n}, q_2\right)$ jest jedynym rozwiązaniem równania $s_{n,p}(2) = s_{n,p}(3)$ względem zmiennej p.

W sformułowaniu oraz dowodzie głównego wyniku używamy następujących oznaczeń. Dla $n \geq 3$ oznaczmy

$$M = M(n) = \left\lfloor \frac{n-1}{2} \right\rfloor, \quad N = N(n) = \left\lceil \frac{n+1}{2} \right\rceil.$$

Wtedy $M < \frac{n}{2} < N$ oraz $N = M + 1$ jeżeli n jest nieparzyste, oraz $M = \frac{n}{2} - 1$, $N = \frac{n}{2} + 1$ jeżeli n jest parzyste. Zdefiniujmy również podzbiory przedziału $(0, 1)$

$$\mathcal{P}_n = \bigcup_{j=0}^{M-1} \left(a_{j,n}, \frac{j+1}{n}\right) \cup \left(a_{M,n}, \frac{1}{2}\right), \quad \mathcal{Q}_n = \bigcup_{j=1}^{M} \left(\frac{j}{n}, a_{j,n}\right).$$

Ponadto, niech

$$\mathcal{P}'_n = \left(\frac{1}{2}, a_{N,n}\right) \cup \bigcup_{j=N}^{n-1} \left(\frac{j}{n}, a_{j+1,n}\right), \quad \mathcal{Q}'_n = \bigcup_{j=N}^{n-1} \left(a_{j,n}, \frac{j}{n}\right).$$

Biorąc pod uwagę symetrię $a_{n-j,n} = 1 - a_{j,n}$ możemy łatwo zauważyć, że $p \in \mathcal{P}'_n$ wtedy i tylko wtedy, gdy $1-p \in \mathcal{P}_n$, i podobne stwierdzenie zachodzi dla zbiorów $\mathcal{Q}_n$ oraz $\mathcal{Q}'_n$.

Twierdzenie 4.4. *Niech $n \geq 3$ oraz $p \neq \frac{1}{2}$. Optymalnym estymatorem kwantyla x_p rzędu p jest:*

(a) zarówno $X_{\lfloor np \rfloor:n}$, jak i $X_{\lceil np \rceil:n}$, jeżeli $p \in \{a_{1,n}, \ldots, a_{M,n}\}$, lub $X_{\lceil np \rceil:n}$ oraz $X_{\lceil np \rceil+1:n}$ jeżeli $p \in \{a_{N,n}, \ldots, a_{n-1,n}\}$;

(b) $X_{\lfloor np \rfloor:n} = X_{\lceil np \rceil:n}$, jeżeli $p \in \left\{\frac{1}{n}, \ldots, \frac{M}{n}\right\}$ oraz $X_{\lfloor np \rfloor+1:n} = X_{\lceil np \rceil+1:n}$ jeżeli $p \in \left\{\frac{N}{n}, \ldots, \frac{n-1}{n}\right\}$;

(c) $X_{\lceil np \rceil:n}$ jeżeli $p \in \mathcal{P}_n \cup \mathcal{P}'_n$, $X_{\lfloor np \rfloor:n}$ jeżeli $p \in \mathcal{Q}_n$ oraz $X_{\lceil np \rceil+1:n}$ jeżeli $p \in \mathcal{Q}'_n$.

Dowód. Dla $p \in (0, \frac{1}{2})$ z Twierdzeń 4.2 oraz 4.3 dostajemy

$$k(n,p) = \begin{cases} \{\lfloor np \rfloor, \lceil np \rceil\}, & \text{jeżeli } p \in \{a_{j,n} : 1 \leq j \leq M\}, \\ \lceil np \rceil = \lfloor np \rfloor, & \text{jeżeli } p \in \{\frac{j}{n} : 1 \leq j \leq M\}, \\ \lceil np \rceil, & \text{jeżeli } p \in \mathcal{P}_n, \\ \lfloor np \rfloor, & \text{jeżeli } p \in \mathcal{Q}_n. \end{cases}$$

Ponadto, z Lematu 4.2 oraz symetrii $\lfloor n(1-p) \rfloor = n - \lceil np \rceil$ możemy otrzymać wartości $k(n,p)$ dla $p \in \left(\frac{1}{2}, 1\right)$. Teza twierdzenia wynika teraz z definicji funkcji $k(n,p)$. □

Przypadek $p = \frac{1}{2}$ był rozważany w Uwadze 4.1. Ponadto, w szczególności przypadek $p = \frac{j}{n}$ otrzymamy z następującego wniosku.

Wniosek 4.2. *Optymalnym estymatorem kwantyla $x_{j/n}$ jest $X_{j:n}$ jeżeli $1 \leq j < \frac{n}{2}$ lub $X_{j+1:n}$ jeżeli $\frac{n}{2} < j < n$. Jeżeli $n \geq 2$ jest parzyste oraz $j = \frac{n}{2}$, to obie statystyki $X_{\frac{n}{2}:n}$ oraz $X_{\frac{n}{2}+1:n}$ są optymalnymi estymatorami kwantyla mediany.*

Dowód. Pokażemy, że dla $1 \leq j \leq \frac{n}{2}$ mamy $k(n, j/n) = j$. Dla $j = 1, 2$ wynika to z Twierdzenia 4.3. Dla $j \geq 3$ mamy $\frac{j}{n} \in (\theta_{j+1}, \xi_j)$ z Wniosku 4.1, gdyż $\frac{j}{n} < q_j$. Przypadek $j > \frac{n}{2}$ wynika z Lematu 4.2(a). □

Otrzymany wynik jest inny niż zazwyczaj podawany w literaturze. Najczęściej jako estymator kwantyla $x_{j/n}$ polecany jest $X_{j:n}$ dla wszystkich $1 \leq j \leq n$. Zaletą naszego podejścia jest własność symetrii, czyli optymalnym estymatorem kwantyla rzędu $p = \frac{n-1}{n}$ jest $X_{n:n}$, a nie $X_{n-1:n}$. Zauważmy, że $X_{n:n}$ jest symetrycznym wyborem do $X_{1:n}$ dla $p = \frac{1}{n}$. Własność symetrii jest bardzo pożądana w estymacji kwantyli (zob. [13]).

4.4 Równoważne kryterium: minimalizacja maksymalnego obciążenia

Inna motywacja wyboru naszego kryterium jest następująca. Jeżeli $F \in \mathcal{F}_2$ oraz x_p jest dowolną ustaloną wartością rzędu kwantyla p rozkładu F, to z definicji oszacowań (2.1) oraz (2.2) dla każdego ustalonego $j \in \{1, \ldots, n\}$

$$\left| \frac{\mathbb{E}_F X_{j:n} - x_p}{\sigma_F} \right| \leq \max\left(\overline{B}_{n,p}(j), -\underline{B}_{n,p}(j) \right) =: w_{p,n}(j). \tag{4.11}$$

Wartości funkcji $w_{p,n}(j)$ mogą być rozważane jako maksymalne oszacowania obciążenia statystyki porządkowej $X_{j:n}$ jako estymatora kwantyla x_p, mierzonego w jednostkach odchylenia standardowego rozkładu F, względem wszystkich dystrybuant $F \in \mathcal{F}_2$. Idea polega na wyborze indeksu j takiego, że maksymalne oszacowanie jest najmniejsze z możliwych. W ten sposób minimalizujemy największy możliwy błąd estymacji kwantyla rzędu p przez pojedyncze statystki porządkowe. W Twierdzeniu 4.5 pokażemy, że dla większości wartości rzędu kwantyla p, funkcje $s_{n,p}$ oraz $w_{n,p}$ przyjmują wartość najmniejszą dla dokładnie tego samego indeksu statystyki porządkowej. Stąd oba te kryteria są równoważne. Nie jest to zaskakujące, gdyż oba kryteria dążą do równowagi pomiędzy górnym i dolnym oszacowaniem, jednakże równoważność kryteriów nie jest oczywista.

W tym podrozdziale główną ideą jest wybór takiej wartości indeksu j, który minimalizuje maksymalne wartości obciążenia, czyli chcemy zminimalizować funkcję $w_{n,p}$ (zob. (4.11)) względem zmiennej j. Niech $\ell(n,p)$ oznacza zbiór wszystkich indeksów j, dla których funkcja $w_{n,p}$ przyjmuje najmniejszą wartość. Następne twierdzenie mówi, że dla $p \in (q_2, q_{n-2})$ funkcja $w_{n,p}$ jest minimalizowana dla tych samych wartości j, dla których funkcja $s_{n,p}$ przyjmuje wartości najmniejsze.

Twierdzenie 4.5. *Dla $n \geq 7$ oraz $p \in (q_2, q_{n-2})$ mamy $\ell(n,p) = k(n,p)$.*

Numeryczne obliczenia sugerują, że to twierdzenie jest prawdziwe dla wszystkich $n \geq 3$ oraz $p \in (0,1)$, jednakże nie mogliśmy znaleźć formalnego dowodu takiego stwierdzenia. W dowodzie twierdzenia używamy elementarnego lematu, który w łatwy sposób wynika z nierówności $\max(x,y) \leq z$ dla $x \leq z$ oraz $y \leq z$.

Lemat 4.7. *Dla dowolnych liczb rzeczywistych a, b, c oraz d takich, że $a < c$ oraz $b > d$ nierówność $\max(a,b) \leq \max(c,d)$ zachodzi wtedy i tylko wtedy, gdy $b \leq c$. Ponadto, $\max(a,b) \geq \max(c,d)$ jeśli $b \geq c$ oraz $\max(a,b) = \max(c,d)$ jeśli $b = c$.*

Dowód Twierdzenia 4.5. Ponieważ n oraz p są ustalone, to dla przejrzystości dowodu będziemy pisać $\overline{B}_j$ oraz $\underline{B}_j$ zamiast odpowiednio $\overline{B}_{n,p}(j)$ oraz $\underline{B}_{n,p}(j)$. Dla $1 \leq j \leq n-1$

n	10	20	50	100
S_n	0.0587	0.0694	0.0769	0.0798

Tabela 4.1: Niektóre przybliżone wartości S_n

oznaczmy

$$v_{n,p}(j) = \overline{B}_{j+1} + \underline{B}_j.$$

Na mocy Wniosku 3.3 funkcja $v_{n,p}$ jest ściśle rosnąca względem zmiennej j. Przyjmując $a = \overline{B}_j$, $b = -\underline{B}_j$, $c = \overline{B}_{j+1}$ oraz $d = -\underline{B}_{j+1}$, z Lematu 4.7 otrzymujemy

$$w_{n,p}(j) < w_{n,p}(j+1) \quad \Leftrightarrow \quad v_{n,p}(j) > 0,$$

dla $1 \le j \le n-1$. Jeśli odwrócimy pierwszą nierówność, to druga również zostanie odwrócona. Zatem $w_{n,p}(j) = w_{n,p}(j+1)$ wtedy i tylko wtedy, gdy $v_{n,p}(j) = 0$.

Używając symetrii (2.15) otrzymujemy, że $w_{n,1-p}(j) = w_{n,p}(n-j+1)$. Zatem wystarczy rozpatrzeć jedynie przypadek $p \le \frac{1}{2}$. Rozważamy dwa przypadki:

- $p \in (q_j, q_{j+1})$ dla pewnego $2 \le j \le \frac{n}{2}$;
- $p = q_j$ dla pewnego $3 \le j \le \frac{n}{2}$;

W pierwszym przypadku mamy $k(n,p) = j+1$ oraz $p \in (q_j, q_{j+1}) \subset [\theta_{j+1}, \xi_{j+1}]$. Przypomnijmy, że $Q_{j,n}(p) = 2(F_{j:n}(p) + F_{j+1:n}(p) - 1)$ dla $p \in [p_j, p_{j+1}]$ oraz $Q_{j,n}(p)$ ma wartości ujemne dla $p < q_j$ oraz dodatnie dla $p > q_j$. Łącząc wartości dolnych i górnych oszacowań obciążenia danych w Twierdzeniu 3.1 z (3.17) otrzymujemy

$$v_{n,p}(j) \le -\frac{Q_{j,n}(p)}{2\sqrt{p(1-p)}} \quad \text{oraz} \quad v_{n,p}(j+1) \ge -\frac{Q_{j+1,n}(p)}{2\sqrt{p(1-p)}}.$$

To implikuje, że $v_{n,p}(j) < 0$ gdy $p > q_j$, oraz $v_{n,p}(j+1) > 0$ gdy $p < q_{j+1}$. Dlatego $v_{n,p}(i)$ jest ujemna dla $i \le j$ oraz dodatnia dla $i \ge j+1$. Zatem $w_{n,p}(i)$ jest ściśle malejąca dla $1 \le i \le j$ i ściśle rosnąca dla $j+1 \le i \le n$, oraz $\ell(n,p) = j+1 = k(n,p)$.

W drugim przypadku mamy $k(n,q_j) = \{j, j+1\}$ oraz $p = q_j \in [\theta_{j+1}, \xi_j]$. Zauważmy, że dla wszystkich p w tym przedziale mamy $2\sqrt{p(1-p)}v_{n,p}(j) = -Q_{j,n}(p)$, więc $v_{n,q_j}(j) = 0$ oraz wartości $v_{n,q_j}(i)$ są ujemne dla $i < j$ oraz dodatnie dla $i > j$. Zatem $w_{n,q_j}(j) = w_{n,q_j}(j+1)$, $w_{n,q_j}(i)$ jest ściśle malejąca dla $i \le j$, oraz ściśle rosnąca dla $i \ge j+1$. Więc w_{n,q_j} przyjmuje wartość najmniejszą jednocześnie dla j oraz $j+1$ oraz $\ell(n,q_j) = \{j, j+1\} = k(n,q_j)$. □

4.5 Podsumowanie oraz przykłady numeryczne

W tym rozdziale zastosowaliśmy nowe kryterium do wyboru pojedynczej statystyki porządkowej jako estymatora nieznanego kwantyla x_p ustalonego rzędu $p \in (0,1)$ na

k	1	2	3	4	5	6	7	8
p_k	0.0451	0.1093	0.1743	0.2393	0.3045	0.3696	0.4348	0.5
k/n	0.0667	0.1333	0.2	0.2667	0.3333	0.4	0.4667	0.5333
q_k	0.0749	0.1407	0.2062	0.2715	0.3368	0.4021	0.4674	0.5326
θ_k	0	0.0051	0.0336	0.0779	0.1321	0.1934	0.2605	0.3326
ξ_k	0	0.1256	0.2332	0.3314	0.4229	0.5091	0.5905	0.6674

Tabela 4.2: Wartości liczb p_k, k/n, q_k, θ_k oraz ξ_k dla $n = 15$ oraz $1 \leq k \leq 8$.

podstawie losowej próby ustalonego rozmiaru n.

Główny wynik uzyskany przy użyciu tego podejścia (Twierdzenie 4.4) mówi, że w przeciwieństwie do tradycyjnego wyboru jako estymatora statystyki $X_{\lceil np \rceil:n}$ dla $p \in \mathcal{Q}_n$ oraz $p \in \mathcal{Q}'_n$ powinniśmy użyć odpowiednio statystyk $X_{\lfloor np \rfloor:n}$ oraz $X_{\lceil np \rceil+1:n}$. Ponieważ długość każdego przedziału będącego składową zbioru $\mathcal{Q}_n$ zbiega do 0 jeżeli n jest coraz większe, to może się wydawać, że nasze kryterium jest niepotrzebne. Jednakże, długość zbioru $\mathcal{Q}_n \cup \mathcal{Q}'_n$ wynosi

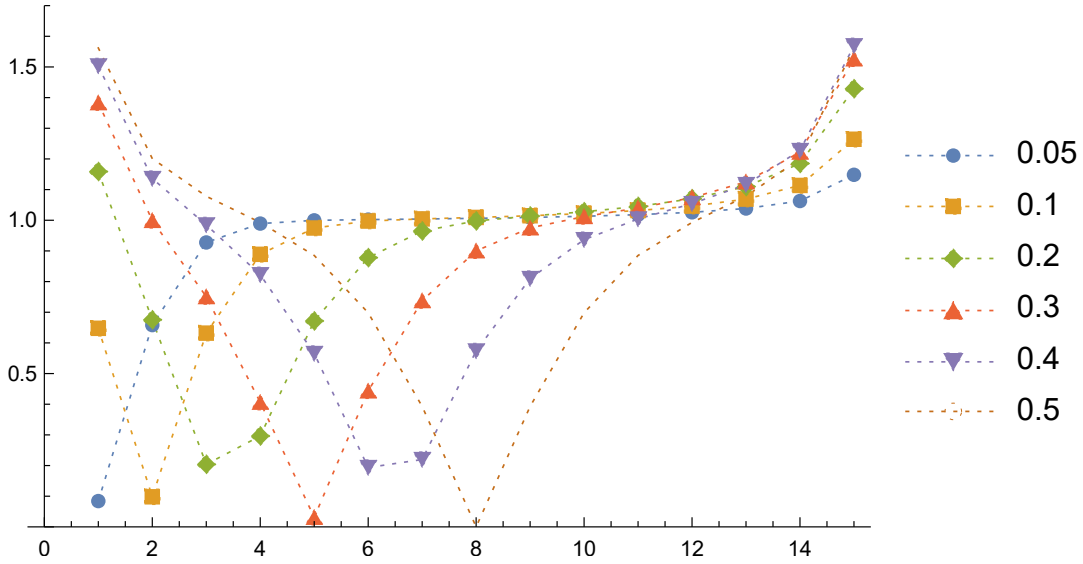
$$S_n = 2 \sum_{k=1}^{N} \left(a_{k,n} - \frac{k}{n} \right).$$

Dość trudno analizować analitycznie graniczne zachowanie ciągu S_n, ale numeryczne obliczenia pokazują, że wartości S_n rosną wraz ze wzrostem n, zob. Tabela 4.1. Niektóre przybliżone wartości zaprezentowane w tabeli pokazują, że proporcje rzędów kwantyli, dla których nasze przybliżenie jest lepsze niż tradycyjne znajduje się w zakresie 6-7% dla n od 10 do 100, więc nie powinno być pomijane.

