ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 227, 2009

*Tomasz Jurkiewicz**

O WYZNACZANIU KWANTYLI ROZKŁADU W SZEREGACH ROZDZIELCZYCH PRZEDZIAŁOWYCH

1. WPROWADZENIE

Analiza statystyczna w oparciu o dane wtórne często ograniczana jest przez materiał statystyczny w postaci szeregów rozdzielczych przedziałowych. Niesie to za sobą w konsekwencji większe błędy oszacowania parametrów rozkładu niż w przypadku, gdy dysponuje się danymi szczegółowymi. Wynika to z nieposiadania przez badacza części informacji o badanej zbiorowości, w szczególności niewiedzy o tym, jaki był rzeczywisty rozkład zbiorowości w poszczególnych przedziałach.

Znaczący wpływ na uzyskiwane wyniki ma również poprawność przeprowadzanego grupowania na etapie analizy danych pierwotnych. Celem grupowania w większości przypadków jest przedstawienie danych w przejrzystej dla odbiorcy formie. Tym samym naturalnie głównym kryterium staje się uzyskanie łatwego w odbiorze szeregu rozdzielczego. Stąd też często dąży się do tego, aby np. rozpiętości przedziałów były jednakowe, by rozpiętości przyjmowały wartości „nominałowe", zaleca się likwidowanie przedziałów zawierających jedną bądź niezawierających żadnych jednostek. Często też, przy publikacji danych okresowych, budowa przedziałów jest taka sama, jak we wcześniejszych okresach. Podejście to skutkować może w praktyce gorszą jakością danych wtórnych z punktu widzenia użytkownika, dla którego są one podstawą analizy.

Kolejnym problemem, z którym w praktyce spotkać się można przy analizie wtórnego materiału statystycznego, jest fakt przedstawiania danych w postaci szeregów rozdzielczych o otwartych przedziałach klasowych. Istnieje wprawdzie możliwość arbitralnego określenia granic przedziałów otwartych, ale ich określenie ma wpływ na uzyskiwane wyniki analizy. Stąd też zaleca się w takich przypadkach rezygnację z miar klasycznych i stosowanie miar pozycyjnych.

* Dr, Katedra Statystyki, Wydział Zarządzania, Uniwersytet Gdański.

W gospodarce globalnej niezwykle ważnym czynnikiem konkurencyjności jest posiadanie aktualnych i wiarygodnych danych będących podstawą do podejmowania decyzji. Rozwój metod ilościowych pozwala na uzyskiwanie coraz lepszych informacji z posiadanych zbiorów danych, jednakże w przypadku danych wtórnych możliwości takich jest znacznie mniej. Celem artykułu jest przedstawienie innego niż tradycyjny sposobu interpolacji kwantyli rozkładu na podstawie danych pogrupowanych oraz ocena jakości tej metody.

2. SZACOWANIE KWANTYLI ROZKŁADU

Kwantyle rozkładu cechy (czy zmiennej losowej) są podstawowymi pozycyjnymi miarami położenia rozkładu. Na bazie kwantyli zbudowane są pozostałe pozycyjne miary opisujące rozkład, jego zróżnicowanie, asymetrię czy spłaszczenie. Kwantyle pełnią fundamentalną rolę w statystyce, kwantylami są bowiem np. wartości krytyczne w testowaniu hipotez, czy granice przedziałów ufności, funkcję kwantyli wykorzystuje się także w symulacjach do generowania zmiennych losowych o rozkładach innych niż równomierny (Kotz, Johnson 1986, s. 424) (metoda odwracania dystrybuanty).

Zgodnie z definicją kwantylem rzędu p nazywamy taką wartość cechy x_p, która dzieli zbiorowość na pN wartości niewiększych od kwantyla i $(1 - p)N$ wartości niemniejszych od kwantyla. W przypadku wnioskowania statystycznego kwantylem jest wartość x_p spełniająca nierówności $P(X \leq x_p) = p$; $P(X \geq x_p) = 1 - p$.

