Mykhaylo DOROZHOVETS
POLITECHNIKA RZESZOWSKA, KATEDRA METROLOGII I SYSTEMÓW DIAGNOSTYCZNYCH
ul. W. Pola 2, 35-959 Rzeszów

Badania korelacji między podstawowymi estymatorami parametru położenia dla serii obserwacji nieskorelowanych

Prof. dr hab. inż. Mykhaylo DOROZHOVETS

Jest absolwentem (1975) Katedry Techniki Informacyjno-Pomiarowej Politechniki Lwowskiej, w 2001 r. obronił pracę habilitacyjną. Obecnie jest zatrudniony na stanowisku profesora zwyczajnego w Katedrze Metrologii i Systemów Diagnostycznych Politechniki Rzeszowskiej. W pracy naukowo-badawczej zajmuje się analizą i oceną niepewności wyników pomiarów, zagadnieniami pomiarów tomograficznych oraz problemami przetwarzania sygnałów pomiarowych. Opublikował ponad 240 prac naukowych.

e-mail: michdor@prz.edu.pl

Streszczenie

W pracy przedstawiono wyniki badań korelacji pomiędzy wartością średnią, medianą oraz środkiem rozstępu nieskorelowanych wyników obserwacji o wybranych rozkładach prawdopodobieństwa: Laplace'a, normalnym, trójkątnym, trapezowym, jednostajnym oraz arksinusoidalnym. Stwierdzono, że istnieje silna korelacja pomiędzy wartością średnią a medianą, mediana i środek rozstępu są najmniej skorelowane, a korelacja pomiędzy średnią i środkiem rozstępu przyjmuje wartości pośrednie. Przy wzroście liczby obserwacji korelacja pomiędzy wartością średnią a medianą stabilizuje się, natomiast korelacja pomiędzy środkiem rozstępu i wartością średnią oraz medianą monotoniczne zmniejsza się.

Słowa kluczowe: korelacja, wartość średnia, mediana, środek rozstępu.

Investigations of correlation between the main location parameter estimators of random uncorrelated observations

Abstract

The results of studies of the correlation between the main location parameter estimators (mean, median and midrange) of uncorrelated random observations are presented. Analytical calculations of the correlation coefficients are based on preliminary determination of the location parameter joint distributions. The joint distributions of the median and midrange are described by simplest formula (7), based on the distribution of order statistics (6). But analytical calculations of the joint distribution of other pairs of position parameters are very difficult and can only be realized by numerical procedures. Formulas (11) - (15) for determining the asymptotical values of all correlation coefficients for a large numbers of observations are presented. The main studies of the correlation coefficients for the number observation from $n = 2$ to $n = 100$ are realised by the Monte Carlo method. The results of simulation investigations are shown in Fig. 1. The median and mean are most correlated, and the asymptotical value of the correlation coefficient is equal to the inverse value of a form factor of the sample probability density function (PDF). The value of the correlation coefficient between the median and midrange does not practically depend on the type of PDF and decreases approximately proportionally to the square root of the number of observations (13, Fig. 1). The value of the correlation coefficient between the mean and midrange also decreases monotonically with an increase in the number of observations, but the rate of decrease depends on the amplitude factor of a sample (15, Fig. 1).

Keywords: correlation, mean, median, midrange.

1. Wstęp

W praktyce opracowania losowych obserwacji zakłada się, że wyniki obserwacji nie są skorelowane pomiędzy sobą oraz, że rozkład prawdopodobieństwa populacji (RPP), z której zostały pobrane obserwacje jest znany *a priori*. Z pośród różnych estymatorów parametru położenia najczęściej wykorzystywane są wartość średnia $\bar{x}$, mediana x_{med} oraz środek rozstępu x_M, które są wyznaczane na podstawie zbioru uporządkowanych obserwacji $x_{(1)}, x_{(2)}, .., x_{(i)}, .., x_{(n-1)}, x_{(n)}$. Rzadziej wykorzystuje się wartości średnie innych statystyk pozycyjnych.

Dla każdego RPP $p(x)$, z której zostały pobrane obserwacje, można obliczyć najbardziej efektywny (z najmniejszą standardową niepewnością) estymator parametru położenia jako ocenę wyniku pomiaru. Przykładowo wartość średnia jest najlepszym parametrem obserwacji o rozkładzie normalnym, natomiast środek rozstępu jest najlepszym parametrem obserwacji o rozkładzie jednostajnym [1].

