ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 227, 2009

Tomasz Jurkiewicz[*], *Arkadiusz Kozłowski*[**]

O WYZNACZANIU DOMINANTY ROZKŁADU CECHY CIĄGŁEJ W SZEREGACH SZCZEGÓŁOWYCH

1. WPROWADZENIE

Parametryczna analiza rozkładu cechy w zbiorowości jest jedną z najczęściej wykonywanych analiz, tak dla samego opisu badanej zbiorowości jak również w przypadkach np. doboru zmiennych do modelu. Do głównych własności rozkładu należy jego położenie (tendencja centralna). Podstawowymi miarami tendencji centralnej, określającymi położenie środka rozkładu, są miary przeciętne tj. klasyczna średnia arytmetyczna oraz wśród miar pozycyjnych mediana i dominanta. O ile średnia arytmetyczna jest wrażliwa na wartości skrajne pojawiające się na krańcach rozkładu, o tyle mediana i dominanta są na wartości skrajne odporne. Często też, np. w naukach medycznych, więcej informacji dostarcza wartość najbardziej prawdopodobna niż pozostałe miary przeciętne (Bickel 2002, s. 154). Dodatkowo moda ma tę zaletę, że może być wykorzystywana do estymacji asymetrii rozkładu (Rousseeuw, Leroy 1987, za Bickel 2002, s. 154).

Szacowanie miar przeciętnych w przypadku cechy dyskretnej nie jest skomplikowane. Również w przypadku zmiennej ciągłej oraz danych w postaci szeregu szczegółowego dysponujemy dobrymi estymatorami zarówno średniej arytmetycznej jak i mediany. Problematyczne jest w tym przypadku określenie wartości dominanty, brak bowiem jednego, najlepszego estymatora. Dominanta w myśl ogólnej definicji jest wartością najbardziej prawdopodobną. Dla zmiennej skokowej może to być jednoznacznie utożsamione z wartością najczęściej występującą, oczywiście pod warunkiem, że wielkość próby będzie odpowiednia do ilości wariantów cechy. W przypadku zmiennej ciągłej takiego utożsamienia nie można dokonać chociażby z tego powodu, że dwie wartości losowe pobrane z rozkładu ciągłego praktycznie nigdy nie będą sobie równe. Dodatkowe problemy

[*] Dr, Katedra Statystyki, Uniwersytet Gdański.
[**] Mgr, Katedra Statystyki, Uniwersytet Gdański.

pojawiają się w sytuacji, w której występuje więcej niż jedno maksimum, a więc w przypadku rozkładów wielomodalnych.

Definicja, która jest zgodna z intuicyjnym rozumieniem dominanty, również w przypadku zmiennej ciągłej, stwierdza, że moda „jest to wartość zmiennej odpowiadająca punktowi maksimum idealnej krzywej, możliwie najlepiej dopasowanej do rozkładu rzeczywistego" (Yule, Kendall 1966, s. 135). Yule i Kendall jako sposób wyznaczania dominanty sugerują „pewien proces wygładzania nieregularności występujących w rozkładzie rzeczywistym", a jako najlepszy sposób wygładzania (uwzględniający wszystkie obserwacje) wskazują proces „dopasowania idealnej krzywej liczebności o danym z góry równaniu do danych rzeczywistych" i za modę, zgodnie z definicją, przyjęcie maksimum tej funkcji. Problemem jest tu jednak prawidłowe określenie teoretycznego rozkładu cechy. Jednym z rozwiązań może być wykorzystanie wiedzy a priori, wymaga to jednak dość zaawansowanej wiedzy zarówno statystycznej, jak i znajomości badanego zjawiska.

Jedną z popularnych metod wyznaczania dominanty jest pogrupowanie obserwacji w przedziały, a następnie – wykorzystując przedział najgęstszy i dwa sąsiednie – oszacowanie wartości modalnej za pomocą wzoru interpolacyjnego. Jej niewątpliwą zaletą jest prostota obliczeń i łatwość zastosowania w praktyce. W literaturze przedmiotu można spotkać wiele różnych metod estymacji dominanty. Niektóre z nich, np. metoda półprób HSM i półrozstępów HRM, opierają się na prostym postępowaniu iteracyjnym, w którym ciąg obserwacji dzielony jest na mniejsze podpróby, tak aby wybrana podpróba zawierała wartość modalną. Innym podejściem jest stosowana np. w metodach standardowej parametrycznej SPM i odpornej parametrycznej RPM transformacja pierwotnej zmiennej w zmienną o przybliżonym rozkładzie normalnym. Następnie na podstawie oszacowania parametrów rozkładu nowej zmiennej wyznacza się dominantę w rozkładzie pierwotnym. Rozwój elektronicznych technik obliczeniowych powoduje, że wciąż pojawiają się nowe, lepsze ale i bardziej skomplikowane i czasochłonne, estymatory dominanty.