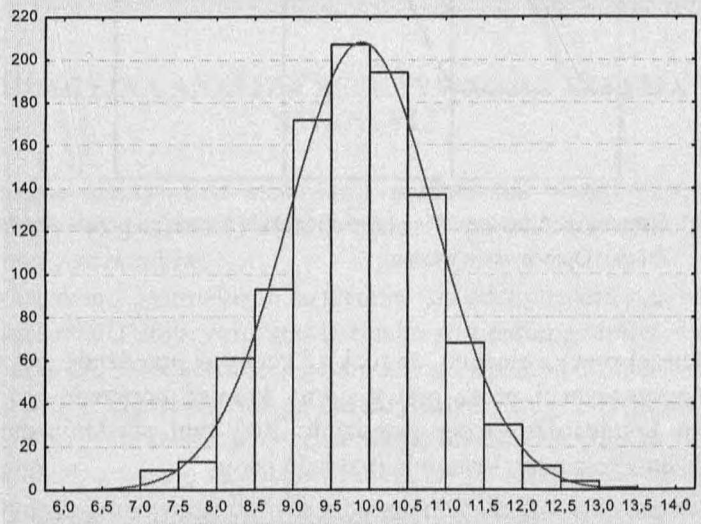
Wyznaczanie (estymacja) kwantyli w szeregu szczegółowym o liczebności N polega na znalezieniu wartości cechy znajdującej się na $p(N + 1)$-ej pozycji w uporządkowanym niemalejąco szeregu (por. Kot, Jakubowski, Sokołowski 2007, s. 168; Luszniewicz, Słaby 2008, s. 29; Ostasiewicz, Rusnak, Siedlecka 1999, s. 57; Sobczyk 2006, s. 38). W przypadku gdy pozycja kwantyla $p(N + 1)$ jest wartością niecałkowitą przyjmuje się jako wartość kwantyla średnią, średnią ważoną lub bliższą z dwóch wartości znajdujących się najbliżej pozycji kwantyla.

Dla szeregu rozdzielczego przedziałowego wartość kwantyla wyznacza się najczęściej za pomocą interpolacji wartości kwantyla w pierwszym przedziale (o numerze i), w którym liczebność skumulowana co najmniej równa się pozycji kwantyla. Wzór interpolacyjny można przedstawić jako (por. Kot, Jakubowski, Sokołowski 2007, s. 174; Luszniewicz, Słaby 2008, s. 51; Ostasiewicz, Rusnak, Siedlecka 1999, s. 57; Sobczyk 2006, s. 39):

$$x_p = x_{i0} + \left(pN - \sum_{j=1}^{i-1} n_j \right) \frac{c_i}{n_i} \tag{1}$$

gdzie c_i jest rozpiętością przedziału, x_{i0} dolną granicą, a n_i liczebnością tego przedziału.

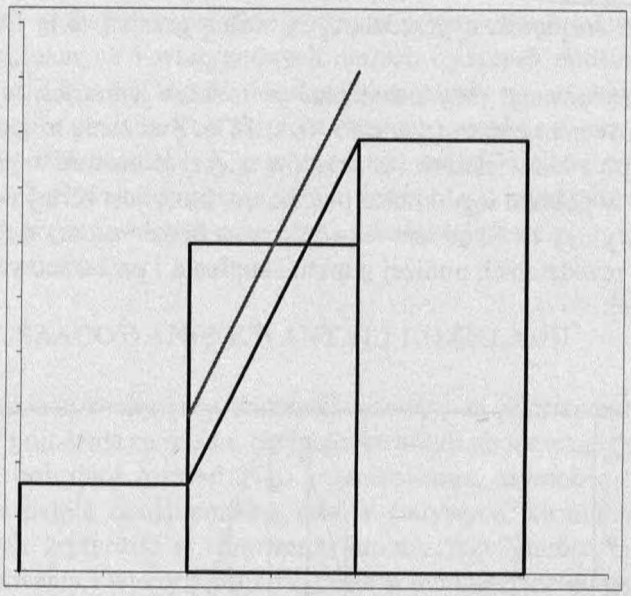
Wzór ten powstał przy założeniu, że rozkład jednostek w przedziale jest rozkładem równomiernym (Szulc 1976, s. 187). Założenie to zazwyczaj nie jest spełnione, gdyż dla większości rozkładów więcej jednostek w przedziale będzie skupionych w pobliżu tego krańca przedziału, który jest bliżej wartości dominującej (por. wykres 1). Skutkiem tego założenia będzie niedoszacowanie wartości kwantyli w przedziałach poniżej punktu skupienia i przeszacowanie w przedziałach powyżej.



Wykres 1. Histogram dla próby z rozkładu $N(10,1)$

Źródło: Opracowanie własne.

