Zygmunt Lech WARSZA [1], Marian Jerzy KORCZYŃSKI [2]
[1] PRZEMYSŁOWY INSTYTUT AUTOMATYKI i POMIARÓW PIAP, Al. Jerozolimskie 202, 02-486 Warszawa
[2] POLITECHNIKA ŁÓDZKA, KATEDRA PRZYRZĄDÓW PÓŁPRZEWODNIKOWYCH I OPTOELEKTRONICZNYCH, ul. Wólczańska 211/215, 90-924 Łódź

Statystyki skośności i kurtozy małych próbek pomiarowych z populacji o rozkładzie normalnym i kilku innych

Doc. dr inż. Zygmunt Lech WARSZA

Studia 1959, doktorat 1967 i praca 1960-70 na Wydz. Elektrycznym Politechniki Warszawskiej oraz w Instytucie Elektrotechniki 1958-63. Docent od 1970. Zorganizował i prowadził: Wydział Transportu Pol. Świętokrzyskiej 1970-76, Ośrodek Aparatury Pomiarowej IMGW 1978-81, Zakład Techniki Pomiarowej Instytutu Chemii Przemysłowej 1983-91. Obecnie: główny specjalista w Przemysłowym Instytucie Automatyki i Pomiarów PIAP Warszawa. Autor około 160 publikacji, 4 monografii, kilku-dziesięciu prac badawczych.

e-mail: zlw@op.pl

Dr inż. Marin Jerzy KORCZYŃSKI

Ukończył wydział Elektryczny Politechniki Łódzkiej w 1973 r., doktorat w 1981. Adiunkt od 1981 r. Zainteresowania naukowe i dydaktyczne: systemy pomiarowe, cyfrowe przetwarzanie sygnałów, niepewność pomiaru, metody matematyczne i techniki informatyczne w metrologii. Autor i współautor przeszło 100 publikacji, 3 książek w języku angielskim, kilku patentów i szeregu opracowań naukowych oraz badawczych i konstrukcyjnych dla przemysłu.

e-mail: jerzykor@p.lodz.pl

Streszczenie

Przedstawiono podstawowe właściwości statystyczne rozkładów skośności i kurtozy dla zbioru próbek o określonych małych liczbach elementów. Rozkłady te wyznaczono je metodą Monte Carlo. Próbki wielokrotnie pobierano losowo z populacji o rozkładzie normalnym oraz dla porównania z populacji o kilku innych prostych rozkładach. Znajomość statystyk skośności i kurtozy umożliwi bardziej wiarygodne oszacowanie odchylenia standardowego i niepewności estymatora wartości mensurandu z próbek pomiarowych o małej liczbie obserwacji, gdy znany jest rozstęp populacji rozrzutu ich wartości.

Słowa kluczowe: próbka pomiarowa, skośność, kurtoza.

Statistics of skewness and kurtosis of small measurement samples from populations of normal and few other distributions

Abstract

Statistics of skewness and kurtosis distributions and their basic parameters for a set of samples of certain small numbers of elements are find. These distributions were determined using the Monte Carlo method. The samples were repeatedly taken at random from a normally distributed population and for comparison from the population of a few other simple distributions. Knowledge of skewness and kurtosis statistics allow a more reliable estimate of the standard deviation and the uncertainty of the value of the measurand estimator from the measurement samples of a small number of observations, when the range of their value distribution is known.

Keywords: Measurement sample, skewness, kurtosis.