W wielu podręcznikach statystyki problem wyznaczania dominanty w przypadku cechy ciągłej i danych szczegółowych jest całkowicie pomijany. Również w najpopularniejszych pakietach statystycznych czy arkuszu Excel dominanta wyznaczana jest na podstawie częstości wystąpień, tak jak dla cech dyskretnych. Celem pracy jest przedstawienie różnych możliwości wyznaczania dominanty cechy ciągłej w szeregach szczegółowych.

W praktyce statystycznej, jak i w dydaktyce, często zależy badaczowi na stosowaniu w miarę prostych metod, które jednak nie będą znacząco gorsze pod względem efektywności od najlepszych. Stąd też potrzeba wiedzy, w jakich okolicznościach metody proste są wystarczająco efektywne, a kiedy ich stosowanie prowadzi do dużych błędów. Równoległym celem pracy jest porównanie efektywności metod szacowania dominanty dla rozkładów o różnym stopniu asymetrii.

2. METODY WYZNACZANIA DOMINANTY

2.1. Grupowanie i interpolacja

Jedną z potencjalnych metod wyznaczania dominanty jest wykorzystanie wzoru interpolacyjnego dla szeregów rozdzielczych przedziałowych, znanego z wielu podręczników statystyki (Sobczyk 2002, s. 42, Szulc 1976, s. 175).

W tym celu należy pogrupować dane w przedziały klasowe, a następnie po ustaleniu przedziału zawierającego najwięcej obserwacji (najgęstszego), skorzystać ze wzoru interpolacyjnego:

$$\hat{D}_{Grup} = x_{i0} + \frac{n_i - n_{i-1}}{(n_i - n_{i-1}) + (n_i - n_{i+1})} \cdot c_i \qquad (1)$$

Jeżeli przedziały mają różne rozpiętości należy zamiast liczebności posługiwać się gęstościami przedziałów – o czym rzadko wspomina się w podręcznikach z zakresu statystyki. Jednym z aspektów, które należałoby rozważyć przed wykorzystaniem tej metody, jest kwestia liczby przedziałów. Wskazania określające optymalną liczbę przedziałów przedstawiane w większości podręczników, przeważnie jako funkcje liczebności próby, dotyczą raczej prezentacji tabelarycznej materiału statystycznego. Niekoniecznie muszą być one optymalne dla wyznaczenia dominanty.

2.2. Dominanta jako funkcja średniej i mediany

Do estymacji dominanty można wykorzystać regułę „kciuka" podaną przez Karla Pearsona. Zakłada ona, że w rozkładach jednomodalnych między średnią i medianą, które jest łatwo estymować, oraz dominantą występuje w przybliżeniu stała relacja. Bez względu na kierunek asymetrii mediana, wg tej reguły, usytuowana jest między średnią i dominantą w odległości od średniej równej 1/3 odległości między średnią i dominantą. Przekształcając tę zależność ze względu na dominantę można otrzymać wzór (Yule, Kendall 1966, s. 136):

$$\hat{D}_{Odl} = \bar{x} - 3 \cdot (\bar{x} - Me) = 3Me - 2\bar{x} \qquad (2)$$

Należy przy tym zaznaczyć, że wzór ten daje dokładne wyniki w rozkładach o umiarkowanej asymetrii (Yule, Kendall 1966).

2.3. Metoda półprób – Half-Sample Mode (HSM)

Metoda HSM opiera się na następującym algorytmie iteracyjnym (Bickel 2006, s. 3502–3504):

1. Z całej próby o liczebności n rozważa się podpróby o liczebności $n' = n / 2$ (jeżeli n jest nieparzyste $(n / 2)$ zaokrągla się w górę). Podpróbami są kolejne fragmenty uporządkowanej niemalejąco próby, od 1 do n', od 2 do $n' + 1$, 3 do $n' + 2$ itd.
2. Do kolejnej iteracji wchodzi podpróba, która charakteryzuje się najmniejszym rozstępem $(x_{max} - x_{min})$.
3. W wybranej podpróbie ponownie rozważa się mniejsze podpróby o połowie liczebności (analogicznie do p.1) i wybiera tę, która cechuje się najmniejszym rozstępem.
4. Procedura wykonywana jest do momentu, w którym zostanie próba co najwyżej 2-elementowa. Estymatorem dominanty jest średnia arytmetyczna z elementów tej próby.