Bardzo często, jeśli RPP obserwacji różni się od pewnego modelu, na przykład normalnego, wykorzystują się uproszczone metody obliczania wyniku pomiaru, bazujące na obliczeniu średniej ważonej kilku prostych estymatorów położenia [2, 3]. Przykładowo, w [2] zaproponowano wykorzystanie estymatorów złożonych zdefiniowanych jako średnie ważone z wartości średniej, mediany i środku rozstępu:

$$\hat{x}_3 = k_1\,\bar{x} + k_2\,x_{med} + k_3\,x_M\,, \tag{1a}$$

$$\text{lub} \quad \hat{x}_2 = k_1\,\bar{x} + k_2\,x_M = k_1\,\bar{x} + (1-k_1)\,x_M\,, \tag{1b}$$

gdzie k_1, k_2, k_3 są współczynnikami wagowymi, dla których ma być spełniony warunek unormowania: $k_1 + k_2 + k_3 = 1$. Przy opracowaniu wyników obserwacji o trapezowym rozkładzie prawdopodobieństwa o nieznanych parametrach oraz z innymi rozkładami o ograniczonych zakresach zmian wartości obserwacji, zalecane są estymatory dwuelementowe $\hat{x}_2$ (1b) [2, 3]. Oprócz warunku unormowania, wartości współczynników wagowych we wzorach (1a) i (1b) wynikają z warunku minimalizacji niepewności standardowej estymatorów złożonych $\hat{x}_2$ lub $\hat{x}_3$. Dla standardowych niepewności wartości średniej $u(\bar{x})$, mediany $u(x_{med})$ oraz środku rozstępu $u(x_M)$ złożona standardowa niepewność wyniku (1a) wynosi:

$$\begin{aligned} u_c(\hat{x}_3) = [\; & k_1^2\,u^2(\bar{x}) + k_2^2\,u^2(x_{med}) + k_3^2\,u^2(x_M) \\ & + 2\,\rho(\bar{x}, x_{med})\,k_1\,k_2\,u(\bar{x})\,u(x_{med}) + \\ & + 2\,\rho(\bar{x}, x_M)\,k_1\,k_3\,u(\bar{x})\,u(x_M) + \\ & + 2\,\rho(x_{med}, x_M)\,k_2\,k_3\,u(x_{med})\,u(x_M)\;]^{\frac{1}{2}}, \end{aligned} \tag{2}$$

gdzie $\rho(\bar{x}, x_{med})$, $\rho(\bar{x}, x_M)$, $\rho(x_{med}, x_M)$ to współczynniki korelacji pomiędzy odpowiednio: średnią i medianą, średnią i środkiem rozstępu oraz medianą i środkiem rozstępu.

Zarówno średnia jak i mediana oraz środek rozstępu wyznaczane są na podstawie tych samych obserwacji. Dlatego należy oczekiwać korelacji pomiędzy tymi parametrami, tym większej, im mniejsza jest liczba zarejestrowanych obserwacji. (W przypadku tylko dwóch obserwacji ($n = 2$), średnia, mediana i środek rozstępu są tożsamościowo równe, to znaczy, że współczynniki korelacji między nimi przyjmują wartość 1). Nieuwzględnienie korelacji wzajemnej przy obliczeniu wartości współczynników wagowych może istotnie zniekształcić zarówno samą ocenę wyniku pomiaru, jak i ocenę jego standardowej niepewności (2).

Celem przedstawionych w pracy badań jest zbadanie zależności współczynników korelacji pomiędzy średnią i medianą, średnią i środkiem rozstępu oraz medianą i środkiem rozstępu od liczby obserwacji dla wybranych modeli rozkładów prawdopodobieństwa populacji.