1. Wprowadzenie

Dane eksperymentalne, w tym wartości obserwacji pomiarowych, są zwykle niesymetrycznie rozproszone losowo i różnie skoncentrowane. Dla przyrządu lub systemu pomiarowego z czujnikami różnych wielkości jest to zbiór odczytów lub wartości sygnału wyjściowego otrzymywanych poprzez próbkowanie. Nierównomierność rozrzutu dotyczy nie tylko próbek z populacji o niesymetrycznych rozkładach prawdopodobieństwa, ale też i niezbyt dużych próbek z populacji o rozkładzie normalnym i o innych rozkładach symetrycznych. Asymetria takich próbek wzrasta przy zmniejszaniu sie ich liczby elementów. Jednakże w wielu przypadkach występujących w praktyce dysponuje się tylko małą liczbą obserwacji pomiarowych i z różnych przyczyn nie ma możliwości jej zwiększenia. Ograniczenie to spowodowane jest:

- brakiem większej liczby obiektów badanych, (np. przy walidacji metody stosowanej tylko w kilku akredytowanych laboratoriach),
- dużym kosztem badań lub ograniczonym czasem ich wykonania,
- niemożnością ponownego przeprowadzenia pomiarów np. w badaniach odległych terenowo i w medycynie, w badaniach procesów jednorazowych i badaniach niszczących obiekt lub zmieniających nieodwracalnie jego właściwości.

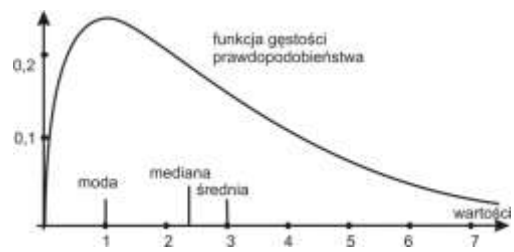
We wszystkich powyższych przypadkach można jedynie wykorzystać otrzymaną małą próbkę danych i dołożyć starań, by przy ich przetwarzaniu jak najlepiej oszacować wynik pomiarów. Dotychczas przy wyznaczaniu tego wyniku wg rekomendacji Przewodnika GUM [2] traktuję się próbkę danych tak, jakby pochodziła z populacji o rozkładzie normalnym. Estymatorem wartości mensurandu jest jej wartość średnia, a niepewność typu A, powstającą wskutek rozrzutu danych, opiera się na wariancji próbki. Należało by zbadać: na ile istotna może być też znajomość innych parametrów statystycznych małych i bardzo małych próbek, w tym jej skośności i kurtozy w przypadkach, gdy rodzaj rozkładu populacji danych nie jest znany. Poniżej, jako pierwszy etap tych badań, metodą symulacji Monte Carlo (MC) wyznaczy się statystyki skośności i kurtozy małych i bardzo małych próbek pobranych losowo z populacji o rozkładzie normalnym i dla porównania - o dwu innych prostych rozkładach.

2. Współczynniki skośności populacji i próbki

Rysunek 1 podaje przykład rozkładu gęstości prawdopodobieństwa (pdf) o asymetrii prawostronnej (wydłużone prawe ramię). Zaznaczono położenie podstawowych parametrów statystycznych tego rozkładu: mody, mediany i średniej. Asymetrię rozkładu opisują się różnymi współczynnikami skośności. Proste miary skośności (ang. *skewness*) są następujące [1]:

$$A_d = \frac{\mu - d}{\sigma} \quad (1a) \qquad A_m = 3\frac{\mu - m}{\sigma} \quad (1b) \qquad A_Q = \frac{Q_1 + Q_3 - 2m}{2Q} \quad (1c)$$

gdzie: ■ – średnia arytmetyczna, m – mediana, d – dominanta (moda), σ – odchylenie standardowe, ■ - pierwszy i trzeci kwartyl, ■ - odchylenie ćwiartkowe.



Rys. 1. Parametry rozkładu o dystrybuancie (pdf) asymetrycznej prawostronnie
Fig. 1. Parameters of the right-asymmetric distribution function (pdf)

Miary (1a-c) nadają się jedynie do porównywania asymetrii jednego rodzaju rozkładów. Jednolity opis asymetrii różnych rozkładów umożliwia podany przez Pearsona współczynnik skośności γ_1 [1, 3]. Dla populacji X jest on stosunkiem momentu

centralnego μ_3 i odchylenia standardowego w trzeciej potędze σ^3 - wzór (2) w Tabeli 1. Współczynnik γ_1 ma wartość zero dla rozkładów symetrycznych, jest dodatni dla rozkładów o asymetrii prawostronnej i ujemny dla rozkładów o asymetrii lewostronnej.

