UNIWERSYTET SZCZECIŃSKI
WYDZIAŁ NAUK EKONOMICZNYCH I ZARZĄDZANIA

Studia i Prace WNEiZ US
nr 45/2 2016

DOI:10.18276/sip.2016.45/2-05

Jan Purczyński*
Uniwersytet Szczeciński

ESTYMATOR NIEOBCIĄŻONY CZY ESTYMATOR MINIMALIZUJĄCY BŁĄD ŚREDNIOKWADRATOWY

Streszczenie

Celem artykułu była odpowiedź na pytanie, czy należy stosować estymator nieobciążony (EN), czy też estymator zapewniający minimum błędu średniokwadratowego (EMSE). W tym celu rozpatrzono proste przykłady wyznaczania estymatorów parametrów rozkładu wykładniczego oraz rozkładu Laplace'a. Jako kryteria rozpatrzono wartości statystyki testów chi-kwadrat i testu Kołmogorowa. Uzyskane wyniki nie dają jednoznacznej odpowiedzi na postawione pytanie.

Słowa kluczowe: estymator nieobciążony, estymator minimalizujący błąd średniokwadratowy

Wstęp

Jednym z ważniejszych zagadnień statystyki jest estymacja parametrów rozkładu zmiennej losowej aproksymującej rozkład danych empirycznych. Teoria estymacji wyróżnia między innymi takie cechy estymatorów, jak nieobciążoność oraz efektywność. Estymator nieobciążony jest zmienną losową o wartości przeciętnej równej nieznanej wartości parametru. Właściwość nieobciążoności zapewnia otrzymanie ocen parametrów nieobciążonych błędem systematycznym. Im mniejsza jest

* Adres e-mail: jan.purczynski@wzieu.pl.

wariancja estymatora, tym bardziej skupione są jego możliwe wartości wokół wartości przeciętnej, a więc tym większe jest prawdopodobieństwo, że oszacowanie nieznanego parametru będzie bliskie jego prawdziwej wartości. Wskazane jest zatem, aby wariancja estymatora była możliwie mała. Tę właściwość posiada estymator najefektywniejszy. Fakt, że dany nieobciążony estymator jest bardziej efektywny od innego nieobciążonego estymatora, oznacza, że jego wartości są bardziej skupione wokół szacowanego parametru, niż w przypadku drugiego estymatora. W literaturze przedmiotu można zauważyć, że wyznaczając postać estymatora, zwraca się uwagę na jego nieobciążoność. Natomiast efektywność estymatora jest rozpatrywana w dalszej kolejności. Istnieje prosty sposób uwzględnienia obydwu właściwości, a mianowicie błąd średniokwadratowy stanowiący sumę wariancji oraz kwadratu obciążenia estymatora.

W świetle dotychczasowych rozważań przyjęto, że celem artykułu jest stwierdzenie, czy należy stosować estymator nieobciążony (EN), czy też estymator zapewniający minimum błędu średniokwadratowego (EMSE). W związku z tym zostaną rozpatrzone proste przykłady wyznaczania estymatorów parametrów rozkładu wykładniczego oraz rozkładu Laplace'a.

1. Estymacja parametru rozkładu wykładniczego

Zadanie estymacji polega na wyznaczeniu oszacowania $\hat{\theta}$ nieznanego parametru θ. Jedną z ważniejszych właściwości estymatora $\hat{\theta}$ jest jego obciążenie $b(\hat{\theta})$:

$$b(\hat{\theta}) = E(\hat{\theta}) - \theta \tag{1}$$

O jakości estymatora decyduje również błąd średniokwadratowy (*Mean Squared Error* – MSE):

$$MSE = E(\hat{\theta} - \theta)^2 \tag{2}$$

który spełnia zależność (Krzyśko, 1997, s. 17):

$$MSE = V(\hat{\theta}) + \left(b(\hat{\theta})\right)^2 \tag{3}$$

gdzie:
$b(\hat{\theta})$ – obciążenie estymatora [wzór (1)];
$V(\hat{\theta})$ – wariancja estymatora.