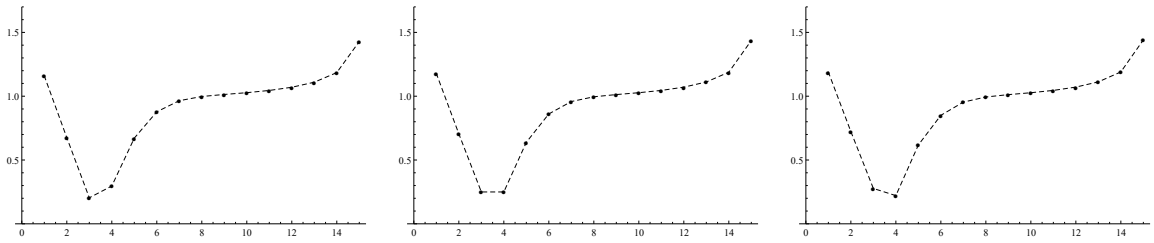
Przedstawiamy teraz obliczenia numeryczne, które ilustrują wyniki przedstawione w tym rozdziale. Najpierw zauważmy, że jawne wartości liczb θ_k, ξ_k, p_k oraz q_k są trudne do znalezienia, ale łatwo możemy wyliczyć ich przybliżone wartości w sposób numeryczny. Istotnie, są one pierwiastkami równań wielomianowych zawierających wielomiany Bernsteina, zob. (A.1), (3.5) oraz (A.4).

Na przykład, w Tabeli 4.2 prezentujemy wartości liczb p_k, $\frac{k}{n}$, q_k, θ_k oraz ξ_k dla $n = 15$ oraz $1 \leq k \leq 8$. Odpowiadające im wartości mogą być otrzymane dla $9 \leq k \leq 15$ przez skorzystanie z odpowiednich symetrii dotyczących liczb p_k, q_k oraz wzoru (3.5). Te wyniki potwierdzają większość wniosków z Lematów 4.5 oraz 4.6.

Wykres 4.1 prezentuje wartości funkcji $s_{15,p}(j)$, $1 \leq j \leq 15$ dla $p = 0.05, 0.1, \ldots, 0.5$. Wykres ten potwierdza tezy z Twierdzenia 4.4. Dokładniej mówiąc, mamy w tym przypadku $k(15, 0.05) = 1$, $k(15, 0.1) = 2$, $k(15, 0.2) = 3$, $k(15, 0.3) = 5$, $k(15, 0.4) = 6$ oraz $k(15, 0.5) = 8$. Na przykład, optymalnym estymatorem kwantyla $x_{0.4}$ z próby rozmiaru 15 jest $X_{6:15}$.



Rysunek 4.1: Wykresy $s_{15,p}(j)$, $1 \leq j \leq 15$ dla $p = 0.05, 0.1, 0.2, 0.3, 0.4, 0.5$.



Rysunek 4.2: Wykresy $s_{15,p}(j)$, $1 \leq j \leq 15$ dla $p = 3/15$ (lewy), $p = q_3(15)$ (środkowy) and $p = 0.21$ (prawy).

Wykres 4.2 pokazuje porównanie wartości funkcji $s_{n,p}(j)$, $1 \leq j \leq 15$ dla $p = 3/15$, $p = q_3(15) \approx 0.2062$ oraz $p = 0.21$. Wykres ten ilustruje, że w tym przypadku mamy $k(15, 0.2) = 3$, $k(15, q_3) = \{3, 4\}$ oraz $k(15, 0.21) = 4$.

Rozdział 5

Wybór kombinacji liniowej dwóch statystyk porządkowych

W tym rozdziale ponownie zastosujemy kryterium optymalności z podrozdziału 2.1, ale tym razem poszukujemy najlepszych estymatorów kwantyli postaci kombinacji liniowej dwóch sąsiednich statystyk porządkowych jak we wzorze (1.10).

Wyznaczymy wartości liczb $b_{k,n}$, $1 \leq k \leq n$, takie że jeśli $p \in (b_{k,n}, b_{k+1,n})$, to optymalnym estymatorem kwantyla x_p jest $(1-\alpha_{n,p})X_{k:n} + \alpha_{n,p}X_{k+1:n}$. Współczynnik $\alpha_{n,p}$ jest wyznaczony jednoznacznie tak, że $\alpha_{n,p} \in \left(0, \frac{1}{2}\right)$ wtedy i tylko wtedy, gdy $p \in (b_{k,n}, a_{k,n})$, gdzie liczby $a_{k,n}$ są opisane w poprzednim rozdziale. W szczególności, jeśli $p = a_{k,n}$, to optymalnym wyborem jest $\frac{1}{2}(X_{k:n} + X_{k+1:n})$. Jeżeli natomiast $p = b_{k,n}$ to $\alpha_{n,p} = 0$ i najlepszym wyborem jest pojedyncza statystyka porządkowa $X_{k:n}$. Zatem nasz wynik jest zgodny na przykład z klasyczną definicją mediany z próby (zob. Lemat 5.4 oraz Uwaga 5.3). Należy podkreślić, że w przeciwieństwie do znakomitej większości estymatorów znanych w literaturze (zobacz podrozdział 1.3), uzyskany poniżej estymator nie jest liniową interpolacją funkcji kwantylowej (zobacz Uwaga 5.4).

5.1 Kryterium optymalności

Oznaczmy przez

$$D_{j:n}^{(\alpha)} = (1-\alpha)X_{j:n} + \alpha X_{j+1:n}$$

gdzie $1 \leq j \leq n-1$ oraz $\alpha \in [0, 1)$, kombinację liniową dwóch sąsiednich statystyk porządkowych. W dalszym ciągu rozważamy klasę rozkładów $\mathcal{F}_2$ rozkładów o skończonej wariancji. Zauważmy teraz, że $D_{j:n}^{(\alpha)}$ jest L–statystyką $L(\mathbf{c})$ gdzie $\mathbf{c} \in \mathcal{C}_n^{(2)}$ (zob. podrozdział 1.3). W szczególności oszacowania (2.1) oraz (2.2) przyjmują postać

$$\overline{B}_{n,p}^{(\alpha)}(j) = \sup_{F\in\mathcal{F}_2} \frac{\mathbb{E}_F D_{j:n}^{(\alpha)} - F^{\leftarrow}(p)}{\sigma_F} \tag{5.1}$$

oraz

$$\underline{B}_{n,p}^{(\alpha)}(j) = \inf_{F\in\mathcal{F}_2} \frac{\mathbb{E}_F D_{j:n}^{(\alpha)} - F^{\rightarrow}(p)}{\sigma_F}, \tag{5.2}$$

gdzie $1 \le j < n$. Funkcje $r_p(\mathbf{c})$ oraz $s_p(\mathbf{c})$ przyjmują postać

$$r_{n,p}^{(\alpha)}(j) = (1-\alpha) r_{n,p}(j) + \alpha r_{n,p}(j+1). \tag{5.3}$$

oraz

$$s_{n,p}^{(\alpha)}(j) = \sqrt{p(1-p)}\left|r_{n,p}^{(\alpha)}(j)\right|, \quad 1 \le j < n.$$

Definicja 2.1 przyjmuje zatem następującą postać.

Definicja 5.1. Powiemy, że $D_{j_0:n}^{(\alpha_0)}$ jest optymalnym estymatorem kwantyla rzędu p rozkładu $F \in \mathcal{F}_2$ jeżeli

$$s_{n,p}^{(\alpha_0)}(j_0) = \min_{\alpha\in(0,1),\, 1\le j<n} s_{n,p}^{(\alpha)}(j).$$

W Dodatku A.3 pokażemy, że dla $\alpha \in (0,1)$, $n \ge 2$, $1 \le j \le n$ oraz większości wartości p mamy

$$\overline{B}_{n,p}^{(\alpha)}(j) = (1-\alpha)\overline{B}_{n,p}(j) + \alpha\overline{B}_{n,p}(j+1), \tag{5.4}$$

i analogiczna własność zachodzi również dla dolnych oszacowań. Innymi słowy, w tych przypadkach dla $\mathbf{c} \in \mathcal{C}_n^{(2)}$ w nierównościach (2.5) i (2.6) zachodzą równości, a więc

$$r_{n,p}^{(\alpha)}(j) = \overline{B}_p(\mathbf{c}) + \underline{B}_p(\mathbf{c}). \tag{5.5}$$

Następny Lemat podaje własności pomocniczej funkcji $r_{n,p}^{(\alpha)}(j)$.

Lemat 5.1. *Funkcja $r_{n,p}^{(\alpha)}(j)$ ma następujące własności:*

(a) $r_{n,p}^{(\alpha)}(j)$ jest ściśle rosnąca względem zmiennej $j \in \{1,\ldots,n-1\}$;

(b) $r_{n,p}^{(\alpha)}(j)$ jest ściśle malejąca i ciągła względem zmiennej $p \in (0,1)$;

(c) dla $1 \le j \le n$ oraz $p \in (0,1)$ mamy

$$r_{n,1-p}^{(\alpha)}(j) = -r_{n,p}^{(1-\alpha)}(n-j).$$

Lemat ten jest prostą konsekwencją Wniosku 3.3, a punkt (c) wynika z Lematu 2.6(b). Zauważmy teraz, że estymator wyznaczony przez nasze kryterium ma własność symetrii (zob. np. [13] własność P4), która łatwo wynika z Lematu 2.7.

Lemat 5.2. *Jeżeli $D_{j:n}^{(\alpha)}$ jest optymalnym estymatorem kwantyla rzędu p, to optymalnym estymatorem kwantyla rzędu $1-p$ jest po prostu kombinacja $D_{n-j:n}^{(1-\alpha)}$.*

Lemat 5.3. *Dla ustalonych wartości n oraz p takich, że $r_{n,p}(1) \leq 0 < r_{n,p}(n)$, istnieje dokładnie jedna para liczb (j, α) taka, że zachodzi równość $r_{n,p}^{(\alpha)}(j) = 0$. Zatem estymator $D_{j:n}^{(\alpha)}$ kwantyla x_p optymalny w sensie Definicji 5.1 jest jednoznacznie wyznaczony.*

Dowód. Przypomnijmy, że dla ustalonych wartości n oraz p funkcja $r_{n,p}(j)$ jest ściśle rosnąca względem zmiennej j. Stąd, ponieważ $r_{n,p}(1) \leq 0$ oraz $r_{n,p}(n) > 0$ możemy wybrać jednoznacznie wyznaczone $j = k_0$ takie, że

$$r_{n,p}(k_0) \leq 0 < r_{n,p}(k_0 + 1). \tag{5.6}$$

Teraz zdefiniujmy

$$\alpha_0 = -\frac{r_{n,p}(k_0)}{r_{n,p}(k_0 + 1) - r_{n,p}(k_0)}. \tag{5.7}$$

Wtedy para (k_0, α_0) jest jedynym szukanym rozwiązaniem. □

Definicja 5.2. Dla ustalonych $n \geq 3$ oraz $p \in (0, 1)$ zakładamy, że $r_{n,p}(1) \leq 0$ oraz $r_{n,p}(n) > 0$. Definiujemy wtedy $j(n, p) = k_0$ oraz $\alpha(n, p) = \alpha_0$, które są zdefiniowane przez wzory (5.6) oraz (5.7).

Uwaga 5.1. W dalszym ciągu używając oznaczeń $j(n, p)$ oraz $\alpha(n, p)$ milcząco zakładamy, że n oraz p są takie, że zachodzi nierówność $r_{n,p}(1) \leq 0 < r_{n,p}(n)$.

Rozważymy teraz przypadek gdy $p = \frac{1}{2}$. Pokażemy, że nasza definicja prowadzi do klasycznej definicji estymatora mediany z próby, a więc optymalnym estymatorem mediany z próby rozmiaru $n \geq 2$ jest $X_{\frac{n+1}{2}:n}$ jeśli n jest nieparzyste, lub $\frac{1}{2}\left(X_{\frac{n}{2}:n} + X_{\frac{n}{2}+1:n}\right)$ jeśli n jest parzyste.

Lemat 5.4. *Załóżmy, że $p = \frac{1}{2}$.*

(a) Jeżeli n jest liczbą parzystą, to $j\left(n, \frac{1}{2}\right) = \frac{n}{2}$ oraz $\alpha\left(n, \frac{1}{2}\right) = \frac{1}{2}$.

(b) Jeżeli n jest liczbą nieparzystą, to $j\left(n, \frac{1}{2}\right) = \frac{n+1}{2}$ oraz $\alpha\left(n, \frac{1}{2}\right) = 0$.

Dowód. Łatwo sprawdzić, że jeśli n jest parzyste oraz

$$\mathbf{c} = \frac{1}{2}\boldsymbol{\delta}_{\frac{n}{2}} + \frac{1}{2}\boldsymbol{\delta}_{\frac{n}{2}+1}$$

to $\underline{\mathbf{c}} = \mathbf{c}$, a więc $r_{1/2}(\mathbf{c}) = 0$ na mocy Lematu 2.6(c). Jeśli n jest nieparzyste, to dla $\mathbf{c} = \boldsymbol{\delta}_{\frac{n+1}{2}}$ mamy $\underline{\mathbf{c}} = \mathbf{c}$, a więc ponownie $r_{1/2}(\mathbf{c}) = 0$. Teza wynika z jednoznaczności wyboru $j(n, p)$ i $\alpha(n, p)$. □

Ostatni wynik w tym podrozdziale mówi, że suma $j(n, p) + \alpha(n, p)$ jest ściśle rosnąca względem zmiennej $p \in (0, 1)$. Razem z poprzednim lematem oznacza to, że jeżeli $p < \frac{1}{2}$, to optymalny indeks $j(n, p)$ jest nie większy niż $\frac{n}{2}$.

Lemat 5.5. *Dla* $0 < p < q < 1$ *mamy* $j(n,p) + \alpha(n,p) < j(n,q) + \alpha(n,q)$.

Dowód. Zauważmy najpierw, że jeżeli $p < q$, to $j(n,p) \leq j(n,q)$. Istotnie, jeżeli $k = j(n,p)$ oraz $\ell = j(n,q)$, to stosując rozumowanie z dowodu Lematu 4.1 otrzymujemy nierówność $k \leq \ell$.

Teraz rozpatrzymy dwa możliwe przypadki: $j(n,p) < j(n,q)$ oraz $j(n,p) = j(n,q)$. W pierwszym z nich teza jest oczywiście prawdziwa ponieważ $\alpha(n,p) \in [0,1)$. Zatem wystarczy rozważyć tylko drugi przypadek.

Zatem załóżmy, że $j(n,p) = j(n,q) = k$, a więc wystarczy udowodnić, że zachodzi nierówność $\alpha(n,p) < \alpha(n,q)$. Łącząc definicję liczby k z faktem, że funkcja $r_{n,p}(k)$ jest ściśle malejąca ze względem p, otrzymujemy

$$r_{n,q}(k) < r_{n,p}(k) \leq 0 < r_{n,q}(k+1) < r_{n,p}(k+1).$$

Jeżeli $r_{n,p}(k) = 0$, to teza jest ponownie oczywista, zatem zakładamy teraz, że $r_{n,p}(k) < 0$. Wtedy ostatnie równanie daje nam

$$\frac{r_{n,p}(k)}{r_{n,q}(k)} < 1 < \frac{r_{n,p}(k+1)}{r_{n,q}(k+1)}.$$

Stąd otrzymujemy $r_{n,p}(k)\, r_{n,q}(k+1) > r_{n,q}(k)\, r_{n,p}(k+1)$. Odejmując stronami iloczyn $r_{n,p}(k)\, r_{n,q}(k)$ i stosując definicję (5.7) kończymy dowód lematu. □

5.2 Rozwiązanie problemu optymalnego wyboru

W tym podrozdziale przedstawimy sformułowanie oraz dowód głównych wyników dotyczących algorytmu wyboru optymalnych wartości liczb $j(n,p)$ oraz $\alpha(n,p)$ dla ustalonego rozmiaru próby n oraz rzędu kwantyla p. Najpierw wprowadzimy oznaczenia, które będą bardzo pomocne w zaprezentowaniu głównego twierdzenia w zwięzły sposób.