Dla próbki o n elementach x_i z populacji generalnej X współczynnik skośności Pearsona g_1 określa się przez jej momenty centralne $\bar{x}$, $m_2 = s^2$ i m_3 - wzór (3) w tabeli 1, które są estymatorami momentów centralnych μ, μ_2, μ_3 populacji generalnej.

Tab. 1. Współczynniki skośności wg Pearsona dla rozkładu i próbki
Tab. 1. Pearson skewness coefficients of distribution and sample

Dla populacji		(2)
Dla próbki		(3)

3. Skośność próbek z populacji o rozkładzie normalnym

Współczynnik skośności próbki g_1 wg wzoru (3) jest obciążony. Wprowadzono więc zmodyfikowaną jego postać:

$$g = \frac{\sqrt{n(n-1)}}{n-2} g_1 \tag{4}$$

W statystycznych programach obliczeniowych używa się też współczynnika skośności próbki o innej postaci [3], tj.:

$$g = \frac{n\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3}{(n-1)(n-2)s^3} \tag{5}$$

oraz standaryzowanego współczynnika skośności

$$SSKE = g\sqrt{\frac{n}{6}} \ . \tag{6}$$

Dla n >150 i populacji symetrycznych ma on rozkład normalny.

Różnice wartości pomiędzy różnymi postaciami współczynnika skośności nie są duże i występują dla bardzo małych próbek.

4. Wariancja współczynnika skośności

Wariancja współczynnika skośności g dla próbki o n elementach z populacji o rozkładzie normalnym wynosi[1]:

$$D(g) = \frac{6n(n-1)}{(n-2)(n+1)(n+3)} \tag{7}$$

W [4] zamieszczono nieco inny wzór podany przez Smirnova

$$D(g) = \frac{6(n-2)}{(n+1)(n+3)} = \frac{6}{n}\left[1 - \frac{12}{2n+7} + O\left(\frac{1}{n^3}\right)\right] \tag{8}$$

gdzie: $O(\cdot)$ - *człon resztkowy* $1/n^3$

Wzór (8) dotyczy próbek o średniej liczności począwszy od n>25 elementów. Przy dużej liczbie n człon resztkowy $O(\cdot)$ staje się pomijalny i dla próbek o n>150 wariancja $D(g) \rightarrow 6/n$.

5. Kurtoza małych próbek

Kurtoza jest miarą spłaszczenia (smukłości) rozkładu. Kurtoza próbki informuje o koncentracji jej danych. Oblicza się ją na podstawie wartości momentów m_4, $m_2 = s^2$ próbki, tj.

$$Kurtoza = \frac{m_4}{m_2^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{(n-1)s^4}$$

Do porównywania innych rozkładów z rozkładem normalnym (Gaussa) o kurtozie 3, stosuje się współczynnik *excesu kurtozy* $K = Kurtoza$ - 3. Dla próbek o $n \geq 4$ w [3] podano dla K wzór

$$K = \frac{n(n+1)\sum_{i=1}^{n}(x_i - \bar{x})^4}{(n-1)(n-2)(n-3)s^4} - \frac{3(n-1)^2}{(n-2)(n-3)s^4}$$

oraz zdefiniowano standaryzowany współczynnik kurt ozy

$$SK(n) = K\left(\frac{24}{n}\right)^{-\frac{1}{2}}$$

6. Modelowanie skośności i kurtozy małych próbek metodą MonteCarlo

W literaturze nie znaleziono szczegółowych informacji o rozkładach skośności g i ekscesu kurtozy K dla małych próbek o liczbie elementów: $3 < n < 25$. Autorzy postanowili te rozkłady wyznaczyć metodą symulacji MonteCarlo i przeanalizować ich właściwości statystyczne dla różnych liczb elementów n próbki.