2. Badania analityczne

Teoretyczne wyznaczanie współczynnika korelacji wzajemnej pomiędzy zadanymi estymatorami parametru położenia obserwacji dla dowolnego RPP $p(x)$ opiera się na znajomości rozkładów $p_1(\bar{x})$, $p_2(x_{med})$, $p_2(x_M)$ oraz łącznych rozkładów prawdopodobieństwa $p_{1,2}(\bar{x},x_{med})$, $p_{1,3}(\bar{x},x_M)$, $p_{2,3}(x_{med},x_M)$ tych estymatorów. W ogólnej postaci rozkłady te oznaczymy jako $p_j(y_j)$, $p_{j,k}(y_j,y_k)$, $j,k=1,2,3;\ k\neq j$, gdzie $y_1=\bar{x}$ $y_2=x_{med}$ $y_3=x_M$. Poszukiwane współczynniki korelacji obliczane są jako:

$$\rho(y_j,y_k)=\frac{R(y_j,y_k)}{\sigma_j\cdot\sigma_k}, \tag{3}$$

gdzie $R(y_j,y_k)$ oznacza kowariancję pomiędzy odpowiednimi parametrami położenia, wyznaczaną z zależności:

$$R(y_j,y_k)=\int_{y_j}\int_{x_k} y_j y_k p_{j,k}(y_j,y_k)dy_j dy_k-\mu_j\mu_k. \tag{4}$$

Symbole $\mu_j=E[y_j]$ oraz $\sigma_j^2=\int_{y_j}(y_j^2-\mu_j)^2 p_j(y_j)dy_j$ oznaczają wartości oczekiwane oraz wariancji użytych estymatorów.

Następnie, analizowane są obserwacje pobrane z populacji o unormowanych rozkładach $p(x)$ symetrycznych względem jednakowych wartości oczekiwanych $\mu=0$ i jednakowych wariancji $\sigma^2=1$. Wtedy wartości oczekiwane użytych estymatorów równe są wartościom oczekiwanym populacji: $E[\bar{x}]=E[x_{med}]=E[x_M]=\mu=0$, a wariancja średniej wyznaczana jest jako $\sigma_1^2=\sigma_{\bar{x}}^2=\sigma^2/n=1/n$.

Łączny rozkład mediany i środku rozstępu. Najłatwiejsze jest wyprowadzenie wyrażenia dla łącznego rozkładu mediany i środku rozstępu. Dla zadanego RPP $p(x)$ (dystrybuanta $F(x)$) przy nieparzystej liczbie obserwacji n łączny rozkład prawdopodobieństwa mediany $x_{med}=x_{((n+1)/2)}$ oraz środku rozstępu $x_M=(x_{(1)}+x_{(n)})/2$ można wyznaczyć na podstawie łącznego rozkładu statystyk pozycyjnych [4], mianowicie centralnej (x_{med}) i skrajnych ($x_{(1)}$, $x_{(n)}$) wartości obserwacji:

$$p_{1,\frac{n+1}{2},n}(x_{(1)},x_{med},x_{(n)})=\frac{n!}{((n-3/2)!)^2}p(x_{med})p(x_{(1)})p(x_{(n)})\times \times\left[(F(x_{med})-F(x_{(1)}))(F(x_{(n)})-F(x_{med}))\right]^{\frac{n-3}{2}}, \tag{5}$$

Ponieważ $x_{(n)}=2x_M-x_{(1)}$ łączny rozkład mediany i środku rozstępu jest wyznaczany poprzez całkowania zależności (5)

$$p_{2,3}(x_{med},x_M)=2\int_{-\infty}^{x_z} p_{1,\frac{n+1}{2},n}(x_{(1)},x_{med},2x_M-x_{(1)})dx_{(1)}, \tag{6}$$

gdzie $x_z=\begin{cases}2x_M-x_{med}, & x_{med}\geq x_M,\\ x_{med}, & x_{med}<x_M.\end{cases}$

W przypadku parzystej liczby obserwacji mediana jest wyznaczana jako średnia dwóch obserwacji centralnych. Dlatego przy wyprowadzeniu wzoru dla łącznego rozkładu mediany i środka rozstępu (podobnie łącznego rozkładu mediany i wartości średniej) należy dodatkowo obliczyć rozkład mediany (przyjmując do uwagi, że $x_{(n/2+1)}=2x_{med}-x_{(n/2)}$):

$$p_{med}(x_{med})=2\int_L p(x_{(n/2)})p(2x_{med}-x_{(n/2)})dx_{(n/2)}, \tag{7}$$

gdzie granice całkowania $L: x_{(n/2-1)}<x_{(n/2)},\ x_{(n/2+1)}<x_{(n/2+2)}$.

