SYMULACYJNE BADANIE SZYBKOŚCI ZBIEŻNOŚCI ROZKŁADU STATYSTYK DO ROZKŁADU NORMALNEGO

ŚLĄSKI PRZEGLĄD STATYSTYCZNY
Nr 11 (17)

Janusz L. Wywiał, Małgorzata Krzciuk, Michał Mierzwa
Uniwersytet Ekonomiczny w Katowicach

ISSN 1644-6739

Streszczenie: Z centralnych twierdzeń granicznych wiadomo, że rozkład średniej z próby jest zbieżny do rozkładu normalnego. Problemem jest jednak ocena szybkości tej zbieżności. Tym zagadnieniem zajmują się autorzy podręczników z rachunku z prawdopodobieństwa, por. np. Krzyśko [2000]. W niniejszej pracy zaproponowano symulacyjne badanie tego problemu. W tym celu generowano rozkład średniej arytmetycznej z próby losowanej z populacji o rozkładzie wykładniczym. Rozbieżność między dystrybuantą empiryczną i teoretyczną była oceniana m.in. za pomocą znanej statystyki Kołmogorowa. Otrzymane w pracy wyniki mają stanowić przyczynek do metod wnioskowania o parametrach populacji na podstawie statystyk wyznaczanych z prób, które niekoniecznie są prostymi próbami losowanymi zwrotnie.

Słowa kluczowe: rozkład średniej z próby, szybkość zbieżności, rozkład normalny.

1. Wstęp

Zbiór (populację) oznaczamy przez U, a próbę dobieraną z niego bezzwrotnie − przez s. Rozmiary populacji i próby oznaczamy odpowiednio przez N i n. W populacji są obserwowane wartości zmiennej x, które oznaczamy przez x_i, $i = 1, \ldots, N$.

Rozważymy rozkłady prawdopodobieństwa prób prostych losowanych bezzwrotnie ze skończonej i ustalonej populacji. Nasze zadanie polega na zbadaniu szybkości zbieżności rozkładu następującej statystyki do rozkładu normalnego standardowego oznaczanego symbolem $N(0,1)$.

$$z_s = \frac{\bar{x}_s - \bar{x}}{\sqrt{v_s(x)}}\sqrt{n}\,, \tag{1}$$

gdzie: $\bar{x} = \frac{1}{N}\sum_{k\in U} x_k$, $\bar{x}_s = \frac{1}{n}\sum_{k\in s} x_k$, $v_s(x) = \frac{1}{n-1}\sum_{k\in s}\left(x_k - \bar{x}_s\right)^2$.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY Nr 11 (17)

Dodajmy, że statystyka z_s jest nazywana studetyzowanym odchyleniem średniej z próby od średniej w populacji.

Symulacyjna procedura analizy stopnia zgodności rozkładu statystyki z_s z rozkładem normalnym standardowym w zależności od rozmiaru próby przebiega następująco. Z populacji U losujemy niezależnie, za pomocą tego samego schematu losowania, M prób, przy czym wybrana za i-tym razem próba ma stały rozmiar n i oznaczamy ją przez s_i, $i = 1,\ldots, M$. Następnie, na podstawie próby s_i, jest wyznaczana statystyka zgodnie ze wzorem (3), która ma postać:

$$z_{s_i} = \frac{\bar{x}_{s_i} - \bar{x}}{\sqrt{v_{s_i}(x)}}\sqrt{n}\,,\ i = 1,\ldots, M, \tag{2}$$

gdzie: $\bar{x}_{s_i} = \frac{1}{n}\sum_{k\in s_i} x_k$, $\quad v_{s_i}(x) = \frac{1}{n-1}\sum_{k\in s_i}\left(x_k - \bar{x}_{s_i}\right)^2$.

Teraz ciąg obserwacji z_{s_i} , $i = 1, \ldots, M$, porządkujemy od najmniejszej do największej, co daje ciąg uporządkowany o postaci: $z_{(1)}\le z_{(i)}\le\ldots\le z_{(M)}$. Na jego podstawie jest wyznaczana wartość dystrybuanty empirycznej za pomocą wzoru:

$$F_{M|n}(z) = \begin{cases} 0 & \text{dla} \quad z \le z_{(1)} \\ \dfrac{i-1}{M} & \text{dla} \quad z_{(i)} < z \le z_{(i+1)}, i = 1,\ldots,M-1\,. \\ 1 & \text{dla} \quad z > z_{(M)} \end{cases} \tag{3}$$