Definicja 5.3. Dla $1 \leq k \leq n$ niech $b_k = b_k(n)$ oznaczają jedyne rozwiązanie równania $r_{n,p}(k) = 0$ względem zmiennej $p \in (0,1)$.

Jednoznaczność liczb b_k wynika z następujących faktów:

(i) $\lim_{p \to 0^+} r_{n,p}(1) = +\infty$, i jest to również prawda jeśli 1 zostanie zastąpione przez dowolną z wartości $k = 2, \ldots, n$;

(ii) $\lim_{p \to 1^-} r_{n,p}(n) = -\infty$, i jest to również prawda jeśli n zostanie zastąpione przez dowolną z wartości $k = 1, \ldots, n-1$;

(iii) $r_{n,p}(k)$ jest ściśle malejącą i ciągłą funkcją względem zmiennej $p \in (0,1)$.

To również implikuje, że $0 < b_1 < \ldots < b_n < 1$. Te liczby są dla nas ważne, gdyż jeżeli $p \in (b_k, b_{k+1})$ to

$$r_{n,p}(k) < r_{n,b_k}(k) = 0 = r_{n,b_{k+1}}(k+1) < r_{n,p}(k+1). \tag{5.8}$$

Zatem dla $p \in [b_k, b_{k+1})$ mamy $j(n,p) = k$. Ponadto jeżeli $p = b_k$, to $\alpha(n, b_k) = 0$ lub innymi słowy optymalnym estymatorem kwantyla rzędu b_k postaci $D_{j:n}^{(\alpha)}$ jest po prostu pojedyncza statystyka porządkowa $X_{k:n}$. Teraz pokażemy, że w większości przypadków wartości b_k również spełniają dosyć proste równanie wielomianowe. Mianowicie, Lemat 4.6 implikuje następujące proste relacje między liczbami p_k oraz b_k.

Lemat 5.6.

(a) Dla $2 \le k \le n-1$ *mamy* $b_k(n) = p_k(n)$.

(b) Dla $n \ge 3$ *mamy* $b_1(n) < p_1(n) = 1 - \frac{1}{\sqrt[n]{2}}$ *oraz* $b_n(n) > p_n(n) = \frac{1}{\sqrt[n]{2}}$.

Dowód. Z Lematu 4.6(d) mamy $\theta_{k+1} < p_k < p_{k+1}$ dla $1 \le k \le n-4$. Zatem, $\theta_k < p_k$ dla $2 \le k \le n-3$. Łącząc Lemat 4.6(a) ze wzorem (4.10) możemy zauważyć, że nierówność ta zachodzi również dla $k = n-2$ oraz $n-1$.

Z drugiej strony z Lematu 4.6(c) mamy $p_k < p_{k+1} < \xi_k$ dla $4 \le k \le n-1$. Ponadto, z punktu (a) tego samego lematu mamy $p_2 < \xi_2$ oraz $p_3 < \xi_3$ dla $n \ge 5$. Zatem dla $2 \le k \le n-1$ mamy $p_k \in (\theta_k, \xi_k)$. Zatem, korzystając z Wniosku 3.4(b) oraz z definicji liczb p_k otrzymujemy

$$r_{n,p_k}(k) = \frac{1 - 2F_{k:n}(p_k)}{\sqrt{p_k(1-p_k)}} = 0.$$

Z definicji liczb b_k otrzymujemy punkt (a) lematu.

Aby udowodnić punkt (b) musimy udowodnić, że $r_{n,p_1}(1) < 0$. Ale wynika to łatwo z Wniosku 3.4(a) i definicji liczby p_1. □

Teraz jesteśmy gotowi, aby udowodnić twierdzenie dotyczące wyboru optymalnego estymatora kwantyla danego rzędu p w postaci kombinacji liniowej dwóch sąsiednich statystyk porządkowych. Problem sprowadza się do wyboru odpowiednich wartości parametrów $j(n,p)$ oraz $\alpha(n,p)$ przedstawionych w Definicji 5.2.

Twierdzenie 5.1. *Ustalmy* $n \ge 3$ *oraz* $p \in [p_2, p_{n-1}]$. *Optymalne wartości* $j(n,p)$ *oraz* $\alpha(n,p)$ *są następujące:*

(a) jeżeli $p = p_k$ *dla* $2 \le k \le n-1$, *to* $j(n,p_k) = k$ *oraz* $\alpha(n,p_k) = 0$;

(b) jeżeli $p \in (p_k, p_{k+1})$ *dla pewnego* $4 \le k \le n-4$, *to* $j(n,p) = k$ *oraz*

$$\alpha(n,p) = \frac{2F_{k:n}(p) - 1}{2B_{k,n}(p)}. \tag{5.9}$$

(c) taka sama konkluzja zachodzi jeżeli

(i) $p \in (p_k, \xi_k]$ *dla* $k = 2$ *lub* $k = 3$,

(ii) $p \in [\theta_{k+1}, p_{k+1})$ *dla* $k = n-3$ *lub* $k = n-2$.

Dowód. Wspomnieliśmy już, że z definicji liczb b_k mamy $j(n, b_k) = k$ oraz $\alpha(n, b_k) = 0$. To w połączeniu z Lematem 5.6(a) dowodzi części (a).

Ze wzoru (5.8) oraz Lematu 5.6(a) zauważamy, że dla $p \in (p_k, p_{k+1})$ mamy $j(n,p) = k$ jeżeli $2 \le k \le n-2$. Następnie z Lematu 4.6(f) oraz (g) mamy $\theta_{k+1} < p_k < p_{k+1} < \xi_k$ dla $4 \le k \le n-4$, a więc

$$(p_k, p_{k+1}) \subset [\theta_k, \xi_k] \cap [\theta_{k+1}, \xi_{k+1}]. \tag{5.10}$$

Zatem dla $p \in (p_k, p_{k+1})$ wartości $r_{n,p}(k)$ oraz $r_{n,p}(k+1)$ są dane prostymi wzorami podanymi we Wniosku 3.4(b). Teraz z Lematu 5.3 otrzymujemy $\alpha(n,p)$ w żądanej postaci.

Jeżeli $k = 2$ lub $k = 3$, to $p_k < \xi_k < p_{k+1}$, więc zamiast (5.10) możemy tylko wnioskować, że

$$(p_k, \xi_k] \subset [\theta_k, \xi_k] \cap [\theta_{k+1}, \xi_{k+1}],$$

co dowodzi części (i) podpunktu (c). Podobnie, jeżeli $k = n-3$ lub $k = n-2$, to używając wzoru (4.10) oraz Lematu 4.6 możemy tylko otrzymać

$$[\theta_{k+1}, p_{k+1}) \subset [\theta_k, \xi_k] \cap [\theta_{k+1}, \xi_{k+1}]$$

co dowodzi części (ii). □

Uwaga 5.2. Jeżeli $p \in (b_1, p_2) \cup (\xi_2, p_3)$ dla $n \ge 3$ lub $p \in (b_1, p_2) \cup (\xi_2, p_3) \cup (\xi_3, p_4)$ dla $n \ge 10$, to wyrażenie na współczynnik $\alpha(n,p)$ jest bardziej skomplikowane. Jest to spowodowane tym, że są to wartości p, dla których funkcja $r_{n,p}(k)$ jest podana we Wniosku 3.4(c), pomimo tego, że $(p_k, p_{k+1}) \subset (\theta_{k+1}, \xi_{k+1})$, więc $r_{n,p}(k+1)$ ma prostą postać podaną w podpunkcie (b) wniosku. Tabela 5.1 podaje numeryczne przybliżenia wartości liczb b_1, p_2, ξ_2 oraz p_3 dla $3 \le n \le 9$.

Wniosek 5.1. *Jeżeli* $p \in (p_k, p_{k+1})$, *gdzie* $4 \le k \le n-4$ *lub*

$$p \in (p_2, \xi_2] \cup (p_3, \xi_3] \cup [\theta_{n-2}, p_{n-2}) \cup [\theta_{n-1}, p_{n-1}),$$

to optymalnym estymatorem kwantyla x_p *postaci* $D_{j:n}^{(\alpha)}$ *jest*

$$\frac{1 - 2F_{k+1:n}(p)}{2B_{k,n}(p)} X_{k:n} + \frac{2F_{k:n}(p) - 1}{2B_{k,n}(p)} X_{k+1:n}.$$

n	b_1	p_2	ξ_2	p_3
3	0.205	0.500	-	-
4	0.158	0.386	-	-
5	0.128	0.313	0.415	0.500
6	0.108	0.264	0.338	0.421
7	0.093	0.228	0.284	0.364
8	0.082	0.201	0.245	0.320
9	0.073	0.179	0.216	0.286

Tabela 5.1: Wartości $b_1(n)$, $p_2(n)$, $\xi_2(n)$ oraz $p_3(n)$ dla $3 \leq n \leq 9$.

Praktyczne zastosowanie wyników przedstawionych w tym rozdziale jest bardzo proste. Załóżmy, że szukamy estymator kwantyla rzędu p nieznanego rozkładu na podstawie losowej próby rozmiaru n postaci $D_{j:n}^{(\alpha)}$. Zgodnie z naszym wynikiem na początek powinniśmy obliczyć wartości liczb $p_1, \ldots, p_n$. Z wyjątkiem kilku trywialnych przypadków, wymaga to obliczeń numerycznych Następnie znajdujemy największą wartość p_k, która jest mniejsza bądź równa zadanemu rzędowi kwantyla p. Wtedy optymalny estymator jest kombinacją liniową k–tej oraz $(k+1)$–szej statystyki porządkowej $X_{k:n}$ oraz $X_{k+1:n}$. Współczynniki tej kombinacji liniowej możemy ponownie obliczyć numerycznie. Dla p spełniającego założenia podane w ostatnim wniosku $\alpha(n,p)$ jest dana prostym wzorem (5.9). Dla pozostałych wartości musimy niestety użyć bardziej skomplikowanego wzoru (5.7) połączonego ze wzorami z Wniosku 3.4. Na zakończenie podamy dwa ważne komentarze.

Uwaga 5.3. Przypomnijmy, że w podrozdziale 4.2 dla $1 \leq j \leq n-1$ zdefiniowaliśmy liczby q_j jako jednoznacznie wyznaczone rozwiązanie równania

$$F_{j:n}(p) + F_{j+1:n}(p) = 1.$$

Na mocy Lematu 4.5 mamy $q_j \in (p_j, p_{j+1})$ dla każdego j. Co więcej, $q_{n/2} = \frac{1}{2}$ jeżeli n jest liczbą parzystą oraz $q_{(n-1)/2} < \frac{1}{2} < q_{(n+1)/2}$ jeżeli n jest liczbą nieparzystą. Wtedy $\alpha(n, q_j) = \frac{1}{2}$, a więc optymalnym estymatorem kwantyla x_{q_j} jest średnia arytmetyczna

$$\frac{1}{2}(X_{j:n} + X_{j+1:n}).$$

Jest to uogólnienie klasycznego estymatora mediany na podstawie próby rozmiaru parzystego. Z drugiej strony, jeżeli rozmiar próby jest liczbą nieparzystą, to $p_{(n+1)/2} = \frac{1}{2}$, a więc stwierdzenie, że $X_{j:n}$ jest optymalnym estymatorem kwantyla x_{p_j} jest uogólnieniem klasycznego estymatora mediany w przypadku gdy rozmiar próby jest liczbą nieparzystą.

Uwaga 5.4. Jeżeli $p = \frac{k}{n}$, to $p \in (p_k, p_{k+1})$ na mocy Lematu 4.5. Zatem, dla $4 \leq k \leq n-4$ optymalnym estymatorem kwantyla $x_{k/n}$ rzędu $\frac{k}{n}$ jest

$$\frac{1 - 2F_{k+1:n}\left(\frac{k}{n}\right)}{2B_{k,n}\left(\frac{k}{n}\right)} X_{k:n} + \frac{2F_{k:n}\left(\frac{k}{n}\right) - 1}{2B_{k,n}\left(\frac{k}{n}\right)} X_{k+1:n}. \tag{5.11}$$

Na mocy Lematu 4.6 mamy również $\frac{3}{n} \in (\theta_4, \xi_3)$ dla $n \geq 6$. Zatem, powyższy wniosek zachodzi również dla $k = 3$ lub $k = n-3$ dla $n \geq 6$. Powyższa uwaga pokazuje, że są to inne wartości niż tradycyjnie przyjmowane dla kwantyla z próby $X_{k:n}$. Jest to również inny estymator niż interpolacja liniowa postaci

$$\left(1 - \frac{k}{n}\right) X_{k:n} + \frac{k}{n} X_{k+1:n}. \tag{5.12}$$

Taka postać estymatora również jest bardzo popularnym wyborem w estymacji kwantyli. Jest to estymator Q_6 opisany w podrozdziale 1.3.1. Na przykład z Tabeli 6.4 dostajemy, że jeżeli $n = 15$ oraz $p = 0.25$, to $k = 4$ i estymator (5.11) przyjmuje postać

$$0.8281 X_{4:15} \ + \ 0.1719 X_{5:15}.$$

Z drugiej strony estymator (5.12) jest równy

$$0.7333 X_{4:15} \ + \ 0.2667 X_{5:15}.$$

Rozdział 6

Analiza błędu średnio-kwadratowego

W poprzednich rozdziałach nasze podejście do problemu wyboru optymalnych estymatorów kwantyla bazowało jedynie na analizie obciążenia rozważanych L–statystyk. W tym rozdziale przeanalizujemy dodatkowo błąd średnio–kwadratowy otrzymanych estymatorów.

W statystyce matematycznej bardzo ważną klasę estymatorów stanowią estymatory nieobciążone o minimalnej wariancji. Jest to spowodowane tym, że dla dowolnego estymatora $\hat{\theta}$ parametru $\theta \in \mathbb{R}$ jego błąd średnio–kwadratowy wynosi

$$MSE(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \theta)^2 = (\mathbb{E}\hat{\theta} - \theta)^2 + \operatorname{Var}\hat{\theta}.$$

Jeżeli $\hat{\theta}$ jest estymatorem nieobciążonym parametru θ, to $\mathbb{E}\hat{\theta} = \theta$, a więc $\mathbb{E}(\hat{\theta} - \theta) = 0$. Zatem estymator nieobciążony o minimalnej wariancji ma najmniejszy możliwy błąd średnio–kwadratowy. W naszym przypadku otrzymane estymatory są obciążone, ale ich obciążenie jest ograniczone z góry i z dołu symetrycznie względem 0. Dlatego do porównania efektywności otrzymanych estymatorów będziemy porównywać ich błędy średnio–kwadratowe. Niestety, z uwagi na bardzo skomplikowane wyrażenia analityczne, większość wyników tego rozdziału to jedynie przykłady numeryczne.

W podrozdziale 6.1 podajemy wyniki teoretyczne dotyczące oszacowania wariancji L–statystyk uzyskane przez Kozyrę [18]. W następnych podrozdziałach zastosujemy ten wynik do analizy numerycznej błędu średnio–kwadratowego estymatorów kwantyli opartych o jedną lub dwie statystyki porządkowe. W podrozdziale 6.4 przeanalizujemy zachowanie tych estymatorów na danych symulowanych, a w podrozdziale 6.5 przeanalizujemy błąd średnio–kwadratowy wybranych estymatorów opisanych w podrozdziale 1.3.1.