Wyprowadzenie wzorów analitycznych na łączne rozkłady dwóch innych par parametru położenia jest istotnie skomplikowanym zagadnieniem.

Łączny rozkład średniej i mediany. Ten rozkład może być wyznaczony w dwóch etapach. Najpierw wyznaczany łączny rozkład $p3(\bar{x}_1,x_{med},\bar{x}_2)$ mediany x_{med} oraz 2-ch wartości średnich dla $(n-1)/2$ obserwacji mniejszych od mediany: $\bar{x}_1=\frac{2}{n-1}\sum_{i=1}^{(n-1)/2}x_{(i)}$ oraz większych od mediany obserwacji: $\bar{x}_2=\frac{2}{n-1}\sum_{i=(n+3)/2}^{n}x_{(i)}$. Na drugim etapie łączny rozkład $p_{1,2}(\bar{x},x_{med})$ wartości średniej $\bar{x}=\frac{(n-1)(\bar{x}_1+\bar{x}_2)+x_{med}}{n}$ i mediany wyznaczany jest na drodze całkowania łącznego rozkładu $p3(\bar{x}_1,x_{med},\bar{x}_2)$ przy podstawieniu $\bar{x}_2=2\bar{x}_{n-1}-\bar{x}_1$ (gdzie $\bar{x}_{n-1}=(\bar{x}_1+\bar{x}_2)/2$ jest wartością średnią tych średnich):

$$p_{1,2}(\bar{x}_{n-1},x_{med})=\frac{n\cdot n!}{n-1}p(x_{med})\int_{-\infty}^{x_z} p3(\bar{x}_1,x_{med},2\bar{x}_{n-1}-\bar{x}_1)d\bar{x}_1, \tag{8}$$

przy czym granica całkowania x_z wyznaczana jest z warunku:

$$x_z=\begin{cases}2\bar{x}_{n-1}-x_{med}, & x_{med}\geq\bar{x}_{n-1},\\ x_{med}, & x_{med}<\bar{x}_{n-1}.\end{cases}$$

Następnie, do otrzymanego wzoru zamiast $\bar{x}_{n-1}$ podstawiano jego wartość: $\bar{x}_{n-1}=\frac{n\bar{x}_n-x_{med}}{n-1}$.

Obliczanie rozkładu wartości każdej średniej $\bar{x}_1$ oraz $\bar{x}_2$ wiąże się z obliczaniem $(n-1)/2-1=(n-3)/2$ całek. Po uwzględnieniu jeszcze jednej całki we wzorze (7) do obliczania łącznego rozkładu średniej i mediany potrzebne jest obliczanie $n-2$ całek związanych z rozkładami prawdopodobieństwa. Z tego wynika, że do wyznaczania współczynnika korelacji z wzorów (3), (4) należy obliczyć n całek.

Łączny rozkład wartości średniej i środku rozstępu. Ten rozkład, podobnie jak poprzednio, też można wyznaczyć dwuetapowo, na pierwszym etapie wyznaczany jest łączny rozkład $p_{n-2}(x_{(1)},\bar{x}_{n-2},x_{(n)})$ obserwacji skrajnych $x_{(1)}$ i $x_{(n)}$ oraz wartości średniej $\bar{x}_{n-2}=\frac{1}{n-2}\sum_{i=2}^{n-1}x_{(i)}$ z $n-2$ pozostałych obserwacji. Dalej realizuje się całkowanie łącznego rozkładu $p_{n-2}(x_{(1)},\bar{x}_{n-2},x_{(n)})$ po wprowadzeniu zamiany $x_{(n)}=2x_M-x_{(1)}$:

$$p_{1,3}(\bar{x}_{n-2},x_M)=\frac{2n\cdot n!}{n-2}\int_{-\infty}^{x_z} p_{1,\,n-2,n}(x_{(1)},\bar{x}_{n-2},2x_M-x_{(1)})dx_{(1)}, \tag{9}$$

gdzie $x_z=\begin{cases}2x_M-\bar{x}_{n-2}, & \bar{x}_{n-2}\geq x_M,\\ \bar{x}_{n-2}, & \bar{x}_{n-2}<x_M.\end{cases}$

Poszukiwany łączny rozkład uzyskuje się po podstawieniu do otrzymanego wzoru zamiast $\bar{x}_{n-2}$ wartości $\bar{x}_{n-2} = (n\bar{x} - 2x_M)/(n-2)$.