Opiszemy bliżej eksperyment symulacyjny przeprowadzony dla współczynnika skośności Pearsona. Z numerycznego modelu badanej populacji X pobierano wielokrotnie losowo próbki o określonej liczbie elementów n i dla każdej z tych próbek obliczono współczynnik skośności g wg wzoru wynikającego z (3)

$$g = \frac{m_3}{s^3} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n\left(\sqrt{\sum_{i=1}^{n}\frac{(x_i - \bar{x})^2}{n-1}}\right)^3} \tag{9}$$

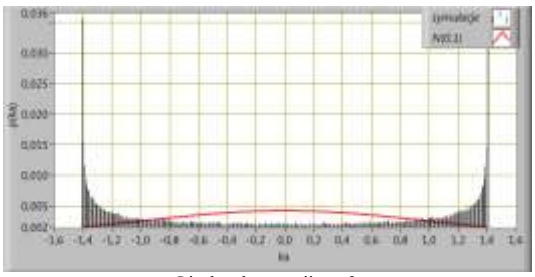
gdzie: m_3 - moment centralny trzeciego rzędu dla próbki, s - jej nieobciążone odchylenie standardowe.

Badania w pierwszej kolejności dotyczyły próbek z populacji generalnej o rozkładzie normalnym. Wybrano z niej losowo serie po N=100 000 próbek o liczbie elementów n=(3, 4, 5, ... 50). Dla każdego n wyznaczono histogramy współczynnika skośności g. Niektóre z histogramów przedstawiono na rysunku 2. Wraz ze wzrostem n próbki kształt histogramu g zbliża się do funkcji Gaussa. Na histogramach narysowano też pdf-y ich najlepszych rozkładów Gaussa.

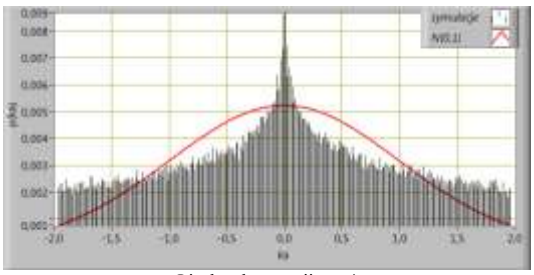
Ponadto dla zbioru próbek o liczbach elementów $n \geq 3$ obliczono odchylenia standardowe współczynnika skośności $s(g)$, średnią wartość jego modułu $\overline{|g|}$, wariancję $D\overline{|g|}$ i standardowe odchylenie $\Delta|g|$ od tej wartości oraz współczynnik kurtozy K. Wyniki tych obliczeń, otrzymane przy stosowaniu różnych wzorów, podano w postaci wykresów na kolejnych rysunkach.

Na rysunku 3 przedstawiono przebiegi obliczonych parametrów skośności w funkcji liczby elementów n próbki po wygładzeniu metodą najmniejszych kwadratów oraz odchylenie standardowe $s(g) = \sqrt{D(g)}$,tj. wg wzoru (8) Smirnowa.

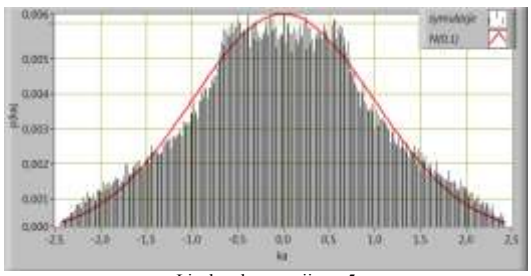
Rys. 4 podaje w powiększeniu zależność średniego modułu skośności $\overline{|g|}$ i jego odchylenia standardowego $\sqrt{D\overline{|g|}}$ dla początkowych liczb elementów n. Osiągają one maksimum około n = 6 i następnie maleją ze wzrostem liczby n elementów próbki.