Rozkład wartości cechy w większości przedziałów wydaje się być bardziej zbliżony do rozkładu wyznaczonego przez prostą równoległą do punktów wyznaczanych przez liczebności sąsiednich przedziałów, skorygowaną tak, aby pole pod prostą w danym przedziale było równe polu prostokąta obrazującego liczebność danego przedziału co ilustruje rysunek 1.



Rysunek 1. Aproksymacja gęstości rozkładu w szeregu przedziałowym
Źródło: Opracowanie własne.

W niniejszej pracy założono, że rozkład cechy w przedziale jest równoległy do prostej wyznaczanej przez prawy górny kraniec poprzedniego przedziału i lewy górny kraniec następnego przedziału. Przy tym założeniu można przedstawić wzór interpolacyjny kwantyla rozkładu jako:

$$x_p^* = x_{i0} + \frac{\sqrt{n_i'^2 + 4\left(n_i - n_i'\right)\left(pN - \sum_{j=1}^{i-1} n_j\right)} - n_i'}{2\left(n_i - n_i'\right)} c_i \qquad (2)$$

gdzie: $n_i' = n_i - \frac{n_{i+1} - n_{i-1}}{2}$.

Z praktycznego punktu widzenia stosowanie wzoru (2) zamiast (1) niesie ze sobą tylko nieznaczne skomplikowanie obliczeń, natomiast sama procedura wyznaczania kwantyla, tzn. znalezienie na podstawie liczebności skumulowanych przedziału kwantyla pozostaje bez zmian.

Wzór (2) nie może i nie powinien być stosowany mechanicznie. Jednym z ograniczeń jest sytuacja, gdy liczebności przedziałów sąsiadujących z przedziałem kwantyla są jednakowe. W takim przypadku $n'_i = n_i$, a więc mianownik

byłby równy 0. Również sytuacja, gdy układ liczebności przedziałów kwantyla, poprzedniego i następnego powodują, że aproksymowany rozkład w przedziale znajdowałby się częściowo poniżej osi OX, prowadzić może w efekcie do dużych błędów oszacowań. Stąd też proponowany algorytm postępowania:

1. Jeżeli $n_{i-1} = n_i$ lub $n_{i+1} = n_i$ lub $(n_{i-1} > n_i$ i $n_{i+1} > n_i)$ lub $n_{i-1} = n_{i+1}$ należy zastosować wzór (1).
2. Jeżeli $n_{i-1} + n_{i+1} > 2n_i$ to należy przyjąć:
   - jeżeli $n_{i-1} < n_{i+1}$ to $n'_i = n_i$ oraz $n_{i+1} = 2n_i - n_{i-1}$;
   - jeżeli $n_{i-1} > n_{i+1}$ to $n'_i = 2n_i - n_{i+1}$ oraz $n_{i-1} = 2n_i - n_{i+1}$

   i zastosować wzór (2).
3. Dla pozostałych przypadków zastosować wzór (2).

3. SYMULACYJNA ANALIZA EFEKTYWNOŚCI ESTYMATORÓW KWANTYLI

Aby ocenić efektywność stosowania modyfikacji wzoru interpolacyjnego przeprowadzono eksperymenty symulacyjne. W pojedynczej symulacji generowano wartości z rozkładów:

- w wariancie 1. normalnego ze średnią 10 i odchyleniem standardowym 1;
- w wariancie 2. logarytmiczno-normalnego z parametrami $\mu = 0$ i $\sigma = 0.3$;
- w wariancie 3. logarytmiczno-normalnego z parametrami $\mu = 0$ i $\sigma = 0.7$.

Pierwszy rozkład jest rozkładem symetrycznym, drugi i trzeci prawostronnie asymetrycznymi z momentem trzecim względnym wynoszącym odpowiednio 0,94953 i 2,8884, a więc o stosunkowo niedużej i o skrajnej asymetrii. Z uwagi na automatyczne grupowanie danych, aby uniknąć błędów powodowanych przez wartości skrajnie odstające, wygenerowane wartości ograniczono do ±5 odchyleń standardowych dla rozkładu normalnego. Dla rozkładów logarytmiczno-normalnych przyjęto analogiczne ograniczenie z prawej strony rozkładu, z analogicznym prawdopodobieństwem wystąpienia wartości odstających. Dla wariantu drugiego ograniczeniem górnym była wartość 5, dla trzeciego wartość 33.