Obliczanie rozkładu wartości średniej $\bar{x}_{n-2}$ wymaga obliczenia $n-3$ całek. Jak i w poprzednim przypadku, do wyznaczania współczynnika korelacji należy obliczyć też n całek.

W ogólnym przypadku dla dowolnego $p(x)$ bezpośrednie analitycznie obliczanie wymienionych wyżej n całek praktycznie nie jest możliwie. W tym celu można zastosować metody numeryczne, jednak i w tym przypadku wyznaczanie współczynników korelacji $\rho(\bar{x}, x_{med})$ oraz $\rho(\bar{x}, x_M)$ na podstawie łącznych rozkładów (7) i (9) tych parametrów jest bardzo czasochłonne.

3. Wartości asymptotyczne dla dużych *n*

Stosunkowo łatwo można otrzymać wzory dla asymptotycznych zależności współczynników korelacji przy $n \to \infty$. Dla wyprowadzenia tych wzorów wykorzystamy właściwości parametrów statystyk pozycyjnych [4]. Widomo jest [4], że kwantyle $x_{(k1)}$ i $x_{(k2)}$ z próby prostej, pobranej z populacji o rozkładzie $p(x)$ (dystrybuanta $F(x)$), przy $n \to \infty$ mają rozkład asymptotycznie normalny o parametrach: $m_1 = x_{(\lambda,1)}$, $m_2 = x_{(\lambda,2)}$, $\sigma_1^2 = \dfrac{\lambda_1(1-\lambda_1)}{n\left(p\left(x_{(\lambda_1)}\right)\right)^2}$, $\sigma_2^2 = \dfrac{\lambda_2(1-\lambda_2)}{n\left(p\left(x_{(\lambda_2)}\right)\right)^2}$, $\rho_{1,2} = \sqrt{\dfrac{\lambda_1(1-\lambda_2)}{\lambda_2(1-\lambda_1)}}$, gdzie $k_1 = [n\lambda_1]+1$, $k_2 = [n\lambda_2]+1$, x_{λ_1}, x_{λ_2} - kwantyle zmiennej losowej ($0 < \lambda_1 < \lambda_2 < 1$). Dla ciągłego rozkładu $p(x)$ kwantyl rzędu λ jest wartością $x_{(\lambda)}$ spełniającą równość $\lambda = F\left(x_{(\lambda)}\right)$, lub inaczej $x_\lambda = qF(\lambda)$ (gdzie $qF(z)$ - odwrotna dystrybuanta). Wtedy dla $1 \le i \le n$ oraz $\lambda_i = i/(n+1)$ mamy: $x_{(\lambda,i)} = qF\left(i/(n+1)\right)$ i dalej: $\sigma_i^2 = \dfrac{i\,(n+1-i)}{n\,(n+1)^2\left(p\left(x_{(\lambda,i)}\right)\right)^2}$, $\sigma_j^2 = \dfrac{j\,(n+1-j)}{n\,(n+1)^2\left(p\left(x_{(\lambda,j)}\right)\right)^2}$, $\rho_{i,j} = \sqrt{\dfrac{i\,(n+1-j)}{j\,(n+1-i)}}$, $1 \le i < j \le n$.

Wykorzystując te zależności we wzorze (3) otrzymujemy asymptotyczne wartości współczynników korelacji dla dużych *n*:

$$\rho_a(\bar{x}, x_{med}) \approx \frac{1}{n\,(n+1)}\left[\sum_{i=1}^{n} \frac{T(i)}{p\left(x_{(\lambda,i)}\right)}\right], \tag{11}$$

$$\rho_a(\bar{x}, x_M) \approx \sqrt{\frac{2}{n+1}} \cdot \frac{1}{2n} \sum_{i=1}^{n} \frac{1}{p\left(x_{(\lambda,i)}\right)}, \tag{12}$$

$$\rho_a(\bar{x}, x_M) \approx \sqrt{\frac{2}{n+1}}\,. \tag{13}$$

gdzie $T(i) = \begin{cases} i, & 1 \le i \le [n/2], \\ n+1-i, & [n/2]+1 \le i \le n. \end{cases}$ – funkcja trójkątna.