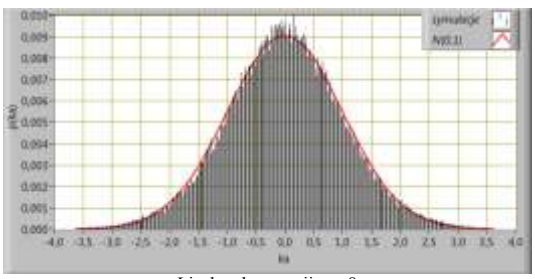
Liczba obserwacji $n = 3$



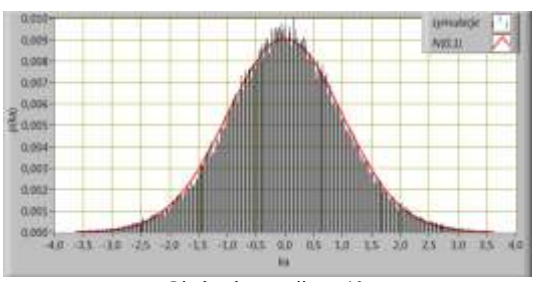
Liczba obserwacji $n = 4$



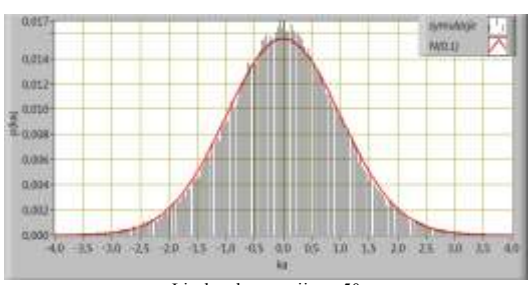
Liczba obserwacji $n = 5$



Liczba obserwacji $n = 8$



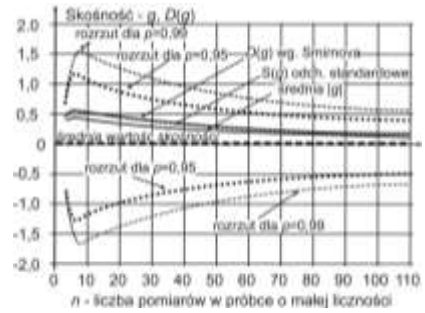
Liczba obserwacji $n = 10$



Liczba obserwacji $n = 50$

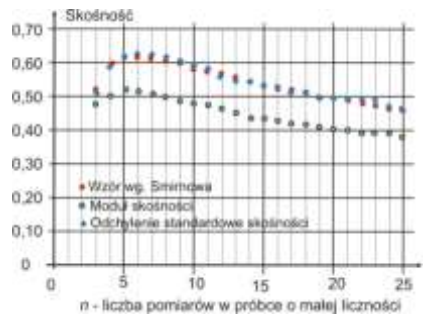
Rys. 2. Przykłady rozkładów współczynnika skośności g dla próbek o małej liczbie elementów n

Fig. 2. Examples of the skewness coefficient g distributions of samples with low number of elements n



Rys. 3. Parametry skośności g w funkcji liczby elementów n próbek z populacji Gaussa

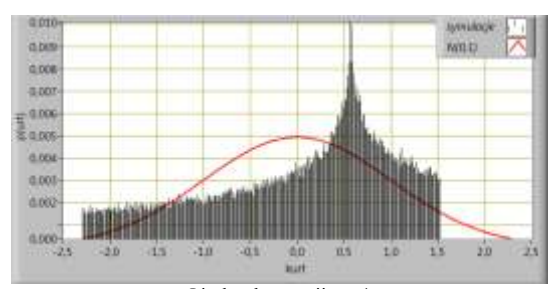
Fig. 3. Parameters of skewness g of n element small samples of Gauss population