2.4. Metoda półrozstępów – Half-Range Mode (HRM)

Metoda HRM opiera się na postępowaniu iteracyjnym analogicznym do metody HSM (Bickel 2002, s. 154–155):

1. Z całej próby o liczebności n rozważa się podpróby, których rozstęp równy jest połowie rozstępu całej próby.
2. Do dalszej iteracji wchodzi podpróba, która zawiera w sobie najwięcej elementów.
3. Z wybranej podpróby ponownie rozważa się mniejsze próby, których rozstęp równy jest połowie poprzednio branego pod uwagę rozstępu.
4. Procedura wykonywana jest do momentu, w którym zostanie próba co najwyżej 2-elementowa. Analogicznie jak w metodzie HSM estymatorem dominanty jest średnia arytmetyczna elementów tej próby.

2.5. Estymatory standardowy parametryczny i odporny parametryczny – Standard Parametric Mode (SPM) i Robust Parametric Mode (RPM)

Estymatory SPM oraz RPM opierają się na transformacji pierwotnej zmiennej X w zmienną $Y = g(X)$, która miałaby rozkład możliwie najbardziej zbliżony do rozkładu normalnego. Przyjmując założenie, że Y ma rozkład normalny oraz znając postać funkcji transformującej g, można wyznaczyć funkcję gęstości zmiennej X. Wzór na dominantę uzyskuje się wyznaczając maksimum z tej

funkcji gęstości. Argumentami tego wzoru są parametry zmiennej Y oraz funkcji g. Algorytm postępowania jest następujący:

1. Transformacja zmiennej pierwotnej X w zmienną Y o przybliżonym rozkładzie normalnym. Do transformacji w metodzie SPM i RPM wykorzystuje się funkcję potęgową postaci $Y = X^{\alpha}$. Kryterium doboru parametru α jest maksymalizacja współczynnika korelacji pomiędzy obserwacjami zmiennej Y, a teoretycznymi wartościami skumulowanego standaryzowanego rozkładu normalnego odwrotnego $z_i = \Phi^{-1}\left(\frac{i-0,5}{n}\right)$. W metodzie SPM obliczany jest współczynnik korelacji liniowej Pearsona, który może być wyrażony jako: $r(\alpha) = \frac{s_+^2(\alpha) - s_-^2(\alpha)}{s_+^2(\alpha) + s_-^2(\alpha)}$, gdzie $s_{\pm}^2(\alpha) = \sigma\left(\frac{y_i(\alpha)}{\sigma(y_i(\alpha))} \pm \frac{z_i}{\sigma(z_i)}\right)$, zaś σ jest nieobciążonym odchyleniem standardowym z próby.

W metodzie RPM obliczany jest analogiczny tzw. odporny współczynnik korelacji wyrażony jako: $R(\alpha) = \frac{S_+^2(\alpha) - S_-^2(\alpha)}{S_+^2(\alpha) + S_-^2(\alpha)}$, gdzie $S_{\pm}^2(\alpha) = \Delta\left(\frac{y_i(\alpha)}{\Delta(y_i(\alpha))} \pm \frac{z_i}{\Delta(z_i)}\right)$, zaś Δ jest standaryzowanym absolutnym odchyleniem mediany (ang. *standardized median absolute deviation, MAD*) definiowanym jako $\Delta(x_i) = \left\lceil \frac{1}{\Phi^{-1}(0,75)} \right\rceil Me\left|x_i - Me(x_i)\right|$.

Iteracyjnie wyszukiwane jest takie α, dla którego współczynnik korelacji osiąga maksimum (dla rozkładów jednomodalnych jest to zawsze tylko jedno maksimum)[1].

2. Wyznaczenie optymalnej wartości α pozwala na przyjęcie założenia, że realizacje zmiennej Y pochodzą z rozkładu normalnego o gęstości danej wzorem:

$$f_Y(y,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right).$$

Dla funkcji transformacji $y_i(\alpha) = x_i^{\alpha}$ funkcja gęstości zmiennej X ma zatem postać:

$$f_X(x,\mu,\sigma,\alpha) = f_Y(x^{\alpha},\mu,\sigma)\left|\frac{\partial y}{\partial x}\right| = \frac{|\alpha|\, x^{\alpha-1}}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x^{\alpha}-\mu)^2}{2\sigma^2}\right).$$

[1] Szczegółowy algorytm wyznaczania α znaleźć można w: Bickel 2003, s. 911–912.

3. Estymator dominanty wyznacza się jako maksimum funkcji gęstości zmiennej X czyli (z uwagi na zakładaną normalność rozkładu) rozwiązując ze względu na x warunek:

$$\left[\frac{\partial f_X(x,\mu,\sigma,\alpha)}{\partial x}\right]_{x=D} = 0,$$

a w konsekwencji uzyskuje się następującą postać wzoru:

$$\hat{D}_{SPM/RPM} = \left[\frac{1}{2}\left(\mu + \sqrt{\mu^2 + \frac{4\sigma^2(\alpha-1)}{\alpha}}\right)\right]^{1/\alpha}. \tag{3}$$

4. W metodzie SPM jako estymatory μ i σ w powyższym wzorze wykorzystywane są wartości odpowiednio średniej arytmetycznej i odchylenia standardowego zmiennej Y z próby. W metodzie RPM jako estymatory μ i σ wykorzystywane są wartości odpowiednio mediany oraz standaryzowanego absolutnego odchylenie mediany (MAD) zmiennej Y. W obu przypadkach $\alpha = \alpha_0$.