Oznaczając przez $F(z)$ dystrybuantę rozkładu normalnego standardowego, jej odległość od dystrybuanty empirycznej określamy następująco:

$$d_{M|n}(z) = \max_{-\infty<z<\infty}\left|F_{M|n}(z) - F(z)\right| \tag{4}$$

Można wykazać, że

$$d_{M|n}(z) = \max_{i=1,\ldots,M}\left|F_{M|n}(z_{(i)}) - F(z_{(i)})\right| \tag{5}$$

Statystyka $d_{M/n}(z)$ jest sprawdzianem testu zgodności Kołmogorowa (por. [Domański 1990, s. 65]), który również podaje tablice warto-

ści krytycznych tego testu za Millerem [1956]. Obszar krytyczny testu jest prawostronny. W szczególności dla M =100 wartości krytyczne testu wynoszą 0,1207, 0,1340 i 0, 1608 dla poziomów istotności odpowiednio: 0, 1, 0, 5 i 0, 01.

2. Analizy symulacyjne

Użyto danych wygenerowanych z rozkładu wykładniczego z parametrem równym przeciętnej wartości faktur rzeczywistych. Liczebność każdej populacji wynosiła 12 000.

Pierwszym etapem analizy było symulacyjne wyznaczenie, z użyciem programu R, dystrybuanty zestandaryzowanych średnich z próbek dla danych zarówno rzeczywistych, jak i wygenerowanych. Wykorzystano w tym celu samodzielnie zaprogramowaną funkcję pozwalającą na zastosowanie metody wkładów prostokątnych. Druga procedura, wykorzystująca funkcję ecdf(), napisana w języku R, wyznacza empiryczną funkcję dystrybuanty określoną wzorem (3). Jako liczbę iteracji M przyjęto 5000. Do obliczenia zestandaryzowanych średnich wartości z próbek wykorzystano wzór (1).

W pierwszej z wymienionych metod wyznaczenie dystrybuanty poprzedzała aproksymacja funkcji gęstości za pomocą wkładów prostokątnych. Przybliżenie nieznanej funkcji $f(x)$ badanej zmiennej rozpoczęto od wstępnego przyjęcia założenia, że rozkład zmiennej X jest rozkładem jednostajnym na przedziale (a, b), gdzie za krańce przedziału przyjęto minimalną oraz maksymalną wartość w ciągu obserwacji zmiennej X. W kolejnych krokach wyznaczany był wkład poszczególnych obserwacji zmiennej w kształtowanie przybliżenia funkcji gęstości, z wykorzystaniem następującego wzoru, pochodzącego z pracy *Perceptron − sistiema raspoznawanija obrazow,* co podajemy za [Kolonko 1980, s. 74]:

$$f_n(x) = \frac{1}{n+1}\left[\frac{1}{b-a} + \sum_{l=1}^{n} c_l\right], \qquad (6)$$

gdzie:

$$c_l = \begin{cases} \dfrac{1}{h} & \text{dla} \quad x \in \left(x_l - \dfrac{h}{2}, x_l + \dfrac{h}{2} \right) \\ 0 & \text{dla} \quad x \notin \left(x_l - \dfrac{h}{2}, x_l + \dfrac{h}{2} \right) \end{cases},$$

$x_1, x_2, \ldots, x_n$ – ciąg obserwacji zmiennej X,

$f(x)$ – nieznana funkcja gęstości zmiennej X,

$f_l(x)$ – przybliżenie funkcji f(x) w l-tym kroku,

h – arbitralnie ustalony parametr; parametr h powinien spełniać warunek $\min\limits_{\substack{l.k \\ l \neq k}} d\left(x_l, x_k\right) < h < b - a$ – (d oznacza odległość między punktami), por. [Kolonko 1980].

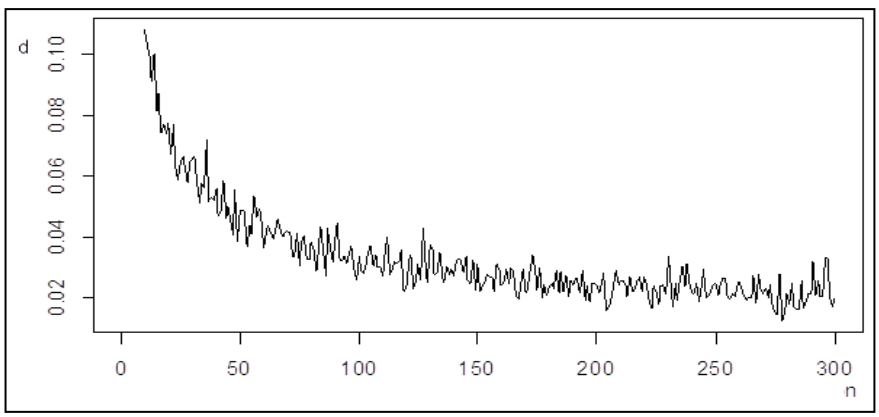
Określenie wartości funkcji dystrybuanty dla badanej zmiennej w tym podejściu oparte było na skumulowanych wartościach pól prostokątów wyznaczonych przez przybliżenie funkcji $f(x)$.