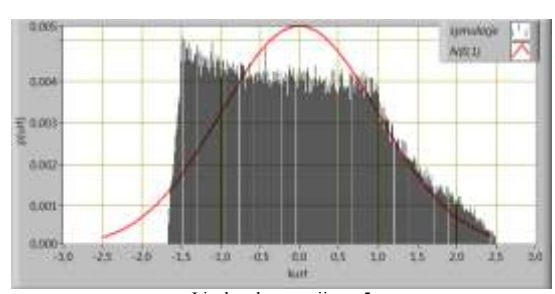
Rys. 4. Parametry skośności g dla próbek o bardzo małych n <25

Fig. 4. Parameters of skewness g of the very small samples (n<25)

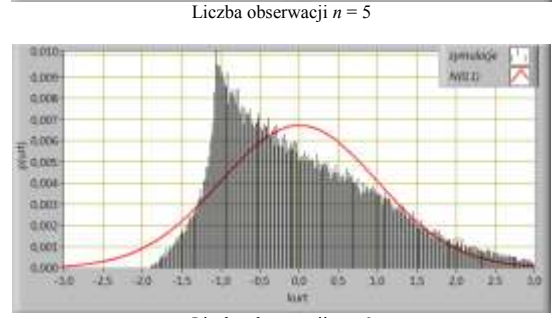
Dla zbiorów próbek pobranych 100 000 razy z populacji o rozkładzie normalnym, korzystając z symulacji metodą MC, wyznaczono też rozkłady kurtozy (rys. 5) oraz przebiegi ich podstawowych parametrów statystycznych w funkcji liczby elementów n próbki (rys. 6 i rys. 7).