Analizę przeprowadzono dla pięciu wielkości zbiorowości wynoszących kolejno: 200, 500, 1000, 5000 i 10000. Dla wygenerowanej próby obliczano najpierw w celach porównawczych wartości 19 kolejnych kwantyli rzędu 0,05; 0,10; 0,15; ...; 0,95 na podstawie szeregu szczegółowego. Za wartość kwantyla przyjmowano wartość w uporządkowanej próbie wskazaną przez pozycję kwantyla lub średnią z dwóch wartości najbliższych pozycji kwantyla.

W kolejnym kroku próbę grupowano w k przedziałów, gdzie przyjmowano podział na $k = 3, 4, \ldots, 20$ przedziałów. Z uwagi na automatyzację obliczeń do grupowania założono podział na przedziały o równych rozpiętościach, gdzie

dolny kraniec pierwszego przedziału był wyznaczany przez wartość minimalną w próbie, a górny kraniec ostatniego przedziału przez wartość maksymalną.

Dla pogrupowanych danych, dla wszystkich 18 wariantów grupowania, wyznaczano kwantyle rzędu 0,05; 0,10; 0,15; ...; 0,95 przy pomocy wzorów interpolacyjnych (1) i (2).

Na podstawie wyników 10 000 symulacji porównywano uzyskiwane trzema metodami oszacowania wartości kwantyli z wartością rzeczywistą dla rozkładu.

4. WYNIKI ANALIZY EFEKTYWNOŚCI ESTYMATORÓW

W klasycznym wnioskowaniu statystycznym efektywność estymatora, czyli miara bliskości uzyskiwanych oszacowań od prawdziwej wartości parametru, jest oceniana na podstawie jego własności próbkowych. Do własności tych należą, m. in. wartość oczekiwana, wariancja, błąd średniokwadratowy i inne charakterystyki rozkładu estymatora. Miarami efektywności estymatora określanymi w badaniu symulacyjnym mogą być przykładowo obciążenie, czyli różnica między średnią uzyskiwanych wyników a wartością rzeczywistą, wariancja uzyskiwanych wyników, średni błąd kwadratowy (MSE), czyli średni kwadrat różnic między uzyskiwanymi oszacowaniami a prawdziwą wartością, czy też przeciętne bezwzględne odchylenie wyników od wartości rzeczywistej parametru. W artykule jako miarę jakości oszacowania przyjęto pierwiastek ze średniego błędu kwadratowego (RMSE). Średni błąd kwadratowy uwzględnia błędy oszacowania powstałe zarówno w wyniku systematycznego zaniżania czy też zawyżania wartości parametru (obciążenie) jak i w wyniku zmienności uzyskiwanych oszacowań (wariancja). Pierwiastek ze średniego błędu kwadratowego jest łatwiejszy w interpretacji, jest to średnie odchylenie uzyskanego oszacowania od rzeczywistej wartości parametru.

Wyniki symulacji dla wybranego przypadku przedstawiono w tablicy 1. Przy podziale na pięć przedziałów i liczebności zbiorowości wynoszącej 1000 elementów zastosowanie zmodyfikowanej wersji wzoru interpolacyjnego daje, zwłaszcza dla rozkładu normalnego, w części przypadków nawet lepsze wyniki niż oszacowanie na podstawie szeregu szczegółowego. Jednocześnie wersja zmodyfikowana w zdecydowanej większości przypadków jest znacznie lepsza od tradycyjnej.

Porównanie efektywności wzorów interpolacyjnych (1) i (2) przedstawiono na wykresach 2, 3 i 4. Przy niewielkiej liczebności zbiorowości i symetrycznym rozkładzie cechy wzór (2) jest efektywniejszy przy małych ilościach przedziałów, w których pogrupowano dane. Przy dużej ilości przedziałów jest tylko nieznacznie mniej efektywny. Wraz ze wzrostem liczebności zbiorowości przewaga efektywności wzoru (2) rośnie, jedynie dla najmniejszej liczby przedziałów w okolicach drugiego i ósmego decyla różnica efektywności ponownie się zmniejsza.