Poziom α jednocześnie określa jaka jest asymetria rozkładu zmiennej X. Dla rozkładów symetrycznych $\alpha = 1$ co implikuje $D = \mu$. Jeżeli α jest bardzo małe, tak że argument pierwiastka jest ujemny (co może się zdarzyć dla małych prób z silną asymetrią), wtedy za ocenę D przyjmuje się najmniejszą wartość z próby.

2.6. Estymator Grenandera

Nieparametryczny estymator dominanty zaproponowany przez Grenandera ma następującą postać (Grenander 1965, s. 138):

$$\hat{D}_{Gre\_p\_k} = \frac{\frac{1}{2}\sum_{i=1}^{n-k}\frac{(x_{i+k}+x_i)}{(x_{i+k}-x_i)^p}}{\sum_{i=1}^{n-k}\frac{1}{(x_{i+k}-x_i)^p}} \tag{4}$$

gdzie: p i k to liczby całkowite, spełniające warunek: $1 < p < k$.

Dla $k > 2p$ rozkład tego estymatora jest zbieżny do rozkładu normalnego (Hall 1982, s. 994). Estymator Grenandera, podobnie jak estymator SPM, jest wrażliwy na obserwacje nietypowe.

3. SYMULACYJNA ANALIZA EFEKTYWNOŚCI ESTYMATORÓW

Aby ocenić efektywność estymatorów dominanty przeprowadzono szereg symulacji na próbach generowanych z rozkładów teoretycznych o znanej dominancie. Próby generowane były z rozkładu normalnego N(10, 2) oraz sześciu rozkładów logarytmiczno-normalnych o rosnącej asymetrii.

Z każdego rozkładu generowano próby losowe o liczebnościach 200, 500 oraz 1000 jednostek. Dla każdej próby obliczano wartość przedstawionych wyżej 7 estymatorów dominanty, przy czym:

- do wzoru interpolacyjnego z danych pogrupowanych zastosowano 13 wariantów grupowania: od 3 do 15 przedziałów, przy czym grupowanie przeprowadzono dla równych przedziałów klasowych,
- estymator Grenandera obliczano dla 50 wariantów parametrów, dla każdej wartości parametru $p = \{2, 4, 6, 8, 10\}$ oraz parametru $k = \{3, 5, 7, 9, 11, 13, 15, 17, 19, 21\}$.

Liczba symulacji wynosiła w każdym przypadku $N = 10000$.

Do oceny jakości szacunków wykorzystano trzy miary:

- Średni błąd bezwzględny j-tego estymatora:

$$d^j = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{D}_i^j - D\right| \tag{5}$$

- Obciążenie j-tego estymatora:

$$b^j = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{D}_i^j - D\right) \tag{6}$$

- Pierwiastek błędu średniokwadratowego j-tego estymatora (ang. *root mean square error,* RMSE):

$$RMSE^j = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{D}_i^j - D\right)^2} \tag{7}$$

gdzie:

$\hat{D}_i^j$ – ocena j-tego estymatora dominanty w i-tej symulacji,

D – rzeczywista wartość dominanty, wynikająca z rozkładu teoretycznego.

Ze względu na różne wartości rzeczywistej dominanty dla każdego rozkładu, obliczone zostały błędy relatywne:

$$d\% = \frac{d}{D}100\%, \qquad b\% = \frac{b}{D}100\%, \qquad RMSE\% = \frac{RMSE}{D}100\%.$$

4. WYNIKI BADANIA SYMULACYJNEGO

W tablicy 1. przedstawione zostały wyniki symulacji dla prób generowanych z rozkładu normalnego. Z 13 wariantów grupowania przedstawiono tylko wariant najlepszy pod względem RMSE%. W przypadku estymatorów Grenandera, zaprezentowano wariant najlepszy oraz najgorszy pod względem RMSE%, a także średnie oceny błędów ze wszystkich 50 wariantów.