6.1 Oszacowania błędu średnio–kwadratowego

Niech $L(\mathbf{c})$ będzie L–statystyką rozważaną jako ustalony estymator kwantyla rzędu p rozkładu F. Wtedy dla dowolnej wartości x_p kwantyla rzędu p mamy

$$\begin{aligned}\frac{MSE_F\big(L(\mathbf{c})\big)}{\sigma_F^2} &= \frac{\mathbb{E}_F\big(L(\mathbf{c})-x_p\big)^2}{\sigma_F^2}\\ &= \left(\frac{\mathbb{E}_F L(\mathbf{c})-x_p}{\sigma_F}\right)^2 + \frac{\operatorname{Var}_F L(\mathbf{c})}{\sigma_F^2}\\ &\le \max\left\{\big(\overline{b}_p(\mathbf{c})\big)^2, \big(\underline{b}_p(\mathbf{c})\big)^2\right\} + \sup_{F\in\mathcal{F}_2}\frac{\operatorname{Var}_F L(\mathbf{c})}{\sigma_F^2}.\end{aligned} \tag{6.1}$$

Jeżeli przez $G_p(\mathbf{c})$ oznaczymy prawą stroną nierówności, czyli

$$G_p(\mathbf{c}) = \max\left\{\big(\overline{b}_p(\mathbf{c})\big)^2, \big(\underline{b}_p(\mathbf{c})\big)^2\right\} + \sup_{F\in\mathcal{F}_2}\frac{\operatorname{Var}_F L(\mathbf{c})}{\sigma_F^2}, \tag{6.2}$$

to dla dowolnego $F\in\mathcal{F}_2$ mamy

$$\frac{MSE_F\big(L(\mathbf{c})\big)}{\sigma_F^2} \le G_p(\mathbf{c}).$$

Zatem dla danych n oraz p powinniśmy wybrać taki wektor $\mathbf{c}$, który minimalizuje funkcję $G_p(\mathbf{c})$. Oczywiście, jeżeli $G_p(\mathbf{c})$ jest mniejsze niż $G_p(\mathbf{d})$, to nie możemy wnioskować, że dla każdej dystrybuanty $F\in\mathcal{F}_2$ mamy $MSE_F(L(\mathbf{c})) \le MSE_F(L(\mathbf{d}))$. Jednakże, jeżeli $G_p(\mathbf{c}) \le G_p(\mathbf{d})$ i nie mamy żadnej wiedzy o rozkładzie F, to używając jako estymatora $L(\mathbf{c})$ otrzymamy średnio mniejszy błąd estymacji. Z drugiej strony, taki wybór wymusza jak najmniejszą wartość w pierwszym wyrażeniu we wzorze (6.2), więc zarówno dolne jaki górne oszacowanie $\overline{b}_p(\mathbf{c})$ oraz $\underline{b}_p(\mathbf{c})$ powinny być mniej więcej równe wartości bezwzględnej, więc wartość $r_p(\mathbf{c})$ jest bliska 0. Jest to kolejne uzasadnienie naszego wyboru wektora $\mathbf{c}$ w Definicji 2.1.

Ostatnia własność może być traktowana jako odpowiednik nieobciążoności. Własność minimalnej wariancji zastąpimy przez minimalizację oszacowania błędu średnio–kwadratowego. Przypomnijmy, że w Definicji 2.1 nazwaliśmy L–statystykę $L(\mathbf{c}_0)$ optymalną w klasie $\mathcal{C}$, jeśli $\mathbf{c}_0$ minimalizuje funkcję $s_p(\mathbf{c}_0)$ w zbiorze $\mathcal{C}$. Jednocześnie zauważyliśmy, że $\mathbf{c}_0$ nie musi być wyznaczone jednoznacznie. Oznaczmy przez $\mathcal{D}_p$ zbiór wszystkich wektorów, dla których minimum funkcji s_p jest osiągnięte, a więc

$$\mathcal{D}_p = \left\{\mathbf{d}\in\mathcal{C} : s_p(\mathbf{d}) = \min_{\mathbf{c}\in\mathcal{C}} s_p(\mathbf{c})\right\}.$$

W sytuacji, gdy $\mathcal{D}_p$ ma co najmniej dwa elementy, dodatkowo wybieramy $\mathbf{c}_0\in\mathcal{D}_p$ tak, aby

$$G_p(\mathbf{c}_0) = \min_{\mathbf{c}\in\mathcal{D}_p} G_p(\mathbf{c}).$$

Zatem ze wszystkich L–statystyk optymalnych w sensie Definicji 2.1 wybieramy te, które minimalizują błąd średnio–kwadratowy, a dokładniej jego górne ograniczenie. Odpowiada to minimalizacji wariancji dla estymatorów nieobciążonych.

Oczywiście pierwsze wyrażenie w $G_p(\mathbf{c})$ zależy od wartości oszacowań obciążeń określonych w Rozdziale 2, a wiec ich wartości są znane. Naszym kolejnym zadaniem jest zatem znalezienie wartości drugiego wyrażenia. Została ona wyznaczona przez Kozyrę [18], który uzyskał oszacowania wariancji dla dowolnej kombinacji liniowej statystyk porządkowych.

Przypomnijmy, że dla ustalonego niezerowego wektora $\mathbf{c} = (c_1, \ldots, c_n) \in \mathcal{C}_n$ określiliśmy

$$F_{\mathbf{c}}(u) = \sum_{i=1}^{n} c_i F_{i:n}(u), \quad u \in [0,1].$$

Ponadto, niech $C_j = \sum_{i=1}^{j} c_i$ dla $1 \leq j \leq n$. Wtedy korzystając ze wzoru (1.8) po prostych przekształceniach otrzymamy dla $u \in [0,1]$

$$F_{\mathbf{c}}(u) = \sum_{j=1}^{n} C_j B_{j,n}(u).$$

Ponadto, oznaczmy dwuwymiarowe wielomiany Bernsteina wzorem

$$B_{i,j,n}(u,v) = \binom{n}{i, j-i} u^i (v-u)^{j-i} (1-v)^{n-j}, \quad 0 < u \leq v < 1,$$

gdzie

$$\binom{n}{a,b} = \frac{n!}{a!b!(n-a-b)!}$$

oznacza współczynnik wielomianowy. Zatem $B_{i,j,n}(u,u) = B_{i,n}(u)$ dla $i = j$ oraz 0 dla $i < j$. Oznaczmy dla $0 \leq u \leq v \leq 1$

$$H_{\mathbf{c}}(u,v) = \sum_{i=1}^{n} \sum_{j=i}^{n} C_i (1 - C_j) B_{i,j,n}(u,v),$$

a więc dla $u \in [0,1]$ mamy

$$H_{\mathbf{c}}(u,u) = \sum_{j=1}^{n} C_j (1 - C_j) B_{j,n}(u).$$

Określmy pomocnicze funkcje

$$\Phi_{\mathbf{c}}(u,v) = \frac{1}{u(1-v)} \left[F_{\mathbf{c}}(u)\big(1 - F_{\mathbf{c}}(v)\big) - H_{\mathbf{c}}(u,v) \right]$$

dla $0 \leq u \leq v \leq 1$, oraz

$$\Psi_{\mathbf{c}}(u) = \Phi_{\mathbf{c}}(u,u) = \frac{1}{u(1-u)} \left[F_{\mathbf{c}}(u)\big(1 - F_{\mathbf{c}}(u)\big) - H_{\mathbf{c}}(u,u) \right]$$

dla $0 \leq u \leq 1$.

Twierdzenie 6.1. *Ustalmy $n \geq 2$ oraz $\mathbf{c} = (c_1, \ldots, c_n) \in \mathcal{C}_n$. Wtedy dla dowolnego $F \in \mathcal{F}_2$ mamy*

$$\frac{\operatorname{Var}_F L(\mathbf{c})}{\sigma_F^2} \leq \sup_{0<u\leq v<1} \Phi_{\mathbf{c}}(u, v). \tag{6.3}$$

Jeżeli

$$\sup_{0<u\leq v<1} \Phi_{\mathbf{c}}(u, v) = \sup_{0<u<1} \Psi_{\mathbf{c}}(u) \tag{6.4}$$

to w nierówności (6.3) *zachodzi równość dla odpowiednio dobranego rozkładu dwupunktowego F.*

Uwaga 6.1. Kozyra [18] podaje dokładną postać rozkładu, dla którego oszacowanie (6.3) jest osiągnięte, ale nie ma ona znaczenia w dalszych rozważaniach. Ponadto, zauważmy, że lewa strona nierówności (6.4) nie może być mniejsza niż prawa strona. Jednakże, numeryczne obliczenia opisane w dalszym ciągu tego rozdziału sugerują, że w naszym przypadku warunek (6.4) jest zawsze spełniony. Niestety nie udało się nam udowodnić tego analitycznie, co było spowodowane skomplikowaną postacią funkcji $\Phi_{\mathbf{c}}$ i $\Psi_{\mathbf{c}}$.

6.2 Przypadek pojedynczej statystyki porządkowej

Twierdzenie 4.4 podaje metodę wyboru optymalnego estymatora kwantyla rzędu p w postaci pojedynczej statystyki porządkowej. Wybór ten oparty jest jedynie o analizę obciążenia estymacji, więc przeanalizujemy teraz jej błąd średnio–kwadratowy. Ze wzoru (6.1) otrzymujemy

$$\frac{MSE_F(X_{j:n})}{\sigma_F^2} \leq (w_{n,p}(j))^2 + \tau_n^2(j),$$

gdzie $w_{n,p}$ jest dane wzorem (4.11) oraz

$$\tau_n^2(j) = \sup_{F\in\mathcal{F}} \frac{\operatorname{Var}_F X_{j:n}}{\sigma_F^2}.$$

Wartości $\tau_n^2(j)$ zostały wyznaczone przez Papadatosa [23] i wynoszą one

$$\tau_n^2(j) = \sup_{0<u<1} \frac{F_{j:n}(u)(1 - F_{j:n}(u))}{u(1-u)}$$

dla $1 \leq j \leq n$. W szczególności $\tau_n^2(1) = \tau_n^2(n) = n$ oraz liczby $\tau_n^2(j)$ są symetryczne w tym sensie, że

$$\tau_n^2(n - j + 1) = \tau_n^2(j).$$

Ponadto liczby $\tau_n^2(j)$ maleją dla $1 \leq j < \frac{n}{2}$ i rosną dla $\frac{n}{2} < j \leq n$. Niestety poza przypadkami $j = 1$ i $j = n$ wartości $\tau_n^2(j)$ nie są znane w postaci jawnej, ale mogą być łatwo wyznaczone numerycznie. W Tabeli 6.1 podajemy przybliżone wartości $\tau_n^2(j)$ dla $n = 15$ oraz $1 \leq j \leq 8$.

Niech $G_{n,p}(j) = G_p(\mathbf{c})$, gdzie $\mathbf{c} \in \mathcal{C}_n^{(1)}$, a więc

$$G_{n,p}(j) = (w_{n,p}(j))^2 + \tau_n^2(j).$$

Rozważmy problem doboru j tak, aby zminimalizować $G_{n,p}(j)$. Z uwagi na symetrię załóżmy, że $p \in \left(0, \frac{1}{2}\right)$ czyli $1 \leq j \leq \frac{n}{2}$. Wiemy już, że wybór j opisany w Twierdzeniu 4.4, minimalizuje również $(w_{n,p}(j))^2$ (zob. Twierdzenie 4.5). Mianowicie dla większości p optymalnym estymatorem x_p jest $X_{\lceil np \rceil:n}$, ale dla $p \in \mathcal{Q}_n$ powinniśmy wybrać jako estymator statystykę $X_{\lfloor np \rfloor:n}$. Dzięki temu pierwszy składnik staje się mniejszy, ale drugi staje się większy. Jest to przykład efektu znanego w literaturze anglojęzycznej jako bias–variance tradeoff, czyli kompromis pomiędzy obciążeniem a wariancją. Polega on na niemożliwości jednoczesnej minimalizacji obciążenia i wariancji estymatora. Jednakże, jeżeli $p = a_{j,n}$ dla pewnego $1 \leq j < \frac{n}{2}$, to lepszym wyborem jest $j = \lceil na_{j,n} \rceil$ niż $j = \lfloor na_{j,n} \rfloor$. Jest to ulepszenie Twierdzenia 4.4(a), które mówi, że dla $p = a_{j,n}$ obydwie statystyki $X_{j:n}$ oraz $X_{j+1:n}$ są równie dobre.

j	1	2	3	4	5	6	7	8
$\sigma_{15}^2(j)$	15	3.1235	1.8951	1.4334	1.2049	1.0818	1.0193	1

Tabela 6.1: Przybliżone wartości $\sigma_{15}^2(j)$ dla $1 \leq j \leq 8$.

6.3 Przypadek kombinacji liniowych dwóch statystyk porządkowych

W tym podrozdziale porównujemy znane estymatory kwantyli będące kombinacjami liniowymi dwóch sąsiednich statystyk porządkowych z estymatorem określonym w Definicji 5.2. Dla rozważanych estymatorów porównamy zarówno ich obciążenie mierzone wartościami funkcji $r_{n,p}^{(\alpha)}$ (wzór (5.3)) oraz ich błędu średnio–kwadratowego mierzonego wartościami odpowiedniej funkcji $G_p(\mathbf{c})$ (zob. wzór (6.7) poniżej). Niestety, z uwagi na bardzo skomplikowane wyrażenia analityczne porównywanych wielkości, podamy jedynie wyniki obliczeń numerycznych.

Jako znane estymatory rozważamy statystyki $Q_4, \ldots, Q_9$ opisane w podrozdziale 1.3.1. Wszystkie te estymatory mają postać

$$Q_i(p) = (1 - \{\ell\})X_{\lfloor \ell \rfloor:n} + \{\ell\}X_{\lfloor \ell \rfloor+1:n}.$$

Wartości ℓ w zależności od i zostały podane w Tabeli 1.1, której skróconą wersję przedstawiamy w Tabeli 6.2. Zauważmy, że wszystkie te estymatory są interpolacjami liniowymi pomiędzy punktami wykresu kwantyli. Dodatkowo do porównania dodamy es-

i	ℓ	p	i	ℓ	p
4	np	$\left[\frac{1}{n}, 1\right)$	7	$(n-1)p+1$	$[0,1)$
5	$np+0.5$	$\left[\frac{1}{2n}, 1-\frac{1}{2n}\right)$	8	$(n+\frac{1}{3})p+\frac{1}{3}$	$\left[\frac{2}{3n+1}, \frac{n-1}{3n+1}\right)$
6	$(n+1)p$	$\left[\frac{1}{n+1}, \frac{n}{n+1}\right)$	9	$(n+\frac{1}{4})p+\frac{3}{8}$	$\left[\frac{5}{2(4n+1)}, \frac{8n-3}{2(4n+1)}\right)$

Tabela 6.2: Wartości $\ell(i)$ dla Q_i, $4 \leq i \leq 9$.

$b_1(n)$	< 0.1	< 0.05	< 0.02	< 0.01	< 0.005
n	≥ 7	≥ 14	≥ 34	≥ 69	≥ 138

Tabela 6.3: Wartości n, dla których liczba $b_1(n)$ jest mniejsza niż określony poziom.

tymatory Q_1, Q_2, Q_3 (zob. podrozdział 1.3.1), które są w istocie oparte na odpowiednio dobranej pojedynczej statystyce porządkowej. Jednakże, podobnie jak Hyndman i Fan [13] będziemy je traktować jako odpowiednie kombinacje liniowe dwóch statystyk porządkowych.

Ponadto oznaczamy przez Q_{10} estymator postaci kombinacji liniowej dwóch statystyk porządkowych zaproponowany w Rozdziale 5. Zatem dla $p \in (b_1, b_n)$ wartość estymatora $Q_{10}(p)$ jest dana wzorem

$$D_{k:n}^{(\gamma)} = (1-\gamma)X_{k:n} + \gamma X_{k+1:n} \tag{6.5}$$

z wartościami $k = k_0$ oraz $\gamma = \alpha_0$ podanymi w Definicji 5.2 i wyznaczonymi jawnie w Twierdzeniu 5.1. Przypomnijmy, że dla $1 \leq k \leq n$ liczba p_k jest medianą statystyki porządkowej $U_{k:n}$ z rozkładu jednostajnego. Na mocy punktu (a) Twierdzenia 5.1 mamy $Q_{10}(p_k) = X_{k:n}$ i dlatego w tym przypadku możemy traktować liczby p_k jako odpowiedniki punktów wykresu kwantyli. Zatem definicja estymatora Q_{10} jest najbliższa estymatorowi Q_8, ale interpolacja pomiędzy punktami $(p_k, X_{k:n})$ przez estymator Q_{10} nie jest liniowa.

Zauważmy, że pierwszą przewagą naszego estymatora Q_{10} jest możliwość użycia go dla $p \in (b_1(n), b_n(n))$. Łatwo można pokazać, że $b_1(n)$ jest ciągiem malejącym oraz numeryczne obliczenia pokazują na przykład, że $b_1(n) < 0.02$ dla $n \geq 34$. Zatem, jeżeli chcemy estymować kwantyl rzędu $p = 0.02$, to wystarczy użyć próby rozmiaru $n = 34$, podczas gdy klasyczne podejście sugeruje użycie próby rozmiaru $n \geq 50$. W związku z symetrią pomiędzy b_1 oraz b_n, ten sam rozmiar próby będzie wystarczający do estymacji kwantyla rzędu $p = 0.98$. Podobnie dla kwantyla rzędu $p = 0.01$ wystarczy użyć próby rozmiaru $n \geq 69$, zamiast próby rozmiaru $n \geq 100$, jaki byłby wymagany przy klasycznym podejściu. W Tabeli 6.3 przedstawiamy podobne wnioski dla najbardziej popularnych małych rzędów kwantyli.