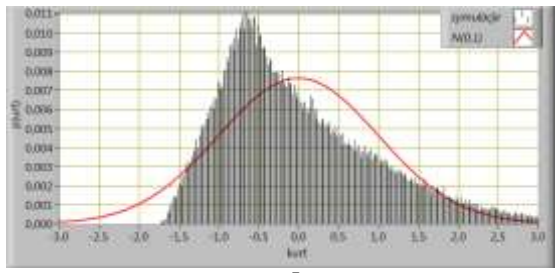
Liczba obserwacji $n = 4$



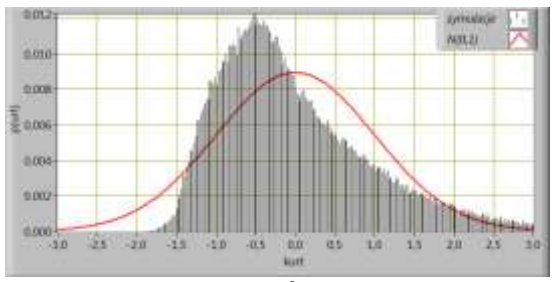
Liczba obserwacji $n = 5$



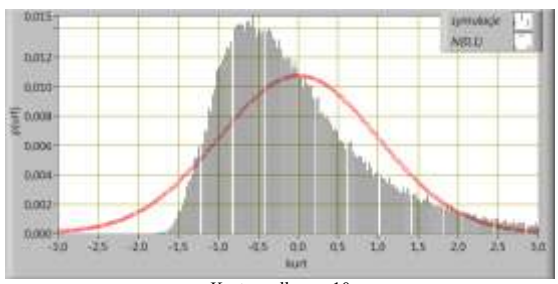
Liczba obserwacji $n = 6$



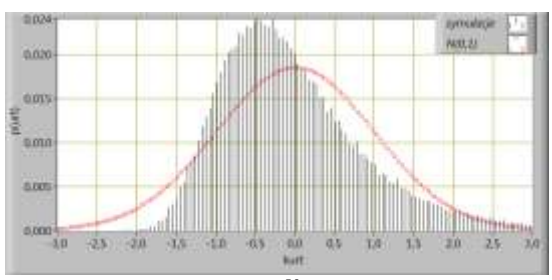
n = 7



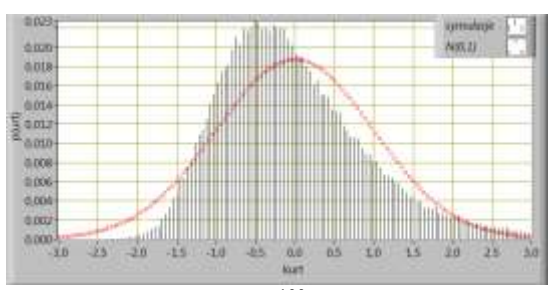
n = 8



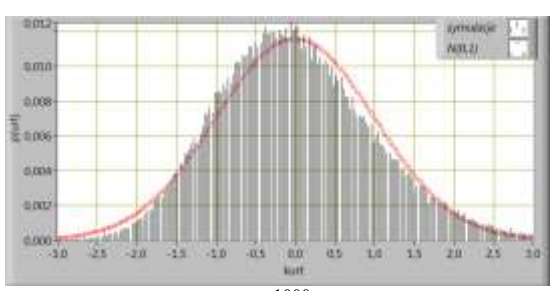
Kurtoza dla n = 10



n = 50

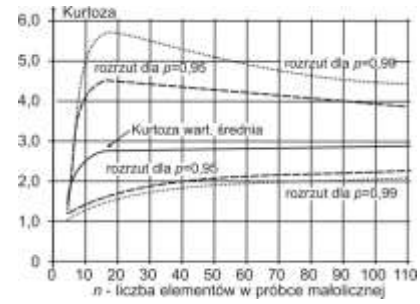


n = 100

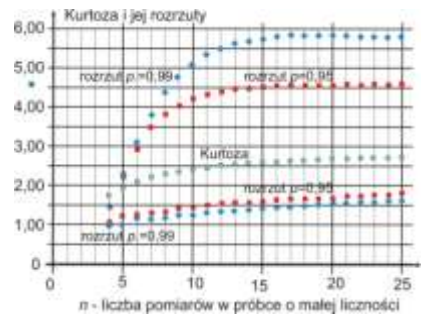


n = 1000

Rys. 5. Histogramy kurtozy kilku próbek o liczności n z populacji o rozkładzie normalnym (dla serii N=100 000)
Fig. 5. Histograms of kurtosis for samples of observations n is taken from population of Normal distribution (N=100 000 trials)



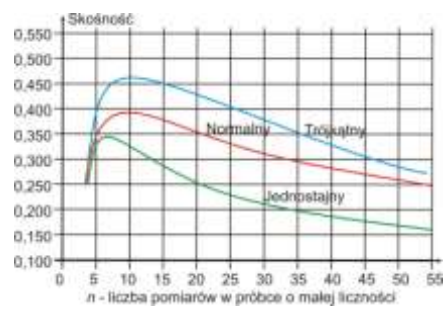
Rys. 6. Parametry statystyczne kurtozy w funkcji liczby elementów n próbek
Fig. 6. Statistical parameters of kurtosis as function of number n of elements of sample from Gauss population



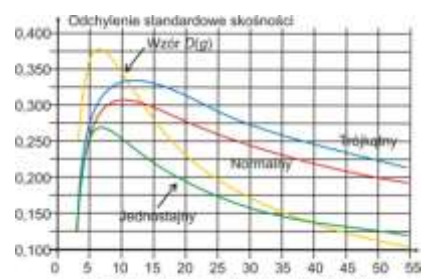
Rys. 7. Parametry statystyczne histogramów kurtozy dla bardzo małych próbek N(0,1)
Fig. 7. Statistical parameters of kurtozis histograms of very small samples of N(0,1)

7. Porównanie skośności próbek z rozkładu normalnego, jednostajnego i trójkątnego

Na rys. 8, dla próbek z populacji o trzech rozkładach: normalnym, jednostajnym i trójkątnym, podano zależności od n średniego modułu skośności $k_a \equiv \overline{|g|}$ oraz wariancję D wg Smirnowa -wzór (8), a na rys. 9 - ich odchylenia standardowe STD.