Tablica 1. Względny średni błąd, względne obciążenie oraz względny pierwiastek błędu średniokwadratowego dla rozkładu N(10,2)

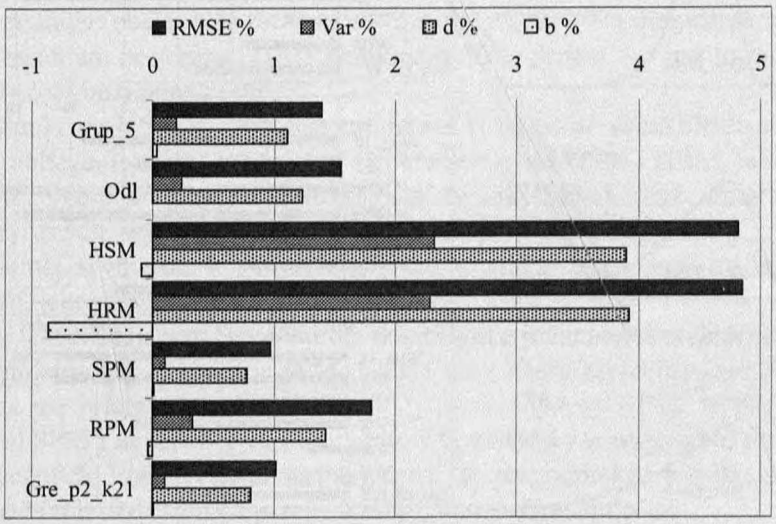
Estymator	n =200			n = 500			n = 1000		
	d %	b %	RMSE %	d %	b %	RMSE %	d %	b %	RMSE %
Grup_5*	2,07	–0,04	2,63	1,29	0,02	1,63	1,11	0,03	1,39
Odl	2,79	–0,03	3,49	1,80	0,03	2,25	1,24	–0,01	1,55
HSM	5,38	–0,01	6,66	4,50	0,12	5,57	3,90	–0,09	4,81
HRM	5,52	–2,04	6,84	4,51	–1,16	5,57	3,92	–0,86	4,86
SPM	1,75	–0,06	2,19	1,11	0,01	1,40	0,78	0,01	0,97
RPM	2,73	0,01	3,33	1,95	0,02	2,44	1,43	–0,03	1,81
Gre_p2_k21**	1,75	–0,01	2,20	1,15	1,15	1,44	0,81	0,01	1,01
Gre_średnia	3,65	0,03	4,61	3,00	0,03	3,83	2,59	0,00	3,33
Gre_p10_k3***	7,40	0,02	9,40	7,09	0,01	9,01	6,94	0,09	8,76

* dla n=200: Grup_3; ** najlepszy pod względem RMSE %; *** najgorszy pod względem RMSE %.

Źródło: Opracowanie własne.

W przypadku rozkładu normalnego najlepsze wyniki uzyskano dla estymatorów SPM, RPM oraz Grenandera (p = 2, k = 21). Wartości RMSE% dla tych estymatorów są podobne, przy czym estymator SPM dla każdej liczebności próby jest nieznacznie efektywniejszy. W przypadku wzoru interpolacyjnego dla prób 200-elementowych najefektywniejszy okazał się podział na 3 przedziały, natomiast dla prób 500- i 1000-elementowych, efektywniejsze było grupowanie w 5 przedziałów klasowych. Tylko dla trzech wymienionych wyżej estymatorów

uzyskano lepsze wyniki od klasycznego wzoru interpolacyjnego. Na wykresie 1 przedstawione zostały względne miary dobroci estymatorów dla prób 1000-elementowych.