Tablica 1. Wartość RMSE oszacowań kwantyli[a)] dla $n = 1000$ i $k = 5$

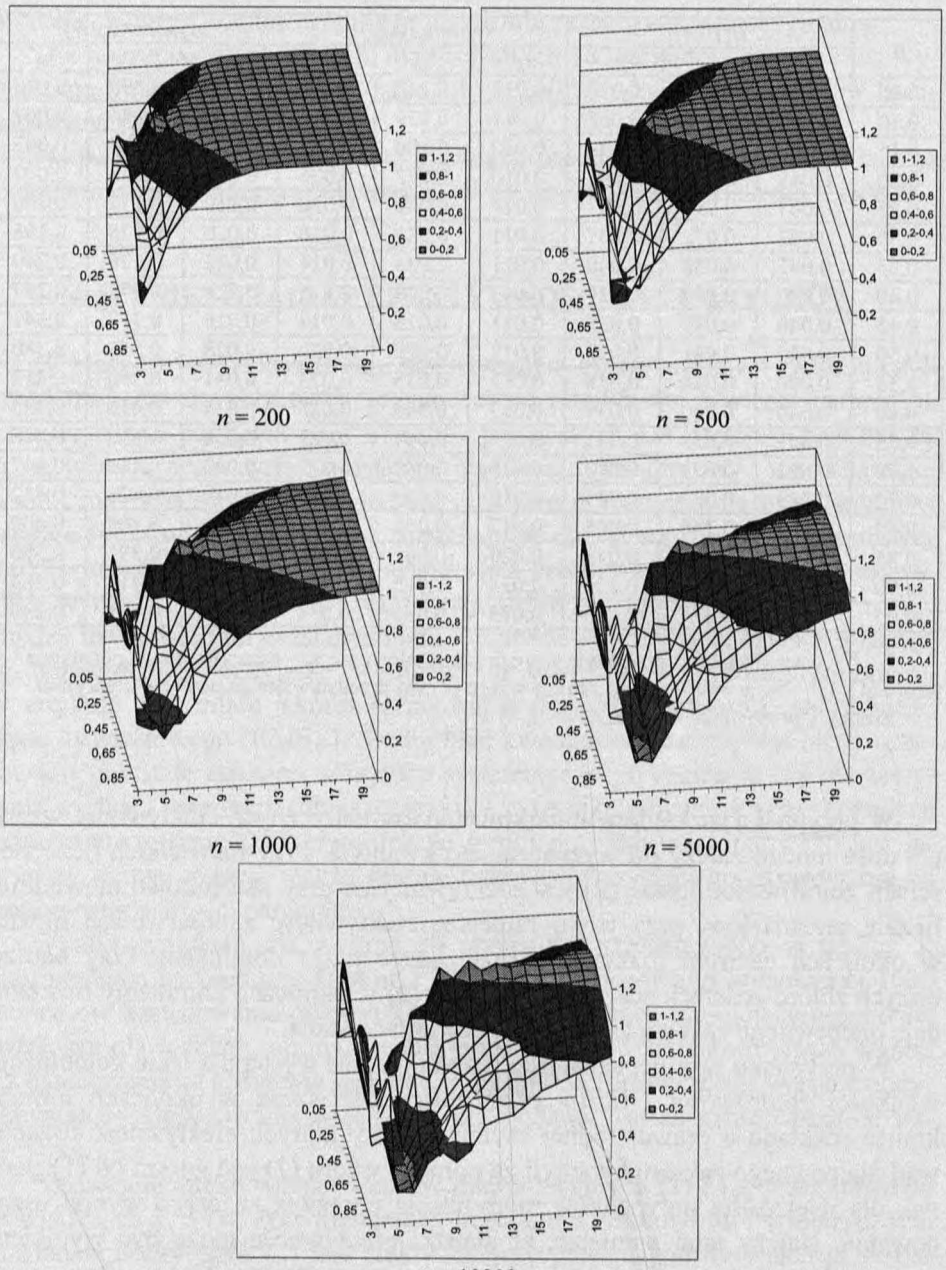
p	$N(10,1)$			lognorm(0,0.3)			lognorm(0,0.7)		
	x_p^{sz}	x_p	x_p^*	x_p^{sz}	x_p	x_p^*	x_p^{sz}	x_p	x_p^*
0,05	0,066	0,211	0,076	0,012	0,145	0,059	0,015	0,098	0,088
0,10	0,053	0,228	0,057	0,011	0,129	0,047	0,015	0,084	0,067
0,15	0,048	0,191	0,049	0,010	0,100	0,035	0,016	0,075	0,063
0,20	0,045	0,141	0,045	0,010	0,072	0,026	0,017	0,094	0,097
0,25	0,043	0,099	0,043	0,011	0,049	0,020	0,019	0,132	0,146
0,30	0,042	0,072	0,041	0,011	0,033	0,016	0,020	0,178	0,198
0,35	0,041	0,058	0,040	0,011	0,023	0,014	0,022	0,226	0,249
0,40	0,040	0,049	0,039	0,011	0,020	0,015	0,024	0,273	0,297
0,45	0,040	0,043	0,038	0,011	0,023	0,017	0,026	0,317	0,341
0,50	0,040	0,041	0,038	0,012	0,029	0,020	0,028	0,357	0,380
0,55	0,040	0,044	0,038	0,012	0,035	0,023	0,031	0,390	0,412
0,60	0,040	0,050	0,039	0,013	0,040	0,026	0,034	0,416	0,435
0,65	0,041	0,059	0,041	0,014	0,044	0,027	0,038	0,432	0,448
0,70	0,042	0,073	0,042	0,015	0,046	0,027	0,043	0,436	0,447
0,75	0,043	0,099	0,043	0,016	0,047	0,026	0,049	0,426	0,432
0,80	0,045	0,140	0,045	0,017	0,052	0,024	0,058	0,408	0,400
0,85	0,048	0,190	0,048	0,020	0,062	0,023	0,070	0,395	0,350
0,90	0,054	0,229	0,056	0,024	0,075	0,026	0,093	0,412	0,286
0,95	0,067	0,214	0,076	0,033	0,080	0,034	0,149	0,439	0,230