Przy dużych n $p\left[x_{(\lambda,i)}\right] \cdot \left(x_{(\lambda,i+1)} - x_{(\lambda,i)}\right) \approx \int_{x_{(\lambda,i)}}^{x_{(\lambda,i+1)}} p(x)dx = 1/(n+1)$

dlatego $\sum_{i=1}^{n} \dfrac{T(i)}{p\left(x_{(\lambda,i)}\right)} \approx (n+1)\left[\sum_{i=n/2+1}^{n} x_{(\lambda,i)} - \sum_{i=1}^{n/2} x_{(\lambda,i)}\right] = n\,(n+1)\overline{\left|x_{(\lambda)}\right|}$

oraz $\sum_{i=1}^{n} \dfrac{1}{p\left(x_{(\lambda,i)}\right)} \approx 2n \cdot x_a$, gdzie $\overline{\left|x_{(\lambda)}\right|} = \dfrac{1}{n}\sum_{i=1}^{n}\left|x_{(\lambda,i)} - x_{(1/2)}\right|$ jest wartością średnią modułów odchyleń kwantyli od mediany, a $x_a = \left|x_{(n)} - \mu\right|$ jest odchyleniem maksymalnego kwantyla od wartości oczekiwanej.

Po podstawieniu tych wartości do wzorów (11) oraz (12) otrzymujemy wartości asymptotyczne współczynników korelacji:

$$\rho_a(\bar{x}, x_{med}) \approx \overline{\left|x_{(\lambda)}\right|}, \tag{14}$$

$$\rho_a(\bar{x}, x_M) \approx \sqrt{\frac{2}{n+1}} \cdot x_a\,. \tag{15}$$

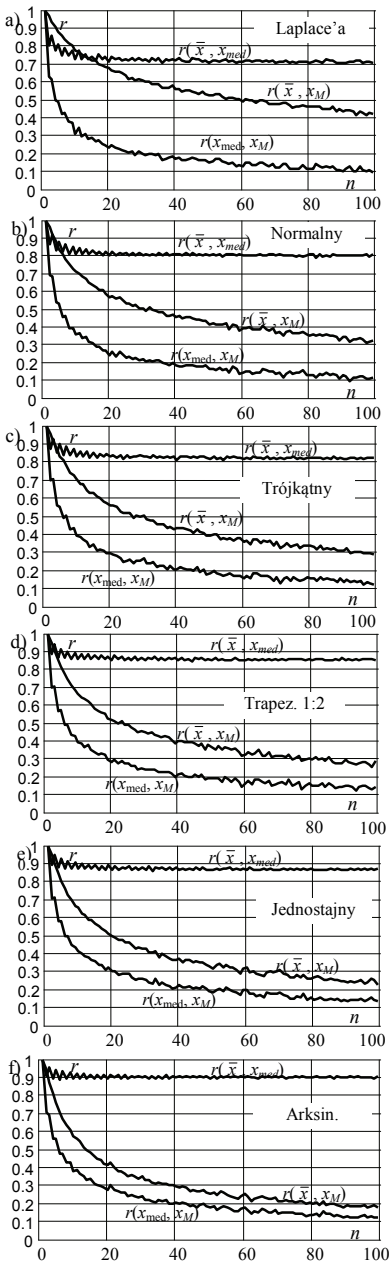
Przy dowolnych wartościach parametrów położenia $\mu \neq 0$ i wariancji $\sigma^2 \neq 1$ w mianownikach tych wzorów ma być wartość standardowego odchylenia σ. Ponieważ stosunek $\sigma \Big/ \overline{\left|x_{(\lambda)}\right|} = k_{ksz}$ jest współczynnikiem kształtu sygnału (przebiegu losowego), dlatego asymptotyczna współczynnika korelacji $\rho_a(\bar{x}, x_{med})$ we wzorze (14) jest odwrotnością współczynnika kształtu. Dla zbadanych rozkładów wartości asymptotyczne $\rho_a(\bar{x}, x_{med})$ następujące: Laplace'a: $1/\sqrt{2} \approx 0{,}707$, normalnego: $\sqrt{2/\pi} \approx 0{,}798$, trójkątnego: $2/\sqrt{6} \approx 0{,}816$, trapezowego ze stosunkiem podstaw $m = 1:2$: $\left(1+m+m^2\right)\sqrt{6/\left(1+m^2\right)}\Big/3(1+m) \approx 0{,}852$, jednostajnego $\sqrt{3}/2 \approx 0{,}866$ oraz arksinusoidalnego $2\sqrt{2}/\pi \approx 0{,}900$. Z analizy uzyskanych danych wynika, że w przypadku rozkładów o wyraźnie ograniczonych wartościach obserwacji, takich jak arksinusoidalny oraz jednostajny, asymptotyczne wartości współczynnika $\rho_a(\bar{x}, x_{med})$ są największe, odpowiednio $\approx 0{,}90$ oraz $\approx 0{,}87$. W przypadku rozkładów Laplace'a oraz normalnego (teoretycznie nieograniczone wartości obserwacji), asymptotyczne wartości tego współczynnika są mniejsze i wynoszą $\approx 0{,}71$ oraz $\approx 0{,}80$. W przypadku rozkładu trójkątnego i trapezowego poziom wartości $\rho_a(\bar{x}, x_{med})$ przybiera wartości pośrednie, odpowiednio $\approx 0{,}82$ oraz $\approx 0{,}85$.