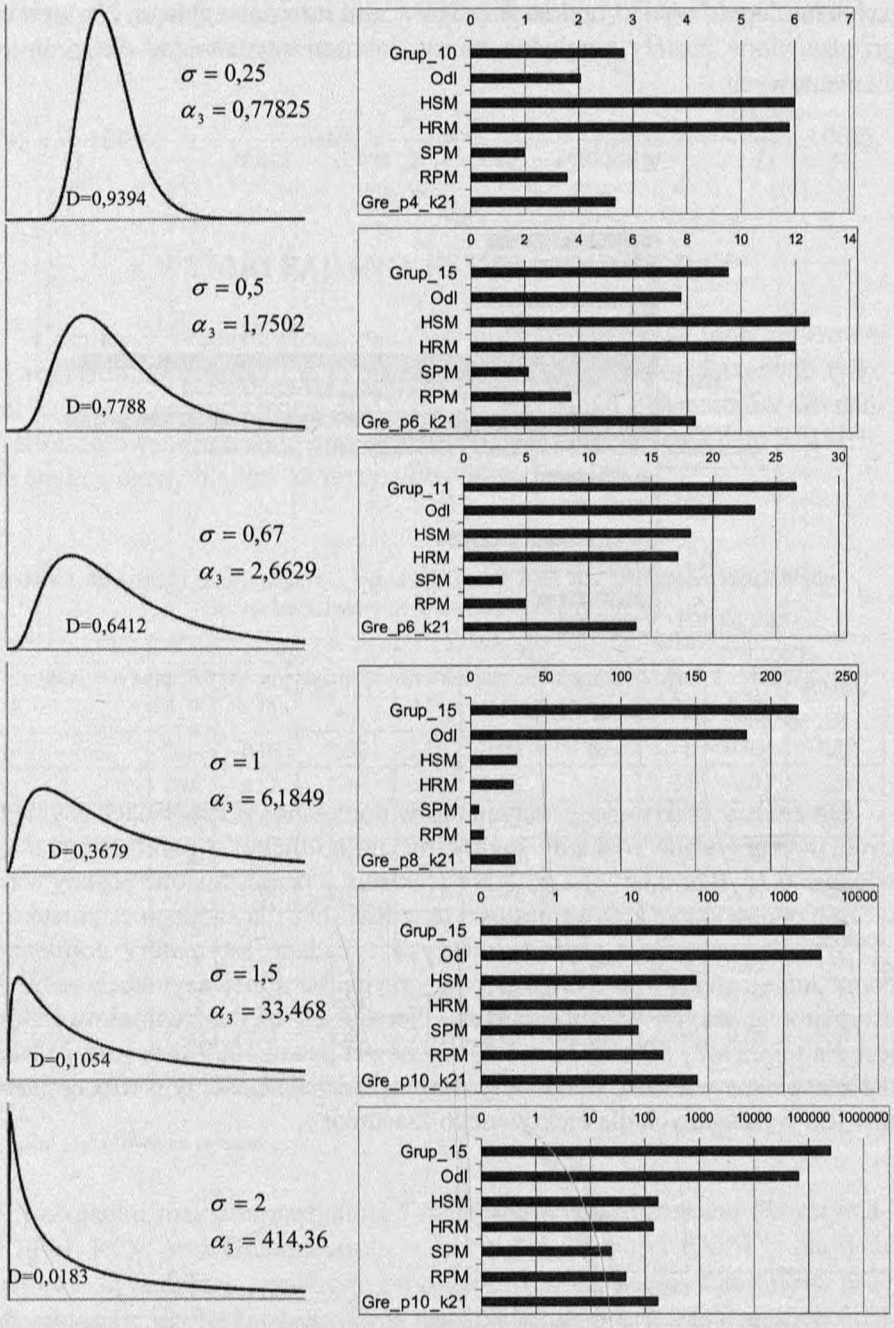


Wykres 1. Błędy oszacowania estymatorów dla rozkładu N(10,2) oraz n = 1000
Źródło: Opracowanie własne.

Do analizy efektywności estymatorów dominanty w rozkładach asymetrycznych, wykorzystano rozkłady logarytmiczno-normalne z parametrem σ odpowiednio 0,25; 0,5; 0,67; 1; 1,5; 2. Na wykresie 2 przedstawione zostały schematycznie odpowiednie krzywe gęstości oraz RMSE% dla każdego estymatora.

Wraz ze wzrostem asymetrii wszystkie badane estymatory dominanty są coraz mniej efektywne. Najgorzej, w przypadku silnej asymetrii, radzi sobie estymator z danych pogrupowanych. Pierwiastek błędu średniokwadratowego jest dla tej metody 10 razy ($\sigma = 1,5$) czy nawet ponad 1000 razy ($\sigma = 2$) większy niż szacowana wielkość dominanty. Im większa skośność, tym więcej przedziałów jest wymaganych dla efektywnego szacunku[2].

[2] Wielkość błędów dla rozkładów o silnej i skrajnej asymetrii wynika najprawdopodobniej z przyjętego w symulacjach grupowania w równe przedziały klasowe. W przypadku skrajnej asymetrii zdecydowana większość jednostek znajdzie się w pierwszym przedziale, a ponieważ liczebność w następnym będzie minimalna, więc oszacowanie dominanty wypadnie w okolicy środka przedziału, tym samym będzie ono mocno zawyżone.



Wykres 2. RMSE% dla rozkładów log-normalnych*, $n = 1000$

*Dla ostatnich dwóch wariantów skala błędów logarytmiczna

Źródło: Opracowanie własne.

Podobnie złe wyniki w przypadku silnej asymetrii ma estymator będący funkcją średniej i mediany, co zresztą wynika z własności rozkładu logarytmiczno-normalnego. Korzystając z prawdziwych wartości miar przeciętnych w rozkładzie logarytmiczno-normalnym można stwierdzić, że do wartości parametru $\sigma = 0{,}5$ różnica (obciążenie estymatora) między rzeczywistą dominantą rozkładu a oszacowaniem ze wzoru (2) nie przekracza 10%, przy $\sigma = 1$ jest to już 180%, a dla $\sigma = 2$ wynosi ponad 600%.