Rys. 8. Skośności g próbek o n elementach z trzech różnych populacji (N = 1 000 000)
Fig. 8. Skewness g of n element samples from 3diferent populations (N = 1 000 000)



Rys. 9. Wariancja D(g) i odchylenie standardowe STD skośności g próbek o n elementach z populacji o trzech różnych rozkładach pdf (1 000 000 powtórzeń)
Fig. 9. Variance and standard deviation STD of the skewness g of n element samples from population of three different pdf. (values from 1 000 000 trials)

Z zależności tych wynika, że próbki z populacji o rozkładzie jednostajnym mają mniejszą skośność niż dla normalnego, a o rozkładzie trójkątnym - większą. Liczba obserwacji n_{max} dla maksimum średniej skośności jest również największa dla rozkładu trójkątnego.

8. Wnioski i podsumowanie

Dla zbioru małych próbek z populacji o rozkładzie normalnym wartość średnia modułu współczynnika skośności odbiega znacznie od zera, a średnia kurtoza - od wartości 3 dla całej populacji. Skośność osiąga maksimum dla liczby n elementów próbki około 6, a kurtoza - dla n ok. 20. Następnie, wraz ze wzrostem n oba parametry bardzo powoli maleją do wartości 0 i 3 dla populacji. Kształty rozkładów obu parametrów dla małych n również zdecydowanie różnią się od krzywej Gaussa.

Skośność i kurtoza nie są dotychczas uwzględniane przy wyznaczaniu zarówno wartości średniej próbki jako estymatora wartości mensurandu, jak i jego niepewności jako miary oceny dokładności (precyzji) wyniku pomiaru.

Dla próbek z populacji o rozkładach niegaussowskich, np. równomiernego, trapezowego i trójkątnego,wartość średnia nie jest najlepszym estymatorem mensurandu [5 - 7], a przebiegi skośności i kurtozy w funkcji n są inne niż rozkładu normalnego.

Zamierza się jeszcze zbadać czy rozmnażając dane małych próbek z kilku różnych populacji metodą resamplingu i uwzględniając ich skośność i kurtozę można wyznaczyć lepszy estymator wartości mierzonej niż średnia wg GUM i jego niepewność rozszerzoną, w tym metodą MC podano w Suplemencie 1 [8].

9. Literatura

[1] Klonecki W.: Statystyka dla inżynierów PWN Warszawa (1999).

[2] Guide to the Expression of Uncertainty in Measurement GUM. ISO/IEC/OIML/BIPM, first ed. 1992, last ed. BIPM JCGM 100 (2008).

[3] Dobosz M.: Statystyczna analiza wyników badań Exit Warszawa (2004).

[4] Bolshev L. N., Smirnov N. V.: Tablicy matematicheskoj statistiki. Nauka (1983) s. 416 (in Russian).

[5] Warsza Z. L.: Jednoelementowe estymatory wartości menzurandu próbek o kilku niegaussowskich rozkładach prawdopodobieństwa – przegląd. Pomiary Automatyka Kontrola, nr 1 (2011) s.101 - 104.

[6] Warsza Z. L.: Dwuelementowe estymatory wartości menzurandu próbek pomiarowych o trapezowych rozkładach prawdopodobieństwa -przegląd prac. Pomiary Automatyka Kontrola nr 1 (2011) s.105 -108.

[7] Kubisa S. Warsza Z: Środek rozstępu jako estymator menzurandu dla próbek z populacji o rozkładzie jednostajnym i płasko-normalnym. PAK vol. 60, nr 6 (2014) s. 398 – 401.

[8] Guide to the Expression of Uncertainty in Measurement (GUM), OIML ed. 2008 Supplement: Propagation of distributions using a Monte Carlo method, G 1 – 101, (2007).

otrzymano / received: 18.09.2014
przyjęto do druku / accepted: 03.11.2014 *artykuł recenzowany / revised paper*