Stosunek $x_a/\sigma = qF(n/n+1)/\sigma = k_a$ we wzorze (15) jest współczynnikiem amplitudy, wartość którego bezpośrednio decyduje o wartości współczynnika $\rho(\bar{x}, x_M)$ i jego zmniejszeniu się ze wzrostem liczby obserwacji. Dla zbadanych rozkładów przy $n = 100$ wartości k_a są równe odpowiednio: $\sqrt{2}\ln(n+1/2) \approx 2{,}773$; $qFnorm(n/(n+1)) \approx 2{,}330$; $\sqrt{6} \approx 2{,}449$; $\sqrt{6/\left(1+m^2\right)} \approx 2{,}191$; $\sqrt{3} \approx 1{,}732$; $\sqrt{2} \approx 1{,}414$.

4. Wyniki badań symulacyjnych oraz ich analiza

Aby uniknąć operacji matematycznych ze złożonymi wyrażeniami, badanie korelacji wzajemnej pomiędzy podstawowymi parametrami położenia przy $n = 2 \div 100$ zostały przeprowadzone metodą Monte Carlo (MC). Liczba eksperymentów statystycznych w metodzie MC wynosiła $M = 10^5$ na każdą próbę. Zależności wyznaczonych metodą MC współczynników korelacji pomiędzy odpowiednimi parametrami położenia prób o różnych RPP od liczby obserwacji przedstawione są na rysunkach 1a – 1f.

Korelacja pomiędzy wartością średnią i medianą. Analiza przedstawionych na rys. 1 zależności pokazuje, że dla wszystkich zbadanych RPP w największym stopniu skorelowane są mediana i środek rozstępu. Przy nawet niedużej liczbie obserwacji (od około $n = 20$), wartość $r(\bar{x}, x_{med})$ szybko stabilizuje się na poziomie bliskim do odpowiedniej wartości asymptotycznej $\rho_a(\bar{x}, x_{med})$: od $\approx 0{,}71$ (rozkład |Laplace'a) do ≈ 90 (rozkład arksinusoidalny).



Rys. 1. Zależności ~~wartości~~ współczynników korelacji $r(\bar{x}, x_{med})$, $r(\bar{x}, x_M)$, $r(x_{med}, x_M)$ od liczby obserwacji wyznaczonych metodą Monte Carlo

Fig. 1. Dependences~~s~~ of the correlation coefficients $r(\bar{x}, x_{med})$, $r(\bar{x}, x_M)$, $r(x_{med}, x_M)$ on the number of observations calculated by the Monte Carlo method

Korelacja pomiędzy medianą i środkiem rozstępu. Dla prób pobranych ze wszystkich zbadanych populacji najmniej skorelowane są mediana i środek rozstępu. Wartości $r(x_{med}, x_M)$ nie zależą od kształtu rozkładu i w pierwszym przybliżeniu zmniejszają się proporcjonalnie do pierwiastka z liczby obserwacji: $r(x_{med}, x_M) \approx \rho_a(x_{med}, x_M) \approx \sqrt{2/(n+1)}$. Taki charakter zależności można wytłumaczyć tym, że mediana i środek rozstępu wyznaczane są najbardziej oddalonymi od siebie obserwacjami: mediana – centralnymi, a środek rozstępu – skrajnymi, które w najmniejszym stopniu są zależne od siebie.