Stosunkowo łatwe w wyznaczaniu, nawet bez zastosowania elektronicznych technik obliczeniowych, estymatory nieparametryczne HSM i HRM, jakkolwiek mało efektywne w przypadku rozkładów symetrycznych oraz umiarkowanie asymetrycznych, wypadają znacznie lepiej, na tle estymatorów (1) i (2), w przypadku silnej asymetrii. Ich przeciętny błąd również rośnie wraz ze wzrostem asymetrii, jednak wzrost ten jest znacznie wolniejszy.

Najefektywniejszym estymatorem w każdym z prezentowanych przypadków jest parametryczny estymator SPM. Nawet przy silnej asymetrii jego RMSE% zwiększa się relatywnie nieznacznie. Estymator odporny RPM, wypada nieco gorzej od SPM i najlepszego z estymatorów Grenandera w przypadku rozkładów symetrycznych. Przy występującej asymetrii, co jest poniekąd wynikiem odporności, jest on lepszy od najlepszego z estymatorów Grenandera.

Estymator Grenandera, podobnie jak w przypadku rozkładu symetrycznego, jest najefektywniejszy dla wartości parametru $k = 21$. Wraz ze wzrostem siły asymetrii rozkładu efektywniejsze są warianty estymatora dla rosnącej wartości parametru p. Warto zaznaczyć, że wartość parametru $k = 21$ była najwyższą z testowanych, być może przy wyższych wartościach wyniki estymatora charakteryzowałyby się mniejszymi błędami. W tablicy 2. przedstawione zostały szczegółowe wyniki symulacji dla prób generowanych z rozkładów log-normalnych.

Prezentację wyników dla rozkładów logarytmiczno-normalnych ograniczono do prób 1000-elementowych. W stosunku do prób 200- i 500-elementowych wyniki były zbliżone, aczkolwiek można zauważyć, że wzrost efektywności estymatorów wraz ze zwiększaniem wielkości próby jest różny i zależny od stopnia asymetryczności rozkładu. Największy wzrost występował w przypadku estymatorów SPM i RPM, najmniejszy w przypadku grupowania, w przypadku silnej asymetrii rozkładu zaobserwowano nawet spadek efektywności tego estymatora. Jednakże można to wytłumaczyć większym prawdopodobieństwem pojawienia się wartości skrajnych, co automatycznie powodowało, że rozpiętości przedziałów były większe a oszacowania dominanty gorsze. Estymator dominanty jako funkcji średniej i mediany również w przypadku silnej asymetrii bardzo mało zyskuje na efektywności, co być może jest efektem wrażliwości średniej na obserwacje odstające. W przypadku estymatorów HSM i HRM wzrost efektywności był większy przy bardzo silnej asymetrii.

Tablica 2. Błędy estymatorów dla rozkładów log-normalnych, $n = 1000$

Estymator	d %	b %	RMSE %	Estymator	d %	b %	RMSE %
$\sigma = 0{,}25$				$\sigma = 0{,}5$			
Grup_10	2,4	1,7	2,9	Grup_15	8,1	7,5	9,6
Odl	1,7	–0,3	2,1	Odl	6,5	–5,7	7,9
HSM	4,9	0,6	6,1	HSM	10,0	2,5	12,6
HRM	4,8	–0,4	6,0	HRM	9,7	0,3	12,1
SPM	0,8	0,4	1,1	SPM	1,8	0,7	2,2
RPM	1,4	0,8	1,9	RPM	2,8	1,2	3,8
Gre_p4_k21	2,3	1,9	2,8	Gre_p6_k21	6,9	5,9	8,4
Gre_średnia	3,9	2,0	4,9	Gre_średnia	10,7	8,1	13,0
Gre_p10_k3	8,9	2,3	11,5	Gre_p10_k3	19,9	10,1	26,7
$\sigma = 0{,}67$				$\sigma = 1$			
Grup_11	24,8	24,8	26,7	Grup_15	191,4	191,4	218,3
Odl	21,5	–	23,4	Odl	180,2	–180,2	183,4
HSM	14,4	4,9	18,2	HSM	24,1	12,5	31,1
HRM	13,8	1,6	17,3	HRM	22,4	6,8	28,6
SPM	2,5	0,7	3,2	SPM	4,4	0,3	5,6
RPM	3,9	1,1	5,0	RPM	6,9	–1,3	8,7
Gre_p6_k21	11,7	10,9	13,9	Gre_p8_k21	24,2	22,3	29,4
Gre_średnia	17,9	15,2	21,3	Gre_średnia	40,7	38,3	47,4
Gre_p10_k3	30,2	19,6	41,3	Gre_p10_k3	60,1	50,0	86,0
$\sigma = 1{,}5$				$\sigma = 2$			
Grup_15	4907	4907	6447	Grup_15	161523	161523	250457
Odl	3094	–	3140	Odl	64427	–64427	66679
HSM	52	39	70	HSM	125	114	172
HRM	45	25	60	HRM	98	78	140
SPM	10	–2	12	SPM	19	–8	24
RPM	19	–11	25	RPM	36	–23	45
Gre_p10_k21	58	55	72	Gre_p10_k21	140	138	173
Gre_śr	115	113	133	Gre_śr	319	318	370
Gre_p10_k3	158	150	243	Gre_p10_k3	428	423	739

Źródło: Opracowanie własne.