[a)] x_p^{sz} – oszacowanie na podstawie szeregu szczegółowego; x_p – oszacowanie na podstawie wzoru (1); x_p^* – oszacowanie na podstawie wzoru (2) wg przedstawionego powyżej algorytmu.

Źródło: Opracowanie własne.

W przypadku rozkładów umiarkowanie asymetrycznych efektywność wzoru (2) dość mocno zależy od wyznaczanego kwantyla. Przy niewielkich liczebnościach zbiorowości postać (2) jest efektywniejsza przy stosunkowo niewielkiej liczbie przedziałów, przy czym mniejszą efektywność zaobserwować można w okolicach centrum rozkładu wyznaczanego przez dominantę. Przy bardzo dużych zbiorowościach zdarza się, że wzór (2) w okolicach dominanty rozkładu daje dla pewnego zakresu kwantyli gorsze oszacowania.

W przypadku skrajnej skośności rozkładu także występują takie kombinacje wielkości zbiorowości i liczby przedziałów, zwłaszcza w okolicach lewego krańca rozkładu o prawostronnej asymetrii, przy których efektywność oszacowań dla pewnego zakresu kwantyli za pomocą wzoru (2) jest gorsza od (1). Jednak dla większości przypadków modyfikacja poprawia znacząco wyniki oszacowania. Należy tutaj pamiętać, że słabsza efektywność może być wynikiem zastosowanego w symulacji podziału na równe przedziały klasowe. Przy silnej asymetrii dane wtórne przedstawiane są zazwyczaj w formie szeregu rozdzielczego o nierównych rozpiętościach poszczególnych klas, co powinno poprawiać efektywność wzoru (2).