Chcemy teraz porównać zachowanie estymatorów Q_i, $1 \leq i \leq 10$, dla ustalonych wartości n oraz p. Dla każdego i w tym zakresie obliczamy wartości $k = k(i)$ oraz $\gamma = \gamma(i)$ zgodnie z powyższymi definicjami. Następnie dla każdego i obliczamy odpowiadające im wartości funkcji $r_{n,p}^{(\gamma)}(k)$ danej wzorem (5.3) oraz $G_{n,p}^{(\gamma)}(k) = G_p(\mathbf{c})$ gdzie $\mathbf{c} = (c_1, \ldots, c_n)$ jest postaci

$$c_i = \begin{cases} 1-\gamma, & \text{dla } i = k, \\ \gamma, & \text{dla } i = k+1, \\ 0, & \text{poza.} \end{cases} \tag{6.6}$$

Zauważmy, że wtedy

$$G_{n,p}^{(\gamma)}(k) = \max\left\{ \left(\overline{b}_{n,p}^{(\gamma)}(k)\right)^2, \left(\underline{b}_{n,p}^{(\gamma)}(k)\right)^2 \right\} + \sup_{F \in \mathcal{F}_2} \frac{\operatorname{Var}_F D_{k:n}^{(\gamma)}}{\sigma_F^2}, \tag{6.7}$$

gdzie $\overline{b}_{n,p}^{(\gamma)}(k) = \overline{b}_p(\mathbf{c})$ oraz $\underline{b}_{n,p}^{(\gamma)}(k) = \underline{b}_p(\mathbf{c})$. Wartość pierwszego składnika obliczamy stosując Twierdzenie 3.1, Wniosek 3.1 oraz wzór (5.3). Aby wyznaczyć wartość drugiego składnika korzystamy z Twierdzenia 6.1. W tym celu zauważmy najpierw, że dla wektora $\mathbf{c}$ postaci (6.6) mamy

$$C_j = \begin{cases} 0, & \text{dla } j < k, \\ 1-\gamma, & \text{dla } j = k, \\ 1, & \text{dla } j > k. \end{cases}$$

Zatem $C_i(1 - C_j) \neq 0$ tylko gdy $i = j = k$, a więc $H_{\mathbf{c}}(u, v) = \gamma(1-\gamma)B_{k,k,n}(u, v)$. Ostatecznie funkcje $F_{\mathbf{c}}$, $\Phi_{\mathbf{c}}$ i $\Psi_{\mathbf{c}}$ przyjmują postać

$$F_{k:n}^{(\gamma)}(u) = (1-\gamma)B_{k,n}(u) + \sum_{i=k+1}^{n} B_{i,n}(u),$$

dla $\quad 0 \leq u \leq 1$,

$$\Phi_k^{(\gamma)}(u, v) = \frac{1}{u(1-v)} \left[F_{k:n}^{(\gamma)}(u)\big(1 - F_{k:n}^{(\gamma)}(v)\big) - \binom{n}{k} \gamma(1-\gamma) u^k (1-v)^{n-k} \right],$$

dla $0 \leq u \leq v \leq 1$, oraz

$$\Psi_k^{(\gamma)}(u) = \frac{1}{u(1-u)} \left[F_{k:n}^{(\gamma)}(u)\big(1 - F_{k:n}^{(\gamma)}(u)\big) - \gamma(1-\gamma)B_{k,n}(u) \right],$$

dla $0 \leq u \leq 1$. Oczywiście $\Psi_k^{(\gamma)}(u) = \Phi_k^{(\gamma)}(u, u)$. Ponadto obie funkcje zależą od n, ale dla uproszczenia notacji pomijamy ten parametr w zapisie.

Z powodu bardzo skomplikowanej postaci analitycznej funkcji $r_{n,p}^{(\gamma)}(k)$ oraz $G_{n,p}^{(\gamma)}(k)$ teoretyczne porównanie estymatorów $Q_1, \ldots, Q_{10}$ wydaje się bardzo trudne. Jednakże,

	$p = 0.05$				$p = 0.25$			
i	k	γ	$r^{(\gamma)}_{15,0.05}(k)$	$G^{(\gamma)}_{15,0.05}(k)$	k	γ	$r^{(\gamma)}_{15,0.25}(k)$	$G^{(\gamma)}_{15,0.25}(k)$
1	0	1.	-0.3860	21.3089	3	1.	-0.1788	2.9812
2	0	1.	-0.3860	21.3089	3	1.	-0.1788	2.9812
3	0	1.	-0.3860	21.3089	3	1.	-0.1788	2.9812
4	-	-	-	-	3	0.75	-0.4389	3.1131
5	1	0.25	0.4654	14.9159	4	0.25	0.0812	2.5693
6	-	-	-	-	4	0.	-0.1788	2.9812
7	1	0.7	1.9979	13.5338	4	0.5	0.3413	2.7612
8	1	0.1	-0.0455	17.6210	4	0.1667	-0.0054	2.5612
9	1	0.1375	0.0822	16.7116	4	0.1875	0.0162	2.5519
10	1	0.1134	0.	17.1558	4	0.1719	0.	2.5493

Tabela 6.4: Wartości funkcji $r^{(\gamma)}_{15,p}(k)$ oraz $G^{(\gamma)}_{15,p}(k)$, gdzie $k = j(15, p)$ oraz $\gamma = \alpha(15, p)$ dla $p = 0.05$ oraz $p = 0.25$.

łatwo jest porównać ich zachowanie wykonując obliczenia numeryczne. W Tabeli 6.4 przedstawiamy wyniki otrzymane dla $n = 15$ oraz $p = 0.05$ lub 0.25. Zauważmy, że w tym drugim przypadku $G^{(\gamma)}_{15,0.25}(k)$ ma najmniejszą wartość dla $i = 10$, czyli dla estymatora zaproponowanego w Rozdziale 5. Niestety sytuacja jest inna dla kwantyla rzędu $p = 0.05$. Wtedy najmniejsza wartość $G^{(\gamma)}_{15,0.05}(k)$ jest uzyskana dla estymatora Q_7, ale w jego przypadku wartość funkcji $r^{(\gamma)}_{15,0.05}(k)$ wynosi 1.9978. To oznacza, że estymator Q_7 znacznie przeszacowuje wartość kwantyla rzędu 0.05. Z drugiej strony, dla estymatora Q_{10} mamy $r^{(\gamma)}_{15,0.05}(k) = 0$, zaś jego błąd średnio–kwadratowy MSE jest ograniczony przez wartość 17.1, która nie jest znacząco większa niż wartość 13.53 otrzymana dla estymatora Q_7.

Podobne wyniki otrzymaliśmy dla innych wartości n oraz p. Rozważaliśmy wszystkie możliwe kombinacje rozmiaru próby $n = 7$, 15, 35, 50, 70, 100, 150 oraz rzędu kwantyla $p = 0.01$, 0.02, 0.05, 0.1, 0.25, 0.4, biorąc pod uwagę ograniczenia dotyczące wartości p podane w Tabelach 6.2 oraz 6.3. Wyniki wszystkich obliczeń są przedstawione w zbiorczej Tabeli 6.5. Puste komórki odpowiadają tym kombinacjom liczb n oraz p, dla których estymator Q_{10} nie jest zdefiniowany. Komórki zawierające znak + odpowiadają przypadkom gdy $G^{(\gamma)}_{n,p}(k)$ ma najmniejszą wartość dla estymatora Q_{10}, podczas gdy znak $-$ odpowiada przeciwnej sytuacji. Zatem dla typowych wartości rzędu kwantyli (odseparowanych od 0 lub 1) estymator Q_{10} ma najlepsze zachowanie w porównaniu do estymatorów $Q_1, \ldots, Q_9$.

Następnie, dla ustalonych wartości n oraz p podanych jak powyżej znaleźliśmy numeryczne wartości liczb l oraz β, które minimalizują wartości funkcji $G^{(\beta)}_{n,p}(l)$ względem

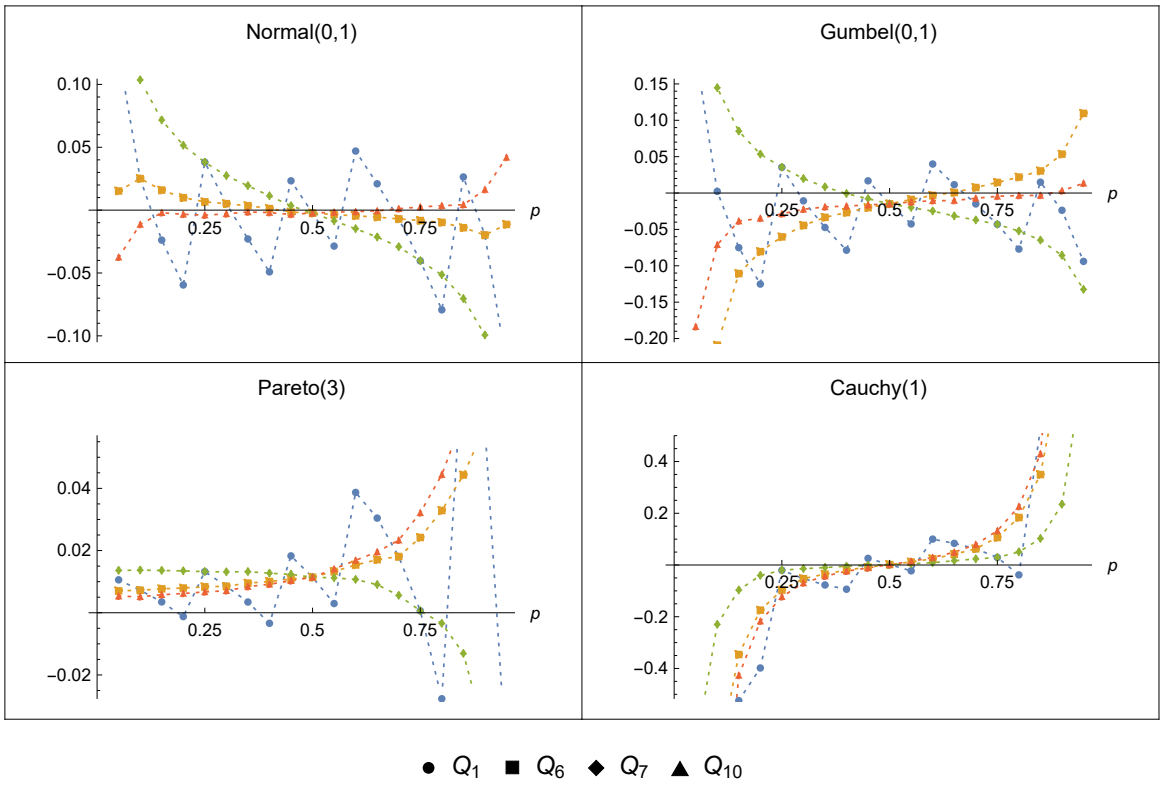
$n\backslash p$	0.01	0.02	0.05	0.1	0.25	0.4
7				−	−	+
15			−	+	+	+
35		−	−	+	+	+
50		−	+	+	+	+
70	−	+	+	+	+	+
100	−	−	+	+	+	+
150	+	+	+	+	+	+

Tabela 6.5: Przypadki dla jakich n oraz p wybór $j(n,p)$ oraz $\alpha(n,p)$ minimalizuje wartość funkcji $G^{(\alpha)}_{n,p}(j)$ są zaznaczone znakiem + .

wszystkich wartości $1 \leq l \leq n$, $\beta \in [0,1)$ jednocześnie. Zaskakująco okazało się, że w przypadkach ze znakiem + wartości l oraz β są dokładnie takie same jak wartości $k = j(n,p)$ oraz $\gamma = \alpha(n,p)$, które zostały zdefiniowane dla naszego estymatora Q_{10}. Jest to ciekawy i niespodziewany wniosek, ponieważ naszym podstawowym celem było zminimalizowanie jedynie wartości funkcji $r^{(\alpha)}_{n,p}(j)$, a jednocześnie przy okazji zminimalizowaliśmy wartości funkcji $G^{(\alpha)}_{n,p}(j)$. Niestety, nie byliśmy w stanie znaleźć ścisłego sformułowania i dowodu takiego stwierdzenia. Sytuacja jest inna dla wartości n oraz p, przy których występuje znak −. Na przykład, dla $n = 15$ oraz $p = 0.05$ najmniejsza wartość funkcji $G^{(\beta)}_{15,0.05}(l)$ wynosi 12.46 i jest uzyskana dla wartości $l = 1$ oraz $\beta = 0.532$. Jednakże, okazuje się, że dają one wartość funkcji $r^{(\beta)}_{15,0.05}(l) = 1.425$. Z drugiej strony, nasz wybór liczb $k = j(15, 0.05)$ oraz $\gamma = \alpha(15, 0.05)$ daje $k = 1$, $\gamma = 0.1133$ oraz $G^{(\gamma)}_{15,0.05}(k) = 17.155$, ale z definicji funkcji otrzymujemy $r^{(\gamma)}_{15,0.05}(k) = 0$.

Podkreślmy, że wyniki numerycznych obliczeń sugerują również, że warunek (6.4) jest zawsze spełniony. Faktycznie, jest tak dla wszystkich przypadków n oraz p używanych przez nas w obliczeniach. To implikuje, że drugie wyrażenie w definicji $G^{(\gamma)}_{n,p}(k)$ jest najmniejsze z możliwych dla wszystkich $F \in \mathcal{F}_2$.

Innym ważnym wnioskiem wyciągniętym z naszych obliczeń jest to, że estymatory Q_5, Q_8, Q_9 są niemal tak samo dobre jak estymator Q_{10}. Można to zobaczyć w Tabeli 6.4 dla $n = 15$ oraz $p = 0.25$. Mianowicie dla $i = 5, 8, 9$ odpowiadające im wartości funkcji $r^{(\gamma)}_{15,0.25}(k)$ są bardzo bliskie 0 oraz wartości funkcji $G^{(\gamma)}_{15,0.25}(k)$ są bardzo bliskie możliwie najmniejszej wartości wynoszącej 2.5493, która została otrzymana dla estymatora Q_{10}. Dla pozostałych wartości i różnice są znacząco większe. Podobne wyniki otrzymujemy dla wszystkich przypadków n oraz p z zaznaczonym znakiem +. W szczególności estymatory Q_5, Q_8 oraz Q_9, których definicje oparte są o intuicję, zachowują się bardzo dobrze w porównaniu z estymatorem Q_{10}, który jest precyzyjnie zdefiniowany przez wprowadzone przez nas kryterium.



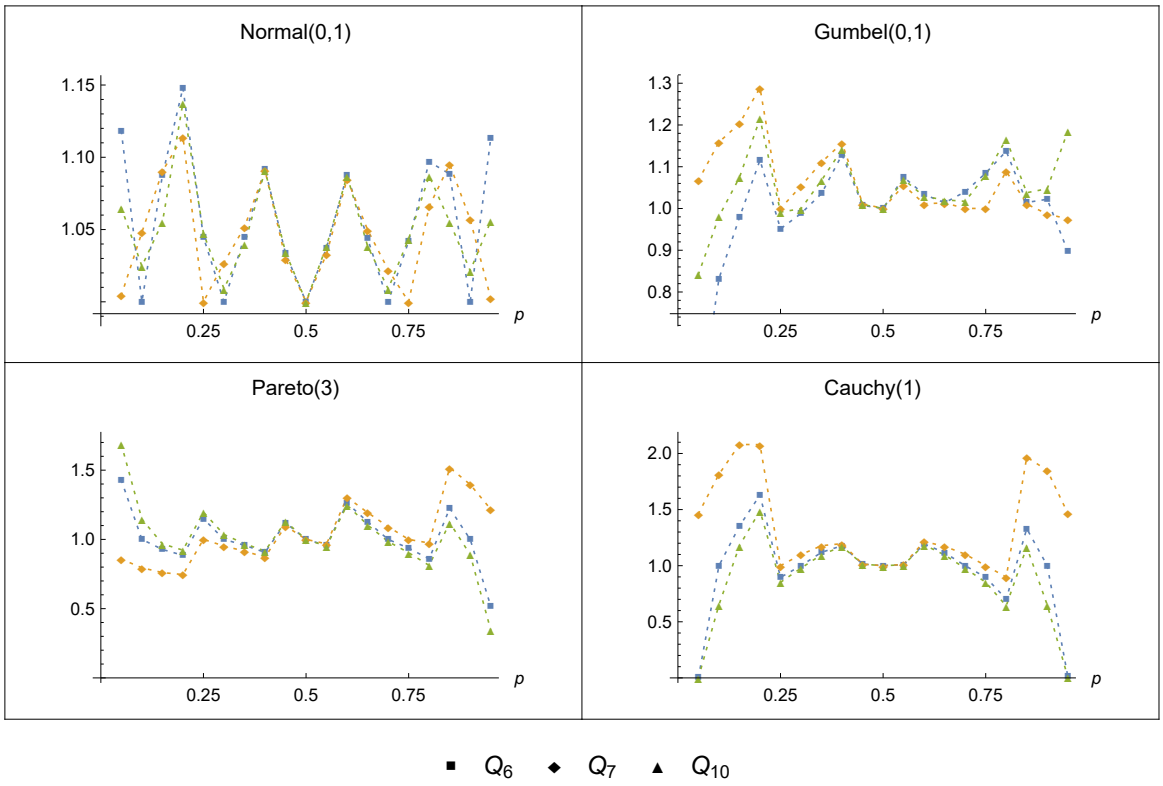
Rysunek 6.1: Wykresy obciążeń estymatorów Q_1, Q_6, Q_7 oraz Q_{10}.