W dalszej części opracowania będzie wykorzystana wartość obciążenia względnego:

$$bL = \frac{b(\hat{\theta})}{\theta} \tag{4}$$

oraz wartość względnego błędu średniokwadratowego:

$$bse = \frac{MSE}{\theta^2} \tag{5}$$

Właściwości estymatora zostaną przetestowane na przykładzie szacowania parametru rozkładu wykładniczego. Wzór (6) określa jedną z dwóch postaci (wariant A) gęstości rozkładu:

$$f(x) = \frac{1}{\beta} \exp\left(-\frac{x}{\beta} \right) \tag{6}$$

Stosując Metodę Największej Wiarygodności (MNW), uzyskuje się następującą postać oszacowania parametru β

$$\hat{\beta}_2 = \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{7}$$

gdzie:

x_i – wartości obserwacji; $i = 1, 2, \ldots, N$.

Estymator ten jest nieobciążony (Housila, Sarjinder, Jong-Min, 2012, s. 301). Natomiast estymator wyrażający się wzorem:

$$\hat{\beta}_3 = \frac{N}{N+1} \bar{x} \tag{8}$$

zapewnia najmniejszą wartość błędu średniokwadratowego (Housila i in., 2012).

Dla zróżnicowania wyników zostanie rozpatrzony estymator opisany wzorem:

$$\hat{\beta}_1 = \frac{N}{N-1} \bar{x} \tag{9}$$

Bardziej popularny jest wzór (10), również określający gęstość rozkładu wykładniczego (wariant B):

$$f(x) = \lambda \exp(-\lambda x) \tag{10}$$

Estymator parametru λ, uzyskany MNW, ma następującą postać:

$$\hat{\lambda}_1 = \frac{1}{\bar{x}} \tag{11}$$

Można wykazać, że estymator ten jest obciążony (El-Sayyad, 1967, s. 525). Nieobciążony estymator wyraża się zależnością:

$$\hat{\lambda}_2 = \frac{N-1}{N \cdot \bar{x}} \tag{12}$$

Natomiast estymator EMSE wyraża się wzorem (El-Sayyad, 1967):

$$\hat{\lambda}_3 = \frac{N-2}{N \cdot \bar{x}} \tag{13}$$

W celu rozstrzygnięcia, którą postać estymatora powinno się stosować, przeprowadzono symulacje komputerowe z użyciem generatora liczb losowych o rozkładzie wykładniczym dla próbki liczącej $N = 21$ elementów. Jako wynik podano uśrednione wartości względnego obciążenia *bL* [wzór(4)] i względnego błędu średniokwadratowego *bse* [wzór (5)] dla K = 100000 powtórzeń, zamieszczonych w kolumnie 2 i 3 tabeli 1.

Następnie wykonano test chi-kwadrat. Jako wynik testu przyjęto unormowaną statystykę *h* stanowiącą stosunek wartości statystyki testu chi-kwadrat do wartości krytycznej. W przypadku wartości unormowanej statystyki *h* większej od 1 należało odrzucić hipotezę o zgodności rozkładu wykładniczego z rozkładem empirycznym.

W tabeli 1 zamieszczono wartość średnią unormowanej statystyki oznaczoną jako *hS* (kolumna 4) oraz procentową wartość przypadków *negh* negatywnego wyniku testu (kolumna 5). Opisane postępowanie dotyczące testu chi-kwadrat powtórzono w odniesieniu do testu Kołmogorowa. Wykorzystano unormowaną statystykę *k*, na podstawie której określono uśrednioną wartość *kS* (kolumna 6 tabeli 1) oraz procentową wartość negatywnych przypadków testu *kneg* (kolumna 7 tabeli 1).

Tabela 1. Wyniki symulacji komputerowych dotyczących estymacji parametru rozkładu wykładniczego

Estymator	*bL*	*bse*	*hS*	*hneg* [%]	*kS*	*kneg* [%]
$\hat{\beta}_1$	0,0505	0,0553	0,417	5,02	0,525	0,490
$\hat{\beta}_2$	0,00046	0,0479	0,413	4,73	0,530	0,498
$\hat{\beta}_3$	–0,0450	0,0456	0,419	4,64	0,547	0,748
$\hat{\lambda}_1$	0,0497	0,0607	0,413	4,73	0,530	0,498
$\hat{\lambda}_2$	–0,00029	0,0528	0,417	5,02	0,525	0,490
$\hat{\lambda}_3$	–0,0503	0,0502	0,432	5,51	0,537	0,636

Źródło: opracowanie własne.