$n = 200$ $n = 500$

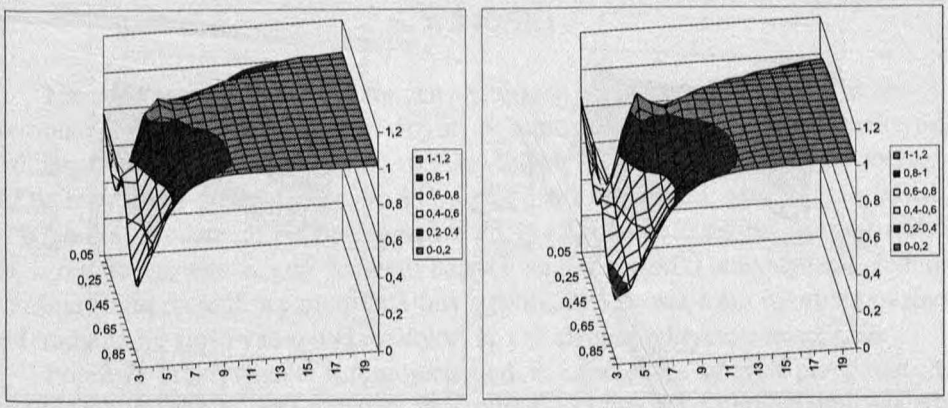
$n = 1000$ $n = 5000$

$n = 10000$

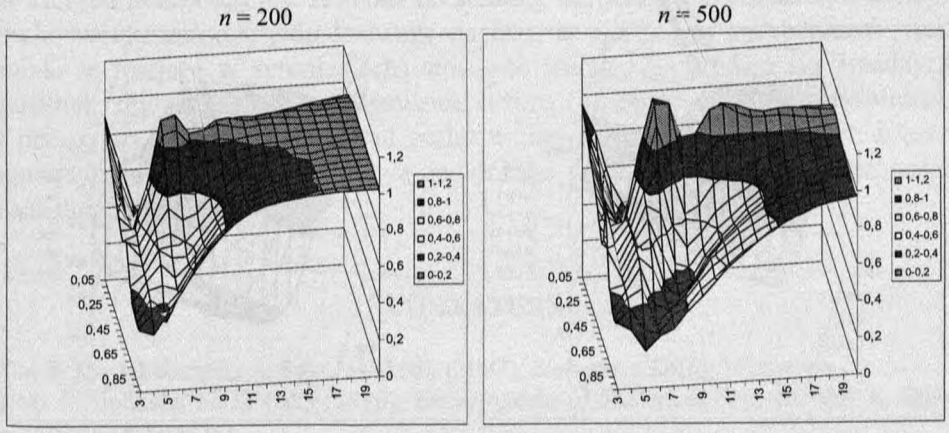
Wykres 2. RMSE(x_p^*) / RMSE(x_p) dla rozkładu normalnego

Oś OX – liczba przedziałów k; oś OY – rząd kwantyla; oś OZ – iloraz RMSE

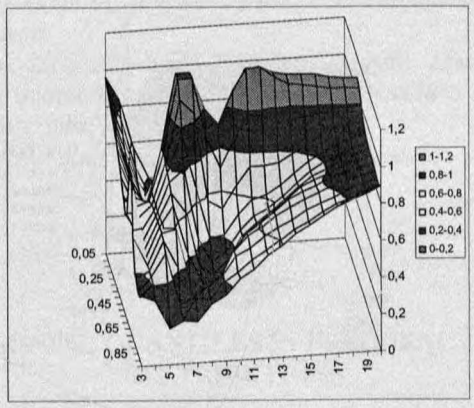
Źródło: Opracowanie własne.



$n = 200$ $n = 500$



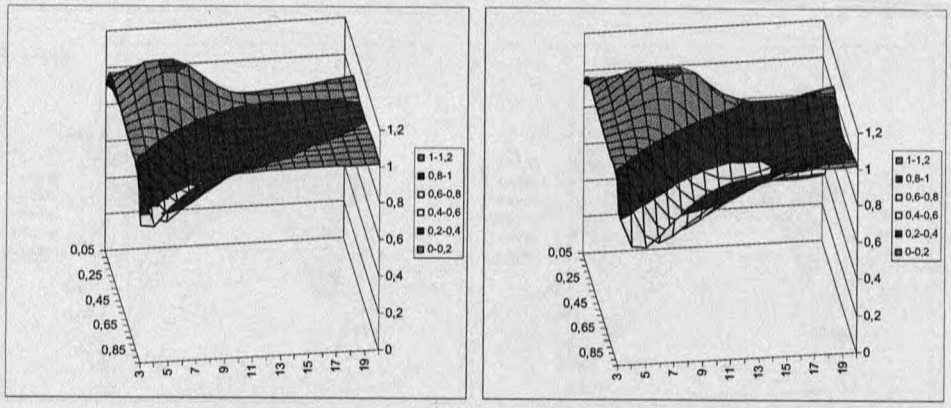
$n = 1000$ $n = 5000$



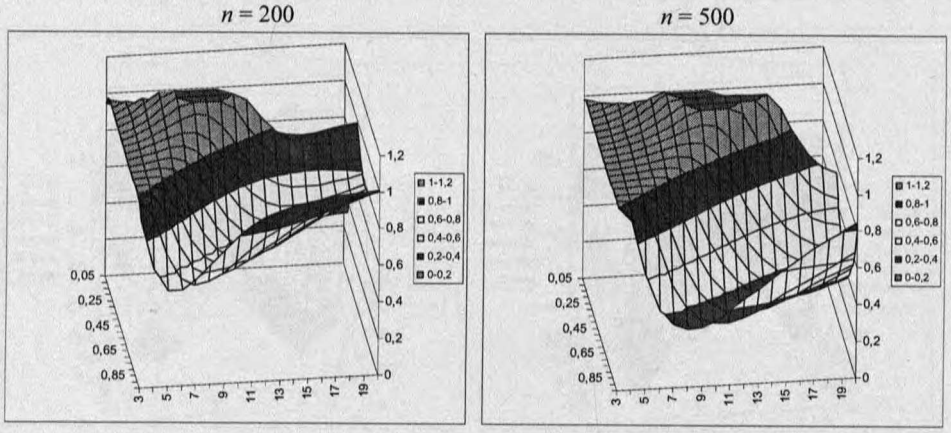
$n = 10000$

Wykres 3. RMSE(x_p^*) / RMSE(x_p) dla rozkładu logarytmiczno-normalnego z $\sigma = 0.3$
Oś OX – liczba przedziałów k; oś OY – rząd kwantyla; oś OZ – iloraz RMSE