6.4 Wyniki dla danych symulowanych

W tym podrozdziale chcemy zbadać jak się zachowują wybrane estymatory postaci kombinacji liniowej dwóch sąsiednich statystyk porządkowych na danych symulowanych. Porównujemy obciążenia oraz błąd średnio–kwadratowy estymatorów Q_1, Q_6, Q_7 z estymatorem Q_{10}, który zaproponowaliśmy. Estymator Q_1 został przez nas wybrany jako najbardziej popularny z estymatorów kwantyli pojawiających się w literaturze ze względu na jego prostotę. Estymator Q_6 wybraliśmy, aby zweryfikować jedną z tez artykułu [19], którego autor stwierdza, że ten estymator ma najlepiej dobrane punkty wykresu kwantyli.

Z twierdzenia Gliwienki-Cantelli'ego wiemy, że zachowanie wszystkich estymatorów kwantyli jest bardzo podobne, w przypadku prób dużego rozmiaru. Dlatego ustalamy $n = 25$, ponieważ interesuje nas w szczególności zachowanie estymatora dla małych prób. Następnie wygenerowaliśmy próby rozmiaru $n = 25$ dla czterech wybranych rozkładów: standardowego normalnego, Gumbela, Cauchy'ego, oraz rozkładu Pareto z parametrem kształtu równym 3. Zauważmy, że rozkład Cauchy'ego nie należy do klasy rozkładów $\mathcal{F}_2$, ale wybraliśmy go w celu sprawdzenia jak estymator Q_{10} będzie zachowywać się dla przykładowego rozkładu z nieskończoną wariancją.

Nasz eksperyment polega na przeprowadzeniu $N = 10000$ symulacji Monte Carlo



Rysunek 6.2: Wykresy względnego błędu średnio–kwadratowego estymatorów Q_6, Q_7 oraz Q_{10} względem estymatora Q_1.

podobnie jak na przykład w artykule [32]. Dla każdej wygenerowanej próby rozmiaru 25 z ustalonego rozkładu o dystrybuancie F obliczyliśmy wartości $Q_i(p)$ dla $i = 1, 6, 7, 10$ oraz dla 19 wartości parametru $p = 0.05, 0.1, \ldots, 0.95$. Następnie porównaliśmy odpowiednie średnie z symulowanych danych ze znanymi prawdziwymi wartościami kwantyli $x_p(F)$. Wtedy obliczyliśmy dwie miary: (a) błąd estymacji, który jest różnicą pomiędzy wartością $Q_i(p)$ uzyskaną metodą Monte Carlo a prawdziwą wartością kwantyla $x_p(F)$; (b) względny błąd średnio–kwadratowy estymatorów Q_6, Q_7 oraz Q_{10} względem estymatora Q_1, który jest ilorazem $MSE(Q_1)/MSE(Q_i)$ dla $i = 6, 7, 10$.

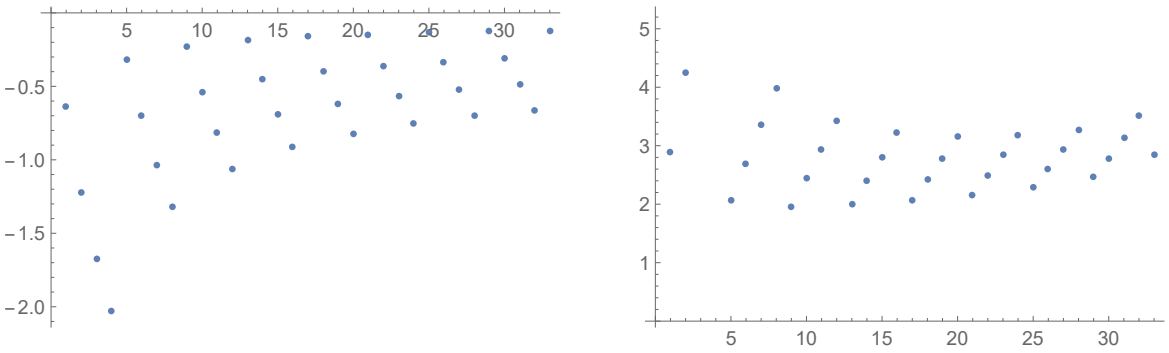
Wyniki powyższych obliczeń dla czterech wybranych rozkładów są zaprezentowane na wykresach 6.1 oraz 6.2. Ogólnym wnioskiem jest, że nasz estymator zachowuje się raczej dobrze bez względu na rozważany rozkład. Jednakże, zachowuje się on raczej słabo w przypadku niektórych wartości p oraz rozkładów F, na przykład dla rozkładu Pareto dla $p \geq 0.8$. Jednakże nie jest to nic zaskakującego, ponieważ nasz estymator jest skonstruowany bez jakiejkolwiek uprzedniej wiedzy o badanej dystrybuancie. Zatem nasza rekomendacja stosowania w praktyce estymatora Q_{10} wydaje się dobrze uzasadniona. Porównanie z estymatorem Q_6 pokazuje, że wnioski z artykułu [19] nie są w pełni uzasadnione. Mimo tego, że Q_6 ma całkiem dobre własności dotyczące błędu średnio–kwadratowego, to okazuje się że jest bardziej obciążony niż pozostałe estyma-

tory. Sprawdzaliśmy również zachowania estymatorów Q_5 oraz Q_8, pominęliśmy wyniki na wykresach 6.1 oraz 6.2, aby były one łatwiejsze do odczytania. Okazuje się, że estymator Q_8 zachowuje się niemal identycznie jak estymator Q_{10}, zaś estymator Q_5 dawał wartości bardzo bliskie do estymatorów Q_8 oraz Q_{10}. Zatem naszym ostatecznym wnioskiem jest rekomendacja stosowania jako estymatorów kwantyla jednego z estymatorów Q_8 lub Q_{10}. Jednakże, jeśli precyzja otrzymanych wyników jest dla nas ważniejsza od szybkości i prostoty obliczeń, zdecydowanie rekomendujemy użycie zaproponowanego przez nas estymatora Q_{10}.

6.5 Porównanie innych znanych estymatorów

W ostatnim podrozdziale wykorzystamy podejście opisane w podrozdziale 6.1 do porównania wybranych L–statystyk jako estymatorów kwantyli. Dla ustalonego rozmiaru próby n i rzędu kwantyla p obliczymy wartości funkcji $r_p(\mathbf{c})$ i $G_p(\mathbf{c})$ dla wektorów $\mathbf{c}$ odpowiadających estymatorom Kaigha-Lachenbrucha, Harella-Davisa, Bernsteina i quasikwantylom (zob. podrozdział 1.3). Aby zilustrować zachowanie powyższych estymatorów ustalamy rozmiar próby $n = 35$, i zakładamy, że chcemy szacować kwantyl rzędu $p = 0.25$.

Przykład 6.1. Jako pierwszy przykład rozważymy estymator $Q_{KL}^{(k)}$ z wartościami parametru $k \in \{3, \ldots, 35\}$. Zgodnie ze wzorem (1.11) różne wartości k odpowiadają różnym wektorom współczynników $\mathbf{c}$. Porównując wartości $r_p(\mathbf{c})$ i $G_p(\mathbf{c})$ dla różnych k chcemy wybrać najlepszą wartość k.



Rysunek 6.3: Wykres wartości $r_p(\mathbf{c})$ i $G_p(\mathbf{c})$ dla $k = 3, \ldots, 35$.

Na Rysunku 6.3 przedstawiamy wykresy wartości $r_p(\mathbf{c})$ i $G_p(\mathbf{c})$ dla $k = 3, \ldots, 35$. Z wykresów tych widzimy, że funkcja r_p przyjmuje wartości najbliższe 0 dla $k = 23, 27, 31, 35$, a G_p ma najmniejsze wartości dla $k = 7, 11, 15, 19$. Wartości te podajemy w Tabeli 6.6. Z danych zawartych w tabeli możemy wywnioskować, że dla $n = 35$

k	7	11	15	19	23	27	31	35
$r_p(\mathbf{c})$	-0.3214	-0.2302	-0.1868	-0.1619	-0.1459	-0.1347	-0.1260	-0.1187
$G_p(\mathbf{c})$	2.0565	1.9655	1.9909	2.0629	2.1623	2.2914	2.4647	2.8444

Tabela 6.6: Wartości $r_p(\mathbf{c})$ i $G_p(\mathbf{c})$ dla $k = 7, 11, 15, 19, 23, 27, 31, 35$.

najlepsze estymatory Kaigha-Lachenbrucha kwantyla rzędu $p = 0.25$ otrzymujemy dla $k = 11$ i $k = 35$. Jednakże nasze kryteria nie pozwalają jednoznacznie stwierdzić, który z nich jest lepszy.

Przykład 6.2. Jako drugi przykład rozważymy estymator $Q_{R1}^{(m)}$ z dopuszczalnymi wartościami parametru $m \in \{1, 2, 3, 4, 5, 6, 7\}$ oraz estymator $Q_{R2}^{(m)}$ dla $m \in \{1, 2, 3\}$. Odpowiadające wektory $\mathbf{c}$ otrzymujemy ze wzorów (1.12) i (1.13). Porównując wartości $r_p(\mathbf{c})$ i $G_p(\mathbf{c})$ dla różnych m chcemy wybrać najlepszą wartość m. Wartości te podajemy w Tabelach 6.7 oraz 6.8.

m	1	2	3	4	5	6	7
$r_p(\mathbf{c})$	-0.7707	-0.6399	-0.4891	-0.3823	-0.3659	-0.4953	-1.0653
$G_p(\mathbf{c})$	3.3971	2.8166	2.5770	2.6533	3.0755	4.3306	10.1195

Tabela 6.7: Wartości $r_p(\mathbf{c})$ i $G_p(\mathbf{c})$ dla estymatora $Q_{R1}^{(m)}$ z $m = 1, 2, 3, 4, 5, 6, 7$.

m	1	2	3
$r_p(\mathbf{c})$	-0.8177	-0.7752	-0.6606
$G_p(\mathbf{c})$	3.7360	3.3825	2.8818

Tabela 6.8: Wartości $r_p(\mathbf{c})$ i $G_p(\mathbf{c})$ dla estymatora $Q_{R2}^{(m)}$ z $m = 1, 2, 3$.

Zauważmy, że wartości funkcji $r_p(\mathbf{c})$ dla estymatorów $Q_{R1}^{(m)}$ i $Q_{R2}^{(m)}$ podane w tabelach 6.7 oraz 6.8 są ujemne, i podobne wyniki uzyskano dla innych wartości n oraz p. To sugeruje, że estymatory te średnio nie doszacowują kwantyli zadanego rzędu. Jednakże

m	1	2	3	4	5	6	7	8
$r_p(\mathbf{c})$	-0.1187	-0.1153	-0.1051	-0.1050	-0.1380	-0.2271	-0.4222	-1.0269
$G_p(\mathbf{c})$	2.4270	2.0722	1.9565	2.0621	2.3287	2.8818	4.2198	10.0506

Tabela 6.9: Wartości $r_p(\mathbf{c})$ i $G_p(\mathbf{c})$ dla zmodyfikowanego estymatora $Q_{R1}^{(m)}$, gdzie $1 \leq m \leq 8$.

m	1	2	3	4
$r_p(\mathbf{c})$	-0.1192	-0.1187	-0.0926	0.0354
$G_p(\mathbf{c})$	2.6636	2.4242	2.1193	1.7393

Tabela 6.10: Wartości $r_p(\mathbf{c})$ i $G_p(\mathbf{c})$ dla zmodyfikowanego estymatora $Q_{R2}^{(m)}$, gdzie $1 \leq m \leq 4$.

zastępując w ich definicjach $\lfloor np \rfloor$ przez $\lceil np \rceil$ otrzymamy wyniki zawarte w Tabelach 6.9 oraz 6.10. Zauważmy, że zmodyfikowane estymatory typu $Q_{R2}^{(m)}$ mają lepsze własności to znaczy wartości r_p są bliższe 0, a wartości G_p są mniejsze niż dla wersji niezmodyfikowanych.

Przykład 6.3. W ostatnim przykładzie chcemy porównać estymatory Kaigha-Lachenbrucha i quasikwantyle wybrane w powyższych przykładach z estymatorami Harrella-Davisa, Bernsteina i estymatorem Q_{10} postaci (6.5) wyznaczonym w Twierdzeniu 5.1. Odpowiadające wartości $r_p(\mathbf{c})$ i $G_p(\mathbf{c})$ prezentujemy w Tabeli 6.11. Z danych zawartych w tabeli wynika, że obydwa nasze kryteria najlepiej spełnia jednocześnie estymator Harella-Davisa. Podobny wynik otrzymaliśmy dla innych wartości n i p.

Estymator	$Q_{KL}^{(11)}$	$Q_{KL}^{(35)}$	$Q_{R1}^{(4)}$	$Q_{R1}^{(3)}$	$Q_{R2}^{(3)}$	Q_{HD}	Q_B	Q_{10}
$r_p(\mathbf{c})$	-0.2302	-0.1187	-0.4891	-0.3823	-0.6606	-0.0228	0.1229	0
$G_p(\mathbf{c})$	1.9655	2.8444	2.5770	2.6533	2.8818	1.8371	1.9449	2.5738

Tabela 6.11: Wartości $r_p(\mathbf{c})$ i $G_p(\mathbf{c})$ dla wybranych estymatorów.

Dodatek A

Dowody pomocniczych lematów

A.1 Dowody Lematów 4.5 i 4.6

W dowodach używamy dwóch własności wielomianów Bernsteina, które zostały podane poniżej. Pierwsza z nich nosi nazwę *własności zmniejszającej zmienność (ang. variation diminishing property, VDP).*

Lemat A.1 (VDP)**.** *Liczba miejsc zerowych dowolnej kombinacji liniowej* $\sum_{i=0}^{n} \alpha_i B_{i,n}$ *wielomianów Bernsteina na przedziale* $(0,1)$ *nie przekracza liczby zmian znaków w ciągu jego współczynników* $\alpha_0, \alpha_1, \ldots, \alpha_n$ *po usunięciu zer. Ponadto, pierwszy oraz ostatni znak kombinacji są takie same jak odpowiednio pierwszy, i ostatni znak ciągu niezerowych współczynników kombinacji.*

Własność VDP dla wielomianów Bernsteina została udowodniona w artykułach [30] oraz [10]. W istocie jest to prosta konsekwencja dobrze znanej reguły znaków Kartezjusza (zob. np. [17]). W pracy [2] przedstawione zostało daleko idące uogólnienie własności VDP do pewnego specjalnego przypadku G–funkcji Meijer'a.

Druga własność nosi nazwę *nierówność Simmonsa.* W artykule [25] można znaleźć jeden z najnowszych dowodów nierówności Simmonsa oraz referencje do innych znanych dowodów.

Lemat A.2. *Dla* $1 \leq k < \frac{n}{2}$ *mamy*

$$\sum_{i=0}^{k-1} B_{i,n}\left(\frac{k}{n}\right) > \sum_{i=k+1}^{n} B_{i,n}\left(\frac{k}{n}\right)$$

i równość zachodzi wtedy i tylko wtedy, gdy $k = \frac{n}{2}$.

Dowód Lematu 4.5. Nierówność $p_j < q_j < p_{j+1}$ dla $1 \leq j \leq n-1$ wynika z dowodu jednoznaczności liczby q_j. Aby udowodnić, że $p_j < \frac{j}{n} < p_{j+1}$, dla $1 \leq j \leq n-1$,

używamy nierówności (1.9). Zatem, z definicji liczb p_j oraz p_{j+1} otrzymujemy

$$F_{j:n}\left(\frac{j}{n}\right) > \frac{1}{2} = F_{j:n}(p_j) \quad \text{oraz} \quad F_{j+1:n}\left(\frac{j}{n}\right) < \frac{1}{2} = F_{j+1:n}(p_{j+1}).$$

Biorąc pod uwagę, że funkcje $F_{j:n}$ oraz $F_{j+1:n}$ są ściśle rosnące, dostajemy szukaną nierówność.