Wyniki obliczeń zamieszczone w kolumnie 2 tabeli 1 jednoznacznie potwierdzają, że estymatory β_2 i λ_2 są estymatorami nieobciążonymi. Na podstawie wartości *bse* (kolumna 3) stwierdza się, że estymatory β_3 i λ_3 są estymatorami EMSE. Wyznaczając wartość ilorazu $\frac{0,0456}{0,0502}=0,908$, stwierdza się, że nieco mniejszą wartość błędu średniokwadratowego zapewnia estymator β_3. Na podstawie rezultatów zawartych w kolumnach 2 i 3 stwierdza się, że estymatory β_2 i λ_2 są EN oraz estymatory β_3 i λ_3 są EMSE, natomiast uzyskane wyniki nie rozstrzygają o przydatności poszczególnych estymatorów.

Z danych zawartych w tabeli 1 wynika zgodność wyników obydwu testów dla estymatorów β_1 i λ_2, a także dla estymatorów β_2 i λ_1. Przyczyny tego należy upatrywać w postaci wzorów (14) i (15):

$$\hat{\lambda}_1 = \frac{N}{\sum_{i=1}^{N} x_i}; \quad \hat{\beta}_2 = \frac{\sum_{i=1}^{N} x_i}{N} \tag{14}$$

$$\hat{\lambda}_2 = \frac{N-1}{\sum_{i=1}^{N} x_i}; \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{N} x_i}{N-1} \tag{15}$$

z których wynika, że estymator λ_1 (λ_2) jest odwrotnością estymatora β_2 (β_1).

Najmniejszą wartość uśrednionej statystyki testu χ^2 (kolumna 4) w przypadku wariantu A [wzór (6)] zapewnia estymator β_2 (*hS* = 0,413), natomiast dla wariantu B [wzór (10)] najmniejszą wartość *hS* = 0,413 uzyskuje się dla estymatora λ_1, który nie jest EN ani EMSE.

Najmniejszą wartość liczby negatywnych przypadków testu χ^2 (kolumna 5) w przypadku wariantu A odnotowuje się dla estymatora β_3 (*hneg* = 4,64%), natomiast dla wariantu B najmniejszą wartość *hneg* = 4,73% uzyskuje się dla estymatora λ_1. Tym samym wyniki testu χ^2 zamieszczone w kolumnach 4 i 5 nie przesądzają o przewadze określonego estymatora. Podobnie wygląda sytuacja dla wyników testu Kołmogorowa, gdzie dla wariantu A najmniejsze wartości w kolumnie 6 i 7 zapewnia estymator β_1, który nie jest EN ani też EMSE.

Podsumowując wyniki zamieszczone w tabeli 1, stwierdza się, że nie ma podstaw do wskazania jednego z estymatorów EN lub EMSE.

Odwołanie się do wyników testu chi-kwadrat oraz testu Kołmogorowa było wymuszone brakiem rozstrzygnięć na podstawie wcześniejszych kryteriów. Mianowicie estymator EN zapewniał najmniejszą wartość obciążenia względnego *bL*, natomiast estymator EMSE prowadził do najmniejszej wartości błędu średniokwadratowego.

2. Estymacja parametru rozkładu Laplace'a

Jako wariant A rozpatruje się zależność (16) opisującą gęstość rozkładu Laplace'a:

$$f(x) = \frac{1}{2 \cdot \beta} \exp\left(-\frac{|x-\mu|}{\beta} \right) \tag{16}$$

Stosując MNW, uzyskuje się następującą postać estymatorów rozkładu:

$$\hat{\mu} = mediana(x_i) \tag{17}$$

$$\hat{\beta} = \frac{1}{N} \sum_{i=1}^{N} |x_i - \hat{\mu}| \tag{18}$$

Estymator opisany wzorem (18) jest zgodny i asymptotycznie normalny (Kotz, Kozubowski, Podgórski, 2001, s. 71).

W celu uproszczenia wnioskowania zakłada się, że wartość parametru μ jest znana.

W ramach symulacji komputerowych parametr ten był zadeklarowany w generatorze liczb losowych o rozkładzie dwuwykładniczym:

$$X = \mu - \beta \cdot \mathrm{sgn}(R) \ln(1 - 2|R|) \tag{19}$$

gdzie $R \subset (-0{,}5; 0{,}5)$ – liczby losowe o rozkładzie równomiernym.