Korelacja pomiędzy wartością średnią i środkiem rozstępu. Wartość współczynnika korelacji pomiędzy wartością średnią i środkiem rozstępu też monotonicznie zmniejsza się proporcjonalnie do pierwiastka z liczby obserwacji: $r(\bar{x}, x_M) \approx k_a \sqrt{2/(n+1)}$, jednak szybkość zmniejszenia się tego współczynnika zależy od wartości współczynnika amplitudy $k_a = x_a/\sigma$ próby losowej. Korelacja jest mniejsza dla prób pobranych z populacji o wyraźnie ograniczonych wartościach granicznych: arksinusoidalnego ($k_a \approx 1{,}41$) i jednostajnego ($k_a \approx 1{,}73$) oraz istotnie większa w przypadku prób pobranych z populacji o rozkładzie normalnym ($k_a \approx 2{,}33$) i Laplace'a ($k_a \approx 2{,}77$). Nawet przy stosunkowo dużej liczbie obserwacji, na przykład $n = 50$, dla prób o rozkładach normalnym, trójkątnym i trapezowym wartość współczynnika $r(\bar{x}, x_M)$ jest istotna i wynosi około ≈ 0,45 ÷ 0,35.

5. Podsumowanie

Na podstawie przeprowadzonych badań wykazano, że dla wszystkich zbadanych RPP istnieje bardzo istotna korelacja pomiędzy wartością średnią a medianą (od ≈ 0,71 do ≈ 0,90). Wartość współczynnika korelacji z wzrostem liczby obserwacji stabilizuje się na poziomie równym odwrotności współczynnika kształtu próby losowej.

We wszystkich przypadkach najmniej skorelowane są mediana i środek rozstępu. Przy zwiększeniu liczby obserwacji ten współczynnik praktycznie nie zależy od kształtu rozkładu i zmniejsza się proporcjonalnie do pierwiastka z liczby obserwacji. Jednak przy stosunkowo małych liczbach obserwacji ($n = 20 \div 40$) ta korelacja pozostaje istotną i wynosi powyżej 0,2.

Korelacja pomiędzy wartością średnią i środkiem rozstępu przyjmuje wartości pośrednie. Współczynnik korelacji też zmniejsza się odwrotnie proporcjonalnie do pierwiastka z liczby obserwacji, jednak jego wartość zależy od kształtu rozkładu i jest większa o wartość współczynnika amplitudy próby. Wartość współczynnika korelacji nawet przy liczbie obserwacji od 10 do 100 może osiągać w przybliżeniu od 0,3 do 0,6.

W metodach bazujących na obliczaniu dwu - lub trzy - elementowych ocen niepewności pomiaru, jako sumy ważonej z tych parametrów, podczas wyznaczania złożonej standardowej niepewności wyniku należy uwzględniać korelację pomiędzy podstawowymi parametrami położenia próby, nawet przy stosunkowo dużej liczebności zarejestrowanych obserwacji.

6. Literatura

[1] Dorozhovets M.: Wyniki badań korelacji między podstawowymi estymatorami parametru położenia dla serii obserwacji nieskorelowanych. Materiały konferencji: Podstawowe Problemy Metrologii: PPM-2011. Krynica 12-15.06. 2011 s. 97-100.

[2] Warsza Z. L., Galovska M.: About the best measurand estimators of trapezoidal probability distributions. Przegląd Elektrotechnczny - Electrical Review 5, 2009, s. 86-91.

[3] Zakharov I.P., Shtefan N.V.: Algorithms for reliable and effective estimation of type A uncertainty. Measurement Techniques, vol. 48, 5, 2005 p.427-437, www. Springer com. (transl. from Izmieritelnaja Tekhnika (rus.)).

[4] Fisz M.: Probability Theory and Mathematical Statistics. John Willey & Sons, London, 1963.

otrzymano / received: 25.10.2011
przyjęto do druku / accepted: 02.07.2012 *artykuł recenzowany / revised paper*