Teraz załóżmy, że $1 \leq j < \frac{n}{2}$. Ponieważ obydwie wartości $\frac{j}{n}$ oraz q_j należą do przedziału (p_j, p_{j+1}), to na mocy własności funkcji $Q_{j,n}$ danej wzorem (4.5) nierówność $\frac{j}{n} < q_j$ zachodzi wtedy i tylko wtedy, gdy $Q_{j,n}\left(\frac{j}{n}\right) < 0$. Na mocy (4.6) jest to równoważne nierówności

$$F_{j:n}\left(\frac{j}{n}\right) < 1 - F_{j+1:n}\left(\frac{j}{n}\right).$$

To z kolei jest równoważne nierówności Simmonsa, co dowodzi (4.8). Aby udowodnić (4.9) wystarczy połączyć wzory (4.3) oraz (4.7) ze wzorem (4.8). Jeżeli $j = \frac{n}{2}$, to nierówność Simmonsa staje się równością równoważną wzorowi

$$1 - F_{n/2:n}\left(\frac{1}{2}\right) = F_{n/2+1:n}\left(\frac{1}{2}\right).$$

Zatem $Q_{n/2,n}\left(\frac{1}{2}\right) = 0$, a więc $q_{n/2} = \frac{1}{2}$. □

W dowodzie Lematu 4.6 potrzebujemy dwóch pomocniczych funkcji. Dla $2 \leq j \leq n-1$ zdefiniujmy funkcje h_j,$\bar{h}_j : [0,1] \to [0,\infty)$ następująco

$$h_j(x) = \frac{1 - F_{j:n}(x)}{1-x}, \quad 0 \leq x < 1,$$

oraz $h_j(1) = h_j(1^-) = 0$. Ponadto

$$\bar{h}_j(x) = \frac{F_{j:n}(x)}{x}, \quad 0 < x \leq 1,$$

oraz $\bar{h}(0) = \bar{h}(0^+) = 0$. Wtedy na mocy wzoru (3.4) mamy $\bar{h}_j(x) = h_{n-j+1}(1-x)$ oraz

$$h_j'(x) = \frac{1}{(1-x)^2}\left[1 - F_{j:n}(x) - (1-x)f_{j:n}(x)\right].$$

Rozwijając $F_{j:n}$ jako sumę wielomianów Bernsteina (zob. (1.8)) dostajemy

$$h_j'(x) = \frac{1}{(1-x)^2}\left[\sum_{i=0}^{j-2} B_{i:n}(x) - (n-j)B_{j-1:n}(x)\right]. \tag{A.1}$$

Stosując własność VDP do wyrażenia w nawiasach możemy zauważyć, że jest ono najpierw dodatnie, a następnie ujemne (w skrócie $+-$). Zatem równanie (3.3) definiujące θ_j ma dokładnie jedno rozwiązanie. Oczywiście h_j' jest również $+-$. Dlatego funkcja

h_j jest ściśle rosnąca na przedziale $(0, \theta_j)$ od wartości 1 dla $x = 0$ do $h_j(\theta_j) > 1$, a następnie ściśle malejąca na przedziale $(\theta_j, 1)$ do 0 dla $x = 1$. Analogicznie,

$$\bar{h}_j'(x) = \frac{1}{x}\left(f_{j:n}(t) - \bar{h}_j(t)\right) = \frac{1}{x^2}\left[(j-1)B_{j,n}(x) - \sum_{i=j+1}^{n} B_{i,n}(x)\right], \tag{A.2}$$

oraz możemy wywnioskować, że funkcja $\bar{h}_j$ jest ściśle rosnąca na przedziale $(0, \xi_j)$ od wartości 0 dla $x = 0$, i ściśle malejąca na przedziale $(\xi_j, 1)$ od $\bar{h}_j(\xi_j) > 1$ do wartości 1 dla $x = 1$. Z tych własności funkcji h_j oraz $\bar{h}_j$ możemy wywnioskować, że dla $2 \le j \le n-1$

$$\theta_j(n) > p \quad \text{wtedy i tylko wtedy, gdy} \quad h_j'(p) < 0,$$

oraz

$$\xi_j(n) > p \quad \text{wtedy i tylko wtedy, gdy} \quad \bar{h}_j'(p) > 0. \tag{A.3}$$

Dowód Lematu 4.6(a). Ponieważ $\theta_1 = 0 < \theta_2$ oraz $\theta_{n-1} < 1 = \theta_n$, to na mocy symetrii (3.5) widzimy, że wystarczy rozważyć liczby θ_j, $2 \le j \le n-1$. Dla $0 \le x \le 1$ oraz $2 \le j \le n-2$ rozważamy funkcję

$$g_j(x) = h_{j+1}(x) - h_j(x) = \frac{n}{n-j} B_{j,n-1}(x).$$

Wtedy $g_j(0) = g_j(1) = 0$ oraz z własności wielomianów Bernsteina dostajemy, że g_j jest funkcją rosnącą na przedziale $\left(0, \frac{j}{n-1}\right)$ i malejącą na przedziale $\left(\frac{j}{n-1}, 1\right)$. Ponieważ

$$\theta_j \in \left(0, \frac{j-1}{n-1}\right) \subset \left(0, \frac{j}{n-1}\right),$$

to suma $h_j + g_j$ jest ściśle rosnąca na przedziale $(0, \theta_j]$ oraz ściśle malejąca na przedziale $\left[\frac{j}{n-1}, 1\right)$. Ponieważ suma ta jest ciągła i różniczkowalna, to istnieją liczby a oraz b takie, że $\theta_j < a < b < \frac{j}{n-1}$ oraz suma $h_j + g_j$ jest rosnąca na przedziale $(0, a)$ i malejąca na przedziale $(b, 1)$. Z drugiej strony mamy $h_j + g_j = h_{j+1}$, a więc jednoznacznie wyznaczone maksimum sumy jest osiągnięte w punkcie

$$\theta_{j+1} \in [a, b] \subset \left(\theta_j, \frac{j}{n-1}\right).$$

W szczególności $\theta_j < \theta_{j+1}$ dla $2 \le j \le n-2$. □

Dowód Lematu 4.6(b). Na mocy (A.3) nierówność $\xi_j > q_j$ jest równoważna warunkowi $\bar{h}_j'(q_j) > 0$. Z definicji liczb q_j mamy

$$\sum_{i=j+1}^{n} B_{i,n}(q_j) = \sum_{i=0}^{j-1} B_{i,n}(q_j). \tag{A.4}$$

Stosując (A.2) otrzymujemy dla $2 \leq j \leq n-1$

$$\bar{h}_j'(q_j) = \frac{1}{q_j^2}\left[(j-1)B_{j,n}(q_j) - \sum_{i=0}^{j-1} B_{i,n}(q_j)\right].$$

Dla $j = 2, 3$ określamy funkcje

$$\overline{g}_j(p) = (j-1)B_{j,n}(p) - \sum_{i=0}^{j-1} B_{i,n}(p).$$

Wtedy $\bar{h}_j'(q_j) = \frac{1}{q_j^2}\overline{g}_j(q_j)$. Z własności VDP funkcje $\overline{g}_j$ są $-+$ na przedziale $(0,1)$.

Dla $j = 2$ i $n = 6$ mamy $Q_{2,6}\left(\frac{41}{120}\right) > 0$, a więc $q_2(6) < \frac{41}{120}$. Ponadto bezpośrednie sprawdzenie pokazuje, że $\overline{g}_2\left(\frac{41}{120}\right) < 0$, czyli $\overline{g}_2(q_2) < 0$. Zatem $\bar{h}_2'(q_2) < 0$ oraz $\xi_2(6) < q_2(6)$.

Dalej, z Lematu 4.5 mamy $\frac{j}{n} < q_j$ dla $1 \leq j < \frac{n}{2}$, a zatem wystarczy udowodnić, że

$$\overline{g}_j\left(\frac{j}{n}\right) > 0 \tag{A.5}$$

dla $j = 2$, $n = 5$ lub $j = 3$, $n \geq 7$. Jeżeli $n = 5$ oraz $j = 2$ to $\overline{g}_2(\frac{2}{5}) = \frac{3^3}{5^5} > 0$. Ponieważ $q_2(5) > \frac{2}{5}$, to mamy również $\overline{g}_{2,5}(q_2) > 0$ oraz $\bar{h}_2'(q_2) > 0$. Ponadto

$$\overline{g}_3\left(\frac{3}{n}\right) = \frac{1}{n^3}\left(1 - \frac{3}{n}\right)^{n-3} c(n),$$

gdzie $c(n) = 23n^3 - 60n^2 - 9n + 54$. Elementarne obliczenia pokazują, że $c(6) > 0$ oraz $c(n)$ jest ciągiem rosnącym dla $n \geq 6$. To implikuje (A.5) dla $j = 3$ i $n \geq 7$. □

Dowód Lematu 4.6(c). Aby porównać liczby $\frac{2}{n}$ oraz $\xi_2(n)$ zapiszmy funkcję $\bar{h}_2'$ w postaci

$$\bar{h}_2'(p) = \frac{1}{p^2}\left[(1-p)^n + np(1-p)^{n-1} + n(n-1)p^2(1-p)^{n-2} - 1\right].$$

Wtedy

$$\bar{h}_2'\left(\frac{2}{n}\right) = \frac{n^2}{4}\left[\frac{1}{n^2}\left(1 - \frac{2}{n}\right)^{n-2}\left(7n^2 - 12n + 4\right) - 1\right].$$

Dla $n = 6$ mamy $\bar{h}_2'\left(\frac{2}{6}\right) > 0$, a więc $\xi_2(6) > \frac{2}{6}$. Ponadto $\bar{h}_2'\left(\frac{2}{7}\right) < 0$ oraz wyrażenie w nawiasach kwadratowych jest ściśle malejące względem zmiennej $n \geq 7$. Z równania (A.3) otrzymujemy $\xi_2(n) < \frac{2}{n}$ dla $n \geq 7$. □

Dowód Lematu 4.6(d). Chcemy udowodnić, że $p_j < \xi_j$, dla $j = 2, 3$ oraz odpowiedniej wartości n. Potrzebujemy zatem pokazać, że $\bar{h}_j'(p_j) > 0$ dla $j = 2, 3$. Z definicji liczb p_j mamy $F_{j:n}(p_j) = 1 - F_{j:n}(p_j)$, a więc

$$\sum_{i=j}^{n} B_{i,n}(p_j) = \sum_{i=0}^{j-1} B_{i,n}(p_j). \tag{A.6}$$

Zatem ze wzoru (A.2) otrzymujemy

$$\overline{h}_j'(p_j) = \frac{1}{p_j^2}\left[jB_{j,n}(p_j) - \sum_{i=0}^{j-1} B_{i,n}(p_j)\right].$$

Dla $j = 2, 3$ rozważamy pomocnicze funkcje

$$\tilde{g}_j(p) = jB_{j,n}(p) - \sum_{i=0}^{j-1} B_{i,n}(p), \quad p \in (0,1).$$

Na mocy własności VDP funkcja $\tilde{g}_{j,n}$ jest $-+$. Oczywiście $\overline{h}_j'(p_j) > 0$ wtedy i tylko wtedy, gdy $\tilde{g}_k(p_j) > 0$. Dla $j = 2$ mamy $\tilde{g}_2\left(\frac{1}{n}\right) = \frac{1}{n^2}\left(1 - \frac{1}{n}\right)^{n-2}(n-1) > 0$. Ponieważ $\frac{1}{n} < p_2$ oraz $\tilde{g}_2$ jest $-+$, to $\tilde{g}_2(p_2) > 0$. Zatem $p_2 < \xi_2$ dla $n \geq 3$. Dla $j = 3$ mamy $\tilde{g}_3\left(\frac{a}{n}\right) = \frac{1}{n^3}\left(1 - \frac{a}{n}\right)^{n-3} c_a(n)$, gdzie

$$c_a(n) = \frac{1}{2}(a^3 - a^2 - 2a - 2)n^3 + \frac{1}{2}a(6 + 5a - 2a^2)n^2 - \frac{1}{2}a^2(a+6)n + a^3.$$

Podstawiając $a = 2.5$ po żmudnych obliczeniach otrzymujemy $F_{3:n}\left(\frac{2.5}{n}\right) \leq \frac{1}{2}$ dla $n \geq 5$ oraz $\tilde{g}_3\left(\frac{2.5}{n}\right) > 0$ dla $n \geq 4$. Zatem $p_3 > \frac{2.5}{n}$ oraz $\tilde{g}_3(p_3) > 0$ dla $n \geq 5$. □

Aby udowodnić części (e) oraz (f) Lematu 4.6 musimy rozważyć nierówności pomiędzy wartościami ξ_j oraz p_{j+1}. Z warunku (A.3) wiemy, że $p_{j+1} < \xi_j$ wtedy i tylko wtedy, gdy $\bar{h}_j'(p_{j+1}) > 0$. Zastępując j przez $j+1$ we wzorze (A.6) otrzymujemy

$$\overline{h}_j'(p_{j+1}) = \frac{1}{p_{j+1}^2}\left[-\sum_{i=0}^{j-1} B_{i,n}(p_{j+1}) + (k-2)B_{j,n}(p_{j+1})\right].$$

Zdefiniujmy funkcje g_j, $2 \leq j < n$, następująco

$$g_j(p) = -\sum_{i=0}^{j-1} B_{i,n}(p) + (j-2)B_{j,n}(p).$$

Oczywiście $g_j(p) = p_{j+1}^2 \bar{h}_j'(p_{j+1})$, a więc nierówność $p_{j+1} < \xi_j$ jest równoważna warunkowi $g_j(p_{j+1}) > 0$.

Zbadamy teraz wybrane własności funkcji g_j. Dla $j \geq 3$ mamy $g_j(0) = -1$ oraz $g_j(1) = 0$ oraz na mocy własności VDP g_j jest $-+$. Pochodna wielomianu Bernsteina $B_{i,n}$ wyraża się wzorem

$$B_{i,n}'(p) = n(B_{i-1,n-1}(p) - B_{i,n-1}(p)). \tag{A.7}$$

Zatem pochodną funkcji g_j jest

$$g_j'(p) = n[(j-1)B_{j-1,n-1}(p) - (j-2)B_{j,n-1}(p)],$$

a więc g'_j jest $+-$. Ponadto $g'_{j,n}(p) = 0$ dla $p = \rho_k = \frac{j(j-1)}{nj-2n+j}$. Zatem g_j jest rosnąca na przedziale $(0, \rho_j)$ i malejąca na przedziale $(\rho_j, 1)$. Dodatkowo, dla $p \in (0,1)$ mamy $g_j(p) = g_{j+1}(p)$ dla $p = \frac{j+1}{n+1}$ i możemy łatwo pokazać, że $\frac{j+1}{n+1} < \rho_j$. Zatem $\frac{j}{n} < \rho_j$ oraz $\frac{j+1}{n} < \rho_{j+1}$. W szczególności, g_j rośnie na przedziale $\left(\frac{j}{n}, \frac{j+1}{n+1}\right)$ oraz g_{j+1} rośnie na przedziale $\left(\frac{j+1}{n+1}, \frac{j+1}{n}\right)$. Powyższe rozważania implikują, że dla $j \geq 3$ mamy

$$g_j\left(\frac{j}{n}\right) < g_j\left(\frac{j+1}{n+1}\right) = g_{j+1}\left(\frac{j+1}{n+1}\right) < g_{j+1}\left(\frac{j+1}{n}\right). \tag{A.8}$$

Dowód Lematu 4.6(e). Chcemy udowodnić, że $\xi_j < p_{j+1}$ dla $j = 2, 3$, lub równoważnie $g_2(p_3) < 0$ oraz $g_3(p_4) < 0$. Dla $j = 2$ mamy równość $g_2(p) = -B_{0,n}(p) - B_{1,n}(p)$, a więc funkcja g_j jest ujemna dla wszystkich $p \in (0,1)$. Zatem $\xi_2 < p_3$ dla $n \geq 3$.

Dla $j = 3$ mamy $g_3(p) = -B_{0,n}(p) - B_{1,n}(p) - B_{2,n}(p) + B_{3,n}(p)$, więc dla $a \in [3,4]$

$$g_3\left(\frac{a}{n}\right) = \frac{1}{6(n-a)^3}\left(1-\frac{a}{n}\right)^n \left[6a^3 - (18+7a)a^2 n + \right.$$
$$\left. + 3a(6+5a)n^2 + (a^3 - 3a^2 - 6a - 6)n^3\right].$$

Podstawiając $a = \frac{29}{8}$ po prostych, ale żmudnych obliczeniach otrzymujemy $g_3\left(\frac{29}{8n}\right) < 0$ dla $n \geq 11$ oraz $F_{4:n}\left(\frac{29}{8n}\right) > \frac{1}{2}$ dla $n \geq 7$. To implikuje, że $p_4 < \frac{29}{8n}$, a więc $g_3(p_4) < 0$ ponieważ g_3 jest $-+$. Zatem $\xi_3 < p_4$ dla $n \geq 11$. Dla $n = 10$ nierówność $g_3(p_4) < 0$ wynika z bezpośredniego sprawdzenia. □

Dowód Lematu 4.6(f). Musimy udowodnić, że $g_j(p_{j+1}) > 0$ dla $4 \leq j \leq n-1$. Ponieważ g_j jest $-+$ oraz $p_{j+1} > \frac{j}{n}$, zatem wystarczy udowodnić, że $g_j\left(\frac{j}{n}\right) > 0$. Jednakże, udowodnimy, że jest to prawdą dla $j \geq 5$, ale dla $j = 4$ mamy $g_4\left(\frac{4.5}{n}\right) > 0$ oraz $p_5 > \frac{4.5}{n}$ dla $n \geq 11$.