W drugim przypadku wykorzystana została empiryczna funkcja dystrybuanty określona wzorem (3). Pozwoliła ona na wyznaczenie empirycznej funkcji dystrybuanty. Argumentem funkcji był wektor zestandaryzowanych średnich wartości faktur z prób.

Wyznaczenie wartości dystrybuant zestandaryzowanych średnich wartości, rzeczywistych oraz wygenerowanych, faktur z próbek za pomocą obu metod pozwoliło na wyznaczenie zmiennej d_n będącej maksymalną wartością modułów różnic między dystrybuantą teoretyczną a empiryczną dla zadanej wielkości próby n. Wielkość tę można zapisać wzorem (4), $n \in [10, 300]$. W analizowanym przypadku za dystrybuantę teoretyczną przyjęto dystrybuantę rozkładu normalnego standardowego. Uzyskane w ten sposób wartości d_{Mn} zaprezentowane zostały na rys. 1-4.

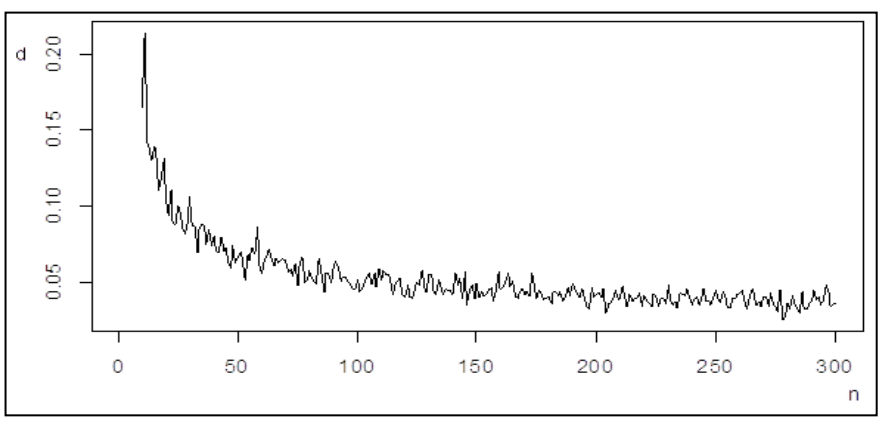
Na rysunku 1 odłożono maksymalne moduły z różnic między dystrybuantą zestandaryzowanych średnich wartości wygenerowanych faktur z próbek a dystrybuantą rozkładu normalnego standardowego. Maksimum d w tym wariancie wyniosło nieco ponad 0,1. W drugim przypadku (rys. 2), gdy za d przyjęto maksymalne wartości bezwzględne dla różnic między dystrybuantą wyznaczoną metodą wkładów prostokątnych z zestandaryzowanych średnich wartości wygene-

rowanych faktur z próbek a dystrybuantą rozkładu normalnego standardowego, wielkość d przyjęła wartość $p \approx 0{,}2$. Poziomy, wokół których oscylowały wartości d dla obu wariantów przy $n \geq 150$, wynoszą odpowiednio $p = 0{,}03$ i $p = 0{,}04$. W związku z tym, traktując wartości p jako p-wartości testu Kołmogorowa, nie ma podstaw do odrzucenia hipotezy głoszącej, iż rozkład badanej statystyki nieistotnie różni się od rozkładu normalnego standardowego przy poziomie istotności równym 0,05.



Rys. 1. Wartości zmiennej *d*, wariant 1

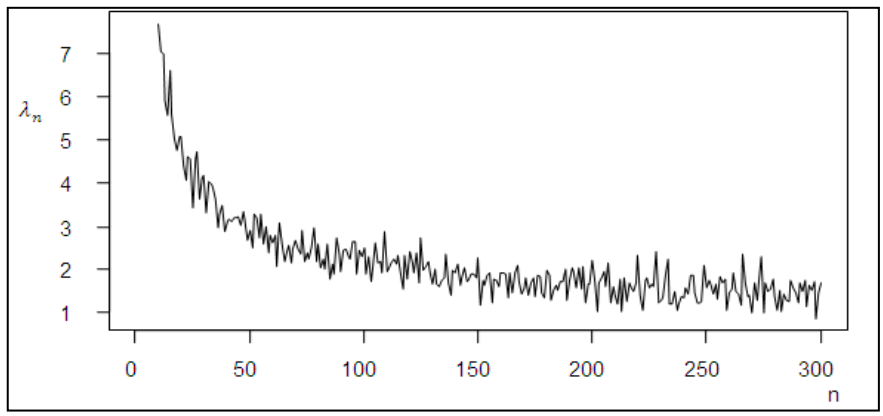
Źródło: opracowanie własne.