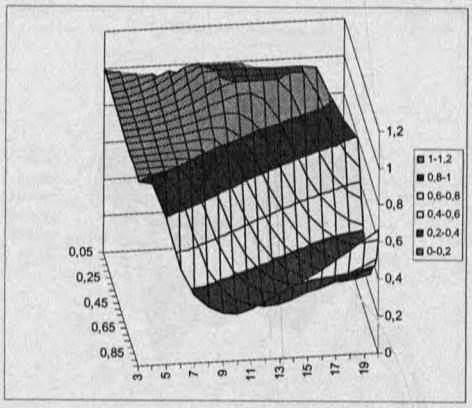
Źródło: Opracowanie własne.



$n = 200$ $n = 500$



$n = 1000$ $n = 5000$



$n = 10000$

Wykres 4. RMSE(x_p^*) / RMSE(x_p) dla rozkładu logarytmiczno-normalnego z $\sigma = 0.7$
Oś OX – liczba przedziałów k; oś OY – rząd kwantyla; oś OZ – iloraz RMSE
Źródło: Opracowanie własne.

5. WNIOSKI

Na podstawie przeprowadzonych symulacji można stwierdzić, że wzór interpolacyjny (2) może być z pewnymi ograniczeniami stosowany w praktyce, ponieważ poprawić może jakość uzyskiwanych wyników. Szczególnie dotyczy to tych sytuacji, gdy zbiorowość, którą przedstawia rozkład, była bardzo liczna, a więc dla np. danych pochodzących z dużej populacji. Z takimi danymi statystyk ma do czynienia, gdy źródłem danych są np. roczniki statystyczne. Jednocześnie sama metoda wyznaczania jest stosunkowo prosta i nie różni się nazbyt od tradycyjnie stosowanej w dydaktyce, co ma swoje praktyczne znaczenie.

Poprawa efektywności interpolacji będzie największa w tych przedziałach, w których liczebność jest zbliżona do średniej liczebności z przedziałów sąsiednich. Należy zwrócić jednak uwagę na fakt, że nie należy mechanicznie (jak miało to miejsce w symulacjach) stosować wersji (2). Wydaje się zasadnym postulat, aby poprzedzić zastosowanie wzoru (2) chociażby analizą graficzną i oceną, czy rzeczywisty rozkład cechy w danym przedziale może być dobrze aproksymowany prostą równoległą do punktów wyznaczanych przez liczebności sąsiednich przedziałów.

LITERATURA

Kot S. M., Jakubowski J., Sokołowski A. (2007), *Statystyka*, Difin, Warszawa.

Kotz S., Johnson N. L. (ed.) (1986), *Encyclopedia of Statistical Science*, Vol 7., John Wiley & Sons Inc.

Luszniewicz A., Słaby T. (2008), *Statystyka z pakietem komputerowym STATISTICA PL*, C.H. Beck, Warszawa.

Ostasiewicz S., Rusnak Z., Siedlecka U. (1999), *Statystyka. Elementy teorii i zadania*, Wyd. Akademii Ekonomicznej im. O. Langego we Wrocławiu, Wrocław.

Sobczyk M. (2006), *Statystyka*, Wyd. UMCS, Lublin.

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

Tomasz Jurkiewicz

ON DETERMINING QUANTILES IN FREQUENCY INTERVAL DISTRIBUTIONS

Interval frequency distributions used as a secondary data do not provide full information about the distribution of the variable of interest. In order to estimate quantiles, it is common to use one of interpolation formulae based on the assumption that the distribution of the variable is uniformly distributed in each interval. This assumption, if not met, may result in large bias.

Author suggests using different formulae, which do not require uniform distribution in intervals, and depend on the frequencies of neighbouring intervals. Simulation experiments were applied for the normal and lognormal distributions to assess efficiency of both kinds of interpolation formulae. Different population sizes and different numbers of intervals in the frequency distribution were also considered.