W celu udowodnienia, że $g_j\left(\frac{j}{n}\right) > 0$ wystarczy udowodnić to dla $j = 5$ i skorzystać ze wzoru (A.8). Po żmudnych obliczeniach otrzymamy, że wyrażenie $g_5\left(\frac{5}{n}\right)$ jest malejące względem zmiennej $n \geq 10$ i jego granicą jest $\frac{51}{4e^5} > 0$. Zatem $g_5\left(\frac{5}{n}\right) > 0$ dla $n \geq 10$.

Podobne obliczenia, pokazują że wyrażenie $g_4\left(\frac{4.5}{n}\right)$ jest malejące i jego granicą jest $\frac{215}{64e^{9/2}} > 0$. Ponadto mamy $F_{5:9}\left(\frac{1}{2}\right) = \frac{1}{2}$ i $F_{5:n}\left(\frac{4.5}{n}\right)$ jest malejące dla $n \geq 9$. Zatem $F_{5:n}\left(\frac{4.5}{n}\right) < \frac{1}{2}$ więc $\frac{4.5}{n} < p_5$. Stąd $g_4\left(\frac{4.5}{n}\right) > 0$, a więc $g_4(p_5) > 0$ dla $n \geq 9$. □

Dowód Lematu 4.6(g). Z punktu (f) i równań (3.5) oraz (4.10) mamy nierówność $p_{k+1} < \xi_k$ dla $4 \leq k \leq n-1$, a więc $\theta_{k+1} < p_k$ dla $1 \leq k \leq n-4$. □

A.2 Dowód Twierdzenia 4.3

Dowód Twierdzenia 4.3(a). Dla $1 \leq j \leq \frac{n}{2}$ na mocy Lematu 4.5(a) otrzymujemy zawieranie $\left[\frac{j}{n}, q_j\right] \subset (p_j, p_{j+1})$. Dlatego dla funkcji $t_{n,p}$ określonej wzorem (3.18) dla

$p \in \left[\frac{j}{n}, q_j\right]$ mamy

$$t_{n,p}(j) < 0 < t_{n,p}(j+1), \tag{A.9}$$

(por. uwagi po wzorze (4.3)). Następnie używając Wniosku 3.4(a), nierówności (A.9) oraz Wniosku 4.1(a) dla $n \geq 3$ oraz $p \in \left[\frac{1}{n}, q_1\right]$ mamy

$$r_{n,p}(1) < t_{n,p}(1) < 0 < t_{n,p}(2) = r_{n,p}(2). \tag{A.10}$$

Zatem $1 \leq k\left(n, \frac{1}{n}\right) \leq k\left(n, q_1\right) \leq 2$ na mocy Lematu 4.1.

Najpierw udowodnimy, że $k(n, q_1) = 2$ lub równoważnie $s_{n,q_1}(1) > s_{n,q_1}(2)$. Dla każdego $p \in \left[\frac{1}{n}, q_1\right]$ ze wzoru (A.10) mamy $s_{n,p}(2) = 1 - 2F_{2:n}(p)$. Ponadto, używając ponownie wzoru (A.10) oraz definicji liczby q_1 otrzymujemy

$$s_{n,q_1}(1) > 2 - F_{1:n}(q_1) - 1 = s_{n,q_1}(2).$$

Aby udowodnić, że $k\left(n, \frac{1}{n}\right) = 1$ musimy pokazać, że dla $n \geq 3$

$$s_{n,\frac{1}{n}}(1) =: a_n < b_n := s_{n,\frac{1}{n}}(2). \tag{A.11}$$

Elementarne ale żmudne obliczenia przy użyciu wzoru (A.10) dla $p = \frac{1}{n}$ pokazują, że ciąg liczb

$$b_n = 2\left[\left(1 - \frac{1}{n}\right)^n + \left(1 - \frac{1}{n}\right)^{n-1}\right] - 1$$

jest ściśle malejący do swojej granicy $\frac{4}{\mathrm{e}} - 1 > \frac{4}{9}$ oraz $a_n < \frac{4}{9}$ dla $n \geq 3$. Zatem dowolny element ciągu $\{b_n\}$ jest większy od każdego elementu ciągu $\{a_n\}$. W szczególności, nierówność (A.11) zachodzi dla wszystkich $n \geq 3$, co kończy dowód (a). □

Dowód Twierdzenia 4.3(b). Załóżmy, że $n \geq 6$. Na mocy Wniosku 4.1(b) oraz (c) mamy $\left[\frac{2}{n}, q_2\right] \subset (\theta_3, \xi_3)$, więc dla $p \in \left[\frac{2}{n}, q_2\right]$ na mocy (A.9) mamy

$$r_{n,p}(3) = t_{n,p}(3) > 0. \tag{A.12}$$

Z drugiej strony, ponieważ funkcja $r_{n,p}$ jest ściśle malejąca względem zmiennej p, to

$$r_{n,q_2}(2) < r_{n,\frac{2}{n}}(2) \leq t_{n,\frac{2}{n}}(2) < 0. \tag{A.13}$$

Tutaj druga nierówność wynika z faktu, że $\frac{2}{n} > \theta_2$, a trzecia nierówność wynika ze wzoru (A.9). Podsumowując, dla $p = \frac{2}{n}$ oraz $p = q_2$ mamy $r_{n,p}(2) < 0 < r_{n,p}(3)$, i dlatego

$$2 \leq k\left(n, \frac{2}{n}\right) \leq k\left(n, q_2\right) \leq 3.$$

Najpierw udowodnimy, że $k\left(n, q_2\right) = 3$ lub równoważnie $s_{n,q_2}(2) > s_{n,q_2}(3)$. Istotnie, ze wzoru (A.12) mamy $s_{n,q_2}(3) = 1 - 2F_{3:n}(q_2)$. Co więcej, z Wniosku 4.1(b)

oraz (c) mamy $q_2 > \xi_2 > \theta_2$, a więc $r_{n,q_2}(2) < t_{n,q_2}(2) < 0$. Zatem $s_{n,q_2}(2) > 2F_{2:n}(q_2) - 1 = s_{n,q_2}(3)$ z definicji liczby q_2.

Teraz udowodnimy, że $k\left(n, \frac{2}{n}\right) = 2$ lub równoważnie $s_{n,\frac{2}{n}}(2) < s_{n,\frac{2}{n}}(3)$. Ponieważ $\frac{2}{n} \in (\theta_3, \xi_3)$, to ze wzoru (A.12) mamy

$$s_{n,\frac{2}{n}}(3) = 1 - 2F_{3:n}\left(\frac{2}{n}\right). \tag{A.14}$$

Dla $n = 6$ mamy również $\frac{2}{6} \in (\theta_3, \xi_2)$, więc ze wzoru (A.13) otrzymujemy

$$s_{6,\frac{2}{6}}(2) = 2F_{2:n}\left(\frac{2}{6}\right) - 1 < s_{6,\frac{2}{6}}(3)$$

na mocy wzoru (A.14) oraz nierówności Simmonsa. Dla $n \geq 7$ definiujemy ciągi

$$c_n := s_{n,\frac{2}{n}}(2), \quad d_n := s_{n,\frac{2}{n}}(3).$$

Żmudne obliczenia pokazują, że d_n jest ciągiem ściśle malejącym do swojej granicy $\frac{10}{\mathrm{e}^2} - 1$ oraz $c_n < \frac{10}{\mathrm{e}^2} - 1$ dla $n \geq 7$. Zatem $c_n < d_n$ dla wszystkich $n \geq 7$, co kończy dowód. □

A.3 Dowód równości (5.5)

Pokażemy, że dla $p \in (0,1)$ spełniających założenia punktów (b) i (c) Twierdzenia 5.1 zachodzi równość (5.5), która przy użyciu oznaczeń (5.1) i (5.2) przyjmuje postać

$$r_{n,p}^{(\alpha)}(j) = \overline{B}_{n,p}^{(\alpha)}(j) + \underline{B}_{n,p}^{(\alpha)}(j).$$

Oczywiście dla $p = p_k$ równość zachodzi, bo wtedy $\alpha = 1$, a więc $r_{n,p}^{(\alpha)}(j) = r_{n,p}(j)$ oraz $\overline{B}_{n,p}^{(\alpha)}(j) = \overline{B}_{n,p}(j)$ i $\underline{B}_{n,p}^{(\alpha)}(j) = \underline{B}_{n,p}(j)$. Skorzystamy z oznaczeń i wyników z podrozdziału 3.3.

Zauważmy, że funkcja $F_{\mathbf{c}}$ przyjmuje wtedy postać

$$F_{j:n}^{(\alpha)}(u) = (1-\alpha)F_{j:n}(u) + \alpha F_{j+1:n}(u)$$

z odpowiadającą jej gęstością $f_{j:n}^{(\alpha)}$. Łatwo teraz zauważyć, że dla $u \in [0,1]$ mamy $F_{j+1:n}(u) \leq F_{j:n}^{(\alpha)}(u) \leq F_{j:n}^{(}u)$. Zatem jeśli $\theta_j^{(\alpha)}$ i $\xi_j^{(\alpha)}$ oznaczają odpowiednio rozwiązania równań

$$1 - F_{j:n}^{(\alpha)}(\theta) = (1-\theta)f_{j:n}^{(\alpha)}(\theta)$$

oraz

$$F_{j:n}^{(\alpha)}(\xi) = \xi f_{j:n}^{(\alpha)}(\xi),$$

to dla $1 \leq j < n$ mamy

$$\theta_j < \theta_j^{(\alpha)} < \theta_{j+1}, \quad \xi_j < \xi_j^{(\alpha)} < \xi_{j+1}. \tag{A.15}$$

Oczywiście θ_j i ξ_j są liczbami zdefiniowanymi w podrozdziale 3.1. W dowodzie Twierdzenia 5.1 pokazaliśmy, że:

(i) dla $4 \leq j \leq n-4$ mamy $(p_j, p_{j+1}) \subset (\theta_{j+1}, \xi_j)$,

(ii) dla $j = 2, 3$ mamy $(p_j, \xi_j) \subset (\theta_{j+1}, \xi_j)$,

(iii) dla $j = n-3$ lub $j = n-2$ mamy $(\theta_{j+1}, p_{j+1}) \subset (\theta_{j+1}, \xi_j)$.

Na mocy (A.15) każdy z powyższych przedziałów po lewej stronie jest zawarty w odpowiednim przedziale postaci $(\theta_j^{(\alpha)}, \xi_j^{(\alpha)})$. Zatem dla p należących do jednego z przedziałów po lewej stronie inkluzji (i), (ii) lub (iii) mamy

$$\overline{B}_{n,p}^{(\alpha)}(j) + \underline{B}_{n,p}^{(\alpha)}(j) = \frac{1 - 2F_{j:n}^{(\alpha)}(p)}{\sqrt{p(1-p)}} = r_{n,p}^{(\alpha)}(j).$$

Pierwsza równość wynika ze wzorów (3.19) i (3.20) , gdyż $\mathbf{c} = (0, \ldots, 0, 1-\alpha, \alpha, 0, \ldots, 0)$ spełnia założenia przypadku (A) w podrozdziale 3.3. Druga równość wynika z dowodu Twierdzenia 5.1.

Bibliografia

[1] B. C. Arnold, N. Balakrishnan, H. N. Nagaraja. *A first course in order statistics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008.

[2] M. Bieniek. Variation diminishing property of densities of uniform generalized order statistics. *Metrika*, 65:297–309, 2007.

[3] M. Bieniek, L. Pańczyk. On the choice of the optimal quantile estimators based on a pair of order statistics. (praca przesłana do druku), 2023.

[4] M. Bieniek, L. Pańczyk. On the choice of the optimal single order statistic in quantile estimation. *Annals of the Institute of Statistical Mathematics*, 75:303–333, 2023.

[5] P. Billingsley. *Probability and measure*. John Wiley & Sons, Inc., New York, 1995.

[6] G. Blom. *Statistical Estimates and Transformed Beta-Variables*. John Wiley & Sons, Inc., New York, 1958.

[7] C. Cheng. The Bernstein polynomial estimator of a smooth quantile function. *Statist. Probab. Lett.*, 24(4):321–330, 1995.

[8] H. A. David, H. N. Nagaraja. *Order statistics*. Wiley-Interscience, Hoboken, NJ, 2003.

[9] I. Frohne, R. J. Hyndman. Sample quantiles. R package documentation available at https://stat.ethz.ch/R-manual/R-devel/library/stats/html/quantile.html, 2004.

[10] L. Gajek, T. Rychlik. Projection method for moment bounds on order statistics from restricted families. II. Independent case. *J. Multivariate Anal.*, 64:156–182, 1998.

[11] E. Gumbel. La probabilité des hypothèses. *Comptes Rendus de l'Académie des Sciences Paris*, 209:645–647, 1939.

[12] F. E. Harrell, C. E. Davis. A new distribution-free quantile estimator. *Biometrika*, 69:635–640, 1982.

[13] R. J. Hyndman, Y. Fan. Sample quantiles in statistical packages. *The American Statistician*, 50:361–365, 1996.

[14] R. Kaas, J. Buhrman. Mean, median and mode in binomial distributions. *Statistica Neerlandica*, 34:13–18, 1980.

[15] W. Kaigh, P. A. Lachenbruch. A generalized quantile estimator. *Communications in Statistics - Theory and Methods*, 11(19):2217–2238, 1982.

[16] J. P. Keating, R. Tripathi. Percentiles, estimation of. *Encyclopedia of Statistical Sciences*, wolumen 9, strony 6054–6060. John Wiley & Sons, Inc., New York, 2006.

[17] V. Komornik. Another short proof of Descartes's rule of signs. *Amer. Math. Monthly*, 113:829–830, 2006.

[18] P. Kozyra. *Sharp bounds on the moments of linear combinations of order statistics and kth records*. Praca doktorska, Instytut Matematyczny, Polska Akademia Nauk, Warszawa, 2017. http://www.math.us.edu.pl/pkozyra/papers/koz dr fin.pdf.

[19] L. Makkonen. Bringing closure to the plotting position controversy. *Communications in Statistics - Theory and Methods*, 37(3):460–467, 2008.

[20] S. Moriguti. A modification of Schwarz's inequality with applications to distributions. *Ann. Math. Statistics*, 24:107–113, 1953.

[21] NumPy Documentation. https://numpy.org/doc/stable/reference/generated/numpy.percentile.html, 2003.

[22] A. Okolewski, T. Rychlik. Sharp distribution-free bounds on the bias in estimating quantiles via order statistics. *Statist. Probab. Lett.*, 52:207–213, 2001.

[23] N. Papadatos. Maximum variance of order statistics. *Ann. Inst. Statist. Math.*, 47:185–193, 1995.

[24] R. S. Parrish. Comparison of quantile estimators in normal sampling. *Biometrics*, 46:247–257, 1990.

[25] O. Perrin, E. Redside. Generalization of simmons theorem. *Statist. Probab. Lett.*, 77:604–606, 2007.

[26] R.-D. Reiss. *Approximate distributions of order statistics*. Springer-Verlag, New York, 1989.

[27] S. I. Resnick. *Extreme values, regular variation, and point processes*. Springer-Verlag, New York, 1987.

[28] W. Rudin. *Principles of mathematical analysis*. McGraw-Hill Book Co., New York, 1976.

[29] T. Rychlik. *Projecting statistical functionals*. Springer-Verlag, New York, 2001.

[30] I. J. Schoenberg. On variation diminishing approximation methods. *On numerical approximation. Proceedings of a Symposium, Madison, April 21-23, 1958*, Edited by R. E. Langer, strony 249–274. The University of Wisconsin Press, Madison, 1959.

[31] R. J. Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, Inc., New York, 1980.

[32] M. E. Sfakianakis, D. G. Verginis. A new family of nonparametric quantile estimators. *Communications in Statistics - Simulation and Computation*, 37(2):337–345, 2008.

[33] S. J. Sheather, J. S. Marron. Kernel quantile estimators. *Journal of the American Statistical Association*, 85(410):410–416, 1990.

[34] W. Weibull. *The Phenomenon of Rupture in Solids*. Ingenjörs Vetenskaps Akademiens Handlingar, 1939.

[35] Wolfram Research. Quantile, Wolfram Language function. https://reference.wolfram.com/language/ref/Quantile.html (updated 2007), 2003.

[36] R. Zieliński. Optimal nonparametric quantile estimators. Towards a general theory. A survey. *Comm. Stat. Theory Methods*, 38(7):980–992, 2009.