Stosując MNW, uzyskuje się:

$$\hat{\beta}_3 = \frac{1}{N} \sum_{i=1}^{N} |x_i - \mu| \tag{20}$$

gdzie: μ jest znane.

Estymator opisany wzorem (18) jest nieobciążony i najefektywniejszy (Kotz i in., 2001).

Estymator opisany wzorem

$$\hat{\beta}_2 = \frac{1}{N+1} \sum_{i=1}^{N} |x_i - \mu| \tag{21}$$

zapewnia minimum błędu średniokwadratowego (EMSE).

Dodatkowo zostaną uwzględnione następujące estymatory:

$$\hat{\beta}_1 = \frac{1}{N+2} \sum_{i=1}^{N} |x_i - \mu| \tag{22}$$

oraz

$$\hat{\beta}_4 = \frac{1}{N-1} \sum_{i=1}^{N} |x_i - \mu| \tag{23}$$

Jako wariant B rozpatrzony został rozkład Laplace'a, którego gęstość wyraża się wzorem:

$$f(x) = \frac{\lambda}{2}\exp(-\lambda|x-\mu|) \tag{24}$$

Podobnie jak dla wariantu A zakłada się znajomość parametru μ. W wyniku zastosowania metody największej wiarygodności otrzymuje się (Purczyński, 2003, s. 135):

$$\hat{\lambda}_1 = \frac{N}{\sum_{i=1}^{N}|x_i-\mu|} \tag{25}$$

gdzie: μ jest znane.

Estymator nieobciążony wyraża się wzorem:

$$\hat{\lambda}_2 = \frac{N-1}{\sum_{i=1}^{N}|x_i-\mu|} \tag{26}$$

Natomiast estymator EMSE opisuje wzór:

$$\hat{\lambda}_3 = \frac{N-2}{\sum_{i=1}^{N}|x_i-\mu|} \tag{27}$$

Dodatkowo rozpatrzono estymator:

$$\hat{\lambda}_4 = \frac{N-3}{\sum_{i=1}^{N}|x_i-\mu|} \tag{28}$$

W tabeli 2 zamieszczono wyniki obliczeń opisanych szczegółowo przy omówieniu tabeli 1. Wyniki zawarte w kolumnie 2 i 3 potwierdzają rozważania teoretyczne, to znaczy: estymatorami nieobciążonymi są estymatory β_3 i λ_2, natomiast β_2 i λ_3 są estymatorami EMSE. Wyznaczając wartość ilorazu $\frac{0,04542}{0,054994}=0,909$, stwierdza się, że nieco mniejszą wartość błędu średniokwadratowego zapewnia estymator β_2.

Tym razem powtórzenie wyników testów występuje dla estymatorów β_3 i λ_1 oraz β_4 i λ_2. Minimalną wartość parametru *hS* (kolumna 4) oraz *hneg* (kolumna 5) uzyskano dla estymatorów β_4 i λ_2 oraz β_3 i λ_1. Należy zauważyć, że estymator β_4 nie jest EN ani też EMSE. W przypadku testu Kołmogorowa minimalną wartość parametru *kS* (kolumna 6) odnotowano dla estymatorów β_4 i λ_2. Natomiast najmniejszą wartość *kneg* (kolumna 7) otrzymano dla estymatorów β_4 i λ_4, które nie są EN ani EMSE.

Tabela 2. Wyniki symulacji komputerowych dotyczących estymacji parametru rozkładu Laplace'a

Estymator	*bL*	*bse*	*hS*	*hneg* [%]	*kS*	*kneg* [%]
$\hat{\beta}_1$	–0,08666	0,0472	0,601	14,8	0,619	4,77
$\hat{\beta}_2$	–0,04514	0,04542	0,583	13,8	0,608	4,37
$\hat{\beta}_3$	0,00033	0,0476	0,571	13,4	0,600	4,06
$\hat{\beta}_4$	0,05034	0,05503	0,569	13,2	0,596	3,82
λ_1	0,04957	0,06036	0,571	13,4	0,600	4,06
λ_2	–0,00041	0,05252	0,569	13,2	0,596	3,82
λ_3	–0,05039	0,04994	0,577	13,7	0,596	3,69
λ_4	–0,10037	0,05262	0,598	14,7	0,601	3,59

Źródło: opracowanie własne.