5. WNIOSKI

Eksperymenty symulacyjne wskazują, że klasyczne postępowanie, czyli grupowanie i korzystanie ze wzoru interpolacyjnego, nie jest najefektywniejszą metodą wyznaczania dominanty. Szczególnie w przypadku rozkładów silnie asymetrycznych szacunki obarczone są stosunkowo dużym błędem. Metody parametryczne SPM oraz RPM, oparte na potęgowej transformacji danych dla

uzyskania w przybliżeniu rozkładu normalnego, w każdym analizowanym przypadku dostarczają efektywniejszych ocen rzeczywistej dominanty. Wadą tych metod jest jednak ich złożoność obliczeniowa, gdyż konieczne jest iteracyjne ustalenie parametru najlepszej funkcji potęgowej. W przeciwieństwie do podejścia klasycznego, zastosowanie tych metod w praktyce wymaga już znacznej wiedzy statystycznej.

Spośród prostszych metod warto wykorzystywać szerzej w dydaktyce zależność pomiędzy dominantą, medianą i średnią. Efektywność takiego sposobu wyznaczania dominanty w przypadku rozkładów umiarkowanie asymetrycznych jest porównywalna z efektywnością tradycyjnego podejścia. Proste iteracyjne metody HSM i HRM nie dają zadawalających wyników w przypadku symetryczności czy niewielkiej asymetryczności rozkładu, lecz mogą być użyteczne, w przypadku bardzo dużej skośności.

Estymator Grenandera przy odpowiednim doborze parametrów, jest efektywniejszy od podejścia klasycznego. Problemem jest jednak odpowiedni dobór parametrów, aczkolwiek na podstawie sprawdzanych w symulacjach rozkładów można wstępnie określić przesłanki doboru parametrów, wysoką wartość parametru k oraz rosnącą wraz z siłą asymetrii wartość parametru p.

W celu oceny jakości zaprezentowanych metod należałoby przeprowadzić dalsze analizy, szczególnie pod kątem odporności na obserwacje nietypowe oraz możliwości wyznaczania przedziału ufności. Warto też zbadać efektywność estymatora opartego na grupowaniu przy zróżnicowanych rozpiętościach przedziałów.

LITERATURA

Bickel D.R. (2002), *Robust Estimators of the Mode and Skewness of Continuous Data*, Computational Statistics & Data Analysis 39.

Bickel D.R. (2003), *Robust and Efficient Estimation of the Mode of Continuous Data: The Mode as a Viable Measure of Central Tendency*, Journal of Statistical Computation and Simulation 73.

Bickel D.R. (2006), *On a Fast, Robust Estimator of the Mode*, Computational Statistics & Data Analysis 12.

Grenander U., (1965), *Some Direct Estimates of the Mode*, Annals of Mathematical Statistics" 36.

Hall P. (1982), *Limit Theorems for Estimators Based on Inverses of Spacings of Order Statistics*, The Annals of Probability 10.

Rousseeuw, P.J., Leroy A.M. (1987) *Robust Regression and Outlier Detection*, Wiley, New York.

Sobczyk M. (2002), *Statystyka*, PWN Warszawa.

Szulc B. (1976), *Statystyka dla ekonomistów*, PWE, Warszawa.

Yule G.U., Kendall M.G. (1966), *Wstęp do teorii statystyki*, PWN, Warszawa.

Tomasz Jurkiewicz, Arkadiusz Kozłowski

ON DETERMINING THE MODE OF A CONTINUOUS VARIABLE IN RAW DATA

One of the main descriptive characteristics is the mode. For continuous variables it is not always easy to properly determine the mode. There are some estimates of the mode provided in literature, however, unlike the median or the arithmetic mean, for the mode there does not exist the estimator which would be commonly considered as the best one. Moreover, in many statistical textbooks and computer packages this problem seems to be ignored.

In this paper authors consider seven different methods of estimation the mode presented in literature. The efficiency of the estimation procedures has been evaluated on the basis simulation experiments for the normal and lognormal distributions with different degrees of skewness. The evaluation criteria of those procedures involve not only the efficiency of estimation but also simplicity of computation, which is an important aspect of teaching statistics.