Rys. 2. Wartości zmiennej *d*, wariant 2

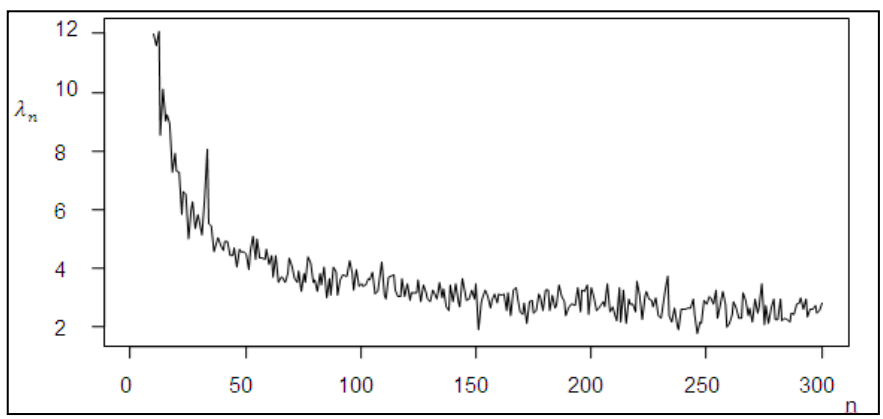
Źródło: opracowanie własne.

ŚLĄSKI PRZEGLĄD STATYSTYCZNY
Nr 11 (17)



Rys. 3. Wartości statystyki Kołmogorowa, wariant 1

Źródło: opracowanie własne.



Rys. 4. Wartości statystyki Kołmogorowa, wariant 2

Źródło: opracowanie własne.

Maksymalna λ_n dla wariantu pierwszego (rys. 3) wynosiła ok. 2 dla rozmiarów prób $n \geq 150$. W przypadku drugiego wariantu, gdy empiryczna dystrybuanta jest wyznaczana za pomocą metody wkładów prostokątnych, maksymalna wartość statystyki λ_n wynosi 3, por. rys. 4. W przypadku analizy danych wygenerowanych również dla prób o liczebności mniejszej niż 150 wartości badanej statystyki testowej osiągały nawet wartości większe od liczby siedem.

3. Wnioski

Podsumowując, na podstawie przeprowadzonych analiz można stwierdzić, że wraz ze wzrostem liczebności próby dla n z przedziału [10,150] wartość maksymalnego modułu z różnic dystrybuanty teoretycznej i empirycznej oraz statystyka Kołmogorowa spadały. Pozwalało to nie odrzucać hipotezy o zgodności rozkładu prawdopodobieństwa badanej statystyki od rozkładu normalnego standardowego. W przypadku $n > 150$ wartości badanej wielkości oscylowały wokół stałego poziomu.

Przedstawiona procedura badania szybkości zbieżności do rozkładu normalnego może być rozszerzana zarówno na przypadki innych statystyk niż średnia arytmetyczna, jak i na przypadki innych prostych schematów losowania próby.

Literatura

Domański Cz., *Testy statystyczne*, PWE, Warszawa 1990.

Iwachnienko A.G. (red.), *Perceptron – sistiema raspoznawanija obrazow*, Naukowa Dumka, Kijów 1975.

Kolonko J., *Analiza dyskryminacyjna i jej zastosowania w ekonomii*, PWE, Warszawa 1980.

Krzyśko M., *Wykłady z teorii prawdopodobieństwa*, Wydawnictwo Naukowo-Techniczne, Warszawa 2000.

Miller L.H., *Table of percentage points of Kolmogorov Statistic*, "Journal of the American Statistical Association" 1956, Vol. 51, s. 111-121.

SIMULATION ANALYSIS OF CONVERGENCE OF SAMPLE MEAN DISTRIBUTION TO NORMAL DISTRIBUTION

Summary: The paper discusses studentized sample mean distribution. The sample is from exponential distribution. On the basis of independent replications of the samples empirical distributions studentized mean was calculated. The distance between the empirical distributions and the standard normal distribution was measured by means well known as statistics of Kolmogorov. Under the appropriate sample sizes the degree of the difference between the empirical and theoretical distributions was evaluated. Moreover, the hypothe-

ŚLĄSKI PRZEGLĄD STATYSTYCZNY
Nr 11 (17)

sis on normality of the empirical distributions was tested by means of the Kolmogorov test.

Keywords: sample mean distribution, normal distribution, convergence of sample mean distribution.