Podsumowanie

Wracając na chwilę do tabeli 1, warto zwrócić uwagę na wartości negatywnej liczby wyników testu. Mianowicie dla kolumny 5 uzyskuje się wartość średnią *hnegsr* = 4,94%, natomiast dla kolumny 7 wartość średnia wynosi *knegsr* = 0,56%. Stosunek tych wartości to 4,94/0,56 = 8,82. Oznacza to, że niemal dziewięciokrotnie częściej uzyskiwano negatywny wynik testu χ^2 niż testu Kołmogorowa. W przypadku tabeli 2 nie ma aż tak znaczącej różnicy: $\frac{hnegsr}{knegsr} = \frac{13,775}{4,0225} = 3,424$, ale w dalszym ciągu częściej uzyskuje się negatywny wynik testu chi-kwadrat niż testu Kołmogorowa. Wynika to z faktu, że test Kołmogorowa odnosi się do przypadku, kiedy znane są parametry rozkładu. Natomiast w podrozdziale 1 oraz 2 wyznaczany był jeden parametr rozkładu. W takim przypadku należy zachować ostrożność w stosunku do uzyskanych wyników (Domański, Pruska, 2000, s. 171).

Tabela 3. Zestawienie estymatorów
zapewniających minimalne wartości uwzględnionych kryteriów

Wariant	Minimum *hS*	Minimum *hneg*	Minimum *kS*	Minimum *kneg*
Tabela 1				
A	β_2 – EN	β_3 – EMSE	β_1	β_3 – EMSE
B	λ_1	λ_1	λ_2 – EN	λ_2 – EN
Tabela 2				
A	β_4	β_4	β_4	β_4
B	λ_2 – EN	λ_2 – EN	λ_2 – EN λ_3 – EMSE	λ_3 – EMSE

Źródło: opracowanie własne.

W tabeli 3 zestawiono estymatory, które realizowały najmniejsze wartości dla poszczególnych kryteriów. Estymator EN występuje 6 razy, a estymator EMSE 4 razy. Najlepiej wypadł estymator β_4, który dla rozkładu Laplace'a (wariant A) był optymalny dla wszystkich kryteriów. Jednak estymator ten nie jest EN ani EMSE. Pytanie postawione na wstępie artykułu, który estymator, EN czy też EMSE, wykazuje przewagę, pozostaje bez odpowiedzi. Uwzględnione testy zgodności rozkładu miały na celu ocenę przydatności wybranych estymatorów. Można też było rozwiązać inny problem, a mianowicie dobrać estymator zapewniający minimum statystyki testu chi-kwadrat. Jednak to wykracza poza ramy artykułu.

Literatura

Domański, C., Pruska, K. (2000). *Nieklasyczne metody statystyczne*. Warszawa: PWE.

El-Sayyad, G.M. (1967). Estimation of the Parameter of an Exponential Distribution. *Journal of the Royal Statistical Society*, *29*, *3*, 525–532.

Housila, P.S., Sarjinder, S., Jong-Min, K. (2012). *Some Alternative Classes of Shrinkage Estimators for a Scale Parameter of the Exponential Distribution*. *The Korean Journal of Applied Statistics*, *25* (2), 301–309.

Kotz, S., Kozubowski, T., Podgórski, K. (2001). *The Laplace Distribution and Generalizations*. Boston: Birkhauser.

Krzyśko, M. (1997). *Statystyka matematyczna*. Cz. 2. Poznań: Wyd. UAM.

Purczyński, J. (2003). *Wykorzystanie symulacji komputerowych w estymacji wybranych modeli ekonometrycznych i statystycznych*. Szczecin: Wyd. Naukowe US.

UNBIASED ESTIMATOR VERSUS MINIMUM MEAN SQUARE ERROR ESTIMATOR

Abstract

The aim of this paper in to answer the question whether the unbiased estimator (UN) or the minimum mean square error estimator (MMSEE). For this purpose simple examples of determining estimators of exponential and Laplace distributions parameters were considered. As criteria, the values of chi-square test statistic and Kolmogorov test statistic were examined. The obtained results do not provide an unequivocal answer to the problem.

Translated by Ewa Stefanowska

Keywords: unbiased estimator, minimum mean square error estimator

JEL Code: C51